STRACHAN, TYLER JEFFREY, Ph.D. A Comparison of Parameter Estimation Algorithms for Estimating a Polytomous Log-Linear Cognitive Diagnosis Model for Polytomous Attributes. (2019)
Directed by Dr. Robert Henson. 146 pp.

Parameter estimation techniques such as an expectation-maximization (EM) algorithm have been used ubiquitously to estimate cognitive diagnosis models (CDM). The primary goal of this study was to utilize polytomous attributes in the polytomous log-linear cognitive diagnosis model (P-LCDM-PA), which is a special case of the general polytomous diagnostic model (GPDM) for polytomous attributes. Then, due to exponentially increasing the number of latent classes, explore the feasibility and efficiency in addition to the quality of parameter estimation of the stochastic expectation-maximization (SEM) and Metropolis-Hastings Robbins-Monro (MH-RM) algorithms relative to the EM algorithm.

The SEM and MH-RM algorithms may be more computationally advantageous over an EM algorithm when there exist many latent classes. As the number of measured attributes increases in a diagnostic assessment, the number of latent classes increases exponentially. The large number of classes is even more problematic when polytomous attribute levels are introduced in the diagnostic assessment. The large number of classes becomes computationally challenging when estimating a model using an EM algorithm because for each respondent, the probability of class membership is computed for every latent class. Simulation experiments were conducted examining item parameter recovery in the P-LCDM-PA, correct classification rates, and computational time between the three algorithms.

A COMPARISON OF PARAMETER ESTIMATION ALGORITHMS FOR ESTIMATING A POLYTOMOUS LOG-LINEAR COGNITIVE DIAGNOSIS MODEL FOR POLYTOMOUS ATTRIBUTES

by

Tyler Jeffrey Strachan

A Dissertation Submitted to
the Faculty of The Graduate School at
The University of North Carolina at Greensboro
in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Greensboro
2019

Approved by

_____
Committee Chair

APPROVAL PAGE

This dissertation written by TYLER JEFFREY STRACHAN has been approved by the following committee of the Faculty of The Graduate School at The University of North Carolina at Greensboro.

Committee Chair _____

Committee Members _____

_____

_____

_____
Date of Acceptance by Committee

_____
Date of Final Oral Examination

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER I

INTRODUCTION

Definition of Key Terms in Parameter Estimation

This section aims to define some of the key terms commonly used in parameter estimation of cognitive diagnosis models (CDM), which are also sometimes referred to as diagnostic classification models. An important property within the framework of parameter estimation in CDM is *conditional independence* (Lord & Novick, 1968). Conditional independence states that the correlation between a set of item responses should be zero after the effect of the latent variables are conditioned out. The set of item responses should only be correlated through the latent variables that the test or survey is measuring. Because responses to a set of items are assumed to be independent within each latent class, the property of conditional independence allows the estimation of item's parameters to be done separately for each item.

Once the property of conditional independence is assumed, estimators such as the *expectation-maximization* (EM; Bock & Aitkin, 1981) algorithm can be used, which is often associated within the frequentist estimation framework. The EM algorithm is an iterative procedure that can be used to obtain maximum likelihood estimates of parameters associated with probabilistic models in the presence of unobserved latent variables (Baker & Kim, 2004). The algorithm consists of two steps; the *expectation step*

and the *maximization step*. The expectation step in the EM algorithm replaces the unknown latent classes with their corresponding expected values, given that the item parameters have been estimated in their previous iteration of the algorithm (Rupp, Templin, & Henson, 2010). The expectation step in the EM algorithm utilizes the probability that any given respondent has the potential of being classified within each latent class. The maximization step involves finding a set of item parameter estimates that maximizes the *expected complete-data likelihood*. Essentially, the algorithm calculates the maximum likelihood estimates for the incomplete-data problem (i.e., where class membership is not known) by utilizing the expected complete-data likelihood instead of the *observed-data likelihood* because the observed-data likelihood might be complicated or numerically impossible to maximize (a more detailed discussion of the algorithm will be provided in a later section). The algorithm assumes a population distribution or *prior distribution* for the latent variables. The prior distribution expresses what is known about the latent variable when the data has not been observed. The prior simplifies the estimation process in two ways. First, the focus shifts from some large number of latent variable parameters to just the parameters of the prior distribution. Second, by imposing a distributional form for the population, the latent variable parameters become temporarily "known" (i.e., forming complete-data) and can then be marginalized out of the estimation process to allow for estimating the item parameters of the model.

Alternatively, an estimator such as a *Markov chain Monte Carlo* (MCMC) algorithm can be used, which is often defined within the Bayesian estimation framework. Bayesian estimation focuses on determining a set of parameter values that maximizes the

joint *posterior distribution* of all parameters (Rupp, et al., 2010). The posterior distribution expresses what is known about model parameters once the data has been observed. However, from a numerical standpoint, directly maximizing the joint posterior distribution of all parameters can be very difficult. MCMC algorithms have been proposed to circumvent this problem by sampling from the posterior distribution, rather than maximizing it (Junker, Patz & VanHoudnos, 2016). MCMC describes a family of algorithms for simulating data i.e., *Monte Carlo* simulation - using a statistical sequence of random draws that is known as a *Markov chain*. By constructing these steps in a particular way, it is possible to simulate values that are from a specific distribution, which is referred to as the *stationary distribution*. When draws of the MCMC are from a stationary distribution that chains are said to have converged (Patz & Junker, 1999; Junker, Patz & VanHoudnos, 2016).

Recently, Cai (2010a; 2010b) introduced a flexible framework for estimating parameters of statistical models by coupling two algorithms to formulate a joint estimation framework that addresses many of the less appealing features of strictly MCMC and maximum likelihood approaches (Chalmers & Flora, 2014). The *Metropolis-Hastings Robbins-Monro* (MH-RM; Cai, 2010a, 2010b) algorithm, like MCMC estimation, jointly estimates both item and ability parameters by utilizing a *stochastically imputed complete-data solution* with some assumed population distribution for the latent variable to exploit on a more manageable complete-data likelihood approach. The stochastically imputed complete-data solution process means that once the latent variables are "known" via a stochastic imputation procedure, the complete-data solution

is formed, which constitutes the observed response data and the now imputed latent variables. The algorithm can be partitioned into three stages: perform burn-in iterations, collect a specified number of iterations for the model parameters and compute the average of this set for each model parameter, and then perform the MH-RM stage until the model coverages with a predetermined tolerance level (a more detailed discussion of the algorithm will be provided in a later section).

The burn-in iterations of the MH-RM utilize an algorithm known as the *stochastic EM* (SEM; Diebolt & Ip, 1994a, 1994b), which involves the following: "fill-in" the latent variable with a single draw from the posterior distribution, thus forming the complete-data solution, then directly maximizing the complete-data likelihood to obtain a maximum likelihood estimate for the model parameters. Alternating between the stochastic imputation step and maximization step generates a Markov chain that coverages to a stationary distribution under mild conditions (Ip, 1994).

Description of the Problem

Over the past several decades, research in educational and psychological assessment has led to the development of CDM which provides a latent profile defining *absolute* or *partial* mastery of a set of predefined attributes (Henson, Templin, & Willse, 2009). Estimation of CDM is a very critical step in research and practice, like any statistical model, when it comes to making inferences about our population of interest. A popular estimation algorithm used to estimate CDM in both research and practice is the EM (Rupp, Templin, & Henson, 2010). However, there can be computation limitations to the EM. For example, if a diagnostic assessment is measuring 12 attributes, the number

of possible latent classes is $2^{12} = 4{,}096$. The estimation algorithm can be computationally intensive in the E-step because the posterior probability of every latent class is computed for each respondent for each step. Instead of using the EM, an MCMC algorithm can be implemented to estimate a diagnostic assessment measure 12 attributes. The MCMC also has its advantages such that it is relatively easy to implement and only requires defining the likelihood function of a CDM. However, there are limitations to using an MCMC because typically, many iterations are needed (i.e., 10,000) for the Markov chain to reach its stationary distribution, and it can be very difficult to assess the accuracy and evaluate convergence, even empirically (Sinharay, 2004). As an alternative to both the EM and MCMC, the SEM or MH-RM algorithms can be used to estimate CDM when there are many attributes present in a diagnostic assessment. Both algorithms have shown their usefulness in estimating high dimensional multidimensional item response theory (MIRT; Reckase, 1997) models. The S-step and M-step in the SEM has shown to generate a Markov chain that converges to its stationary distribution faster than the Markov chains produced by many MCMC algorithms (Diebolt & Ip, 1994a, 1994b), which are typically much longer. The MH-RM utilizes the SEM in the initial portion of the algorithm such that it moves the parameter quickly to the "neighborhood" of the MLE (Stage I and Stage II). From there, an RM update (Stage III) it used to update the parameters estimates until some set convergence criteria are met (Monroe & Cai, 2014). The advantage of SEM and MH-RM from a computational standpoint is that it couples the idea of a stochastic imputation process like MCMC estimation and also includes the maximization process used in the EM. However, when a lower number of dimensions (or

latent classes) is present, the EM can show to be computationally efficient over the SEM and MH-RM. When using CDM in application there are plenty of examples when the number attributes could be considered large. Table 1 below presents examples of various DCM studies that involved analyses of diagnostic assessments measuring many attributes (Sessoms & Henson, 2017). These are examples where the SEM and MH-RM may have provided usefulness over other algorithms.

Table 1. Examples of Diagnostic Assessments Measuring many Attributes

Source	Number of Attributes	Number of Items
Im & Park (2010)	10	43 items
Kim (2015)	10	30 items
Choi, Lee, & Park (2015)	12	43 items
Sen & Arican (2015)	13	31 items
Syetina, Gorin, & Taksuoka (2011)	15	28 items
Lee, Park, & Taylan (2011)	15	25 items
Chen, Ferron, Thompson, Gorin, & Tatsuoka (2010)	23	162 items
Chen (2012)	23	23 items

In addition to the number of attributes, the number of latent classes increases even further when polytomous attributes are used in a diagnostic assessment. Specifically, when using polytomous attributes, the number of latent classes used during the E-step is no longer computed as 2^K, where K represent the number of attributes measured in the diagnostic assessment. Instead the possible latent classes in the E-step, the computation is over $\prod_{k=1}^{K} S_k$ possible latent classes where S_k represents the total number of polytomous attribute levels for the k^{th} attribute. For example, if a diagnostic assessment measures six attributes with each attribute consisting of four polytomous skill levels, the total possible

latent classes are $4^6 = 4{,}096$. This example provides motivation for the need of more efficient estimation algorithms such as the SEM and MH-RM in diagnostic measurement with polytomous attribute levels. Table 2 provides an example of math-based content items taken from Chen & de la Torre (2013) that utilizes polytomous attribute levels. Assuming the math items are open ended questions (i.e., not multiple-choice items with 0/1 score categories), scoring of the math items such as these could include; 0 = incorrect, 1 = partial credit, or 2 = correct.

Table 2. Polytomous Attributes in an Eight-Grade Proportional Reasoning Assessment

Attributes	Level 1	Level 2
Comparing and ordering of fractions	Students should be able to compare two fractions and determine whether one of these fractions is equal to, less than, or greater than the other.	Students should be able to order three or more fractions.
Constructing ratios and proportions from a situation	Given a problem situation involving ratios, students should be able to construct a single ratio to describe the situation.	Given a proportional situation, students should be able to construct an appropriate proportion.

Note: These attributes were adapted from de la Torre, Lam, Rhoads, and Tjoe (2010). Level 0 is defined as lack of attribute mastery.

Purpose of Study

The primary goal of this study was to utilize the polytomous log-linear cognitive diagnosis model (P-LCDM; Hansen, 2013) for polytomous attributes (PA) i.e., P-LCDM-PA, which is a special case of the general polytomous diagnostic model (GPDM; Chen & de la Torre, 2018) for polytomous attributes. Then, due to the potential of exponentially increasing the number of latent classes, explore the feasibility and efficiency in addition to the quality of parameter estimation of the SEM and MH-RM algorithms relative to the

EM algorithm. In general, the EM algorithm and MCMC-based algorithms have been commonly used to estimate the parameters of various CDM in both research and practice. The SEM and MH-RM are two algorithms that have been used to estimate other multidimensional models (e.g., refer to Diebolt & Ip, 1994a, 1994b; Cai, 2010a, 2010b; Monroe & Cai, 2014) but have never been implemented for estimating CDM. The advantage of these two algorithms over EM and MCMC are that they have shown to be computationally more efficient when estimating MIRT models and even more so as the number of dimensions in the MIRT increase. The efficiency is because as the number of latent dimensions increase, algorithms become computationally slow in the EM. For an MCMC, the issue is that such a large number of iterations is needed (e.g., 10,000) to obtain convergence (Sinharay, 2004; Patz & Junker, 1999; Junker, Patz & VanHoudnos, 2016). In addition, it can be very difficult to assess accuracy and evaluate convergence, even empirically resulting in more iterations. The SEM and MH-RM circumvent these issues presented in the EM and MCMC. In the context of CDM, the efficiency of an EM algorithm largely depends on the number of latent classes. In application the number of latent classes can be large because of the number of attributes that are being measured by an instrument. For example, a cognitive diagnostic assessment that measures 10 attributes will have a total of 1,024 latent classes. The computation can be computationally expensive in the EM. In contrast an MCMC does not directly compute this probability at each step, however, as was discussed previously, the number of steps often renders a MCMC algorithm inefficient with respect to time and computation power. Because of these issues it is believed that the use of SEM and MH-RM can provide

improved efficiency with respect to computation time and power in the context of the CDM estimation framework. The improved time efficiency will ultimately increase the feasibility of using diagnostic models in application.

Research Questions

With respect to the comparison of SEM and MH-RM to the EM algorithm, the three research questions motivating this study are:

1) To what extent does the SEM and MH-RM algorithms show to be computationally faster over the EM algorithm as the number of latent classes increases?

2) How accurately are the item parameters of the P-LCDM-PA submodels estimated when comparing the SEM, MH-RM, and EM algorithms for estimation?

3) How accurately are examinees attributes (and attribute patterns) estimated when using the SEM, MH-RM, and EM algorithms to estimate the P-LCDM-PA submodels?

Organization of Study

Chapter II is divided into two main sections including a discussion of current CDM used for handling various types of data and attribute structures and a discussion of various types of estimation algorithms that can be used for estimating CDM. Chapter III discusses the functional relationship between the P-LCDM-PA and various traditional CDM, model assumptions, and derivation of the observed and complete-data likelihood functions for the P-LCDM-PA. Simulation experiments are also conducted to answer the research questions provided in the previous section.

CHAPTER II

REVIEW OF THE LITERATURE

Background

Over the past several decades, research in educational and psychological assessment has led to the development of statistical models that provide a latent profile defining *absolute mastery* or *partial mastery* of a set of predefined attributes (Rupp et al., 2010; Henson, Templin, & Willse, 2009). This set of statistical models are commonly call cognitive diagnosis models (CDM) aka diagnostic classification models (DCM). Given these latent profiles, the conditional probability of a response to a set of items is defined as a function of absolute mastery or partial mastery of the attributes that are measured by those items. CDM rely on multiple latent variables to classify respondents, thus, by definition, can be categorized as multidimensional models (Rupp, et al., 2010). Other statistical models such as MIRT contain continuous latent variables that can lead to a continuous multidimensional profile, while CDM contain categorical latent variables that lead to a discrete multidimensional profile based on statistical classifications. Typically, these types of classifications are commonly used in reporting mechanisms in diagnostic settings e.g., clinical diagnosis (e.g., refer to Templin & Henson, 2006; Jaeger, Tatsuoka, & Berns, 2003; Millon, Millon, Davis, & Grossman, 2006).

Another comparison between MIRT and CDM is the definitional grain size of the constructs and associated response processes that can be investigated (Rupp, et al., 2010).

The focus of the separate dimensions in a MIRT is typically on how different and how broadly defined the constructs relate to one another. However, the latent variables in a CDM can represent attributes that are components of a single more narrowly defined construct. For example, the latent variables in a CDM may represent different arithmetic operations that must be performed by respondents to solve addition, subtraction, multiplication, or division problems on a diagnostic measure. These attributes can be a set of more narrowly defined constructs as opposed to one general construct thought to be basic arithmetic ability.

The majority of CDM are classified as confirmatory in nature based on two perspectives (Rupp, et al., 2010): (1) a *substantive perspective* of use and (2) a *statistical perspective* of model structure. Looking from a substantive perspective of use, CDM are used to confirm or refute a hypothesis related to the relationship between how certain respondents answer to items and their underlying cognitive characteristics. In the context of diagnostic assessments, CDM could be used to test various hypotheses about a set of cognitive attributes that respondents draw on when responding to items on the diagnostic assessment. Looking from the statistical perspective of the modeling structure, CDM are of a confirmatory nature because they typically require a loading structure to be specified *a-priori* in the form of what's referred to as a $\boldsymbol{Q}$ matrix (Tatsuoka, 1983). The specification of loadings from the $\boldsymbol{Q}$ matrix are like confirmatory MIRT models such that certain factor loadings are fixed at 0 and thus the abilities of those dimensions do not influence that items response (i.e., the item does not measure that ability). The $\boldsymbol{Q}$ matrix

can be represented as a $J \times K$ matrix of dichotomous (e.g., 0/1) or polytomous (e.g., 0/1/2/3) elements that specifies which attributes are measured by each item:

$$\boldsymbol{Q}_{[J\times K]} = \begin{pmatrix} q_{11} & \cdots & q_{1K} \\ \vdots & \ddots & \vdots \\ q_{J1} & \cdots & q_{JK} \end{pmatrix}. \tag{2.1}$$

Here, J represents the number of items in the diagnostic assessment and K represents the number of attributes measured by the diagnostic assessment. For dichotomous $\boldsymbol{Q}$ matrices, a value of 1 indicates that the respondent must have mastery of that attribute to have a high conditional probability of response to an item, while a value of 0 in the $\boldsymbol{Q}$ matrix indicates that mastery of the attribute is not related to the conditional probability of response to an item (i.e., the item does not measure that attribute). For polytomous $\boldsymbol{Q}$ matrices, values greater than 0 indicate that the respondent must have that level of mastery or higher for that attribute to have a high conditional probability of response to an item, while a value of 0 follows the same definition of the dichotomous $\boldsymbol{Q}$ matrix. An important advantage of CDM is that they can unfold their theoretical potential for *complex loading structures* where each item can load on multiple attributes. The probability of a correct response depends on several attributes as opposed to absolute mastery or partial mastery of only a single attribute. This type of loading structure has been referred to a *within-item multidimensionality* in contrast to *between-item multidimensionality* in the case for *simple loading structures* (Rupp et al., 2010). Within the CDM literature, specific models make different assumptions with respect to how the

attributes interact to result in the conditional probability of a response. In general, these models can be classified into two groups: noncompensatory and compensatory models.

Basic Cognitive Diagnosis Models

Noncompensatory Models

Noncompensatory CDM reflect the assumption that a deficit in one latent variable cannot be compensated for by a surplus in a different latent variable (Rupp, et al., 2010). The noncompensatory relationship can usually be expressed mathematically, as a sequence of products such that the deficit of one latent variable defines a maximum probability that cannot be surpassed regardless of mastery of other required attributes. The noncompensatory CDM are defined such that the conditional relationship between any attribute and an item's response *depends* on the remaining required attributes that have been mastered or not mastered (Henson, et al., 2009). Due to the nature of this dependency, noncompensatory CDM can be also be referred to as *conjunctive models*. Conjunctive models are defined as a set of models for which the respondent cannot "make-up" for non-mastery of required attributes by mastery of other required attributes (Rupp et al., 2010). To have a high conditional probability of a correct response for an item that is classified under the assumption of a conjunctive model, a respondent must master all required attributes for that items.

The *deterministic input noisy "and" gate* (DINA; Haertel, 1989; Junker & Sijtsma, 2001; de la Torre & Douglas, 2004) model is among one of the simplest conjunctive models. The model utilizes the *conjunctive condensation rule*, which from the deterministic perspective, states that a respondent needs to have mastered all required

attributes for an item to have a high probability of obtaining a correct response. Additionally, the model separates respondents into two groups where one group includes those who have mastered all attributes measured by the item and the other includes all other examinees (i.e., those lacking mastery on at least one attribute measure by the item). Let $\boldsymbol{\Delta}_j = (s_j, g_j)$ be a collection of parameters for the j^{th} item. The conditional probability of a correct response for the i^{th} respondent on the j^{th} item under the DINA is defined by

$$P(X_{ij} = 1|\boldsymbol{\Delta}_j, \xi_{ij}) = (1 - s_j)^{\xi_{ij}} g^{(1-\xi_{ij})} \tag{2.2}$$

where the latent variable $\xi_{ij} \in \{0,1\}$ defines whether the i^{th} respondent has mastered ($\xi_{ij} = 1$) or not mastered $(\xi_{ij} = 0)$ all required attributes for the j^{th} item. This latent variable, ξ_{ij} , is referred to as the *deterministic input* portion of the model. The conjunctive kernel that creates ξ_{ij} is refered to as the *and-gate* because it functions like an output summary based on information about the respondent and item (Rupp, et al., 2010). This parameter is defined as

$$\xi_{ij} = \prod_{k=1}^{K} \alpha_{ik}^{q_{jk}} . \tag{2.3}$$

If the i^{th} item does not measure mastery of the k^{th} attribute, then $q_{jk} = 0$. If the i^{th} item does measure the k^{th} attribute, then $q_{jk} = 1$. Because the product is defined over all possible attributes, $\xi_{ij} = 1$ only occurs when all product terms are equal to 1. The $\xi_{ij} = 1$

indicates the i^{th} respondent had mastered all required attributes for the j^{th} item. The slip parameter s_j in the model is defined as the conditional probability of an incorrect response for the j^{th} item given the i^{th} respondent has mastered all required attributes for that item. The guess parameter g_j in the model is defined as the conditional probability of a correct response for the j^{th} item given the i^{th} respondent has not mastered all required attributes for that item. The s_j and g_j are both expressed as

$$s_j = P\left(X_{ij} = 0 | \xi_{ij} = 1\right) \tag{2.4}$$

and

$$g_j = P\left(X_{ij} = 1 | \xi_{ij} = 0\right). \tag{2.5}$$

When the chances of slipping and guessing are known and when ξ_{ij} is known, the conditional probability of a correct response can be computed. An assumption in CDM is *monotonicity* which is defined as the property such that given any respondent that masters additional skills their conditional probability of a response must be equal to or greater than the conditional probability of a response prior to learning the additional set of skills. The assumption of monotonicity under the DINA is defined as $\left(1 - s_j\right) > g_j$. As a result, a respondent who has mastered all required attributes $\left(\xi_{ij} = 1\right)$ for an item must have a higher conditional probability of responding correctly to that item compared to a respondent who has not mastered all required attributes $\left(\xi_{ij} = 0\right)$ for that item. Furthermore, the conditional probability of a correct response is only expected to be high

when all required attributes have been mastered i.e. $(1 - s_j)$ component of the model. If a subset of attributes has been mastered, the conditional probability of a correct response is expected to be low i.e., the g_j component of the model.

One possible limitation of the DINA is that sometimes it is over restrictive in the assumption that all respondents lacking at least one required attribute have the same conditional probability of a correct response. A conjunctive model that does not make this restrictive assumption is a *reduced* version of the *reparametrized unified model* (RUM, Hartz, 2002). Let $\boldsymbol{\Delta}_j = (\pi_j^*, r_{jk}^*, \ldots, r_{jK}^*)$ be a collection of parameters for the j^{th} item. The conditional probability of a correct response for the i^{th} respondent on the j^{th} item under the reduced RUM (R-RUM; DiBello, Stout, Roussos, 1995; Hartz, 2002) is defined by

$$P(X_{ij} = 1|\boldsymbol{\Delta}_j, \boldsymbol{\alpha}_i) = \pi_j^* \prod_{k=1}^{K} {r_{jk}^*}^{q_{jk}(1-\alpha_{ik})} \tag{2.6}$$

where π_j^* is the conditional probability of a correct response when the i^{th} respondent has mastered all required attributes for the j^{th} item,

$$\pi_j^* = P(X_{ij} = 1|\boldsymbol{\alpha}_i^T \boldsymbol{q}_j = \boldsymbol{q}_j^T \boldsymbol{q}_j) \tag{2.7}$$

otherwise a penalty parameter is imposed r_{jk}^* for every required attribute not mastered for that item. The r_{jk}^* penalty parameters indicate the proportional amount that the conditional probability of a correct response is reduced for each required attribute not

mastered. This pentalty is constrained such that $0 < r_{jk}^{*} < 1$. The formulation in Equation (2.6) states that the i^{th} respondent has a baseline conditional probability of π_j^* for responding correctly to the j^{th} item. The expression ${r_{jk}^{*}}^{q_{jk}(1-\alpha_{ik})}$ following the product term denotes the influence of measured attributes that have not been mastered for the i^{th} respondent. Large values of r_{jk}^{*} mean that the attribute minimally impacts the conditional probability of a correct response, while small values of r_{jk}^{*} mean that the attribute dramatically impacts the conditional probability of a correct response. Note that there is no penalty on the conditional probability of a correct response when $q_{jk} = 0$. However, when $q_{jk} = 1$ the i^{th} respondent's conditional probability of a correct response for the j^{th} item is reduced when $\alpha_{ik} = 0$ and remains unchanged when $\alpha_{ik} = 1$.

Compensatory Models

Compensatory CDM are defined such that the conditional association of any required attribute and an item does not depend on mastery of any other required attributes (Henson et. al., 2009; Rupp, et al., 2010). Thus, the increase in the conditional probability of a correct response when comparing masters to non-masters is constant across all other levels of mastery and non-mastery of the other measured attributes. Compensatory CDM have sometimes be referred to as *disjunctive models*. Disjunctive models are defined as a set of models for which the respondent can master a subset of required attributes for an item and still have a high conditional probability of correctly responding (Rupp et al., 2010). Note that while disjunctive models can be thought of compensatory, they do not satisfy the definition suggested by Henson et al. (2009) and Rupp et al. (2010).

The *deterministic input noisy "or" gate* (DINO; Templin & Henson, 2006; Templin, 2006) model is the compensatory analog to the DINA model. The DINO is classified as a disjunctive model, which assumes that mastery of any additional attributes provides little or no improvement in the conditional probability of a correct response over mastery of a single item measured attribute. Specifically, the DINO model separates respondents into two groups including those who have mastered *at least* one of the measured attributes and those who have not mastered any of the measured attributes. Like the DINA, slipping and guessing parameters are modeled for each item. Let $\boldsymbol{\Delta}_j = (s_j, g_j)$ be a collection of parameters for the j^{th} item. The conditional probability of a correct response for the i^{th} respondent on the j^{th} item is defined by

$$P(X_{ij} = 1|\boldsymbol{\Delta}_j, \omega_{ij}) = (1 - s_j)^{\omega_{ij}} g^{(1-\omega_{ij})} \tag{2.8}$$

where the latent variable $\omega_{ij} \in \{0,1\}$ defines whether the i^{th} respondent has mastered at least one of the required attributes ($\omega_{ij} = 1$) or not mastered any required attributes $(\omega_{ij} = 0)$ for the j^{th} item. This component, ω_{ij}, is referred to as the *deterministic input* portion of the model. The disjunctive kernel that creates ω_{ij} is referred to as the *or-gate* because it utilizes the disjunctive condensation rule that indicates whether or not at least one measured attribute is present and can be expressed as

$$\omega_{ij} = 1 - \prod_{k=1}^{K} (1 - \alpha_{ik})^{q_{jk}} \tag{2.9}$$

If the i^{th} item does not require mastery of the k^{th} attribute, $q_{jk} = 0$, then whether a respondent has an $\alpha_{ik} = 0$ or $\alpha_{ik} = 1$ does not matter. However, if i^{th} item does require mastery of the k^{th} attribute, $q_{jk} = 1$, then whether a respondent has an $\alpha_{ik} = 0$ or $\alpha_{ik} = 1$ does help identify group membership. Because the product is defined over all attributes, $\omega_{ij} = 1$ only occurs when the product term is 0. An $\alpha_{ik} = 1$ for at least one measured attribute must be present for the i^{th} respondent to have a high conditional probability of a correct response. Thus, mastery of any measured attribute for an item can compensate for not mastering any of the other measured attributes for that item. The slip parameter s_j in the model is defined as the conditional probability of an incorrect response for the j^{th} item given the i^{th} respondent has mastered at least one required attribute for that item. The guess parameter g_j in the model is defined as the conditional probability of correct response for the j^{th} item given the i^{th} respondent has not mastered any measured attributes for that item. The s_j and g_j are both expressed similar to the DINA, but are conditional on ω_{ij} and thus are

$$s_j = P\left(X_{ij} = 0 | \omega_{ij} = 1\right) \tag{2.10}$$

and

$$g_j = P\left(X_{ij} = 1 | \omega_{ij} = 0\right). \tag{2.11}$$

When the chances of slipping and guessing are known and when ω_{ij} is known, the conditional probability of a correct response can be computed. The assumption of

monotonicity under the DINO is defined as $(1 - s_j) > g_j$. Thus, a respondent who has at least one measured attribute $(\omega_{ij} = 1)$ for an item must have a higher conditional probability of responding correctly to that item compared to a respondent who has not mastered any measured attributes $(\omega_{ij} = 0)$ for that item. The conditional probability of a correct response is expected to be high when at least one required attribute has been mastered. If no attributes have been mastered, the conditional probability of a correct response is expected to be low.

The *compensatory reparametrized unified model* (C-RUM; Hartz, 2002) model is among one of the simplest compensatory CDM. There are two different types of item parameter components defined in the model including and intercept parameter, $\lambda_{j,(0)}$ for the j^{th} item where $-\infty < \lambda_{j,(0)} < \infty$ and a set of slope parameters $\boldsymbol{\lambda}_j = (\lambda_{j1}, \ldots, \lambda_{jK})$ for each required attribute for the j^{th} item such that $0 < \boldsymbol{\lambda}_j < \infty$. The intercept and slope parameters are combined to form the kernel function of the model such that

$$kernel = -\lambda_{j,(0)} - \boldsymbol{\lambda}_j h(\boldsymbol{\alpha}_i, \boldsymbol{q}_j) \tag{2.12}$$

In Equation (2.12) the function $h(\boldsymbol{\alpha}_i, \boldsymbol{q}_j)$ is a mapping function that defines the linear combination of $\boldsymbol{\alpha}_i$ and $\boldsymbol{q}_j$ such that, for the compensatory RUM,

$$h(\boldsymbol{\alpha}_i, \boldsymbol{q}_j) = \alpha_{i1} q_{j1} + \cdots + \alpha_{ik} q_{jk} + \cdots + \alpha_{iK} q_{jK}. \tag{2.13}$$

Let $\boldsymbol{\Delta}_j = \left(\boldsymbol{\lambda}_j, \lambda_{j,(0)}\right)$ be a collection of parameters for the j^{th} item. Once the kernel element has been defined, the C-RUM is defined such that the conditional probability of a correct response for the i^{th} respondent on the j^{th} item is

$$P\left(X_{ij} = 1 | \boldsymbol{\Delta}_j, \boldsymbol{\alpha}_i\right) = \frac{1}{1 + Exp\left[-\lambda_{j,(0)} - \boldsymbol{\lambda}_j h\left(\boldsymbol{\alpha}_i, \boldsymbol{q}_j\right)\right]}. \tag{2.14}$$

For the C-RUM, the lowest conditional probability of a correct response is defined as a function of $\lambda_{j,(0)}$ which is similar to a guessing parameter but on a different scale. The conditional probability of a correct response is increased for every measured attribute of that item that is mastered, which is defined by $\boldsymbol{\lambda}_j$. Thus, the relationship between any attribute and item performance, as defined by expected change in log-odds, is not conditional on the remaining measured attributes for an item.

Advanced Cognitive Diagnosis Models

Background

The CDM discussed in the previous sections have certain adherent limitations that can inhibit the ability to extract a larger amount diagnostic information from an assessment. Specifically, these models are also only limited to dichotomous response data including multiple-choice diagnostic assessments that are scored as either right or wrong. These basic models essentially limit the possible applications of CDM when psychometricians are wanting to analyze either ordinal or nominal response items. Within the last several years, research has led to the development of more advanced CDM that allowed for more flexibility in handling different types of response data. These advanced

models also allow for larger extractions of diagnostic information from an assessment that would otherwise be ignored by less complex models. The advanced models discussed in the following sections have the capabilities of handling various types of response data including ordinal and nominal responses and allow for the generalization of ordinal attribute levels that go beyond the prototypical definitions of *mastery* and *non-mastery* ubiquitously discussed in the CDM literature. These advanced models also utilize the same definitional structure of compensatory and noncompensatory previously discussed. Table 2 provides a classification summary of the various models used in diagnostic measurement. This classification is based on whether the models are defined as compensatory/noncompensatory or have the capabilities of handling polytomous/dichotomous manifest response variables or latent predictor variables. Looking at Table 3, models such as the *log-linear cognitive diagnosis model* (LCDM; e.g., Henson, et al., 2009) and *generalized deterministic input noisy "and" gate* (G-DINA; e.g., de la Torre, 2011) model are both defined as compensatory/noncompensatory and have the capabilities of handing both dichotomous/ polytomous manifest response variables and latent predictor variables. The *general diagnostic model* (GDM; e.g., von Davier, 2005) is a compensatory model that has the capabilities of handing both dichotomous/ polytomous manifest response variables and latent predictor variables. The compensatory/noncompensatory *general diagnostic classification model for multiple-choice* (GDCM-MC; e.g., Dibello, Henson, &, Stout, 2015) option-based scoring and the noncompensatory *multiple-choice deterministic input*

noisy "and" gate (MC-DINA; e.g., de la Torre, 2009) are both models that allow for dichotomous latent predictor variables and polytomous manifest response variables

Table 3. Classifications of Various CDM

		Latent Predictor Variables		
		Dichotomous	Polytomous	Model Type
Manifest Response Variables	Dichotomous	LCDM G-DINA R-RUM DINA	DINA LCDM G-DINA R-RUM	Noncompensatory
		GDM C-RUM LCDM DINO G-DINA	GDM DINO C-RUM LCDM G-DINA	Compensatory
	Polytomous	GDCM-MC MC-DINA G-DINA LCDM	 LCDM G-DINA	Noncompensatory
		GDCM-MC GDM G-DINA LCDM	GDM LCDM G-DINA	Compensatory

Note. LCDM = log-linear cognitive diagnosis model; G-DINA = generalized deterministic input noisy "and" gate; GDM = general diagnostic model; GDCM-MC = general diagnostic classification model for multiple-choice; MC-DINA = multiple-choice deterministic input noisy "and" gate

General Diagnostic Model

The general diagnostic model (GDM; von Davier, 2005, 2008; von Davier & Yamamoto, 2004) provides a generalized framework for the development of cognitive diagnostic models. In the item response modeling framework, the conditional probability

of a response to the x^{th} category where $X_{ij} \in \{0,1, \ldots, m_j\}$ for the i^{th} respondent on the j^{th} item is defined by

$$P(X_{ij} = x|\boldsymbol{\lambda}_{xj}, \boldsymbol{q}_j, \boldsymbol{\theta}_i) = \frac{\exp[f(\boldsymbol{\lambda}_{xj}, \boldsymbol{q}_j, \boldsymbol{\theta}_i)]}{1 + \sum_{y=1}^{m_j} \exp[f(\boldsymbol{\lambda}_{yj}, \boldsymbol{q}_j, \boldsymbol{\theta}_i)]} \tag{2.15}$$

where $\boldsymbol{\lambda}_{xj}$ represents a collection of item parameters such that $\boldsymbol{\lambda}_{xj} = (\gamma_{xj}, \beta_{xj})$, $\boldsymbol{q}_j$ represents the $\boldsymbol{Q}$ vector for the j^{th} item such that $\boldsymbol{q}_j = (q_{j1}, \ldots, q_{jk}, \ldots, q_{jK})$, and $\boldsymbol{\theta}_i$ represents the vector of continuous, binary, or ordinal skills for the i^{th} respondent such that $\boldsymbol{\theta}_i = (\alpha_{i1}, \ldots, \alpha_{ik}, \ldots, \alpha_{iK})$. The item parameters γ_{xj} and β_{xj} represent the slope and intercept parameters, respectively, for the j^{th} item. Given a nonzero $\boldsymbol{Q}$ matrix element, γ_{xj} in $f(.)$ determines how much the attributes in $\boldsymbol{\theta}_i$ contribute to the conditional response probabilities for the j^{th} item. The mathematical formulation of the model was utilized as a basis for numerous developments such as the LCDM (Henson et al. 2009; Rupp et al., 2010) for binary skill attributes and dichotomous response data, and the linear or partial-credit GDM (pGDM; von Davier, 2005, 2008) for binary and ordinal attributes and dichotomous and polytomous response data. The conditional probability of a response to the x^{th} category where again $X_{ij} \in \{0,1, \ldots, m_j\}$ for the i^{th} respondent on the j^{th} item under the pGDM is defined by

$$P(X_{ij} = x|\boldsymbol{\lambda}_{xj}, \boldsymbol{q}_j, \boldsymbol{\alpha}_i) = \frac{\exp[\beta_{xj} + \sum_{k=1}^{K} x\gamma_{jk} h(\boldsymbol{q}_j, \boldsymbol{\alpha}_i)]}{1 + \sum_{y=1}^{m_j} \exp[\beta_{yj} + \sum_{k=1}^{K} y\gamma_{jk} h(\boldsymbol{q}_j, \boldsymbol{\alpha}_i)]} \tag{2.16}$$

where $\boldsymbol{\alpha}_i = (\alpha_{i1}, \dots, \alpha_{ik}, \dots, \alpha_{iK})$ is a discrete k-dimensional skill vector that harbors binary or ordinal elements and $h(\boldsymbol{q}_j, \boldsymbol{\alpha}_i)$ is the mapping function used to specify the linear combination of $\boldsymbol{\alpha}_i$ and $\boldsymbol{q}_j$ where

$$h(\boldsymbol{q}_j, \boldsymbol{\alpha}_i) = (q_{j1}\alpha_{i1}, \dots, q_{jk}\alpha_{ik}, \dots, q_{jK}\alpha_{iK}) \tag{2.17}$$

so that the k^{th} element of h is specified as $h_k(q_j, \alpha_i) = q_{jk}\alpha_{ik}$. When q_{jk} is defined as 0/1, this is equivalent to

$$h_k(q_j, \alpha_i) = \begin{cases} \alpha_k, & if\ q_{jk} = 1 \\ 0, & else \end{cases} \tag{2.18}$$

such that only the k skills with nonzero $\boldsymbol{Q}$ matrix elements q_{jk} contributes to the conditional response probabilities for the i^{th} respondent on the j^{th} item. If $q_{jk} = 1$ then the total contribution of $\gamma_{jk}h_k(q_j, \alpha_i) = \gamma_{jk}\alpha_{ik}$ in the kernel, else if $q_{jk} = 0$ no contribution is made in the kernel. The imposed definition in Equation (2.18) is appropriate for $\boldsymbol{Q}$ matrices with 0/1 elements in correspondence with various discrete skill level selections such as $\alpha_{ik} \in \{-m, \dots, 0, \dots, m\}$ or dichotomies like $\alpha_{ik} \in \{0,1\}$. However, the choice of the mapping function $h(.)$ does not work well with $\boldsymbol{Q}$ matrices that have elements other than 0/1. This choice of the mapping function is particularly important if the γ_{jk} are to be estimated in the pGDM. In the cases with integer or real-valued $\boldsymbol{Q}$ matrices a useful choice for the mapping function h is

$$h_k(q_j, \alpha_i) = \min(q_{jk}, \alpha_{ik}) \tag{2.19}$$

$$\forall k = 1, \dots, K$$

where $q_{jk} = 0, 1, 2, \dots, m$ in correspondence with skill level $\alpha_{ik} \in \{0, 1, 2, \dots, m\}$. This notation is consistent with the definition in Equation (2.18) when q_{jk} is 0/1 and $\alpha_{ik} \in \{0,1\}$ but differentiates in cases using arbitrary skill levels for q_{jk} and $\alpha_{\boldsymbol{ik}}$ elements. The purpose of choosing the minimum of q_{jk} and $\alpha_{\boldsymbol{ik}}$ is that the pGDM may be used for skill assessments where $\boldsymbol{Q}$ matrix elements represent a sufficient level for the k^{th} skill on the j^{th} item. A higher skill level than q_{jk} will not increase the conditional probability of response to the j^{th} item. However, a skill level lower than q_{jk} decreases the conditional probability of a response to the j^{th} item.

The choice of $h(\boldsymbol{q}_j, \boldsymbol{\alpha}_i)$ in correspondence with $\boldsymbol{Q}$ matrices containing 0/1 elements leads to a model that retains many components of well-known IRT models while generalizing these models to diagnostic applications with multivariate latent skills. In addition, the slope parameter of the model is subject to the constraint $\gamma_{xjk} = x\gamma_{jk}$ such that the resulting instance is a GDM for dichotomous and polytomous pGDM. The discrete scores α_k are determined before estimation of the model and can be specified by the user. These discrete scores are used to assign real numbers to the discrete skill levels. Assuming the number of skill levels is $s_k = 2$ possible choices of dichotomous skill levels could be $\alpha_{ik} \in \{-1, 1\}$ or $\alpha_{ik} \in \{-.5, .5\}$. Generalizing this concept to polytomous, ordinal skill levels with the number of levels being $s_k = m + 1$ while determining the levels can be specified as $\alpha_{ik} \in \{(0 - z), (1 - z), \dots, (m - z)\}$ where $z = \frac{m}{2}$. If $K = 1$ and $s_k = 61$ such that $\alpha_{ik} \in \{-4, \dots, 0, \dots 4\}$ i.e., α is treated as a continuous latent

variable, the pGDM is mathematically equivalent to the generalized partial credit model (Muraki, 1992) in the IRT literature.

Log-Linear Cognitive Diagnosis Model

The log-linear cognitive diagnostic model (LCDM; Henson, et al., 2009) is a flexible model with the capabilities of defining the relationships between categorical variables and an items response using both compensatory and noncompensatory relationships. Let $\boldsymbol{\Delta}_j$ be a collection of effect-level (main and interaction) and intercept parameters $\boldsymbol{\Delta}_j = \left(\boldsymbol{\lambda}_j, \lambda_{j,0}\right)$ for the j^{th} item such that $\boldsymbol{\lambda}_j = \left(\lambda_{j,1,(k)}, \dots, \lambda_{j,2,(k,k')}, \dots, \lambda_{j,K_j,(1,\dots,K_j)}\right)$. The conditional probability of a correct response under the LCDM for the i^{th} respondent on the j^{th} item is defined by

$$P\left(X_{ij} = 1 | \boldsymbol{\Delta}_j, \boldsymbol{\alpha}_i\right) = \frac{1}{1 + exp\left[-\boldsymbol{\lambda}_j^T h\left(\boldsymbol{\alpha}_i, \boldsymbol{q}_j\right) - \lambda_{j,0}\right]}, \tag{2.20}$$

It's assumed that the intercept parameters $-\infty < \lambda_{j,0} < \infty$, main-effect parameters $0 < \lambda_{j,1,(k)}, \dots, \lambda_{j,1,(K_j)} < \infty$ and interaction-effect parameters $-\infty < \lambda_{j,2,(k,k')}, \dots, \lambda_{j,K_j,(1,\dots,K_j)} < \infty$. The $\boldsymbol{\alpha}_i = (\alpha_{i1}, \dots, \alpha_{ik}, \dots, \alpha_{iK})^T$ represents an attribute mastery profile for the i^{th} respondent such that $\alpha_{ik} = 0$ if the respondent has not mastered the k^{th} attribute and $\alpha_{ik} = 1$ if the respondent has mastered the k^{th} attribute. The $\boldsymbol{q}_j = \left(q_{j1}, \dots, q_{jk}, \dots, q_{jK}\right)$ represents elements from the $\boldsymbol{Q}$ matrix for the j^{th} item where a $q_{jk} = 0$ states that the k^{th} attribute is not measured by the j^{th} item and $q_{jk} = 1$

states that the k^{th} attribute is measured by the item. The mapping function $h(\boldsymbol{\alpha}_i, \boldsymbol{q}_j)$ is used to specify the linear combination of $\boldsymbol{\alpha}_i$ and $\boldsymbol{q}_j$,

$$\boldsymbol{\lambda}_j^T h(\boldsymbol{\alpha}_i, \boldsymbol{q}_j) = \sum_{k=1}^{K} \lambda_{j,1,(k)} \alpha_{ik} q_{jk} + \sum_{k=1}^{K-1} \sum_{k'=k+1}^{K} \lambda_{j,2,(k,k')} \alpha_{ik} q_{jk} \alpha_{ik'} q_{jk'} + \cdots + \lambda_{j,K_j,(1,\ldots,K_j)} \prod_{k=1}^{K_j} \alpha_{ik} q_{jk}, \quad (2.21)$$

where $K_j = \sum_{k=1}^{K_j} q_{jk}$ represents the number of required attributes for the j^{th} item defined in the $\boldsymbol{Q}$ matrix. The subscript following the first comma in λ_j represent the effect-level and the parentheses following the second comma include the attribute effect. For example, $\lambda_{j,1,(2)}$ would represent the main effect-level for second attribute and $\lambda_{j,2,(1,2)}$ would represent the two-way interaction effect-level between the first and second attribute.

A property of the LCDM is that there exists a mathematical relationship such that a set of constraints placed on the LCDM can correspond to the natural definition of noncompensatory and compensatory models. Typically, the unconstrained LCDM can be initially used to investigate the nature of the relationship between attribute mastery and the conditional probability of a correct response at an item-by-item basis. This framework allows the LCDM to function as a general model that may be used to suggest specific reduced models that align with the particular patterns of the original model estimates. That is, because the LCDM is defined as a general model that has as many parameters as equivalence classes (i.e., attribute patterns with unique condition probabilities), then it is

possible to place constraints on these parameters such that the LCDM reduces to many of the models familiar in the literature (e.g., DINA, DINO, R-RUM, and C-RUM).

DINA

There exists a mathematical relationship such that a set of constraints placed on the LCDM can correspond to the prototypical formulation of the DINA (Henson, et al., 2009). Recall that the DINA defines only two parameters: slipping parameters s_j that define the conditional probability of an incorrect response for the j^{th} item given a respondent has mastered all required attributes and guessing parameters g_j that define the conditional probability of a correct response given all attributes have not been mastered. In the case of $K = 2$ for simplicity purposes, the reduced LCDM under the DINA can be expressed as

$$P(X_{ij} = 1|\boldsymbol{\Delta}_j, \boldsymbol{\alpha}_i) = \frac{1}{1 + exp[-(0)\alpha_{i1} - (0)\alpha_{i2} - \lambda_{j,2,(1,2)}\alpha_{i1}\alpha_{i2} - \lambda_{j,0}]} \qquad (2.22)$$

where the 0's have been imposed to emphasize the constraint that $\lambda_{j,1,(1)} = \lambda_{j,1,(2)} = 0$ and $\lambda_{j,2,(1,2)} > 0$. Notice that if all required attributes have not been mastered for the j^{th} item, then $\alpha_{i1}\alpha_{i2} = 0$ and the conditional probability of a correct response is only a function of $\lambda_{j,0}$. However, if all required attributes have been mastered for the j^{th} item, then the conditional probability of a correct response increases by a factor of $\lambda_{j,2,(1,2)}$. Because $\lambda_{j,2,(1,2)} > 0$ the reduced LCDM can be defined as a conjunctive model. The functional relationship between the reduced LCDM and prototypical formulation of the DINA is mathematically expressed as

$$\lambda_{j,0} = -\ln\left(\frac{1-g_j}{g_j}\right) \tag{2.23}$$

and

$$\lambda_{j,2,(1,2)} = -\lambda_{j,0} - \ln\left(\frac{s_j}{1-s_j}\right). \tag{2.24}$$

when there are two attributes measured by the j^{th} item. If assuming $K = 4$, and the j^{th} item measured all four attributes, the $\lambda_{j,2,(1,2)}$ in Equation (2.24) would be replaced with $\lambda_{j,4,(1,2,3,4)}$.

DINO

In addition to the DINA, there exists a mathematical relationship such that a set of constraints placed on the LCDM can correspond to the prototypical formulation of the DINO (Henson, et al., 2009). Recall that the DINO also defines only two parameters: slipping parameters s_j which defines the conditional probability of an incorrect response for the j^{th} item given a respondent has mastered at least one required attribute (slips up and misses) and guessing parameter g_j which defines the conditional probability of a correct response given all attributes have not been mastered (guesses the correct response). In the case of $K = 2$ for simplicity purposes, the reduced LCDM under the DINO can be expressed as

$$P(X_{ij} = 1|\boldsymbol{\Delta}_j, \boldsymbol{\alpha}_i) = \frac{1}{1 + exp[-\lambda_j\alpha_{i1} - \lambda_j\alpha_{i2} + \lambda_j\alpha_{i1}\alpha_{i2} - \lambda_{j,0}]} \tag{2.25}$$

where λ_j is a single value estimated along with $\lambda_{j,0}$ for the j^{th} item. Notice that if no required attributes have been mastered for the j^{th} item, the conditional probability of a correct response is only a function of $\lambda_{j,0}$. The conditional probability of response increases by λ_j when at least one attribute has been mastered. The sign in front of each λ_j is determined by

$$(-1)^{(q-1)} = sign\ of\ \lambda \tag{2.26}$$

where q denotes the number of attributes involved in that specific effect. Thus, main effects are positive, two-way interaction effects are negative, three-way interaction effects are positive, and so on. The signed relationship denotes that mastery of any additional required attributes for the j^{th} item does not increase the conditional probability of a correct response. The functional relationship between the reduced LCDM and prototypical formulation of the DINO is mathematically expressed as Equation (2.23) and

$$\lambda_j = -\lambda_{j,0} - \ln\left(\frac{s_j}{1-s_j}\right). \tag{2.27}$$

R-RUM

The DINA and DINO models are considered simple models in that only two parameters are used to model the conditional probability for all attribute patterns. However, it is also possible to define more complex constraints for models such as the R-RUM (Henson, et al., 2009). Specifically, there exists a mathematical relationship such that a set of constraints placed on the LCDM can correspond to the prototypical

formulation of the R-RUM. Recall that the R-RUM defines only two parameters: The π_j^* parameter that defines the conditional probability of a correct response for the j^{th} item given a respondent has mastered all required attributes and penalty parameters r_{jk}^* that penalize the respondent for not mastering the k^{th} attributes on the j^{th} item. Unlike the R-RUM, the LCDM has a single model parameter that is used to describe the probability of a correct response given that all measured attributes have not been mastered (in contrast to the R-RUM π_j^*). As a result, the LCDM typically defines an increase in the the odds of a correct response for each attribute that is mastered as opposed to a penalty for each attribute not mastered. Therefore, an "inverse" R-RUM (Henson, et al., 2009) can be used to define the relationship between the LCDM and prototypical formation of the R-RUM. In the case of $K = 2$ for simplicity purposes, the reduced LCDM under the R-RUM can be expressed as

$$P(X_{ij} = 1|\boldsymbol{\Delta}_j, \boldsymbol{\alpha}_i) = \frac{1}{1 + exp[-\lambda_{j,1,(1)}\alpha_{i1} - \lambda_{j,1,(2)}\alpha_{i2} - \lambda_{j,2,(1,2)}\alpha_{i1}\alpha_{i2} - \lambda_{j,(0)}]} \quad (2.28)$$

The "inverse" R-RUM is mathematically equal to the R-RUM and therefore, differentiates only in the definition of each item parameter and its respective space. Let $\boldsymbol{\Delta}_j = (\pi_j^*, r_{jk}^*, \dots, r_{jK}^*)$ be a collection of parameters for the j^{th} item. The conditional probability of a correct response for the "inverse" R-RUM (Henson, et al., 2009) is expressed as

$$P(X_{ij} = 1|\boldsymbol{\Delta}_j, \boldsymbol{\alpha}_i) = \pi_j^{*'} \prod_{k=1}^{K} \frac{1}{r_{jk}^{* \, q_{jk}\alpha_{ik}}} \quad (2.29)$$

where $\pi_j^{*\prime}$ is the conditional probability of a correct response given no attributes have been mastered for the j^{th} item and r_{jk}^* is defined in the same way as the R-RUM. The functional relationship between the reduced LCDM and prototypical formulation of the "inverse" R-RUM is mathematically expressed as

$$\lambda_{j,0} = -\ln\left(\frac{1-\pi_j^{*\prime}}{\pi_j^{*\prime}}\right) \tag{2.30}$$

and

$$\lambda_{j,2,(1,2)} = -\ln\left(\frac{1+e^{-\lambda_{j,0}}}{1+e^{-\lambda_{j,1,(1)}-\lambda_{j,0}}+e^{-\lambda_{j,1,(2)}-\lambda_{j,0}}-e^{-\lambda_{j,1,(1)}-\lambda_{j,1,(2)}-\lambda_{j,0}}}\right). \tag{2.31}$$

Because $\lambda_{j,2,(1,2)}$ is a function of $\lambda_{j,1,(1)}$ and $\lambda_{j,1,(2)}$, no additional parameters are required when compared to the total number of estimated parameters in the "inverse" R-RUM. Again, this example represents an item where $K = 2$, but it's possible to extend Equation (2.31) to items that measure three or more attributes.

C-RUM

Finally, the easiest of relationships to demonstrate is between the LCDM and the C-RUM (Henson, et al., 2009). There exists a mathematical relationship such that there is a set of constraints placed on the LCDM that relates to the prototypical formulation of the C-RUM. Recall that the C-RUM only defines two different types of parameters for each item: an intercept parameter $\lambda_{j,0}$ and main-effect parameters $\boldsymbol{\lambda}_j = (\lambda_{jk}, \ldots, \lambda_{K_j})$. As is the case with the LCDM, the conditional probability of a correct response when no

required attributes have been mastered is only a function of $\lambda_{j,0}$ when using the C-RUM. Because $0 \leq \boldsymbol{\lambda}_j < \infty$, the conditional probability a correct response increases for each measured attributed that is mastered by the j^{th} item. In the case of $K = 2$ for simplicity purposes, the reduced LCDM under the C-RUM can be expressed as

$$P\left(X_{ij} = 1|\boldsymbol{\Delta}_j, \boldsymbol{\alpha}_i\right) = \frac{1}{1 + exp\left[-\lambda_{j,1,(1)}\alpha_{i1} - \lambda_{j,1,(2)}\alpha_{i2} - \lambda_{j,2,(1,2)}(0)(0) - \lambda_{j,0}\right]}. \quad (2.32)$$

In this example, the C-RUM is simply defined by setting $\lambda_{j,1,(2)} = 0$.

Polytomous Log-Linear Cognitive Diagnosis Model

The *polytomous log-linear cognitive diagnosis model* (P-LCDM; Hansen, 2013) is an extension of the dichotomous LCDM proposed by Henson et al. (2009). The LCDM framework can be adapted to handle polytomous response data in ordinal response categories, which allows the application of these models to a broader range of diagnostic assessments. Following the notation in Hansen (2013), let K represent the total number of ordered categories where $y_i \in \{0,1, \dots, K-1\}$. Following Samejima's (1969) graded response IRT model and subsequent multidimensional extensions (e.g., Muraki & Carlson, 1995; Gibbons et al., 2007), the conditional probability for a given response to the i^{th} item on the k^{th} category may be computed by

$$P(y_i = k|\boldsymbol{x}) = P(y_i \geq k|\boldsymbol{x}) - P(y_i \geq k+1|\boldsymbol{x}). \quad (2.33)$$

where the set of boundary response probabilities are defined as

$$P(y_i \geq 0|\boldsymbol{x}) = 1$$

$$P(y_i \geq 1|\boldsymbol{x}) = \frac{1}{1 + Exp\left[-\left(\alpha_{i,1} + h(\boldsymbol{\gamma}_i, \boldsymbol{q}_i, \boldsymbol{x})\right)\right]}$$

...

$$P(y_i \geq k|\boldsymbol{x}) = \frac{1}{1 + Exp\left[-\left(\alpha_{i,k} + h(\boldsymbol{\gamma}_i, \boldsymbol{q}_i, \boldsymbol{x})\right)\right]} \tag{3.34}$$

...

$$P(y_i \geq K-1|\boldsymbol{x}) = \frac{1}{1 + Exp\left[-\left(\alpha_{i,K-1} + h(\boldsymbol{\gamma}_i, \boldsymbol{q}_i, \boldsymbol{x})\right)\right]},$$

$$P(y_i \geq K|\boldsymbol{x}) = 0.$$

Here $\alpha_{i,1}, \ldots, \alpha_{i,k}, \ldots, \alpha_{i,K-1}$ are defined as the $K-1$ intercept parameters for the i^{th} item, $\boldsymbol{\gamma}_i$ represents a vector of parameters for all main effects and interaction effects for the i^{th} item, and $\boldsymbol{x}$ is the vector of discrete latent variables. The mapping function $h(\boldsymbol{\gamma}_i, \boldsymbol{q}_i, \boldsymbol{x})$ is defined similarly in Equation (2.21).

Because the P-LCDM is an extension of the LCDM and the LCDM is also a general model. In addition, an important property of the P-LCDM is that there exists a mathematical relationship such that a set of constraints placed on the P-LCDM can correspond to the natural definition of noncompensatory and compensatory models for polytomous responses. In fact, it would be possible to first fit the unconstrained P-LCDM as a method to investigate the nature of the relationships between attribute mastery and the conditional probability of a response at an item-by-item basis. Depending on the estimates, the unconstrained P-LCDM may then be used to suggest specific reduced models that align with the particular patterns of the original model

estimates. Specifically, because the P-LCDM is defined as a general model that has as many parameters as equivalence classes (i.e., attribute patterns with unique condition probabilities), then it is possible to place constraints on these parameters such that the P-LCDM reduces to many of the models familiar in the literature (e.g., DINA, DINO, R-RUM, or C-RUM). Another property of the P-LCDM is that the model can be used to define polytomous graded response models that are natural extensions to the DINA, DINO, R-RUM, and C-RUM, which is discussed in Chapter III. Chapter III also includes a discussion about the extension of the P-LCDM to allow for polytomous attributes.

Generalized Deterministic Input Noisy "and" Gate Model

The generalized deterministic input noisy "and" gate (G-DINA; de la Torre, 2011) model is an extension of the LCDM that allows for different link functions other than the logit link function. The G-DINA also can be seen as a generalized version of the DINA model with more relaxed assumptions. Specifically, the model relaxes the DINA assumption that states that the conditional probability of a correct response is equal for all latent classes within each of the groups. Recall the DINA separates respondents into two groups including those who have mastered and not mastered all required attributes for an item. Similar to the DINA, the G-DINA requires specification of a $J \times K$ $\boldsymbol{Q}$ matrix where $q_{jk} = 1$ if mastery of the k^{th} attribute is required to have a higher probability of correctly responding to the j^{th} item, while $q_{jk} = 0$ if mastery of the k^{th} attribute is not measured for the j^{th} item. However, instead of two groups, the G-DINA partitions the latent classes into $2^{K_j^*}$ groups in a similar way as the LCDM. Here, $K_j^* = \sum_{k=1}^{K} q_{jk}$ defines the number of required attributes for the j^{th} item. For notational convenience but without loss of

generality, let the first K_j^* attributes be required for the j^{th} item, and define $\boldsymbol{\alpha}_{lj}^*$ as the reduced attribute vector whose elements are the required attributes for the j^{th} item. For example, if mastery of the first and third attributes are required for the j^{th} item, then the attribute vector $\boldsymbol{\alpha}_{lj}$ reduces to $\boldsymbol{\alpha}_{lj}^* = (\alpha_{l1}, \alpha_{l3})$. What is unique to the G-DINA is the introduction of several link functions. In a similar way as link functions are used in the Generalized Linear Model, the choice of link function can allow for models that were initially expressed in a complex way (e.g., products or requiring interaction terms) to be expressed as linear model. Three of the most commonly discussed link functions discussed here are referred to as *identity*, *logit*, and *log*. The G-DINA is based on the identity link where the conditional probability of a correct response for the l^{th} latent class on the j^{th} is represented mathematically as

$$P(\boldsymbol{\alpha}_{lj}^*) = \delta_{j0} + \sum_{k=1}^{K_j^*} \delta_{jk}\alpha_{lk} + \sum_{k'>k}^{K_j^*} \sum_{k=1}^{K_j^*-1} \delta_{jkk'}\alpha_{lk}\alpha_{lk'} + \cdots + \delta_{j12\ldots K_j^*} \prod_{k=1}^{K_j^*} \alpha_{lk} \qquad (2.35)$$

where δ_{j0} represents the intercept parameter for the j^{th} item where $0 \leq \delta_{j0} < 1$ and δ_{jk} is the main effect parameter due to α_k for the j^{th} item where $0 \leq \delta_{jk} < 1$ (non-negative) if $P\left(\mathbf{0}_{K_j^*}\right) \leq P(\boldsymbol{\alpha}_{lj}^*)$ for $\sum_{k=1}^{K_j^*} \alpha_{lk} = 1$ where $\mathbf{0}_{K_j^*}$ is represented as the null vector of length K_j^*. These equations would imply that mastery of any single attribute required by an item would correspond to an increase in the respondent's conditional probability of a correct response. Like the LCDM, the $\delta_{jkk'}$ is the interaction effect parameter due to α_k and $\alpha_{k'}$ for the j^{th} item. Finally, while interactions can be positive or negative, they must

be defined for the identity link such that $0 \leq P(\boldsymbol{\alpha}_{lj}^{*}) \leq 1$. The logit link results in a generalized model that can be referred to as the *log-odds* CDM, which is equivalent to the LCDM. The logit link function can be mathematically expressed as

$$logit\left[P(\boldsymbol{\alpha}_{lj}^{*})\right] = \lambda_{j0} + \sum_{k=1}^{K_j^{*}} \lambda_{jk}\alpha_{lk} + \sum_{k'>k}^{K_j^{*}} \sum_{k=1}^{K_j^{*}-1} \lambda_{jkk'}\alpha_{lk}\alpha_{lk'} + \cdots + \lambda_{j12\ldots K_j^{*}} \prod_{k=1}^{K_j^{*}} \alpha_{lk}\,. \quad (2.36)$$

In addition to the logit link, a log link can be used. The log link function results in a model that can be referred to as the *log* CDM. The log link function can be mathematically expressed as

$$logP(\boldsymbol{\alpha}_{lj}^{*}) = \upsilon_{j0} + \sum_{k=1}^{K_j^{*}} \upsilon_{jk}\alpha_{lk} + \sum_{k'>k}^{K_j^{*}} \sum_{k=1}^{K_j^{*}-1} \upsilon_{jkk'}\alpha_{lk}\alpha_{lk'} + \cdots + \upsilon_{j12\ldots K_j^{*}} \prod_{k=1}^{K_j^{*}} \alpha_{lk}\,. \quad (2.37)$$

Although the three generalized CDM have similar mathematical formulations, the specifications of these models describe unique phenomena. The G-DINA model and logit CDM uses main effects to describe the additive impact of mastery for a set of attributes on the conditional probability and logit of the conditional probability of a correct response, respectively. Whereas, the log CDM uses main effects to define the multiplicative impact of mastery for a set of attributes on the conditional probability of a correct response as a linear model. Note that this can be seen as an advantage for model specification and estimation. For example, the R-RUM, when using the logit link (i.e., the LCDM) requires complex constraints on all interaction terms whereas the log CDM defines the R-RUM as an additive model with only main effects on the log scale. Noting

this difference in how the models are defined is crucial because applying the same constraints to the different link functions will results in different reduced models. A detailed description of how the G-DINA can be used to model various reduced models (e.g., DINA, DINO, R-RUM, and C-RUM) is further discussed in de la Torre (2011).

General Polytomous Diagnosis Model

Chen & de la Torre (2018) proposed a general cognitive diagnostic model for polytomous responses i.e., the *general polytomous diagnosis model* (GPDM), which combines the G-DINA modeling process for dichotomous responses with the item-splitting process for polytomous responses. The GPDM can also be seen as an extension of the P-LCDM in Hansen (2013). The polytomous items are specified similar to dichotomous items in a $J \times K$ $\boldsymbol{Q}$ matrix where the elements of $\boldsymbol{Q}$ are 0/1. For the j^{th} item, irrelevant attributes (i.e., $q = 0$) can be excluded and the measured attributes are represented by the reduced attribute vector $\boldsymbol{\eta}_{jh} = \left(\eta_{j1}, \ldots, \eta_{jg}, \ldots, \eta_{jG_j}\right)^T$, where $h = 1, \ldots, H_j = 2^{G_j}$ and $G_j = \sum_{k=1}^{K} q_{jk}$. Similar to the G-DINA model presented in de la Torre (2011), the attribute vector $\boldsymbol{\alpha}_l$ is simplified as the reduced attribute vector $\boldsymbol{\eta}_{jh}$. This simplification means that the $L = 2^K$ latent classes of the diagnostic test are simplified to H_j latent classes for the j^{th} item. The latent variable modeling process with polytomous responses involves splitting the polytomous item with C_j response categories into C_j dichotomous sub-items, each of which then can be formulated using a modeling approach for dichotomous responses. The GPDM splits the item indirectly based on the difference of cumulative probability between response categories. The conditional probability of

respondents responding to the c^{th} category is $P(X_j = c|\boldsymbol{\eta}_{jh}) = P_c(\boldsymbol{\eta}_{jh})$, where $\sum_{c=0}^{C_j-1} P_c(\boldsymbol{\eta}_{jh}) = 1$. The boundary response probability can be denoted as $P(X_j \geq c|\boldsymbol{\eta}_{jh}) = P_c^*(\boldsymbol{\eta}_{jh})$. Following the graded response approach (Samejima, 1969), the relationship between the conditional and cumulative probabilities can be defined by

$$P_c(\boldsymbol{\eta}_{jh}) = P_c^*(\boldsymbol{\eta}_{jh}) - P_{c+1}^*(\boldsymbol{\eta}_{jh}) \tag{2.38}$$

where $P_0^*(\boldsymbol{\eta}_{jh}) = 1$ and $P_{C_j}^*(\boldsymbol{\eta}_{jh}) = 0$. The monotonicity assumption for the G-DINA discussed in de la Torre (2011) can be extended to the GPDM. Namely, the cumulative probability of responding in a higher category will increase monotonically for respondents who have mastered more required attributes. Using different link functions, $P_c(\boldsymbol{\eta}_{jh})$ can be linearly transformed into item effects in different saturated forms, as

$$F[P_c(\boldsymbol{\eta}_{jh})] = \beta_{jc0} + \sum_{g=1}^{G_j} \beta_{jcg}\eta_{hg} + \sum_{g'>g}^{G_j} \sum_{g=1}^{G_j-1} \beta_{jcgg'}\eta_{hg}\eta_{hg'} + \cdots + \beta_{jc12\ldots G_j} \prod_{k=1}^{G_j} \eta_{hg} \tag{2.39}$$

where $F(.)$ is a specified linking function and β_{jc0}, β_{jcg}, and $\beta_{jcgg'}$ are the baseline, main effect, and interaction effect parameters for the c^{th} category on the j^{th} item, respectively. Similar to de la Torre (2011), various link functions include; identity, logit, and log links. For dichotomous responses, the formulation is equivalent to the G-DINA model with the identity link or LCDM with the logit link functions. The GPDM is a saturated model, with equivalent saturated forms, and subsumes a variety of reduced CDM for polytomous and dichotomous responses. A detailed description of how the

GPDM can be used to model various reduced models (e.g., DINA, DINO, R-RUM, or C-RUM) for polytomous and dichotomous responses is further discussed in Chen & de la Torre (2018).

Sequential Generalized Deterministic Input Noisy "and" Gate Model

Ma & de la Torre (2016) proposed a general polytomous CDM for a special type of graded responses such that item categories are attained in a sequential manner and affiliated with some attributes explicitly. Similar to Samejima (1995), they define the conditional probability of respondents with latent class $\boldsymbol{\alpha}_c$ responding to the h^{th} category for the j^{th} item correctly provided that the respondents have already completed the $h^{th} - 1$ category successfully as the *processing function* of the h^{th} category, defined as $S_j(h|\boldsymbol{\alpha}_c)$, where its assumed that

$$S_j(h|\boldsymbol{\alpha}_c) = \begin{cases} 1, & if\ h = 0 \\ 0, & if\ h = H_j + 1 \end{cases} \tag{2.40}$$

because respondents are expected to always achieve category 0, but never achieve the $H_j^{th} + 1$ category. Respondents are expected to score $h\ iff$ they answered categories $1, \ldots, h$ correctly, and if h is not defined as the highest category, an incorrect response is expected for the $h^{th} + 1$ category. Therefore, the categorical response function for the j^{th} item can be defined by

$$P(X_j = h|\boldsymbol{\alpha}_c) = \left[1 - S_j(h+1|\boldsymbol{\alpha}_c)\right] \prod_{x=0}^{h} S_j(x|\boldsymbol{\alpha}_c) \tag{2.41}$$

which is subject to the imposed constraints

$$\sum_{h=0}^{H_j} P(X_j = h|\boldsymbol{\alpha}_c) = 1 \; \forall c \tag{2.42}$$

where $P(X_{ij} = h|\boldsymbol{\alpha}_c)$ is the probability of respondents in latent class $\boldsymbol{\alpha}_c$ scoring h on the j^{th} item. The processing function defined in Equation (2.40) is the kernel of the sequential process model and can be formulated using most prototypical formulations for CDM. For example, if solving a step entails the possession of at least one required attribute for the j^{th} item, the DINO model can be used as the processing function. By parameterizing each category independently, the sequential process model allows different cognitive processes to be modeled at different categories within an item.

The G-DINA (de la Torre, 2011) model can be used as the processing function because it allows a generalized framework subsuming several commonly used CDM. The resulting model is referred to as the *sequential G-DINA model*. Similar to the prototypical G-DINA model, 2^K latent classes can be collapsed into $2^{K_j^*}$ latent classes with unique conditional probabilities of success. For the h^{th} category, the possible $2^{K_j^*}$ latent classes can be further collapsed into $2^{K_{jh}^*}$ latent classes, where K_{jh}^* is the number of measured attributes for the h^{th} category on the j^{th} item. The processing function $S_j(h|\boldsymbol{\alpha}_c)$ can be rewritten as $S_j(h|\boldsymbol{\alpha}_{ljh}^*)$ for the sequential G-DINA model using the identity link G-DINA model:

$$S_j(h|\boldsymbol{\alpha}_{ljh}^*) = \phi_{jh0} + \sum_{k=1}^{K_{jh}^*} \phi_{jhk}\alpha_{lk} + \sum_{k'>k}^{K_{jh}^*} \sum_{k=1}^{K_{jh}^*-1} \phi_{jhkk'}\alpha_{lk}\alpha_{lk'} + \cdots + \phi_{jh12\ldots K_j^*} \prod_{k=1}^{K_{jh}^*} \alpha_{lk} \tag{2.43}$$

where ϕ_{jh0} represents the intercept parameter for the j^{th} item where $0 \leq \phi_{jh0} < 1$ and ϕ_{jhk} is the main effect parameter due to α_k for the j^{th} item where $0 \leq \phi_{jhk} < 1$. The $\phi_{jhkk'}$ is the interaction effect parameter due to α_k and $\alpha_{k'}$ for the j^{th} item. Finally, while interactions can be positive or negative, they must be defined for the identity link such that $0 \leq S_j\left(h|\boldsymbol{\alpha}_{lj}^*\right) \leq 1$. The ϕ_{jh0} can be represented as the processing function for the h^{th} category $\forall l = 1, \ldots, 2^{K_j^*}$ who mastered none of the required attributes, ϕ_{jhk} represent the change of the processing function of the h^{th} category when the h^{th} attribute has been mastered, and $\phi_{jhkk'}$ and $\phi_{jh12\ldots K_j^*}$ can be represented as the change in the processing function of the h^{th} category when mastery of a combination of attributes are obtained. Similar to the prototypical G-DINA model, the processing function can also be defined using the log or logit link functions.

The sequential G-DINA model can use either a restricted or unrestricted $\boldsymbol{Q}$ matrix denoted as either RS-GDINA or US-GDINA model, respectively. The use of a restricted $\boldsymbol{Q}$ matrix allows for the modeling of different underlying processes in different response categories. In contrast, the unrestricted $\boldsymbol{Q}$ matrix provides a possible solution to account for the uncertainty in the attribute and category relationships. When the attribute and category relationships are present, the RS-GDINA model may be preferred theoretically because it typically estimates fewer item parameters than the US-GDINA model (refer to Ma & de la Torre (2016) for further details).

Multiple-Choice Deterministic Input Noisy "and" Gate Model

Typically, CDM are applied to dichotomous or dichotomized data, including MC assessments that are scored as either right or wrong. The issue with the dichotomization approach for the analysis of MC data is that if often ignores diagnostic information that can be found in the distractors, which could be diagnostically suboptimal. The multiple-choice deterministic input noisy "and" gate (MC-DINA; de la Torre, 2009) model was introduced to maximize the diagnostic information of MC assessments. The framework for MC data is based on the prototypically formulation of the DINA. Using the notation expressed in de la Torre (2009), the conditional probability of a correct response for the i^{th} respondent on the j^{th} item under the DINA is defined by

$$P\left(X_{ij}=1|\boldsymbol{\alpha}_i\right)=P\left(X_{ij}=1|g_{ij}\right)=P_j(1|0)^{1-g_{ij}}\left[1-P_j(0|1)\right]^{g_{ij}} \tag{2.44}$$

where $P_j(1|0)$ and $P_j(0|1)$ are the guessing and slipping parameters, respectively, for the j^{th} item, $g_{ij}=\prod_{k=1}^{K}\alpha_{ik}^{q_{jk}}$ is represented at the latent group classification for the i^{th} respondent on the j^{th} item where $g_{ij}=1$ if the i^{th} respondent has mastered all required attributes for the j^{th} item and else $g_{ij}=0$. The g_{ij} can alternatively be expressed as

$$g_{ij}=\begin{cases}1, & if\ \boldsymbol{\alpha}_i^T\boldsymbol{q}_j=\boldsymbol{q}_j^T\boldsymbol{q}_j\\ 0, & otherwise\end{cases}. \tag{2.45}$$

The MC format for the DINA can be represent by $X_{ij}\in\{1,2,..,H_j\}$ where each element of the set X_{ij} corresponds to a different response option and H_j represents the total

number of coded response options for the j^{th} item. The conditional probability that the i^{th} respondent selects the h^{th} coded response option is defined as

$$P(X_{ij} = h|\boldsymbol{\alpha}_i) = P(X_{ij} = h|g_{ij} = g) = P_j(h|g) \tag{2.46}$$

where $P_j(h|g)$ is defined as the conditional probability of a respondent in group g choosing the h^{th} coded response option of the j^{th} item and $g \in G_j$, which G_j harbors 0 and a subset of the sample space $\{1,2,\dots,H\}$. For a fixed value of g, $\sum_{h=1}^{H} P_j(h|g) = 1$. Therefore, the MC-DINA has a total of $\sum_{j=1}^{J} H(H_j^* + 1)$ parameters, where $\sum_{j=1}^{J} H_j^* + J$ are not free to vary. The g_{ij} in the MC-DINA can be expressed as

$$g_{ij} = arg\max_{h'}\left\{\boldsymbol{\alpha}_i^T \boldsymbol{q}_{jh'} | \boldsymbol{\alpha}_i^T \boldsymbol{q}_{jh'} = \boldsymbol{q}_{jh'}^T \boldsymbol{q}_{jh'}\right\} \tag{2.47}$$

$\forall h' = 0, \dots, H_j$ for the j^{th} item. The $\boldsymbol{Q}$ vector $\boldsymbol{q}_{jh'}$ is represented as the attribute specification for the h^{th} coded response option on the j^{th} item. The $\boldsymbol{Q}$ vector for any noncoded response options is set to 0. Equation (2.47) signifies that the i^{th} respondents' latent class will be classified as $g_{ij} = 0\ iff$ the latent class does not meet the attribute specification of at least one of the coded response options. Recall the DINA separates respondents into two groups indicating those who have mastered and not mastered all required attributes for an item. The MC-DINA separates respondents into $H_j^* + 1$ groups, which represents the number of response options for the j^{th} item plus one. The original 2^K possible latent classes are classified into $H_j^* + 1$ latent groups. Thus, by coding some of the distractors, latent classes that don't satisfy the specification of the key can be

further distinguished from one another, therefore providing additional diagnostic information for the j^{th} item. The conjunctive property of "all or nothing" still holds in the MC-DINA. However, unlike the prototypical formulation of the DINA, this conjunctive property does not always result in a single undifferentiated group for those respondents who lack a required attribute for the correct response on the j^{th} item. To further illustrate the MC-DINA framework, an example MC item with four coded response options is given in Table 4.

Table 4. MC Item with Four Coded Response Options

	Attribute			
Option	α_1	α_2	α_3	α_4
A		✓		
B		✓	✓	
C	✓		✓	
D	✓	✓	✓	

Let $A = 1$, $B = 2$, $C = 3$ and $D = 4$ where $H_j^* = 4$ representing number of coded response options for the j^{th} item. For the i^{th} respondent who possess $(\alpha_1, \alpha_2, \alpha_3)$ for the j^{th} item such that $\boldsymbol{\alpha}_i^T \boldsymbol{q}_{jh'} = 1\ \forall h$. Because q_{j4} is the $\boldsymbol{Q}$ vector with the largest h where $\boldsymbol{\alpha}_i^T \boldsymbol{q}_{jh'} = 1$, will be classified under latent Group 4. In contrast, for the i^{th} respondent who possess (α_2, α_3) (i.e., $\alpha = \{0,1,1,0\}$) of the three required attributes for the j^{th} item where $\boldsymbol{\alpha}_i^T \boldsymbol{q}_{jh'} = 1$ for only $h = (1,3)$, will be classified under latent Group 2. Finally, for the i^{th} respondent who possess none or only α_1 or α_3 of the three required attributes for the j^{th} item where $\boldsymbol{\alpha}_i^T \boldsymbol{q}_{jh'} = 0\ \forall h$, will be classified under latent Group 0.

The modified $\boldsymbol{Q}$ matrix under the MC-DINA can be represented in the following example

$$\boldsymbol{Q}_{[J\times K]} = \begin{pmatrix} 2 & 1 & 0 & 0 \\ 1 & 3 & 1 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 2 & 1 & 0 & 2 \end{pmatrix} \tag{2.48}$$

where a unique coding option is defined where each entry in a cell represents the number of times an attribute is specified in the options. For example, the modified $\boldsymbol{Q}$ vector for the first row $(2,1,0,0)$ indicates that the correct option requires (α_1, α_2) whereas the only coded distractor requires α_1. We can alternatively express $P_j(h|g)$ in the following example

$$P_j(h|g) = \begin{cases} .20, & if\ g = 0 \\ .80, & if\ g > 0\ and\ g = h \\ .10, & if\ g > 0\ and\ g \neq h \end{cases} \tag{2.49}$$

which proports that for a respondent whose latent class does not meet requirements of any of the coded response options, choices are made at random e.g., with equal probability, $\frac{1}{H_j^*}$. However, respondents who meet the requisites of at least one of the coded response options will choose the expected response options 80% of the time and choose the remaining options randomly for the j^{th} item.

General Diagnostic Classification Model for Multiple-Choice

In addition to the MC-DINA, the generalized diagnostic classification models for multiple-choice (GDCM-MC; DiBello, Henson, & Stout, 2015) option-based scoring was

also created for multiple-choice diagnostic assessments with response options designed to attract particular kinds of respondent thinking and understanding, both desired (correct) thinking and problematic (incorrect or partially correct) thinking. There are four key features of the GDCM-MC including: (1) an expanded latent space that can include both desirable and problematic attributes of thinking. (2) an expanded $\boldsymbol{Q}$ matrix that includes a row for each response option and utilizes a three-value coding mechanism to specify which latent classes are strongly attracted to that option. (3) a guessing component that responds to the forced choice aspect of multiple-choice questions. (4) a general modeling framework that can incorporate the diagnostic modeling functionality many dichotomous CDM.

The respondent's latent class $\boldsymbol{\alpha}$ is expanded in the GDCM-MC to include both problematic (including misconceptions and partially correct thinking) and desirable (including skills and conceptual understanding) attributes of thinking. Let $k = 1, \dots, K$ where each k represents either a desirable or problematic attribute. The latent classes can be defined by the set of all $L = 2^K$ vectors where $\boldsymbol{\alpha}_i = (\alpha_1, \dots, \alpha_k, \dots, \alpha_K)$, where $\alpha_{ik} = 0$ represents the i^{th} respondent who lacks the k^{th} attribute, while $\alpha_{ik} = 1$ represents the i^{th} respondent who possesses the k^{th} attribute. Note there is a shift in terminology from "non-master/master of an attribute" to "lacks/possesses an attribute". Whether $\alpha_k = 1$ is advantageous or not for a respondent depends entirely on whether the k^{th} attribute is a desirable or problematic form of thinking.

The $\boldsymbol{Q}$ matrix for the GDCM-MC is expanded in two different ways from the prototypical dichotomous CDM $\boldsymbol{Q}$ matrix formulation where (1) $\boldsymbol{Q}$ now has a row per

response option instead of one row per item as in the dichotomous scoring method. The $\boldsymbol{Q}$ matrix for a diagnostic assessment with 40 items, each with four options, will have a total of $40 \times 4 = 160$ rows. (2) Each $\boldsymbol{Q}$ entry can be any of the three values: 0/1/*N*. The *link vector* for the h^{th} response option of the j^{th} item is the (j, h) row where $\boldsymbol{q}_{jh} = (q_{jh1}, \ldots, q_{jhk}, \ldots, q_{jhK})$ of $\boldsymbol{Q}$ where each of the elements $q_{jhk} = 0/1/N$ and vector $\boldsymbol{q}_{jh}$, specifies that the latent class $\boldsymbol{\alpha}_i = (\alpha_1, \ldots, \alpha_k, \ldots, \alpha_K)$ is *cognitively most strongly attracted* to the h^{th} response option most satisfy the following condition:

$$\forall k = 1, \ldots, K \text{ for which } q_{jhk} \neq N,\ \alpha_{ik} = q_{jhk}.$$

This statement indicates that the latent class $\boldsymbol{\alpha}_i$ for the i^{th} respondent is cognitively most strong attracted to the h^{th} response option if $\boldsymbol{\alpha}_i$ is satisfied for each of the k^{th} attributes such that

- $\alpha_{ik} = 0$, which defines that the i^{th} respondent *lacks* the k^{th} attribute if $q_{jhk} = 0$;
- $\alpha_{ik} = 1$, which defines that the i^{th} respondent *possesses* the k^{th} attribute if $q_{jhk} = 1$;
- the value of α_{ik} for the i^{th} respondent on the k^{th} attribute does not directly influence the strength of attraction towards the h^{th} option if $q_{jhk} = N$.

Note that $q_{jhk} = 0$ in the GDCM-MC has an alternative definition from that of $q_{jk} = 0$ in the prototypical dichotomous DCM. For example, if the k^{th} attribute is a skill, then the condition $q_{jhk} = 0$ for an incorrect response to the h^{th} response option defines that lacking a skill makes it *more likely* that a respondent will select that response option,

while $q_{jhk} = N$ defines that for the h^{th} response option, whether the respondent lacks or possesses the k^{th} attribute does not cognitively, or directly, affect their attraction to the h^{th} response option. Thus, the $q_{jhk} = N$ in the GDCM-MC has the identical definition that $q_{jk} = 0$ has in the prototypical dichotomous DCM: Neither lacking or possessing the k^{th} attribute is irrelevant cognitively for the h^{th} response option. Finally, note that some response options may be defined as *cognitively neutral* such that $\boldsymbol{q}_{jh} = (N_{jh1}, \ldots, N_{jhk}, \ldots, N_{jhK})$, which indicates that the option was not specifically designed to measure any of the attributes.

Assuming the forced choice response imposed by standard multiple-choice question formats, it can be hypothesized that a typical strategy for responding to a particular item can be classified as one of three types: (1) a *cognitive strategy* that uses the latent class $\boldsymbol{\alpha}_i$ for the i^{th} respondent where $\boldsymbol{\alpha}_i$ can be defined as the pattern of problematic and desirable attributes that are lacked or possessed, (2) *guessing strategy* assumes that selecting each of the h^{th} response options has an equal probability of occurring, and (3) a *hybrid strategy*, which is modeled as an initial cognitive step, then assumes random guessing from the remaining h^{th} response options. The probability of selecting the h^{th} response option for the j^{th} item is modeled as a mixture of cognitive and guessing strategies, conditional on the respondent's latent class $\boldsymbol{\alpha}$. Let $C_j = C$ define the use of a cognitive strategy on the j^{th} item and $G_j = \sim C$ define the use of the complementary guessing strategy on the j^{th} item. Thus, the GDCM-MC mixture model for the j^{th} item is mathematically expressed as

$$P_j(h|\boldsymbol{\alpha}) = P_j(h, C|\boldsymbol{\alpha}) + P_j(h, \sim C|\boldsymbol{\alpha}) = P_j(h|C, \boldsymbol{\alpha})P_j(C|\boldsymbol{\alpha}) + P_j(h|G, \boldsymbol{\alpha})P_j(G|\boldsymbol{\alpha}) \quad (2.50)$$

$$= P_j(h|C, \boldsymbol{\alpha})\boldsymbol{\omega}_{j,\boldsymbol{\alpha}} + \frac{1}{H_j}\left(1 - \boldsymbol{\omega}_{j,\boldsymbol{\alpha}}\right) \quad (2.51)$$

$$\forall i = 1, \dots, N$$

where the conditional probability of applying a cognitive strategy to the j^{th} item is defined by $\boldsymbol{\omega}_{j,\boldsymbol{\alpha}} = P_j(C|\boldsymbol{\alpha})$ and $1 - \boldsymbol{\omega}_{j,\boldsymbol{\alpha}} = P_j(G|\boldsymbol{\alpha})$ is the conditional probability of a guessing strategy to the j^{th} item, where guessing is modeled by $P_j(h|G, \boldsymbol{\alpha}) = \frac{1}{H_j}$. Thus, the formulation of the GDCM-MC is defined by specifying two components for the h^{th} response option on the j^{th} item: (1) the *cognitive potions* $P_j(h|C, \boldsymbol{\alpha})$; (2) the *mixing portions* $\boldsymbol{\omega}_{j,\boldsymbol{\alpha}} = P_j(C|\boldsymbol{\alpha})$.

A set of functions must be selected for the cognitive portions $P_j(h|C, \boldsymbol{\alpha})$ of the GDCM-MC. Let $F_j(h|\boldsymbol{\alpha})$ be defined as a function selected for each h^{th} response option on the j^{th} item that guides the desired modeling functionality with respect to the attractiveness of an option given $\boldsymbol{\alpha}$. The conditional probability, $P_j(h|C, \boldsymbol{\alpha})$ is defined as

$$P_j(h|C, \boldsymbol{\alpha}) = \frac{F_j(h|\boldsymbol{\alpha})}{\sum_{h'=1}^{H_j} F_{jh'}(\boldsymbol{\alpha})} = \frac{F_j(h|\boldsymbol{\alpha})}{S_{j,\boldsymbol{\alpha}}}, \quad (2.52)$$

where $S_{j,\boldsymbol{\alpha}} \equiv \sum_{h=1}^{H_j} F_j(h|\boldsymbol{\alpha})$ and $F_j(h|\boldsymbol{\alpha})$ is defined as the *cognitive kernel modeling function* for the h^{th} response option on the j^{th} item with constraint $F_j(h|\boldsymbol{\alpha}) \geq 0$. Any convenient choice of the cognitive kernel functions $F_j(h|\boldsymbol{\alpha})$ can be imposed for the h^{th}

response option on the j^{th} item. However, DiBello et al. (2015) suggest that using any of the traditional dichotomous DCM could be used as a link function defining the attractiveness of a given option. The GDCM-MC mixture model defined in Equation (2.50) is therefore mathematically expressed for the j^{th} item as

$$P_j(h|\boldsymbol{\alpha}) = P_j(h|C,\boldsymbol{\alpha})\boldsymbol{\omega}_{j,\boldsymbol{\alpha}} + \frac{1}{H_j}\left(1-\boldsymbol{\omega}_{j,\boldsymbol{\alpha}}\right) = \frac{F_j(h|\boldsymbol{\alpha})}{S_{j,\boldsymbol{\alpha}}}\boldsymbol{\omega}_{j,\boldsymbol{\alpha}} + \frac{1}{H_j}\left(1-\boldsymbol{\omega}_{j,\boldsymbol{\alpha}}\right) \quad (2.53)$$

with cognitive portion $P_j(h|C,\boldsymbol{\alpha}) = \frac{F_j(h|\boldsymbol{\alpha})}{S_{j,\boldsymbol{\alpha}}}$, guessing portion $P_j(h|G,\boldsymbol{\alpha}) = \frac{1}{H_j}$, and cognitive mixture probabilities $\boldsymbol{\omega}_{j,\boldsymbol{\alpha}}$ and $1-\boldsymbol{\omega}_{j,\boldsymbol{\alpha}}$.

As building blocks for the cognitive portion $P_j(h|C,\boldsymbol{\alpha}) = \frac{F_j(h|\boldsymbol{\alpha})}{S_{j,\boldsymbol{\alpha}}}$ of the GDCM-MC, a $F_j(h|\boldsymbol{\alpha})$ must be selected for each h^{th} response option, to be *monotonically increasing* as a function of latent class $\boldsymbol{\alpha}$. In this case, monotonically increasing states that $F_j(h|\boldsymbol{\alpha})$ must be larger when $\boldsymbol{\alpha} = \boldsymbol{q}_{jh}$ and smaller as more incongruencies occur between $\boldsymbol{\alpha}$ and $q_{jhk} \neq N$. A feature of the GDCM-MC is that any DCM's mastery of required attributes can be converted to matching the model's response option link vector's required lack or possession of a set attributes. This type of modeling approach is referred to as *penalty-for-mismatch heuristic* for function $F_j(h|\boldsymbol{\alpha})$.

Polytomous Attribute Reparametrized Unified Model

The *polytomous attribute reparametrized unified model* (Templin, 2004) incorporates polytomous attributes (or skills) for the RUM model by defining a set of general functions that relates the item response function to the level of the attribute. The

method for defining the item-attribute relationship as a function of the level of the respondent's α_{ik} and the $\boldsymbol{Q}$ matrix element for the j^{th} item, q_{jk}, is denoted as $f_{jk}(\alpha_{ik}, q_{jk})$. The unique parameterization of the RUM allows for a natural transition from a set of dichotomous attributes $\alpha_{ik} \in \{0,1\}$ to a set of polytomous attributes $\alpha_{ik} \in \{0,1, \dots, p\}$. Let $\boldsymbol{\Delta}_j$ represent a collection of item parameters for the RUM where $\boldsymbol{\Delta}_j = (\pi_j^*, r_{jk}^*, \dots, r_{jK}^*)$. The conditional probability of a correct response under the polytomous attribute formulation of the RUM is defined by

$$P(X_{ij} = 1|\boldsymbol{\Delta}_j, \boldsymbol{\alpha}_i, \theta_i) = \pi_j^* \prod_{k=1}^{K} \left[r_{jk}^{*\, f_{jk}(\alpha_{ik}, q_{jk})} \right] P_{c_j}(\theta_i) \tag{2.54}$$

where $P_{c_j}(\theta_i)$ is the item response function under the one-parameter logistic model (1PL) which is equivalent to the notation in the dichotomous RUM (Hartz, 2002):

$$P_{c_j}(\theta_i) = \frac{\exp[D(\theta_i + c_j)]}{1 + \exp[D(\theta_i + c_j)]}. \tag{2.55}$$

The θ_i represent the continuous latent variable for the i^{th} respondent, c_j is the easiness parameter, and D is the scaling constant set to 1.701. The c_j is bounded within the range $0 < c_j < 3$. The purpose for imposing this constraint on c_j is that the values within this boundary would lead to a maximization of the effect of the non-1PL portion of the item response function on the test data. Similar to the RUM (Hartz, 2002), the c_j can be defined as a measure of the completeness of the $\boldsymbol{Q}$ matrix elements for the j^{th} item. A

$c_j \to 0$ indicates the attributes defined in $\boldsymbol{Q}$ matrix does not fully describe the attributes necessary for correctly responding to the j^{th} item. Thus, the item response function is heavily influenced by θ_i. In contrast, a $c_j \to 3$ indicates that the attributes defined in the $\boldsymbol{Q}$ matrix fully describe the necessary skills for correctly responding to the j^{th} item. Thus, the item response function is weakly influenced by θ_i. Note that the $P_{c_j}(\theta_i)$ in Equation (2.54) is dropped when using the R-RUM.

The general nature of $f_{jk}(\alpha_{ik}, q_{jk})$ in the polytomous attribute RUM allows for any function of attributes and $\boldsymbol{Q}$ matrix element to impact the model likelihood. The model assumes that $\forall k = 1, \dots, K$ for α_{ik} has discrete skill levels $\alpha_{ik} \in \{0,1, \dots, p\}$. The constraints placed on the model are defined by

$$\begin{aligned} f_{jk}(\alpha_{ik} = 0, q_{jk} = 1) &= 1 \\ f_{jk}(\alpha_{ik} = p, q_{jk} = 1) &= 0 \\ f_{jk}(\alpha_{ik} = 1, q_{jk} = 1) > f_{jk}(\alpha_{ik} = 2, q_{jk} = 1) > \cdots > f_{jk}(\alpha_{ik} &= p - 1, q_{jk} = 1) \end{aligned} \tag{2.56}$$

where the first two constraints define the upper and lower limits of $f_{jk}(\alpha_{ik}, q_{jk})$ whereas the third constraint, a monotonic decreasing ordering of $f_{jk}(\alpha_{ik}, q_{jk})$, defines the structure of the relationship between the discrete skill levels $\{0,1, \dots, p\}$ and the item response function. Respondents where $\alpha_{ik} = 0$ have the complete imposition of the penalty parameter r_{jk}^* on the conditional probability of a correct response to the j^{th} item, while respondents with $\alpha_{ik} = p$ do not have the imposition of the penalty parameter r_{jk}^* on the conditional probability of a correct response to the j^{th} item. Respondents who

possess $0 < \alpha_{ik} < p$ have a decreased imposition of the penalty parameter r_{jk}^{*} to the conditional probability of a correct response.

Another beneficial property to the parameterization of the polytomous attribute RUM is the ability to impose an ordered polytomous skill structure by a small number of parameters. The $f_{jk}(\alpha_{ik}, q_{jk})$ requires $p - 1$ additional parameters for each $\boldsymbol{Q}$ matrix element. Depending on the number of elements in the $\boldsymbol{Q}$ matrix, the number of $f_{jk}(\alpha_{ik}, q_{jk})$ parameters can increase exponentially. A constraint can be imposed for providing a method of incorporating polytomous skills using a single parameter per skill level:

$$f_{1k}(\alpha_{ik} = p, q_{jk} = 1) = f_{2k}(\alpha_{ik} = p, q_{jk} = 1) = \cdots = f_{jk}(\alpha_{ik} = p, q_{jk} = 1) \quad (2.57)$$
$$\forall p \neq \{0, l\}$$

where $l \in \boldsymbol{\alpha}$.

Defining Polytomous Skill Levels in Deterministic Input Noisy "and" Gate Model

In the initial stages of learning, student begin to accumulate knowledge about a desired skill, then they learn how to apply this skill in simple settings (Karelitz, 2004). As a student begins to progress in the learning process, they learn more complex concepts in relation to the desired skill and eventually learn how to apply this skill in more complex settings. This learning process continues until the student has mastered every aspect of the desired skill. When a $\boldsymbol{Q}$ matrix defines skills as only mastery or non-mastery, the student's progress through these learning phases of the desired skill cannot be traced. A potential solution to circumvent this limitation is to define each skill as a stage in the

learning of a broader skill. However, this can be viewed as an ineffective method for organizing such information. Finer division of skill into subskills is not a plausible way to improve a DCM diagnostic ability and can introduce more parameters in the model, thus making it more complex (DiBello, et al., 1995).

Karelitz (2004) proposed an ordered-category attribute coding (OCAC) framework that uses ordered categories to represent *mastery levels* of each skill. This OCAC framework allows for the number of skills in the design matrix to stay the same, while the values within the matrix take on a larger range. Thus, increasing the diagnostic power of a model without making it overly complex. The number of mastery states increases from two to m_k, where m_k is defined as the number of finite mastery states for the k^{th} skill. The levels within each skill are qualitatively ordered with respect to their cognitive complexity, or the position in the sequence of learning phases. The various levels within each skill represents an increasingly cognitively demands in the acquisition and application of each skill. The OCAC framework also allows the levels to be defined differently for each skill.

Certain diagnostic assessments may require various skills, and test developers may be interested in diagnosing different aspects of each skill. Consequently, different skills may require different number of levels and the definitional structure that defines a transition to each level may be different between skills. However, enforcing the same definitional structure to all levels across a set of skills may improve the interpretability of the diagnostic results. In the OCAC framework, the number of levels for the k^{th} skill is represented as m_k. To illustrate the how the OCAC framework connects items, attributes,

and respondents consider a simple $\boldsymbol{Q}$ matrix for a diagnostic assessment with 2 skills and 4 items. Each skill consists of three levels 0, 1, and 2.

Table 5. Simple $\boldsymbol{Q}$ Matrix

Item	q_1	q_2
1	0	1
2	2	1
3	1	2
4	2	2

The diagnostic assessment in Table 5 consists of five classes, each with a unique latent response pattern as shown in Table 6.

Table 6. Classes and Latent Response Patterns

Class	Mastery Levels	Latent Response
C_1	$\{0,0\},\{1,0\},\{2,0\}$	$(0,0,0,0)$
C_2	$\{0,1\},\{1,1\}\{0,2\}$	$(1,0,0,0)$
C_3	$\{2,1\}$	$(1,1,0,0)$
C_4	$\{1,2\}$	$(1,0,1,0)$
C_5	$\{2,2\}$	$(1,1,1,1)$

The $\boldsymbol{Q}$ matrix and $\boldsymbol{\alpha}$ can be jointly represented on a grid-like formation (refer to Karelitz, 2004 for more detail).

An implication of the OCAC framework relates to the number of latent classes derived in the $\boldsymbol{Q}$ matrix. Prototypical formulations of DCM that utilize binary skills 0/1 have a total of 2^K possible latent classes. Assuming that all the relationships can be represented as prerequisites between levels of a skill (e.g., refer to Karelitz (2004) for

more detail), the OCAC framework defines the number of latent classes as $\prod_{k=1}^{K} m_k$. For example, assume the number of skills measured in an OCAC model is $K = 3$ and that the number of levels for each skill is $m_k = 3\ \forall k$. The equivalent binary model will need $K = 6$ skills, resulting in $2^6 = 64$ latent classes, where the OCAC model only results in $\prod_{k=1}^{K} m_k = 27$ latent classes. The 64 latent classes in the binary skills model are mostly redundant because there are many prerequisite relations between skills. However, these relationships need to be defined independently from the $\boldsymbol{Q}$ matrix, and the process of class reduction must be implemented in the estimation algorithm. The DINA model can be extended within the OCAC framework (Karelitz, 2004). Let $\boldsymbol{\Delta}_j = (s_j, g_j)$ be a collection of parameters for the j^{th} item. The conditional probability of a correct response for the i^{th} respondent on the j^{th} item under the PS-DINA is defined by

$$P(X_{ij} = 1|\boldsymbol{\Delta}_j, \xi_{ij}^*) = (1 - s_j)^{\xi_{ij}^*} g^{\left(1-\xi_{ij}^*\right)} \tag{2.58}$$

where the latent variable $\xi_{ij} \in \{0,1\}$ and is expressed as

$$\xi_{ij}^* = \prod_{k=1}^{K} I[\alpha_{ik} \geq q_{jk}]. \tag{2.59}$$

where $\alpha_{ik} \in \{0,1,2, \dots, m_k\}$ and the indicator function $I = 1$, when the i^{th} respondent's mastery level is at least as high as the item requires, and $I = 0$ otherwise.

The Generalized Deterministic Input Noisy "and" Gate Model for Polytomous Attributes

Defining polytomous attributes in the test developmental process can provide additional diagnostic information that may not be available when using binary 0/1 attribute levels. Specifically, binary attribute levels may not be quite enough to account for the relationship between items and so more attribute levels are needed but not a fully continuous latent variable like in MIRT. Chen & de la Torre (2013) proposed a polytomous generalization of the G-DINA (i.e., pG-DINA) model to accommodate polytomous attribute levels. Using the notation from de la Torre (2011) & Chen & de la Torre (2013), the pG-DINA involves defining the number of required attributes for the j^{th} item $K_j^* = \sum_{k=1}^{K} I\left(q_{jk} > 0\right)$ where $I(.)$ is the indicator function and q_{jk} can take of values $0, \dots, m, \dots, M_k - 1$. The M_k is the number of levels for the k^{th} attribute. If the k^{th} attribute does not require any level of mastery for the j^{th} item, then $q_{jk} = 0$. The required attributes for the j^{th} item can be denoted by the reduced attribute vector $\boldsymbol{\alpha}_{lj}^* = \left(\alpha_{l1}, \dots, \alpha_{lk}, \dots, \alpha_{lK_j^*}\right)$ where $l = 1,2, \dots, M^{K_j^*}$ and $M^{K_j^*}$ denotes the number of unique latent classes. An item with M level polytomous attributes will divide respondents into M groups, with each group potentially having its own conditional probability of a correct response. The issue with handling such generality would require a model with many parameters, and the complexity of the model groups exponentially as the number of attribute levels increases. To make the pG-DINA formulations for polytomous attributes more manageable, the items are assumed to distinguish between two latent groups: respondents who are *on* or *above* a specific attribute level or *below* this mastery level.

Items with this assumption are referred to as *specific attribute level mastery* (SALM) items. Using SALM items, the substantive definition of each attribute level can be incorporated into the modeling procedure through a modified $\boldsymbol{Q}$ matrix, as done in the ordered-category attribute coding framework (Karelitz, 2004). Specifying $q_{jk} = m$ for the k^{th} attribute on the j^{th} item, the M-level α_{lk} can collapsed into a dichotomous attribute α_{lk}^{**} defined by

$$\alpha_{lk}^{**} = \begin{cases} 0, & if\ \alpha_{lk} < q_{jk} \\ 1, & otherwise \end{cases} \tag{2.60}$$

where $\boldsymbol{\alpha}_{lj}^{**} = \left(\alpha_{l1}^{**}, \dots, \alpha_{lk}^{**}, \dots, \alpha_{lK_j^*}^{**}\right)$ is denoted as the *collapsed attribute vector*, where $l = 1, \dots, 2^{K_j^*}$. Let's assume an item measures four attributes $K =$ with each attribute consisting of three levels $M = 3$ and $\boldsymbol{Q}$ vector defined as $(0,1,2,1)$ where $K_j^* = 3$. The original attribute vector $\boldsymbol{\alpha}_{lj}$ with have a total of $M^K = 3^4 = 81$ unique classes, and the reduce attribute vector $\boldsymbol{\alpha}_{lj}^{**}$ will have a total of $M^{K_j^*} = 2^3 = 8$ unique classes. The conditional probability of a correct response given a collapsed attribute vector $\boldsymbol{\alpha}_{lk}^{**}$ for the j^{th} item is defined as

$$P\left(X_{ij}|\boldsymbol{\alpha}_{lj}^{**}\right) = \delta_{j0} + \sum_{k=1}^{K_j^*} \delta_{jk}\alpha_{lk}^{**} + \sum_{k'>k}^{K_j^*} \sum_{k=1}^{K_j^*-1} \delta_{jkk'}\alpha_{lk}^{**}\alpha_{lk'}^{**} + \cdots + \delta_{j12\dots K_j^*} \prod_{k=1}^{K_j^*} \alpha_{lk}^{**} \tag{2.61}$$

where model parameters δ_{j0}, δ_{jk}, $\delta_{jkk'}$, and $\delta_{j12\dots K_j^*}$ retain the same interpretation as in the G-DINA model.

Polytomous Log-Linear Cognitive Diagnosis Model for Polytomous Attributes

The P-LCDM-PA is an extension of the P-LCDM (Hansen, 2013) and a special case of the GPDM (Chen & de la Torre, 2018) using polytomous attributes. Let $\boldsymbol{\omega}_j$ be a collection of effect (main and interaction) and intercept parameters $\boldsymbol{\omega}_j = \left(\boldsymbol{\lambda}_j, \boldsymbol{\lambda}_{j,0}\right)$ for the j^{th} item such that $\boldsymbol{\lambda}_j = \left(\lambda_{j,1,(k)}, \dots, \lambda_{j,2,(k,k')}, \dots, \lambda_{j,K_j,(1,\dots,K_j)}\right)$ and $\boldsymbol{\lambda}_{j,1} = \left(\lambda_{j1,0}, \dots, \lambda_{jc,0}, \dots, \lambda_{j(C_j-1),0}\right)$. Following the approach of Samejima's (1969) graded response IRT model and subsequent multidimensional extensions (e.g., Muraki & Carlson, 1995; Gibbons et al., 2007), the conditional probability of a response to the c^{th} category such that $c \in \{0,1, \dots, C_j - 1\}$ under the P-LCDM-PA for the i^{th} respondent on the j^{th} item is defined by

$$P\left(X_{ij} = c \middle| \boldsymbol{\omega}_j, \boldsymbol{\alpha}_i\right) = P\left(X_{ij} \geq c \middle| \boldsymbol{\omega}_j, \boldsymbol{\alpha}_i\right) - P\left(X_{ij} \geq c+1 \middle| \boldsymbol{\omega}_j, \boldsymbol{\alpha}_i\right) \tag{2.62}$$

where the set of boundary response probabilities are defined as

$$P\left(X_{ij} \geq 0 \middle| \boldsymbol{\omega}_j, \boldsymbol{\alpha}_i\right) = 1$$

$$P\left(X_{ij} \geq 1 \middle| \boldsymbol{\omega}_j, \boldsymbol{\alpha}_i\right) = \frac{1}{1 + Exp\left[-\lambda_{j1,0} - \boldsymbol{\lambda}_j^T \boldsymbol{\alpha}_{ij}^{**}\right]}$$

...

$$P\left(X_{ij} \geq c \middle| \boldsymbol{\omega}_j, \boldsymbol{\alpha}_i\right) = \frac{1}{1 + Exp\left[-\lambda_{jc,0} - \boldsymbol{\lambda}_j^T \boldsymbol{\alpha}_{ij}^{**}\right]} \tag{2.63}$$

...

$$P\left(X_{ij} \geq C_j - 1 \middle| \boldsymbol{\omega}_j, \boldsymbol{\alpha}_i\right) = \frac{1}{1 + Exp\left[-\lambda_{j(C_j-1),0} - \boldsymbol{\lambda}_j^T \boldsymbol{\alpha}_{ij}^{**}\right]},$$

$$P\left(X_{ij} \geq C_j \middle| \boldsymbol{\omega}_j, \boldsymbol{\alpha}_i\right) = 0.$$

It's assumed that all intercept parameters $-\infty < \lambda_{j1,0}, \ldots, \lambda_{jc,0}, \ldots, \lambda_{j(C_j-1),0} < \infty$, main-effect parameters $0 \leq \lambda_{j,1,(1)}, \ldots, \lambda_{j,1,(k)}, \ldots, \lambda_{j,1,(K_j)} < \infty$, and interaction-effect parameters can be defined as either positive or negative depending on the particular submodel constraints place on the P-LCDM-PA. The $\boldsymbol{\alpha}_{ij}^{**} = \left(\alpha_{ij,1,(k)}^{**}, \ldots, \alpha_{ij,2,(k,k')}^{**}, \ldots, \alpha_{ij,K_i,(1,\ldots,K_i)}^{**}\right)$ represents a vector of indicator variables for the i^{th} respondent on the j^{th} item with respect to either dichotomous or polytomous attributes. Furthermore, note that this vector allows for both main effects and interaction terms in a similar method as the P-LCDM. Specifically, the main effect terms of $\boldsymbol{\vartheta}_{ij}$ are defined as

$$\alpha_{ij,1,(k)}^{**} = \begin{cases} 1, & if\ \alpha_{ik} \geq q_{jk} \\ 0, & else \end{cases} \tag{2.64}$$

$$\forall q_{j1}, \ldots, q_{jk}, \ldots, q_{jK_j} \neq 0$$

Again, note that α_{ik} can be defined as binary $\alpha_{ik} \in \{0,1\}$ or polytomous $\alpha_{ik} \in \{0, 1, 2, \ldots, S_k\}$. attribute levels. With respect to the general notation used for $\boldsymbol{\alpha}_{ij}^{**}$, the subscript following the first comma in α_{ij}^{**} and λ_j represent the effect-level and the parentheses following the second comma include the attribute effect. The remaining values in $\boldsymbol{\alpha}_{ij}^{**}$ are defined as a function of the indicator variables, $\alpha_{ij,1,(k)}^{**}$. Specifically,

once the main effect indicator $\alpha^{**}_{ij,1,(k)}$ is obtained $\forall k = 1, \dots, K_i$, the interaction indicators $\alpha^{**}_{ij,2,(k,k')}, \dots, \alpha^{**}_{ij,K_i,(1,\dots,K_i)}$ are obtained by computing the product of the main effects indicated by the subscript. As a result, $\alpha^{**} = 1$ if the respondent has "mastered" a single attribute or in the case of an interaction the examinee has "mastered" all attributes involved in that interaction. Note that mastery is now used generally such that when the attribute is a polytomous attribute then mastery indicates that the level of mastery is at or above that of what is required by the item. If $\alpha^{**} = 0$, the respondent has not mastered (i.e., at or above the ***Q*** matrix specified level) a single attribute or has not mastered at least one attribute in the case of an interaction involving multiple attributes. Note that K_j represents the number of required attributes for the j^{th} item, where $K_j = \sum_{k=1}^{K} I\left[q_{jk} > 0\right]$ and K_i represents the number attributes for the i^{th} respondent, where $K_i = \sum_{k=1}^{K} I[\alpha_{ik} > 0]$. Table 7 demonstrates how α^{**} is obtained as a function of α and q, where $q = (1,2,1)$:

Table 7. Functional Relationship between α^{**} and α using $q = (1,2,1)$

α	$\alpha^{**}_{1,(1)}$	$\alpha^{**}_{1,(2)}$	$\alpha^{**}_{1,(3)}$	$\alpha^{**}_{2,(1,2)}$	$\alpha^{**}_{2,(1,3)}$	$\alpha^{**}_{2,(2,3)}$	$\alpha^{**}_{3,(1,2,3)}$
(0,0,0)	0	0	0	0	0	0	0
(1,0,0)	1	0	0	0	0	0	0
(1,0,1)	1	0	1	0	1	0	0
(0,2,1)	0	1	1	0	0	1	0
(1,1,2)	1	0	1	0	1	0	0
(0,2,0)	0	1	0	0	0	0	0
(1,2,1)	1	1	1	1	1	1	1
(2,2,2)	1	1	1	1	1	1	1

Given the set of weights, $\boldsymbol{\lambda}_j$, and the set of indictors for main effects and interactions "mastery", $\boldsymbol{\alpha}_{ij}^{**}$, the linear combination of $\boldsymbol{\lambda}_j$ and $\boldsymbol{\alpha}_{ij}^{**}$ can be represented as

$$\boldsymbol{\lambda}_j^T \boldsymbol{\alpha}_{ij}^{**} = \lambda_{j,1,(k)}\alpha_{ij,1,(k)}^{**} + \cdots + \lambda_{j,2,(k,k')}\alpha_{ij,2,(k,k')}^{**} + \ldots + \lambda_{j,K_j,(1,\ldots,K_j)}\alpha_{ij,K_i,(1,\ldots,K_i)}^{**} \quad (2.65)$$

Because the P-LCDM-PA is an extension of the P-LCDM, which is a general model, a property of the P-LCDM-PA is that there exists a mathematical relationship such that a set of constraints placed on the P-LCDM-PA can correspond to the natural definition of noncompensatory and compensatory models for polytomous responses. Furthermore, it would be possible to first fit the unconstrained P-LCDM-PA as a method to investigate the nature of the relationships between attribute mastery (i.e., defined as greater that the $\boldsymbol{Q}$ matrix defined cutoff) and the conditional probability of a response at an item-by-item basis. Depending on the estimates, the unconstrained P-LCDM-PA may then be used to suggest specific reduced models that align with the particular patterns of the original model estimates. Specifically, it is possible to place constraints on these parameters such that the P-LCDM-PA reduces to many of the models familiar in the literature (e.g., DINA, DINO, or C-RUM). Another property of the P-LCDM-PA is that the model can be used to define polytomous graded response models that are natural extensions to the DINA, DINO, and C-RUM for polytomous attributes, which is discussed in Chapter III.

Parameter Estimation Algorithms

Background

Estimation of CDM is a very critical step in research and practice, like any statistical model, when it comes to making inferences about our population of interest. Popular estimation algorithms used to estimate the CDM discussed in the previous sections is the expectation-maximization (EM; Bock & Aitkin, 1981) algorithm and Markov chain Monte Carlo (Junker, Patz & VanHoudnos, 2016) algorithms. The stochastic expectation-maximization (SEM; Diebolt & Ip, 1994a, 1994b) and Metropolis-Hastings Robbins-Monro (MH-RM; Cai, 2010a, 2010b) are two estimation algorithms that have been used to estimate other multidimensional models (e.g., refer to Diebolt & Ip, 1994a, 1994b; Cai, 2010a, 2010b; Monroe & Cai, 2014) but have never been implemented for estimating CDM. The following sections will provide a detailed discussion of the EM algorithm, MCMC algorithm, SEM algorithm, and MH-RM algorithm in the context of estimating CDM.

Expectation-Maximization Algorithm

The EM algorithm is an iterative procedure for finding maximum likelihood estimates of parameters associated with probabilistic models in the presence of unobserved latent variables (Baker & Kim, 2004). The algorithm consists of two steps; the *expectation step* and the *maximization step*. The expectation step in the algorithm involves replacing the unknown latent variables of a model with their corresponding expected values, given that the item parameters have been estimated in the previous iteration of the algorithm (Rupp, et al., 2010). Because the latent variables are classes in

CDM, this step in the algorithm utilizes the probability that any given respondent has the potential of being classified within each latent class given the responses.

Treating response patterns as fixed once observed, the observed-data likelihood function is denoted as $L(\boldsymbol{\omega}|\boldsymbol{X})$ (Cai, 2010a, 2010b; Monroe & Cai, 2014). In the maximization step, instead of maximizing $L(\boldsymbol{\omega}|\boldsymbol{X})$ directly, the observed-data estimation problem can be transformed into a sequence of complete-data estimation problems by iteratively maximizing the conditional expectation of $L(\boldsymbol{\omega}|\boldsymbol{Y})$ over $f(\boldsymbol{\alpha}|\boldsymbol{\omega},\boldsymbol{X})$, where $L(\boldsymbol{\omega}|\boldsymbol{Y})$ is the complete-data likelihood, and $f(\boldsymbol{\alpha}|\boldsymbol{\omega},\boldsymbol{X})$ denotes the posterior predictive distribution of the missing data $\boldsymbol{\alpha}$ given observed-data $\boldsymbol{X}$ and parameters $\boldsymbol{\omega}$ (Dempster, Laird, & Rubin, 1977). In $(\boldsymbol{\omega}|\boldsymbol{Y})$, the variable $\boldsymbol{Y} = (\boldsymbol{X},\boldsymbol{\alpha})$ is denoted as the complete-data. Convergance results (Wu, 1983) show that successive EM iterations will result in a (local) maximizer of $L(\boldsymbol{\omega}|\boldsymbol{X})$. The iterative approach to the algorithm is summarized as follows:

1) Initialize parameters $\boldsymbol{\omega}^{(t-1)} = \boldsymbol{\omega}^{*}$.

2) *Expectation step*: Evaluate the posterior predictive distribution of the latent variable $f\left(\boldsymbol{\alpha}|\boldsymbol{\omega}^{(t-1)},\boldsymbol{X}\right)$ using old parameters $\boldsymbol{\omega}^{(t-1)}$. Then the expected complete-data log-likelihood, under this distribution is defined as

$$Q\left(\boldsymbol{\omega}^{(t)}|\boldsymbol{\omega}^{(t-1)}\right) = \sum_{\forall \boldsymbol{\alpha}} logL\left(\boldsymbol{\omega}^{(t)}|\boldsymbol{Y}\right) f\left(\boldsymbol{\alpha}|\boldsymbol{\omega}^{(t-1)},\boldsymbol{X}\right). \tag{2.66}$$

3) *Maximization step*: Update parameters $\boldsymbol{\omega}$ to maximize the expected complete-data log-likelihood

$$\boldsymbol{\omega}^{(t)} = arg\max_{\boldsymbol{\omega}} Q\left(\boldsymbol{\omega}^{(t)}|\boldsymbol{\omega}^{(t-1)}\right). \tag{2.67}$$

4) Check convergence criterion ϵ such that $\left\|\boldsymbol{\omega}^{(t)} - \boldsymbol{\omega}^{(t-1)}\right\| < \epsilon$, else return to Step 2 if condition not satisfied.

Computing $Q\left(\boldsymbol{\omega}^{(t)}|\boldsymbol{\omega}^{(t-1)}\right)$ is equivalent to finding the following two quantities

$$n_l^{(t-1)} = \sum_{i=1}^{N} f\left(\boldsymbol{\alpha}_l|\boldsymbol{\omega}^{(t-1)}, \boldsymbol{X}_i\right) = \sum_{i=1}^{N} \frac{L\left(\boldsymbol{X}_i|\boldsymbol{\omega}^{(t-1)}, \boldsymbol{\alpha}_l\right) f^{(t-1)}(\boldsymbol{\alpha}_l)}{\sum_{\forall \boldsymbol{\alpha}} L\left(\boldsymbol{X}_i|\boldsymbol{\omega}^{(t-1)}, \boldsymbol{\alpha}\right) f^{(t-1)}(\boldsymbol{\alpha})} \tag{2.68}$$

$$r_{ljc}^{(t-1)} = \sum_{i=1}^{N} I[X_{ij} = c] f\left(\boldsymbol{\alpha}_l|\boldsymbol{\omega}^{(t-1)}, \boldsymbol{X}_i\right) = \sum_{i=1}^{N} I[X_{ij} = c] \frac{L\left(\boldsymbol{X}_i|\boldsymbol{\omega}^{(t-1)}, \boldsymbol{\alpha}_l\right) f^{(t-1)}(\boldsymbol{\alpha}_l)}{\sum_{\forall \boldsymbol{\alpha}} L\left(\boldsymbol{X}_i|\boldsymbol{\omega}^{(t-1)}, \boldsymbol{\alpha}\right) f^{(t-1)}(\boldsymbol{\alpha})} \tag{2.69}$$

$$f^{(t)}(\boldsymbol{\alpha}_l) = \frac{n_l^{(t-1)}}{N} \tag{2.70}$$

where n_l represents the expected frequency of respondents belonging to the l^{th} latent class, r_{ljc} represents the expected frequency of respondents belonging to the l^{th} latent class who responded to the c^{th} category for the j^{th} item, and $f^{(t)}(\boldsymbol{\alpha}_l)$ is the prior distribution for $\forall \boldsymbol{\alpha}$. The indicator function $I[X_{ij} = c]$ evaluates to 1 when $X_{ij} = c$, otherwise 0. The maximization of $Q\left(\boldsymbol{\omega}^{(t)}|\boldsymbol{\omega}^{(t-1)}\right)$ is equivalent to first finding the following quantities once $n_l^{(t-1)}$ and $r_{ljc}^{(t-1)}$ are obtained

$$g\left(\boldsymbol{\omega}_j^{(t)}|\boldsymbol{\omega}^{(t-1)}\right) = \nabla_{\boldsymbol{\omega}} Q\left(\boldsymbol{\omega}^{(t)}|\boldsymbol{\omega}^{(t-1)}\right) = \frac{\partial Q\left(\boldsymbol{\omega}^{(t)}|\boldsymbol{\omega}^{(t-1)}\right)}{\partial \boldsymbol{\omega}_j} \tag{2.71}$$

$$H\left(\boldsymbol{\omega}_j^{(t)}|\boldsymbol{\omega}^{(t-1)}\right) = \nabla_{\omega}^{2} Q\left(\boldsymbol{\omega}^{(t)}|\boldsymbol{\omega}^{(t-1)}\right) = \frac{\partial^2 log Q\left(\boldsymbol{\omega}^{(t)}|\boldsymbol{\omega}^{(t-1)}\right)}{\partial \boldsymbol{\omega}_j \partial \boldsymbol{\omega}_j'} \tag{2.72}$$

where $g\left(\boldsymbol{\omega}_j^{(t)}|\boldsymbol{\omega}^{(t-1)}\right)$ is a gradient vector of $(2^{K_j} - 1) + \left(C_j - 1\right) \times 1$ dimensions for the complete-data log-likelihood and $H\left(\boldsymbol{\omega}_j^{(t)}|\boldsymbol{\omega}^{(t-1)}\right)$ be the complete-data information matrix of $(2^{K_j} - 1) + \left(C_j - 1\right) \times (2^{K_j} - 1) + \left(C_j - 1\right)$ dimensions of for the j^{th} item. Once $g\left(\boldsymbol{\omega}_j^{(t)}|\boldsymbol{\omega}^{(t-1)}\right)$ and $H\left(\boldsymbol{\omega}_j^{(t)}|\boldsymbol{\omega}^{(t-1)}\right)$ have been solved, an iterative *Newton-Raphson* optimization method with a predetermined convergence criterion ϵ can be implemented to obtain a new set of parameters for the t^{th} iteration in the algorithm,

$$\boldsymbol{\omega}_j^{(t)} = \boldsymbol{\omega}_j^{(t-1)} + H\left(\boldsymbol{\omega}_j^{(t)}|\boldsymbol{\omega}^{(t-1)}\right)^{-1} g\left(\boldsymbol{\omega}_j^{(t)}|\boldsymbol{\omega}^{(t-1)}\right) \tag{2.73}$$

where $H\left(\boldsymbol{\omega}_j^{(t)}|\boldsymbol{\omega}^{(t-1)}\right)^{-1}$ is the inverse of the complete-data information matrix for the j^{th} item. It's important to note that a Newton-Raphson optimization method is not always needed in maximum likelihood estimation. There are cases when the maximum value of a parameter can be determined in closed form (i.e., where the first derivative is zero and second derivative is negative). A fundamental property of the algorithm assumes that the updated guess $\boldsymbol{\omega}^{(t)}$ will never be less likely that the previous guess $\boldsymbol{\omega}^{(t-1)}$ (monotonicity). However, there is no guarantee that the sequence of $\{\boldsymbol{\omega}^{(0)}, \boldsymbol{\omega}^{(1)}, \ldots\}$ coverages to a global maximum.

Markov Chain Monte Carlo Algorithm

Maximum likelihood estimation, described in the previous section, involves the determination of a set of parameter estimates that maximize the value of the likelihood function. Bayesian estimation focuses on determining a set of parameter values that maximizes the joint posterior distribution of all parameters (Rupp, et al., 2010). However, from a numerical standpoint, directly maximizing the joint posterior distribution of all parameters be very difficult. MCMC algorithms have been proposed to circumvent this problem by sampling from the posterior distribution, rather than maximizing it. MCMC describes a family of algorithms for simulating data (i.e., *Monte Carlo* simulation) using a statistical sequence of random draws that is known as a *Markov chain* $\{\omega^{(t)}\}$. By imposing these steps in a way, it is possible to simulate values that are from a specific distribution, referred to as the *stationary distribution*, which is the distribution to which the $\{\omega^{(t)}\}$ converges as $t \to \infty$ (Patz & Junker, 1999; Junker, Patz & VanHoudnos, 2016). The essential idea is to define a stationary Markov chain $\mathrm{M}_0, \mathrm{M}_1, \mathrm{M}_2$ ... with states $\mathrm{M}_t = \{\omega^{(t)}\}$ and transition kernel

$$\kappa\left(\omega^{(1)}|\omega^{(0)}\right) = P\left[\mathrm{M}_t = \{\omega^{(1)}\}|\mathrm{M}_t = \{\omega^{(0)}\}\right], \forall t \tag{2.74}$$

where the probability of moving to a new state $\omega^{(1)}$ given the current state $\omega^{(0)}$ with stationary distribution $\Psi(\omega)$ assuming ω is continuous is defined by

$$\Psi(\omega^{(1)}) = \int_{\omega} \kappa\left(\omega^{(1)}|\omega^{(0)}\right)\Psi(\omega^{(0)})d\omega^{(0)}\,. \tag{2.75}$$

Once the transition kernel $\kappa\left(\omega^{(1)}|\omega^{(0)}\right)$ is defined such that $\Psi(\omega) = f(\omega|\boldsymbol{\alpha}, \boldsymbol{X})$, then removing the first m_0 observations – the "burn-in" period before the distribution of M_t has converged to stationary distribution $f(\omega|\boldsymbol{\alpha}, \boldsymbol{X})$ – the "retained" observations

$$\left\{\omega^{(1)}\right\} = \mathrm{M}_{m_0+1}, \left\{\omega^{(2)}\right\} = \mathrm{M}_{m_0+2}, \ldots, \left\{\omega^{(M)}\right\} = \mathrm{M}_{m_0+M} \tag{2.76}$$

can be treated like dependent draws from $f(\omega|\boldsymbol{\alpha}, \boldsymbol{X})$. Given the "retained" observations, an *expected a-posteriori* (EAP) estimate of an integrable function $f(\omega)$ can be obtained simply by

$$E[f(\omega|\boldsymbol{X})] = \int_{\omega} f(\omega|\boldsymbol{\alpha}, \boldsymbol{X}) f(\omega) d\omega \tag{2.77}$$

with convergence as $t \to \infty$. Note that the transition kernel $\kappa\left(\omega^{(1)}|\omega^{(0)}\right)$ can be constructed such that the stationary distribution of the Markov chain is that of the posterior distribution $f(\omega|\boldsymbol{\alpha}, \boldsymbol{X})$. For example, let (ω_1, ω_2) be a disjoint partition block the parameter vector $\boldsymbol{\omega}$ into two blocks of parameters. A short calculation verifying Equation (2.75) shows that

$$\kappa\left(\omega^{(1)}|\omega^{(0)}\right) = f\left(\omega_1^{(1)}|\omega_2^{(0)}, \boldsymbol{\alpha}, \boldsymbol{X}\right) f\left(\omega_2^{(1)}|\omega_1^{(1)}, \boldsymbol{\alpha}, \boldsymbol{X}\right) \tag{2.78}$$

has a stationary distribution $f(\omega|\boldsymbol{\alpha}, \boldsymbol{X})$.

An MCMC algorithm is simplest to implement when the complete conditionals can be written in closed form and can be sampled from directly. In this scenario, the MCMC algorithm is referred to as a *Gibbs sampler* (Patz & Junker, 1999; Junker, Patz &

VanHoudnos, 2016). Assume that $(\omega_1, \omega_2, \dots, \omega_J)$ is a fixed disjoint partition of the parameter vector $\boldsymbol{\omega}$. A Gibbs sampling procedure is defined to move from $\mathrm{M}_{t-1} = \left(\omega_1^{(t-1)}, \omega_2^{(t-1)}, \dots, \omega_J^{(t-1)}\right)$ to $\mathrm{M}_t = \left(\omega_1^{(t)}, \omega_2^{(t)}, \dots, \omega_J^{(t)}\right)$ in the Markov chain

$$
\begin{aligned}
&Sample\ \omega_1^{(t)} \sim f\left(\omega_1 | \omega_2^{(t-1)}, \dots, \omega_J^{(t-1)}, \boldsymbol{\alpha}, \boldsymbol{X}\right) \\
&Sample\ \omega_2^{(t)} \sim f\left(\omega_2 | \omega_1^{(t)}, \omega_3^{(t-1)}, \dots, \omega_J^{(t-1)}, \boldsymbol{\alpha}, \boldsymbol{X}\right) \\
&\dots \\
&Sample\ \omega_j^{(t)} \sim f\left(\omega_j | \omega_1^{(t)}, \dots, \omega_{j-1}^{(t)}, \omega_{j+1}^{(t-1)}, \dots, \omega_J^{(t-1)}, \boldsymbol{\alpha}, \boldsymbol{X}\right) \\
&\dots \\
&Sample\ \omega_J^{(t)} \sim f\left(\omega_J | \omega_1^{(t)}, \dots, \omega_{J-1}^{(t)}, \boldsymbol{\alpha}, \boldsymbol{X}\right)
\end{aligned}
\tag{2.79}
$$

where $f(.)$ is defined as the *full conditional densities* for $\omega_1^{(t)}, \omega_2^{(t)}, \dots, \omega_J^{(t)}$ because the distribution of each partition element ω_j is expressed conditional on all other parameters, latent variables $\boldsymbol{\alpha}$ and data $\boldsymbol{X}$ in the model. For notational clarity, the full conditional densities are defined as $f\left(\omega_j | \omega_{-j}\right)$ where ω_{-j} represents all other parameters except ω_j. An extension of the calculations in Equation (2.78) can show that the kernel density $\kappa\left(\omega^{(t)} | \omega^{(t-1)}\right)$ has $f(\omega | \boldsymbol{\alpha}, \boldsymbol{X})$ as its stationary distribution such that the kernel density consists of the product of the complete conditional densities

$$
\kappa\left(\omega^{(t)} | \omega^{(t-1)}\right) = f\left(\omega_1^{(t)} | \omega_{-j}, \boldsymbol{\alpha}, \boldsymbol{X}\right) \times f\left(\omega_2^{(t)} | \omega_{-j}, \boldsymbol{\alpha}, \boldsymbol{X}\right) \times \dots \times f\left(\omega_J^{(t)} | \omega_{-j}, \boldsymbol{\alpha}, \boldsymbol{X}\right). \tag{2.80}
$$

It's important to note that each of the full conditional densities are proportional to the joint density as a function of its block of parameters e.g.,

$$f\left(\omega_1|\omega_2^{(t-1)},\dots,\omega_J^{(t-1)},\boldsymbol{\alpha},\boldsymbol{X}\right)=\frac{L\left(\boldsymbol{X}|\omega_1^{(t-1)},\omega_2^{(t-1)},\dots,\omega_J^{(t-1)},\boldsymbol{\alpha}\right)f\left(\omega_1^{(t-1)},\omega_2^{(t-1)},\dots,\omega_J^{(t-1)}\right)}{\int_{\omega_1}L\left(\boldsymbol{X}|\omega_1^{(t-1)},\omega_2^{(t-1)},\dots,\omega_J^{(t-1)},\boldsymbol{\alpha}\right)f\left(\omega_1^{(t-1)},\omega_2^{(t-1)},\dots,\omega_J^{(t-1)}\right)d\omega_1}$$
$$\propto f\left(\boldsymbol{X}|\omega_1^{(t-1)},\omega_2^{(t-1)},\dots,\omega_J^{(t-1)},\boldsymbol{\alpha}\right)f\left(\omega_1^{(t-1)},\omega_2^{(t-1)},\dots,\omega_J^{(t-1)}\right) \quad (2.81)$$

as a function of ω_1**,** holding all other blocks $\left(\omega_2,\omega_3,\dots,\omega_J\right)$, latent variables $\boldsymbol{\alpha}$, and data $\boldsymbol{X}$ fixed. Thus, when the likelihood $L\left(\boldsymbol{X}|\omega_1^{(t-1)},\omega_2^{(t-1)},\dots,\omega_J^{(t-1)},\boldsymbol{\alpha}\right)$ and prior $f\left(\omega_1^{(t-1)},\omega_2^{(t-1)},\dots,\omega_J^{(t-1)}\right)$ factor into a product of terms involving separate blocks of the partition it is easy to "pick out" a function proportional to the complete conditional, by simply retaining those terms in the joint density that depend on ω_1.

The full conditional densities are typically difficult to directly sample from, but they can be specified up to a proportionality constant i.e., refer to Equation (2.81). This specification of a proportionality constant suggests the Gibbs sampling procedure can be coupled with the Metropolis-Hastings algorithm (MH; Patz & Junker, 1999; Hastings, 1970; Metropolis et. al., 1953), which utilizes an accept/reject sampling method. The concept of the MH algorithm is such that at each step, sample $\omega_j{}^{(t)}\sim q\left(\omega_j{}^{(t)}|\omega_j{}^{(t-1)}\right)$, where $q\left(\omega_j{}^{(t)}|\omega_j{}^{(t-1)}\right)$ is defined as the *proposal density*. Once $\omega_j{}^{(t)}$ has been sampled, the probability of accepting proposed state $\omega_j{}^{(t)}$ given previous state $\omega_j{}^{(t-1)}$ is defined by

$$\alpha^*\left(\omega_j{}^{(t)}|\omega_j{}^{(t-1)}\right)=min\left\{\frac{f\left(\omega_j{}^{(t)}|\omega_{-j},\boldsymbol{\alpha},\boldsymbol{X}\right)q\left(\omega_j{}^{(t-1)}|\omega_j{}^{(t)}\right)}{f\left(\omega_j{}^{(t-1)}|\omega_{-j},\boldsymbol{\alpha},\boldsymbol{X}\right)q\left(\omega_j{}^{(t)}|\omega_j{}^{(t-1)}\right)},1\right\}. \quad (2.82)$$

The proposal density can be chosen to be any convenient density e.g.,

1) *Normal random walk Metropolis-Hastings* which involves taking the proposal density $q\left(\omega_j{}^{(t)}|\omega_j{}^{(t-1)}\right) = q\left(\omega_j{}^{(t)}\right)$, independent of $\omega_j{}^{(t-1)}$.
2) *Independence Metropolis-Hastings* which involves taking the proposal density $q\left(\omega_j{}^{(t)}|\omega_j{}^{(t-1)}\right) = N\left(\mu = \omega_j{}^{(t-1)}, \sigma^2\right)$, a normal density with mean of $\omega_j{}^{(t-1)}$ and variance of σ^2.

An important note to mention is that if $q(.)$ is symmetric in $\omega_j{}^{(t)}$ and $\omega_j{}^{(t-1)}$, then the $q(.)$ terms in Equation (2.82) cancel, and the algorithm tends to move toward the mode of $f\left(\omega_j{}^{(t)}|\omega_{-j}, \boldsymbol{\alpha}, \boldsymbol{X}\right)$ (Junker, Patz & VanHoudnos, 2016). In addition, if $f\left(\omega_j{}^{(t)}|\omega_{-j}, \boldsymbol{\alpha}, \boldsymbol{X}\right) = q\left(\omega_j{}^{(t)}|\omega_j{}^{(t-1)}\right)$ and $f\left(\omega_j{}^{(t-1)}|\omega_{-j}, \boldsymbol{\alpha}, \boldsymbol{X}\right) = q\left(\omega_j{}^{(t-1)}|\omega_j{}^{(t)}\right)$ then the MH algorithm reduces to a Gibbs sampler such that we are always sampling from $f\left(\omega_j{}^{(t)}|\omega_{-j}, \boldsymbol{\alpha}, \boldsymbol{X}\right)$ and always accepting ω_j with $\alpha^* = 1$, which implies that the Gibbs sampler is a special case of the MH algorithm.

Stochastic Expectation-Maximization Algorithm

The SEM algorithm is a flexible and powerful algorithm with the capabilities to handle complex models, especially those for which the EM is difficult to implement (Diebolt & Ip, 1994). When implementing the SEM, the missing data $\boldsymbol{\alpha}$ with at each t^{th} iteration is "filled-in" with a single draw $f(\boldsymbol{\alpha}|\boldsymbol{\omega}, \boldsymbol{X})$ using the MH algorithm, thus forming a complete-data solution $\boldsymbol{Y} = (\boldsymbol{X}, \boldsymbol{\alpha})$. Note that $\boldsymbol{\alpha}$ is assumed to have some population distribution $f(\boldsymbol{\alpha})$. Using this complete-data set, $logL(\boldsymbol{\omega}|\boldsymbol{Y})$ is directly maximized to obtain a maximum likelihood estimate for $\boldsymbol{\omega}$**.** Note that this maximum is

usually determined by a method such as the Newton-Raphson optimization method. Alternating between the *stochastic imputation step* and maximization step generates a Markov chain $\{\boldsymbol{\omega}^{(t)}\}$ that coverages to a stationary distribution $f(\boldsymbol{\omega}|\boldsymbol{Y})$ under mild conditions (Ip, 1994). The stationary distribution is approximately centered at the maximum likelihood estimates of $\boldsymbol{\omega}^{(t)}$ and has a variance that depends on the rate of change of $\boldsymbol{\omega}^{(t)}$ in the maximization step. Typically, a certain number of iterations, t_o, are required as a "burn-in" period, allowing $\{\boldsymbol{\omega}^{(t)}\}$ to approach its stationary distribution. The iterative approach to the algorithm is summarized as follows (Grünewald, Humphreys, & Hössjer, 2010):

1) Select a starting parameter value $\boldsymbol{\omega}^{(t-1)} = \boldsymbol{\omega}^*$. Set $t = 1$.

2) *Stochastic imputation step* involves simulating $M = 1$ set of missing data $\boldsymbol{\alpha}^{(t)} = \left(\boldsymbol{\alpha}_1^{(t)}, \ldots, \boldsymbol{\alpha}_N^{(t)}\right) \sim \boldsymbol{\alpha}|\boldsymbol{X}, \boldsymbol{\omega}^{(t-1)}$ using a MH algorithm in the same way as described in the previous section discussing MCMC. Set $\boldsymbol{Y}_t = \left(\boldsymbol{X}, \boldsymbol{\alpha}^{(t)}\right)$ and compute $logL\left(\boldsymbol{\omega}^{(t-1)}|\boldsymbol{Y}_t\right)$, which is defined as the Monte Carlo approximation of the complete-data log-likelihood of the observed-data set, using the single imputed sample, $\boldsymbol{\alpha}^{(t)}$. Because $\boldsymbol{\alpha}$ is assumed to be sample independent for each respondent, the probability of accepting proposed state $\boldsymbol{\alpha}_i^{(t)}$ given previous state $\boldsymbol{\alpha}_i^{(t-1)}$ for the i^{th} respondent is defined by

$$\alpha^*\left(\boldsymbol{\alpha}_i^{(t)}|\boldsymbol{\alpha}_i^{(t-1)}\right) = min\left\{\frac{f\left(\boldsymbol{\alpha}_i^{(t)}|\boldsymbol{\alpha}_{-i}, \boldsymbol{\omega}, \boldsymbol{X}\right)q\left(\boldsymbol{\alpha}_i^{(t-1)}|\boldsymbol{\alpha}_i^{(t)}\right)}{f\left(\boldsymbol{\alpha}_i^{(t-1)}|\boldsymbol{\alpha}_{-i}, \boldsymbol{\omega}, \boldsymbol{X}\right)q\left(\boldsymbol{\alpha}_i^{(t)}|\boldsymbol{\alpha}_i^{(t-1)}\right)}, 1\right\}. \quad (2.83)$$

3) *Maximization step* involves obtaining a new parameter estimates that maximizes $logL\left(\boldsymbol{\omega}^{(t-1)}|\boldsymbol{Y}_t\right)$:

$$\boldsymbol{\omega}^{(t)} = \underset{\boldsymbol{\omega}}{arg\,max}\, logL\left(\boldsymbol{\omega}^{(t-1)}|\boldsymbol{Y}_t\right) \tag{2.84}$$

4) Set $t-1=t$. If $t \leq T + t_o$, go to Step 2, otherwise compute:

$$\widetilde{\boldsymbol{\omega}} = \frac{1}{T}\sum_{t=t_o+1}^{T+t_o} \boldsymbol{\omega}^{(t)} \tag{2.85}$$

The T represents the total number of retained ω estimates to be averaged over and recall t_0 represents the burn-in. Finding a new set of parameter estimates $\boldsymbol{\omega}^{(t)}$ in Step 3 that maximizes $logL\left(\boldsymbol{Y}_t|\boldsymbol{\omega}^{(t-1)}\right)$ can be done using the Newton-Raphson optimization method. However, for the SEM, a single iterative approach in the Newton-Raphson optimizer is sufficient enough for obtaining the maximum likelihood estimates (Diebolt & Ip, 1994a, 1994b).

Metropolis-Hastings Robbins-Monto Algorithm

Recently, Cai (2010a; 2010b) introduced a flexible framework for estimating parameters of statistical models by coupling two algorithms to formulate a joint estimation framework that addresses many of the less appealing features of strictly MCMC and maximum likelihood approaches (Chalmers & Flora, 2014). The MH-RM algorithm, like MCMC estimation, jointly estimates both item and ability parameters by utilizing a stochastically imputed complete-data solution with some assumed population

distribution for the latent variable to exploit on a more manageable complete-data likelihood approach. The MH-RM algorithm combines the process of stochastic imputation through a MH algorithm with a RM (Robbins & Monro, 1951) root-finding algorithm for noise-corrupted functions. The iterative algorithm can be partitioned into three stages: 1) perform b burn-in SEM iterations, 2) collect c iterations of $\boldsymbol{\omega}$ and find the average of this set, $\widetilde{\boldsymbol{\omega}}$, and 3) then perform the MH-RM stage until the model converges with a predetermined tolerance level. The MH-RM stage begins estimation with the initial parameter estimates $\boldsymbol{\omega}^{(t-1)} = \widetilde{\boldsymbol{\omega}}$ and recursively completes the following steps:

1) The *stochastic imputation step* involves stochastically impute an $N \times K$ matrix of missing latent classes $\boldsymbol{\alpha}^{(t)} = \left(\boldsymbol{\alpha}_1^{(t)}, \ldots, \boldsymbol{\alpha}_N^{(t)}\right) \sim \boldsymbol{\alpha}|\boldsymbol{X}, \boldsymbol{\omega}^{(t-1)}$ with a MH algorithm M times. Because $\boldsymbol{\alpha}$ is assumed to be sample independent for each respondent, the probability of accepting proposed state $\boldsymbol{\alpha}_i^{(t)}$ given previous state $\boldsymbol{\alpha}_i^{(t-1)}$ for the i^{th} respondent is defined by Equation (2.83).

2) For the *stochastic approximation step*, using Fisher's identity, approximation of the observed-data gradient is done by using the sample average of the complete-data gradients,

$$\widetilde{\boldsymbol{g}}_{j+1} = \frac{1}{M} \sum_{m=1}^{M} g\left(\boldsymbol{\omega}_j^{(t)}|\boldsymbol{Y}\right) \tag{2.86}$$

and a recursive approximation of the conditional expectation of the complete-data information matrix,

$$\boldsymbol{\Gamma}_{j+1} = \boldsymbol{\Gamma}_j + \gamma_t \left\{ \frac{1}{M} \sum_{m=1}^{M} -H\left(\boldsymbol{\omega}_j^{(t)} | \boldsymbol{Y}\right) - \boldsymbol{\Gamma}_j \right\} \tag{2.87}$$

that allows for effectively speeding up convergence. Here, $\boldsymbol{\Gamma}_j$ represents a $d \times d$ positive definite symmetric matrix, where d is the total number of item parameters to be estimated for the j^{th} item. The initial choice of $\boldsymbol{\Gamma}$ can be arbitrary (e.g., identity matrix) in the initial step of the algorithm.

3) In the *RM update step,* a new set of item parameter estimates are proposed,

$$\boldsymbol{\omega}_j^{(t)} = \boldsymbol{\omega}_j^{(t-1)} + \gamma_t \left(\boldsymbol{\Gamma}_{j+1}^{-1} \tilde{\boldsymbol{g}}_j \right). \tag{2.88}$$

4) Check convergence criterion ϵ such that $\left\| \boldsymbol{\omega}^{(t)} - \boldsymbol{\omega}^{(t-1)} \right\| < \epsilon$, else return to Step 1 if condition not satisfied.

The γ_t in Equations (2.87) and (2.88) is referred to as *gain constants*. The gain constants purpose is to scale the updates and serve to slowly average out the noise in the approximation to the complete-data gradients. Thus, γ_t needs to slowly decrease to zero, which can be ensured by the following conditions:

$$\gamma_t \in (0, 1], \sum_{t=1}^{\infty} \gamma_t = \infty, \text{and } \sum_{t=1}^{\infty} \gamma_t^2 < \infty. \tag{2.89}$$

Note if γ_t decreases too quickly, the $\boldsymbol{\omega}$ estimates may stabilize prematurely before the maximum likelihood estimate is reached. Alternatively, if γ_t decreases too slowly, the $\boldsymbol{\omega}$ estimates may never stabilize (Monro & Cai, 2014). The algorithm handles the inherent

noise-corrupted stochastic imputation procedure by using the RM root-finding algorithm to stabilize both the updates and the information matrix (Chalmers & Flora, 2014). In this way, the inaccuracies borne from the MH algorithm are adequately accounted for when attempting to maximize the complete-data log-likelihood, and subsequent standard errors can be computed appropriately using Louis's (1982) complete-data method (refer to Cai, 2010a).

Summary of Research Study

The primary goal of this study was to provide an extension of the P-LCDM GPDM to allow for polytomous attributes, which may have use in the exploration of learning progressions or when dichotomous attributes are an over simplification of the abilities of interest. Then, due to the potential of exponentially increasing the number of latent classes, explore the feasibility and efficiency in addition to the quality of parameter estimation of the SEM and MH-RM algorithms relative to the EM algorithm. Note that while the polytomous attribute case provides one example for when computational burden can be challenging, a similar challenge can occur when the number of dichotomous attributes becomes large. This study introduces the SEM and MH-RM algorithms as computationally faster methods for estimating parameters of a CDM when many latent classes are present in a diagnostic assessment. The EM algorithm can be a computationally slow method for estimating parameters of a CDM when many latent classes are present in a diagnostic assessment. The EM algorithm is typically slow because the E-step involves computing the posterior probability of every latent class for each given respondent. This computation becomes even challenging at a much faster rate

(i.e., even when measuring six to eight attributes) when polytomous attributes are introduced in the model because the number of latent classes is then represented as $\prod_{k=1}^{K} S_k$, where S_k represents the number of attribute levels for the k^{th} attribute. With respect to the comparison of SEM and MH-RM to the EM algorithm, recall the three research questions motivating this study are:

1) To what extent does the SEM and MH-RM algorithms show to be computationally faster when compared to the EM algorithm as the number of latent classes increases?

2) How accurately are the item parameters of the P-LCDM-PA submodels estimated when comparing the SEM, MH-RM, and EM algorithms for estimation?

3) How accurately are examinees attributes (and attribute patterns) estimated when using the SEM, MH-RM, and EM algorithms to estimate the P-LCDM-PA submodels?

CHAPTER III

METHODOLOGY

Background

This study looks to using the P-LCDM for polytomous attributes, which is a special case of the GPDM for polytomous attributes. Recall that the P-LCDM models ordinal responses using the LCDM with multiple intercepts while all effect-level parameters are held constant across all category thresholds. Such an extension to the LCDM for an ordinal response is consistent with the graded response model in IRT literature (i.e., refer to Samejima, 1969), and the subsequent multidimensional extensions (e.g., Muraki & Carlson, 1995; Gibbons et al., 2007). In contrast the GPDM generalized the P-LCDM to allow for multiple intercepts and multiple effect-level parameters to be estimated across each category threshold. Although both approaches are general models that subsume many constrained ordinal response models, which are extensions of models previously introduced in the literature, Hansen (2013) and Chen and de la Torre (2018) did not provide a detailed description of these respective submodels (e.g., ordinal response extension for the DINA, DINO, and C-RUM).

Note that because of the two different parametrization of an ordinal model there are differences related to the number of estimated item parameters, overall fidelity and model fit of each model. For example, a DINA for ordinal responses using the P-LCDM for an item with four categories will have a total of four estimated item parameters (one

interaction and three intercepts), whereas the GPDM would have six (three interactions and three intercepts). Furthermore, while the R-RUM could be defined as a model for ordinal responses using the GPDM, this is not possible using the P-LCDM. It's important to note that the addition of polytomous attributes dramatically increasing the number of possible latent classes, which then naturally allows for the systematic exploration of the efficiency of the SEM and MH-RM for this more complex model.

The estimation algorithms that will be implemented to obtain estimates of the P-LCDM-PA submodels are the EM, SEM, and MH-RM. The MCMC algorithm is not included in the current study because it has been ubiquitously used in estimating CDM over the past several years due to its simplicity of implementation. Also, recall that there are limitations to using an MCMC algorithm due to a large amount of iterations that are typically needed (i.e., 10,000) for the Markov chain to reach its stationary distribution. Note that it is also very difficult to objectively evaluate convergence, even empirically (Sinharay, 2004).

Polytomous DINA for Polytomous Attributes

Given the relationship between the LCDM and DINA and the fact that the P-LCDM-PA is a polytomous extension of the LCDM, then the P-LCDM-PA can be used to define a polytomous attribute version of the DINA for ordinal responses i.e., polytomous DINA for polytomous attributes (P-DINA-PA). Recall that the DINA for dichotomous responses only defines two parameters for each item: a slipping parameter, s_j, and guessing parameter, g_j. Specifically, the conditional probability of a correct response is equal to g_j unless all required attributes for the j^{th} item have been mastered,

in which the conditional probability of a correct response increases to $1 - s_j$. However, for the DINA model to be used for an ordinal scale more slip and guess parameters must be defined.

Let $\boldsymbol{\Delta}_j$ be a collection of slip and guess parameters for the j^{th} item such that $\boldsymbol{\Delta}_j = \left(s_{j1}, \dots, s_{jc}, \dots, s_{j(C_j-1)}, g_{j1}, \dots, g_{jc}, \dots, g_{j(C_j-1)}\right)$. Note that the reduction of the P-LCDM-PA to a P-DINA-PA will still use a similar modeling approach as the graded response model introduced earlier. Thus, the conditional probability of responding to the c^{th} category under the P-DINA-PA for the i^{th} respondent on the j^{th} item can be defined by

$$P\left(X_{ij} = c \middle| \boldsymbol{\Delta}_j, \xi_{ij}^*\right) = P\left(X_{ij} \geq c \middle| \boldsymbol{\Delta}_j, \xi_{ij}^*\right) - P\left(X_{ij} \geq c+1 \middle| \boldsymbol{\Delta}_j, \xi_{ij}^*\right) \tag{3.1}$$

Only now the set of boundary response probabilities are defined based on the corresponding slip and guess parameters for that boundary, which are defined as

$$P\left(X_{ij} \geq 0 \middle| \boldsymbol{\Delta}_j, \xi_{ij}^*\right) = 1$$

$$P\left(X_{ij} \geq 1 \middle| \boldsymbol{\Delta}_j, \xi_{ij}^*\right) = \left(1 - s_{j1}\right)^{\xi_{ij}^*} g_{j1}^{\left(1-\xi_{ij}^*\right)}$$

$$\dots$$

$$P\left(X_{ij} \geq c \middle| \boldsymbol{\Delta}_j, \xi_{ij}^*\right) = \left(1 - s_{jc}\right)^{\xi_{ij}^*} g_{jc}^{\left(1-\xi_{ij}^*\right)} \tag{3.2}$$

$$\dots$$

$$P\left(X_{ij} \geq C_j - 1 \middle| \boldsymbol{\Delta}_j, \xi_{ij}^*\right) = \left(1 - s_{j(C_j-1)}\right)^{\xi_{ij}^*} g_{j(C_j-1)}^{\left(1-\xi_{ij}^*\right)}$$

$$P\left(X_{ij} \geq C_j \middle| \boldsymbol{\Delta}_j, \xi_{ij}^*\right) = 0.$$

Furthermore, the $s_{j1}, \ldots, s_{jc}, \ldots, s_{C_j-1}$ and $g_{j1}, \ldots, g_{jc}, \ldots, g_{C_j-1}$ are defined by

$$
\begin{aligned}
s_{j1} &= P\left(X_{ij} < 1 | \xi_{ij}^{*} = 1\right) \\
&\ldots \\
s_{jc} &= P\left(X_{ij} < c | \xi_{ij}^{*} = 1\right) \\
&\ldots \\
s_{j(C_j-1)} &= P\left(X_{ij} < C_j - 1 | \xi_{ij}^{*} = 1\right)
\end{aligned}
\tag{3.3}
$$

and

$$
\begin{aligned}
g_{j1} &= P\left(X_{ij} \geq 1 | \xi_{ij}^{*} = 0\right) \\
&\ldots \\
g_{jc} &= P\left(X_{ij} \geq c | \xi_{ij}^{*} = 0\right) \\
&\ldots \\
g_{j(C_j-1)} &= P\left(X_{ij} \geq C_j - 1 | \xi_{ij}^{*} = 0\right),
\end{aligned}
\tag{3.4}
$$

respectively. Recall that $\xi_{ij}^{*} \in \{0,1\}$ is a latent variable that defines whether the i^{th} respondent has mastered (i.e., $\alpha \geq q$) all required attributes for the j^{th} item. The s_{jc} parameters under the prototypical formulation of the P-DINA-PA are now defined as the conditional probability of responding below the c^{th} category given all required attributes have been mastered the required level for the j^{th} item. The g_{jc} parameters under the prototypical formulation of the P-DINA-PA are now defined as the conditional

probability of responding to the c^{th} category or higher given no required attributes have been mastered for the j^{th} item. Note that when using the P-LCDM-PA the same constraints are used to define the DINA as when using the LCDM. However, because there are multiple intercepts corresponding to the each of the ordinal levels then it is possible to solve for slip and guess parameters for each level. Defining the P-LCDM-PA parameters as a function of the P-DINA-PA parameters:

$$\lambda_{j1,0} = -\ln\left(\frac{1-g_{j1}}{g_{j1}}\right)$$

...

$$\lambda_{jc,0} = -\ln\left(\frac{1-g_{jc}}{g_{jc}}\right) \tag{3.5}$$

...

$$\lambda_{j(c_j-1),0} = -\ln\left(\frac{1-g_{j(c_j-1)}}{g_{j(c_j-1)}}\right)$$

and

$$\lambda_{j,K_j,(1,\ldots,K_j)} = -\lambda_{j1,0} - \ln\left(\frac{s_{j1}}{1-s_{j1}}\right)$$

...

$$\lambda_{j,K_j,(1,\ldots,K_j)} = -\lambda_{jc,0} - \ln\left(\frac{s_{jc}}{1-s_{jc}}\right) \tag{3.6}$$

...

$$\lambda_{j,K_j,(1,\ldots,K_j)} = -\lambda_{j(c_j-1),0} - ln\left(\frac{s_{j(c_j-1)}}{1-s_{j(c_j-1)}}\right).$$

Table 8 shows the relationship between the prototypical formulation of the P-DINA-PA to the P-LCDM-PA for a single item. It's important to note that the interaction term $\lambda_{j,K_j,(1,\ldots,K_j)}$ is the same across all category thresholds, and as a result the s parameters do have a constrained relationship as opposed to the GPDM. Suppressing the subscripts i and j and assuming $K = 2$, the boundary response probability for the c^{th} category can be computed by

$$P(X \geq c|\boldsymbol{\alpha}) = \frac{1}{1 + Exp\left[-\lambda_{c,0} - \lambda_{1,(1)}\alpha^{**}_{1,(1)} - \lambda_{1,(2)}\alpha^{**}_{1,(2)} - \lambda_{2,(1,2)}\alpha^{**}_{2,(1,2)}\right]} \qquad (3.7)$$

Table 8. Relationship between the P-LCDM-PA and P-DINA-PA

	$P(X \geq c\|\boldsymbol{\alpha})$	
$\boldsymbol{\alpha}^{**}$	P-LCDM-PA	P-DINA-PA
(0,0,0)	$\frac{1}{1 + Exp\left[-\lambda_{c,0} - \lambda_{2,(1,2)}(0)\right]}$	g_c
(1,0,0)	$\frac{1}{1 + Exp\left[-\lambda_{c,0} - \lambda_{2,(1,2)}(0)\right]}$	g_c
(0,1,0)	$\frac{1}{1 + Exp\left[-\lambda_{c,0} - \lambda_{2,(1,2)}(0)\right]}$	g_c
(1,1,1)	$\frac{1}{1 + Exp\left[-\lambda_{c,0} - \lambda_{2,(1,2)}(1)\right]}$	$1 - s_c$

The formulation in Table 7 for the P-LCDM-PA states that there is only a positive conditional relationship between a required attribute and the item when all other attributes that are measured by that item have been mastered. All conditional relationships between an attribute and the item, given at least one attribute has not been

mastered are set to 0. Specifically, there is no increase in conditional probability of a higher response for only knowing a subset of attributes, but only when all attributes measured by the item have been mastered, which is an example of a conjunctive model.

Polytomous DINO for Polytomous Attributes

Given the relationship between the LCDM and DINO and the fact that the P-LCDM-PA is a polytomous extension of the LCDM, then the P-LCDM-PA can be used to define a polytomous version of the DINO for ordinal responses i.e., polytomous DINO for polytomous attributes (P-DINO-PA). Recall that, similar to the DINA, the DINO for dichotomous responses only defines two parameters for each item: a slipping parameter s_j and guessing parameter g_j. Specifically, the conditional probability of a correct response is equal to g_j unless at least one measured attribute for the j^{th} item has been mastered, in which the conditional probability of a correct response increases to $1 - s_j$. For the DINO model to be used for an ordinal scale, more slip and guess parameters must be defined. Similar to the DINA, Let $\boldsymbol{\Delta}_j$ be a collection of slip and guess parameters for the j^{th} item such that $\boldsymbol{\Delta}_j = \left(s_{j1}, \dots, s_{jc}, \dots, s_{j(C_j-1)}, g_{j1}, \dots, g_{jc}, \dots, g_{j(C_j-1)}\right)$. Note that the reduction of the P-LCDM-PA to a P-DINO-PA will still use a similar modeling approach as the graded response model introduced earlier. Thus, the conditional probability of responding to the c^{th} category under the P-DINO-PA for the i^{th} respondent on the j^{th} item can be defined by

$$P\left(X_{ij} = c \middle| \boldsymbol{\Delta}_j, \eta_{ij}^*\right) = P\left(X_{ij} \geq c \middle| \boldsymbol{\Delta}_j, \eta_{ij}^*\right) - P\left(X_{ij} \geq c + 1 \middle| \boldsymbol{\Delta}_j, \eta_{ij}^*\right). \quad (3.8)$$

Only now the set of boundary response probabilities (i.e., the probability of getting a score at or above a given value c) are defined based on the corresponding slip and guess parameters for that boundary. Specifically, they are defined as

$$P\left(X_{ij} \geq 0 \middle| \boldsymbol{\Delta}_j, \eta_{ij}^*\right) = 1$$

$$P\left(X_{ij} \geq 1 \middle| \boldsymbol{\Delta}_j, \eta_{ij}\right) = \left(1 - s_{j1}\right)^{\eta_{ij}^*} g_{j1}^{\left(1-\eta_{ij}^*\right)}$$

$$\ldots$$

$$P\left(X_{ij} \geq c \middle| \boldsymbol{\Delta}_j, \eta_{ij}\right) = \left(1 - s_{jc}\right)^{\eta_{ij}^*} g_{jc}^{\left(1-\eta_{ij}^*\right)} \tag{3.9}$$

$$\ldots$$

$$P\left(X_{ij} \geq C_j - 1 \middle|\, \boldsymbol{\Delta}_j, \eta_{ij}\right) = \left(1 - s_{j(C_j-1)}\right)^{\eta_{ij}^*} g_{j(C_j-1)}^{\left(1-\eta_{ij}^*\right)}$$

$$P\left(X_{ij} \geq C_j \middle| \boldsymbol{\Delta}_j, \eta_{ij}^*\right) = 0.$$

The $s_{j1}, \ldots, s_{jc}, \ldots, s_{jC_j-1}$ and $g_{j1}, \ldots, g_{jc}, \ldots, g_{jC_j-1}$ are defined by

$$s_{j1} = P\left(X_{ij} < 1 | \eta_{ij}^* = 1\right)$$

$$\ldots$$

$$s_{jc} = P\left(X_{ij} < c | \eta_{ij}^* = 1\right) \tag{3.10}$$

$$\ldots$$

$$s_{j(C_j-1)} = P\left(X_{ij} < C_j - 1 | \eta_{ij}^* = 1\right)$$

and

$$g_{j1} = P\left(X_{ij} \geq 1 | \eta_{ij}^{*} = 0\right)$$

$$\ldots$$

$$g_{jc} = P\left(X_{ij} \geq c | \eta_{ij}^{*} = 0\right) \tag{3.11}$$

$$\ldots$$

$$g_{j(C_j-1)} = P\left(X_{ij} \geq C_j - 1 | \eta_{ij}^{*} = 0\right),$$

respectively. The $\eta_{ij}^{*} \in \{0,1\}$ is a latent variable that defines whether the i^{th} respondent has mastered (i.e., $\alpha \geq q$) at least one required attribute for the j^{th} item

$$\eta_{ij}^{*} = \begin{cases} 1, & if\ \exists k\ such\ that\ \alpha_{ik} \geq q_{jk} \\ 0, & if\ \alpha_{ik} < q_{jk}\ \forall k \end{cases} . \tag{3.12}$$

The s_{jc} parameters under the prototypical formulation of the P-DINO-PA are now defined as the conditional probability of responding below the c^{th} category given at least one required attribute has been mastered the required level for the j^{th} item. The g_{jc} parameters under the prototypical formulation of the P-DINO-PA are now defined as the conditional probability of responding to the c^{th} category or higher given no required attributes have been mastered for the j^{th} item. Note that when using the P-LCDM-PA similar constraints are used to define the DINO as when using the LCDM. However, because there are multiple intercepts corresponding to the each of the ordinal levels then it is possible to solve for slip and guess parameters for each level. Defining the P-LCDM-PA parameters as a function of the P-DINO-PA parameters: Equation (3.5) and

$$\lambda_j = -\lambda_{j1,0} - ln\left(\frac{s_{j1}}{1 - s_{j1}}\right)$$

...

$$\lambda_j = -\lambda_{jc,0} - ln\left(\frac{s_{jc}}{1 - s_{jc}}\right) \tag{3.13}$$

...

$$\lambda_j = -\lambda_{j(C_j-1),0} - ln\left(\frac{s_{jC_j-1}}{1 - s_{jC_j-1}}\right).$$

Table 9 shows the relationship between the prototypical formulation of the P-DINO-PA to the P-LCDM-PA for a single item. It's important to note that the interaction term λ_j is the same across all category thresholds, and as a result the s parameters do have a constrained relationship as opposed to the GPDM. Suppressing the subscript i and j and assuming $K = 2$, the boundary response probability for the c^{th} category can be computed by

$$P(X \geq c|\boldsymbol{\alpha}) = \frac{1}{1 + Exp\left[-\lambda_{c,0} - \lambda\alpha^{**}_{1,(1)} - \lambda\alpha^{**}_{1,(2)} - \lambda\alpha^{**}_{2,(1,2)}\right]} \tag{3.14}$$

Table 9. Relationship between the P-LCDM-PA and P-DINO-PA

	$P(X \geq c\|\boldsymbol{\alpha})$	
$\boldsymbol{\alpha}^{**}$	P-LCDM-PA	P-DINO-PA
(0,0,0)	$\frac{1}{1+Exp\left[-\lambda_{c,0}-\lambda(0)-\lambda(0)+\lambda(0)\right]}$	g_c
(1,0,0)	$\frac{1}{1+Exp\left[-\lambda_{c,0}-\lambda(1)-\lambda(0)+\lambda(0)\right]}$	$1-s_c$
(0,1,0)	$\frac{1}{1+Exp\left[-\lambda_{c,0}-\lambda(0)-\lambda(1)+\lambda(0)\right]}$	$1-s_c$
(1,1,1)	$\frac{1}{1+Exp\left[-\lambda_{c,0}-\lambda(1)-\lambda(1)+\lambda(1)\right]}$	$1-s_c$

Here, λ is denoted as a single value that is estimated for the item along with $\lambda_{c,0}$. The formulation in Table 8 for the P-LCDM-PA shows that there is a positive conditional relationship between a required attribute and the item when at least one attribute has been mastered. Given that any other required attributes have been mastered, this conditional relationship between an attribute and the item is set to 0. Specifically, there is an increase in the conditional probability of responding to the c^{th} category or higher for knowing at least one attribute, but there is no additional increase in conditional probability of responding to the c^{th} category or higher when additional attributes have been mastered. A similar strategy can be applied to items measuring more than two attributes where the sign in front of λ is determined by Equation (2.26).

Polytomous C-RUM for Polytomous Attributes

Finally, there exists a mathematical relationship such that there is a set of constraints placed on the P-LCDM-PA that relates to the prototypical formulation of the C-RUM for ordinal responses i.e., polytomous C-RUM for polytomous attributes (P-C-

RUM-PA). In this case the relationship is most obvious because of the general form of the P-C-RUM-PA. Recall that the C-RUM for dichotomous responses only defines two types of parameters for each item: an intercept parameter $\lambda_{j,0}$ and main-effect parameters $\boldsymbol{\lambda}_j = (\lambda_{j,1,(1)}, \ldots, \lambda_{j,1,(k)}, \ldots, \lambda_{j,1,(K_j)})$. In this case, the conditional probability of a correct response when no required attributes have been mastered is $logit^{-1}(\lambda_{j,0})$. Since $0 \leq \boldsymbol{\lambda}_j < \infty$, the conditional probability of a correct response increases for every required attributed mastered by the j^{th} item. Table 10 shows the relationship between the prototypical formulation of the P-C-RUM-PA to the P-LCDM-PA for a single item using Equation (3.7).

Table 10. Relationship between the P-LCDM-PA and P-C-RUM-PA

	$P(X \geq c\|\boldsymbol{\alpha})$	
$\boldsymbol{\alpha}^{**}$	P-LCDM-PA	P-C-RUM-PA
(0,0,0)	$\frac{1}{1+Exp[-\lambda_{c,(0)}-\lambda_{1,(1)}(0)-\lambda_{1,(2)}(0)]}$	$\frac{1}{1+Exp[-\lambda_{c,(0)}-\lambda_{1,(1)}(0)-\lambda_{1,(2)}(0)]}$
(1,0,0)	$\frac{1}{1+Exp[-\lambda_{c,(0)}-\lambda_{1,(1)}(1)-\lambda_{1,(2)}(0)]}$	$\frac{1}{1+Exp[-\lambda_{c,(0)}-\lambda_{1,(1)}(1)-\lambda_{1,(2)}(0)]}$
(0,1,0)	$\frac{1}{1+Exp[-\lambda_{c,(0)}-\lambda_{1,(1)}(0)-\lambda_{1,(2)}(1)]}$	$\frac{1}{1+Exp[-\lambda_{c,(0)}-\lambda_{1,(1)}(0)-\lambda_{1,(2)}(1)]}$
(1,1,1)	$\frac{1}{1+Exp[-\lambda_{c,(0)}-\lambda_{1,(1)}(1)-\lambda_{1,(2)}(1)]}$	$\frac{1}{1+Exp[-\lambda_{c,(0)}-\lambda_{1,(1)}(1)-\lambda_{1,(2)}(1)]}$

The P-C-RUM-PA is simply defined by setting $\lambda_{2,(1,2)} = 0$. Thus, the conditional probability responding to the c^{th} category or higher will increase by a factor of e^{λ_1} when comparing a non-master to a respondent who has mastered the first attribute, which does not depend on mastery or non-mastery of the second attribute.

Model Assumptions

There are several assumptions in the P-LCDM-PA to ensure identifiability of which are the same as what are required for the LCDM (Henson, et al, 2009). The first assumption is determined by the $\boldsymbol{Q}$ matrix. Identifying a $\boldsymbol{Q}$ matrix in the analysis is comparable to a confirmatory factor analysis such that the definition of attributes is identified by the items that measure each attribute. Without the $\boldsymbol{Q}$ matrix, attributes could alternate in their definition, much like a rotational indeterminacy in an exploratory MIRT analysis. The second assumption is to ensure *monotonicity* in the item response function. In the context of CDM, monotonicity is defined as the property such that assuming that any respondent who masters additional attributes will have a conditional probability of a response equal to or greater than the conditional probability of a response prior to learning the additional set of attributes. The property of monotonicity in the LCDM can be expressed as

$$P\left(X_{ij} \middle| \boldsymbol{\omega}_j, \boldsymbol{\alpha}_i^w\right) \geq P\left(X_{ij} \middle| \boldsymbol{\omega}_j, \boldsymbol{\alpha}_i\right), \forall w \tag{3.15}$$

where

$$\alpha_{ij,1,(k)}^{**(w)} = \begin{cases} \alpha_{ij,1,(k)}^{**}, & if\ w \neq k \\ 1, & else \end{cases} \tag{3.16}$$

The third assumption is based on the fact the attributes and $\boldsymbol{Q}$ matrix entries are defined as either $\{0,1\}$ for the dichotomous attributes and $\{0,1,2,\ldots,S_k\}$ for polytomous attributes. By imposing this constraint, a reference group is identified as those

respondents who have not mastered any of the required attributes for an item. Thus, identifying the conditional probability of a response for the respondents who have not mastered any of the required attributes for an item as the $logit^{-1}(\lambda_0)$. The final assumption is based on *conditional independence* (Lord & Novick, 1968). This assumption states that a sequence of item responses $X_{i1}, X_{i2}, \ldots, X_{iJ}$ for the i^{th} respondent is conditionally independent given all item parameters $\boldsymbol{\Delta}_j$ and an individual's mastery profile (i.e., latent class) $\boldsymbol{\alpha}_i$:

$$P\left(\bigcap_{j=1}^{J} X_{ij} \,|\boldsymbol{\omega}_j, \boldsymbol{\alpha}_i\right) = \prod_{j=1}^{J} P\left(X_{ij}\middle|\, \boldsymbol{\omega}_j, \boldsymbol{\alpha}_i\right)$$
$$= P\left(X_{i1}|\boldsymbol{\omega}_j, \boldsymbol{\alpha}_i\right) \times \ldots \times P\left(X_{ij}|\boldsymbol{\omega}_j, \boldsymbol{\alpha}_i\right) \times \ldots \times P\left(X_{iJ}|\boldsymbol{\omega}_j, \boldsymbol{\alpha}_i\right). \quad (3.17)$$

This equation states that the correlation between a set of item responses $X_{i1}, X_{i2}, \ldots, X_{iJ}$ should be zero after the effect of $\boldsymbol{\alpha}_i$ is conditioned out. The set of item responses should only be correlated through the latent variables that the test or survey is measuring.

Observed-Data and Complete-Data Likelihoods

To determine the *observed-data likelihood* function for the P-LCDM-PA, the conditional probability of observing a response for the i^{th} respondent on the j^{th} item must first be defined as

$$P\left(X_{ij}\middle|\boldsymbol{\omega}_j, \boldsymbol{\alpha}_i\right) = \prod_{c=0}^{C_j-1} P\left(X_{ij} = c\middle|\, \boldsymbol{\omega}_j, \boldsymbol{\alpha}_i\right)^{I\left[X_{ij}=c\right]} \quad (3.18)$$

where the indicator function $I\left[X_{ij} = c\right]$ is equal to

$$I[X_{ij} = c] = \begin{cases} 1, & if\ X_{ij} = c \\ 0, & else \end{cases}. \tag{3.19}$$

When conditional independence is assumed as defined in Equation (3.17), the conditional likelihood of the i^{th} respondent's $J \times 1$ response vector $\boldsymbol{X}_i$ is

$$L(\boldsymbol{X}_i|\boldsymbol{\omega}, \boldsymbol{\alpha}_i) = \prod_{j=1}^{J} P(X_{ij}|\boldsymbol{\omega}_j, \boldsymbol{\alpha}_i) \tag{3.20}$$

and the marginal likelihood of response pattern $\boldsymbol{X}_i$ for the i^{th} respondent is then,

$$L(\boldsymbol{X}_i|\boldsymbol{\omega}) = \sum_{\forall \boldsymbol{\alpha}} f(\boldsymbol{\alpha}) \prod_{j=1}^{J} P(X_{ij}|\boldsymbol{\omega}_j, \boldsymbol{\alpha}) \tag{3.21}$$

where $f(\boldsymbol{\alpha})$ is the prior distribution defining the probability of any given respondent randomly being sampled from the population belonging to the latent class $\boldsymbol{\alpha}$. The summation is taken over all possible latent classes $\boldsymbol{\alpha}_l, \ldots, \boldsymbol{\alpha}_L$. The total number latent classes $\boldsymbol{\alpha}$ can be computed as $L = \prod_{k=1}^{K} S_k$. The following constraint is place on $f(\boldsymbol{\alpha})$ such that,

$$\sum_{\forall \boldsymbol{\alpha}} f(\boldsymbol{\alpha}) = 1. \tag{3.22}$$

Recall that imposing a prior distribution on $\boldsymbol{\alpha}$ simplifies the estimation process in two ways. First, the focus shifts from some large number of latent variable parameters to just

the parameters of the prior distribution. Second, by imposing a distributional form for the population, the latent variable parameters become temporarily "known" (i.e., forming complete-data) and can then be marginalized out of the estimation process. Treating response patterns as fixed once observed, the observed-data likelihood function across the $N \times J$ response data matrix can be formulated:

$$L(\boldsymbol{\omega}|\boldsymbol{X}) = \prod_{i=1}^{N}\left[\sum_{\forall \boldsymbol{\alpha}} f(\boldsymbol{\alpha}) \prod_{j=1}^{J} P\left(X_{ij}\middle|\boldsymbol{\omega}_j, \boldsymbol{\alpha}\right)\right]. \tag{3.23}$$

The latent classes $\boldsymbol{\alpha}$ can be thought of as the "missing data" component. Now assuming the missing data $\boldsymbol{\alpha}$ are "filled-in" for each respondent via a stochastic imputation process, the complete-data solution can be formulated $\boldsymbol{Y} = (\boldsymbol{X}, \boldsymbol{\alpha})$. Thus, the *complete-data likelihood* function of the P-LCDM-PA is defined as follows,

$$L(\boldsymbol{\omega}|\boldsymbol{Y}) = \prod_{i=1}^{N}\left[\prod_{j=1}^{J} P\left(X_{ij}\middle|\boldsymbol{\omega}_j, \boldsymbol{\alpha}_i\right) f(\boldsymbol{\alpha}_i)\right] = \left[\prod_{i=1}^{N} f(\boldsymbol{\alpha}_i)\right]\left[\prod_{i=1}^{N}\prod_{j=1}^{J} P\left(X_{ij}\middle|\boldsymbol{\omega}_j, \boldsymbol{\alpha}_i\right)\right] \tag{3.24}$$

and further taking the log of the complete-data likelihood results in sum of two independent parts,

$$logL(\boldsymbol{\omega}|\boldsymbol{Y}) = \sum_{i=1}^{N} logf(\boldsymbol{\alpha}_i) + \sum_{i=1}^{N}\sum_{j=1}^{J}\sum_{c=0}^{C_j-1} I\left[X_{ij} = c\right] logP\left(X_{ij} = c\middle|\, \boldsymbol{\omega}_j, \boldsymbol{\alpha}_i\right) \tag{3.25}$$

$$= \sum_{i=1}^{N} log f(\boldsymbol{\alpha}_i) + \sum_{i=1}^{N} \sum_{j=1}^{J} \sum_{c=0}^{C_j - 1} I\left[X_{ij} = c\right] log\left[P\left(X_{ij} \geq c \middle| \boldsymbol{\omega}_j, \boldsymbol{\alpha}_i\right) - P\left(X_{ij} \geq c + 1 \middle| \boldsymbol{\omega}_j, \boldsymbol{\alpha}_i\right)\right]. \quad (3.26)$$

Simulation Experiment

The primary goal of this study was to utilize polytomous attributes in the polytomous log-linear cognitive diagnosis model (P-LCDM-PA), which is a special case of the general polytomous diagnostic model (GPDM) for polytomous attributes. Then, due to the potential of exponentially increasing the number of latent classes, explore the feasibility and efficiency in addition to the quality of parameter estimation of the SEM and MH-RM algorithms relative to the EM algorithm. Note that although this particular study focuses on polytomous attributes, similar computational challenges exist when the number of attributes become large. Thus, this study will also contribute to applications where the number of attributes is large. For example, Choi, Lee, & Park's (2015) application study involved the analysis of a diagnostic assessment that measured 12 attributes, while other application studies conducted by Chen et al. (2010) and Chen (2012) involved the analysis of diagnostic assessments that measured as many as 23 attributes.

To assess the efficacy of estimation algorithms, a function in R (R, 2017) was created to simulate observed response data from the P-LCDM-PA submodels. There was a total of $1(sample\ size) \times 1(survey\ length) \times 2(item\ quality) \times 2(number\ of\ attributes) \times 2(attribute\ levels) \times 1(response\ categories) \times 1(correlation\ between\ attributes) \times 3(generating\ models) = 24$ joint conditions

with each joint condition being replicated 50 times. Table 11 presents a summary of the simulation experiment.

Table 11. Summary of Simulation Experiment

Design Factor	Number of Levels	Values of Levels
Sample Size	1	5,000
Survey Length	1	50 items
Item Quality	2	Low and High
Number of Attributes	2	4 and 6 attributes
Number of Attribute Levels	2	2 and 3 levels
Number of Response Categories	1	4 categories
Correlation between Attributes	1	$\rho \in [.5, .8]$
Generating Model	3	P-DINA-PA, P-DINO-PA, P-C-RUM-PA
Total Joint Conditions	24	
Replications	50	
Total	1,200	

$\boldsymbol{Q}$ matrices of $J \times K$ dimensions with either dichotomous or polytomous attribute levels were randomly simulated from a categorical distribution with for each given element of the j^{th} item's $\boldsymbol{Q}$ vector. The corresponding probabilities $\boldsymbol{p}$ for each of the sample spaces $\{0,1\}$ and $\{0,1,2\}$ are

$$p(q_{jk}) = \begin{cases} .7, & if\ q_{jk} = 0 \\ .3, & if\ q_{jk} = 1 \end{cases}, \tag{3.27}$$

and

$$p(q_{jk}) = \begin{cases} .7, & if\ q_{jk} = 0 \\ .15, & if\ q_{jk} = 1, \\ .15, & if\ q_{jk} = 2 \end{cases} \tag{3.28}$$

respectively. The corresponding probability mass function of the categorical distribution for the random variable q_{jk} can be defined by

$$f(q_{jk}|\boldsymbol{p}) = \prod_{k=0}^{S_k-1} p(q_{jk})^{I[q_{jk}=k]} \tag{3.29}$$

where $I[q_{jk} = k]$ evaluates to 1 if $q_{jk} = k$, otherwise evaluates to 0. It was assumed that all attributes measured the same number of levels $S_1 = S_2 =, \ldots, = S_K$ which resulted in S^K number of latent classes. The choice for number of polytomous attribute levels was based on prior research done in Chen & de la Torre (2013) who examined three levels in their simulation study. Simulated $\boldsymbol{Q}$ matrices were generated to measure either $K = 4$ or 6 attributes. The choice for number of attributes measured was based on a study completed by Sessoms & Henson (2017). Results from the study show that there was wide range in the number of attributes measured, with as few as four attributes (e.g., Li & Suen, 2013a, 2013b) and as many as 23 attributes (e.g., Chen, 2012). The average number of attributes measured was eight ($M = 8.19$, median = 6.5, $SD = 4.95$). The most frequent number of attributes measured was four and eight. Specifically, 25% of the studies measured four attributes, while 19% measured eight attributes. In addition, when simulating the $\boldsymbol{Q}$ matrices, a set of constraints where placed such that each item measured no more than four attributes and each attribute was measured by no more than 40% of the survey

length. Furthermore, because the model being considered allows for ordinal responses, a survey length of 50 items was generated with each item having four response categories with corresponding sample space $X_{ij} \in \{0,1,2,3\}$. The survey length selected correspond to studies reported in the literature. For example, Henson & Templin (2009) used 40 items and seven attributes. Choi et al. (2010) considered $\boldsymbol{Q}$ matrix with 40 items and four attributes in their simulation experiment. Hansen's (2013) study considered 50 items in the simulation experiment. The choice for number response categories is based on common survey development practices e.g., a four-category case can have the following response options: "strongly disagree", "disagree", "agree", and "strongly agree" (e.g., Nering & Ostini, 2010). The probability mass function of the random variable X_{ij} is represented in Equation (3.18).

To obtain respondent's true latent class profiles, a sample size of $N = 5{,}000$ was first randomly generated from a multivariate normal distribution:

$$\boldsymbol{\alpha}_i' \sim MVN\left(\mathbf{0}, \boldsymbol{\Sigma}_{[K\times K']}\right) \tag{3.30}$$

with a mean vector of $\mathbf{0} = (0_1, \ldots, 0_k, \ldots, 0_K)$, and covariance matrix,

$$\boldsymbol{\Sigma}_{[K\times K']} = \begin{pmatrix} 1 & \cdots & \rho_{1K'} \\ \vdots & \ddots & \vdots \\ \rho_{K1} & \cdots & 1 \end{pmatrix} \tag{3.31}$$

which consisted of unit diagonal elements and correlations, $\rho \sim U(.5, .8)$, between the continuous latent variables $\boldsymbol{\alpha}_i' = (\alpha_{i1}', \ldots, \alpha_{ik}', \ldots, \alpha_{iK}')$. Once $\boldsymbol{\alpha}'$ was drawn, $\alpha_{ik} \in \{0,1\}$ and $\alpha_{ik} \in \{0,1,2\}$ was obtained by using the following thresholds:

$$\alpha_{ik} = \begin{cases} 0, & -\infty < \alpha'_{ik} < 0 \\ 1, & 0 \leq \alpha'_{ik} < \infty \end{cases}, \tag{3.32}$$

and

$$\alpha_{ik} = \begin{cases} 0, & -\infty < \alpha'_{ik} < -.5 \\ 1, & -.5 \leq \alpha'_{ik} < .5 \\ 2, & .5 \leq \alpha'_{ik} < \infty \end{cases}. \tag{3.33}$$

The sample size chosen for this study were based on prior research which has shown that although intercepts and main effects appeared to be estimated consistently in sample sizes of at least 500, higher sample sizes up to 4,000 are required to estimate two-way interactions reliably (Choi et al., 2010). Because the current study sought to additionally estimate three-way and four-way interaction effect parameters, even higher demands on sample size i.e., $N = 5{,}000$ were required for more reliable parameter estimates. The choice for fixing the mean vector to $\mathbf{0}$ was based on a prior simulation study done in Kunina-Habenicht, Rupp, & Wilhelm (2012). The choice for generating correlations $\rho \in [.5, .8]$ was based on typically reported correlations between subscores for subdomains in the national and international educational surveys (e.g., Sinharay, Puhan, & Haberman, 2011). Item parameters for the four submodels were generated in a way that would exhibit either low-quality items or high-quality items. The purpose for exploring item quality is because low-quality items could potentially impact rates of convergence in the estimation procedures and thus, contribute to larger computation times for the estimation algorithms. The item constraints were imposed on the P-LCDM-PA to emulate behavior

of the prototypical formulation corresponding to the P-DINA-PA, P-DINO-PA, or P-C-RUM-PA. For the P-LCDM-PA to reduce to an extension of the submodels, effect-level parameters λ (indicating main or interaction) were randomly generated as followed:

$$\lambda \sim \begin{cases} U(.5,1), & if\ low\ item\ quality \\ U(1,1.5), & if\ high\ item\ quality \end{cases} \tag{3.34}$$

The choice for generating λ this way is based similarly to von Davier (2005) where effects were randomly generated $\lambda \sim N(1,.25)$. The mean $\mu = 1$ for the normal distribution was used as threshold for determining low/high item quality. The lower bound (i.e., .5) for the uniform distribution under the low item quality condition and upper bound (i.e., 1.5) for the uniform distribution under the high item quality condition were used to represent approximately 45% of the generated items given that the standard deviation was set to $\sigma = .25$ in the normal distribution by von Davier (2005). For the intercept parameters $\lambda_{1,0}, \dots, \lambda_{c,0}, \dots, \lambda_{C-1,0}$ the following generating method was implemented where $\lambda_{1,0} \sim N(1,.5)$ and remaining intercept parameters were generated recursively using the following,

$$\lambda_{c,0} = \lambda_{c-1,0} - N(1,.2). \tag{3.35}$$

Software such as flexMIRT (Cai, 2017) has the capability of estimating a P-LCDM using the EM. However, because the SEM or MH-RM does not exist in any available software for estimating CDM, the SEM and MH-RM was implemented using a combination of functions written in both R and Fortran programming languages.

flexMIRT also has the capabilities for defining polytomous attribute levels "experimentally". To allow for a more direct comparison between the EM, SEM and MH-RM, such that biased results could be plausible if implementing the EM from flexMIRT (Cai, 2017), the author also coded the EM estimation algorithm using a combination of functions written in both R and Fortran programming languages. Random starting values for item parameters when using the EM, SEM, and MH-RM algorithms were obtained using the same item parameter generating methods discussed earlier. The convergence criteria for the EM and MH-RM was set to .0005 and .001, respectively. In the EM, the E-step had a maximum of 2,000 cycles while the M-step had a maximum of 500 cycles. Convergence criteria for the M-step was .0001. The SEM was set to have burn-in cycles $b = 800$ and retained post-burn-in used for estimation was $c = 200$. The MH-RM was set to have $b = 800$ burn-in cycles, $c = 200$ cycles to be average over for starting maximum likelihood estimates in the MH-RM, and a maximum of 2,000 cycles for the MH-RM. There was $M_t = 1$ MH draws of the $\boldsymbol{\alpha}$ parameters per cycle (e.g., refer to Cai, 2010b), and the *gain constant* (γ_t) for the MH-RM was computed as follows (Chalmers, 2012),

$$\gamma_t = \left(\frac{\varepsilon_1}{t}\right)^{\varepsilon_2} \tag{3.36}$$

where $\varepsilon_1 = .1$ and $\varepsilon_2 = .75 \;\; \forall t = 1,2,\ldots,2000$. In addition, for estimation using SEM and MH-RM, the probability transition matrices in the MH algorithm, $q\left(\boldsymbol{\alpha}_i^{(t)} | \boldsymbol{\alpha}_i^{(t-1)}\right)$, for each set of attribute levels was defined as follows,

$$q\left(\boldsymbol{\alpha}_i^{(t)}|\boldsymbol{\alpha}_i^{(t-1)}\right) =$$

$\boldsymbol{\alpha}_i^{(t-1)}\backslash\boldsymbol{\alpha}_i^{(t)}$	0	1
0	.7	.3
1	.3	.7

(3.37)

$$q\left(\boldsymbol{\alpha}_i^{(t)}|\boldsymbol{\alpha}_i^{(t-1)}\right) =$$

$\boldsymbol{\alpha}_i^{(t-1)}\backslash\boldsymbol{\alpha}_i^{(t)}$	0	1	2
0	.6	.2	.2
1	.2	.6	.2
2	.2	.2	.6

(3.38)

Tuning each element within $q\left(\boldsymbol{\alpha}_i^{(t)}|\boldsymbol{\alpha}_i^{(t-1)}\right)$ will adjust acceptance rates under the MH algorithm for $\boldsymbol{\alpha}$ accordingly in the SEM and MH-RM algorithms. The elements of $q\left(\boldsymbol{\alpha}_i^{(t)}|\boldsymbol{\alpha}_i^{(t-1)}\right)$ were chosen in a way such that the acceptance rates were between the acceptable values of .2 and .4 (e.g., refer to Junker, Patz & VanHoudnos, 2016). To help bring stability to the estimation procedures, partially informative priors were imposed on the item parameters of the P-LCDM-PA submodels. The prior distribution for λ was assumed to be log-normally distributed (e.g., refer to Junker, Patz & VanHoudnos, 2016): $\lambda \sim lognormal(1.19,1.09)$. The prior distribution for $\lambda_{1,0}$, $\lambda_{2,0}$, and $\lambda_{3,0}$ were all assumed to be normally distributed (e.g., refer to Baker & Kim, 2004): $\lambda_{1,0} \sim normal(1,2)$, $\lambda_{2,0} \sim normal(0,2)$, and $\lambda_{3,0} \sim normal(-1,2)$.

A main objective was to assess the computational efficiency of each approach relative to the amount time required to obtain estimates computation time (in minutes) was reported for the EM, SEM, and MH-RM algorithms. It is also possible that specific algorithms obtain estimates in less time but do not perform as well with respect to

estimation. To evaluate the general performance of each algorithm, the mean absolute difference (MAD) metric was used to assess overall recovery of the item parameters. The general form for calculating MAD is presented as follows,

$$MAD = \frac{\sum_{p=1}^{P} \left| \hat{x}_p - x_p \right|}{P}, \tag{3.39}$$

where x_p and $\hat{x}_p$ are the true and estimated p^{th} parameter, respectively, and P is the total number of item parameters (e.g., λ or λ_0). A maximum *a-posteriori* (MAP; Embretson & Reise, 2000) method was used to estimate respondents' latent classes $\boldsymbol{\alpha}$. A uniform prior distribution, $f(\boldsymbol{\alpha})$ was assumed. Correct classification rates (CCR) or probability of correct classification, $p(CC)$, were examined to determine the overall recovery of respondents estimated latent classes

$$CCR = \frac{\sum(\hat{\boldsymbol{\alpha}} = \boldsymbol{\alpha})}{N}. \tag{3.40}$$

Note that Equations (3.39) and (3.40) are computed within each replication. The results are summerized based on the average MAD and CCR across the 50 replications for each joint condition. An additional set of levels for K were included in a separate simulation experiment when evaluating computational time; $K = 5$, 7, and 8. For this simulation experiement, the smallest number of possible latent classes was $2^4 = 16$ while the largest number of possible latent classes was $3^8 = 6{,}561$. The reason for including these addition K levels was to see at which points did the SEM and MH-RM become computationally faster over the EM. Because some of the simulation conditions required

eight hours to complete a single replication (e.g., when 3^8 latent classes were present), only five replications per joint condition was used in reporting the results. To help bring stability to the estimation procedure, inparticular when 3^8 latent classes were present, informative priors were imposed on the item parameters of the P-LCDM-PA submodels. The prior distribution for λ was assumed to be log-normally distributed: $\lambda \sim lognormal(0, .5)$. The prior distribution for $\lambda_{1,0}$, $\lambda_{2,0}$, and $\lambda_{3,0}$ were all assumed to be normally distributed: $\lambda_{1,0} \sim normal(1,1)$, $\lambda_{2,0} \sim normal(0,1)$, and $\lambda_{3,0} \sim normal(-1,1)$. The simulation experiment was ran on a desktop with a 64-bit operating system using an Intel(R) Core(TM) i7-4790K CPU at 4GHz and 16 GB of installed memory (RAM).

The primary goal of this study was to utilize the P-LCDM-PA, which is a special case of the GPDM for polytomous attributes and then, due to the potential of exponentially increasing the number of latent classes, explore the feasibility and efficiency in addition to the quality of parameter estimation of the SEM and MH-RM algorithms relative to the EM algorithm. The SEM and MH-RM algorithms have been popular choices for estimating higher-dimensional latent variable models (e.g., refer to Diebolt & Ip, 1994a, 1994b; Cai, 2010a, 2010b; Monroe & Cai, 2014) but have never been used to estimate CDM. The SEM and MH-RM algorithms can be useful in the context of estimating CDM when the number of latent classes increases exponentially i.e., 2^K. This exponential increase is even more prominent when polytomous attribute levels are introduced in a diagnostic assessment because the number of latent classes is then represented as $\prod_{k=1}^{K} S_k$. The results of this study will be helpful to researchers and

practitioners who are interested in developing diagnostic assessments that may contain many attributes and/or dichotomous/polytomous attribute levels and are need more computationally efficient estimation algorithms

CHAPTER IV

RESULTS

This chapter will provide a summary of the results from the simulation study proposed in the previous chapter. This first section of this chapter will discuss the MAD results associated with recovery of the main and effect-level parameters. The second section of this chapter will discuss the MAD results associated with recovery of the intercept parameters. The third section of this chapter will discuss the CCR results. Finally, the last section of this chapter will provide a discussion of the results associated with computational time (in minutes) between the EM, SEM, and MH-RM algorithms. Recall, with respect to the comparison of SEM and MH-RM to the EM algorithm, the three research questions motivating this study are:

1) To what extent does the SEM and MH-RM algorithms show to be computationally faster over the EM algorithm as the number of latent classes increases?

2) How accurately are the item parameters of the P-LCDM-PA submodels estimated when comparing the SEM, MH-RM, and EM algorithms for estimation?

3) How accurately are examinees attributes (and attribute patterns) estimated when using the SEM, MH-RM, and EM algorithms to estimate the P-LCDM-PA submodels?

Recovery of Effect-Level Parameters

The results in Table 12 show the average MAD along with the corresponding standard deviations between true and estimate λ parameters across 50 replications, for each joint condition. Figure 1 displays the average MAD along with the corresponding standard deviations between true and estimate λ parameters across 50 replications, for each joint condition. Results from the simulation experiment showed that overall the EM presented better recovery of the λ parameters compared to the SEM and MH-RM. The average MAD indicated that recovery of the λ parameters between the SEM and MH-RM were very similar. The largest *MAD* difference between the SEM and MH-RM compared to the EM was when item quality was low, 3^6 latent classes were present, and the P-DINO-PA was used as the estimated model. The average MAD results indicated that recovery of the λ parameters is better when item quality is high compared to low item quality under the P-DINA-PA and P-DINO-PA. The difference in the MAD results between low and high item quality under the P-C-RUM-PA was minimal. It's speculated that the reason the results between low and high item quality under the P-C-RUM-PA were very similar was because of how effect-level parameters were generated. For example, results may vary more if there was a larger distinction between low and high performing items where $\lambda \sim U(.1, .5)$ for low item quality and $\lambda \sim U(1,1.5)$ for high item quality. As the number of measured attributes increased in the $\boldsymbol{Q}$ matrices, recovery of the λ parameters is poorer. Results from the simulation study showed that when dichotomous attribute levels were used, recovery of the λ parameters were better compared using

polytomous attribute levels. These overall results indicate that as the number of latent classes increases, recovery of the λ parameters is poorer.

Table 12. Average MAD with SD between True and Estimated Effect-level Parameters

Model	S	K	Low Item Quality			High Item Quality		
			EM	MH-RM	SEM	EM	MH-RM	SEM
P-C-RUM-PA	2	4	.12 (.01)	.14 (.02)	.14 (.02)	.11 (.02)	.13 (.03)	.13 (.02)
		6	.18 (.02)	.21 (.04)	.21 (.04)	.18 (.02)	.21 (.04)	.21 (.04)
	3	4	.15 (.01)	.18 (.03)	.18 (.03)	.13 (.02)	.17 (.04)	.17 (.04)
		6	.21 (.02)	.25 (.03)	.25 (.03)	.22 (.02)	.27 (.04)	.26 (.03)
P-DINA-PA	2	4	.10 (.01)	.11 (.01)	.11 (.01)	.07 (.01)	.07 (.01)	.07 (.01)
		6	.18 (.01)	.20 (.02)	.20 (.02)	.12 (.01)	.13 (.01)	.13 (.01)
	3	4	.14 (.02)	.19 (.04)	.19 (.04)	.11 (.01)	.14 (.02)	.14 (.02)
		6	.25 (.03)	.30 (.03)	.30 (.03)	.22 (.02)	.26 (.03)	.26 (.03)
P-DINO-PA	2	4	.09 (.01)	.12 (.03)	.12 (.03)	.07 (.01)	.08 (.01)	.08 (.01)
		6	.16 (.02)	.22 (.04)	.22 (.04)	.11 (.01)	.13 (.02)	.13 (.02)
	3	4	.13 (.01)	.20 (.03)	.21 (.03)	.10 (.01)	.15 (.03)	.15 (.03)
		6	.21 (.02)	.33 (.05)	.33 (.05)	.18 (.02)	.25 (.04)	.25 (.04)

Note. EM = expectation maximization algorithm; SEM = stochastic expectation maximization algorithm; MH-RM = Metropolis-Hastings Robbins-Monro algorithm; P-C-RUM-PA = polytomous compensatory reparameterized unified model for polytomous attributes; P-DINA-PA = polytomous deterministic input noisy "and" gate model; P-DINO-PA = polytomous deterministic input noisy "or" gate model; S = number of attribute levels; K = number of attributes

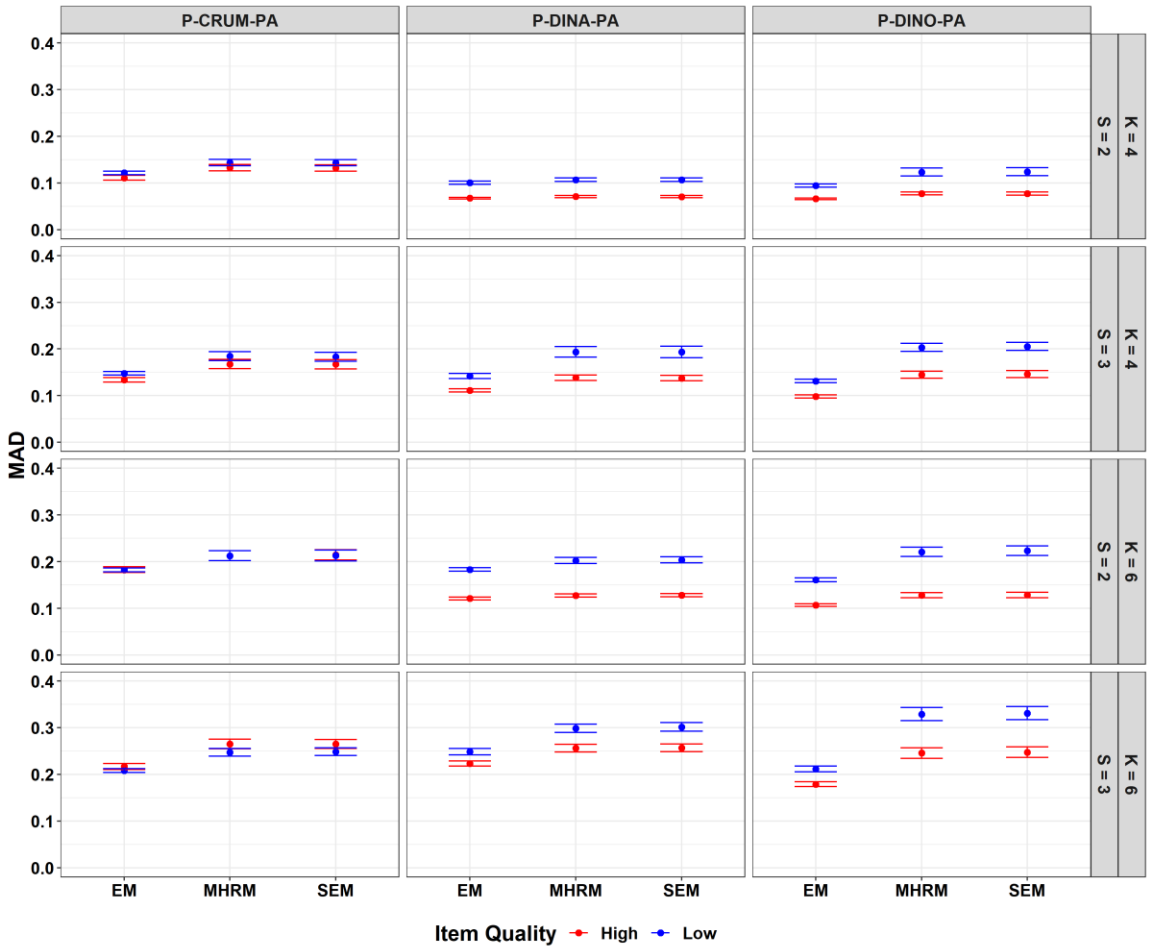

Figure 1. Average MAD with 95% Confidence Intervals between True and Estimated Effect-level Parameters

Recovery of Intercept Parameters

The results in Table 13 show the average MAD along with the corresponding standard deviations between true and estimate λ_0 parameters across the 50 replications, for each joint condition. Figure 2 displays the average MAD along with the corresponding 95% confidence intervals between true and estimate λ_0 parameters across the 50 replications, for each joint condition. Results from the simulation experiment showed that

overall the EM presented slightly better recovery of the λ_0 parameters compared to the SEM and MH-RM. The average MAD results showed that recovery of the λ_0 parameters between the SEM and MH-RM were very similar. The largest *MAD* difference between the SEM and MH-RM compared to the EM was when item quality was low, 3^6 latent classes were present, and the P-DINO-PA was used as the estimated model. The average MAD results indicated that recovery of the λ parameters is better when item quality is high compared to low item quality under the P-DINO-PA. The difference in the MAD results between low and high item quality under the P-DINA-PA and P-C-RUM-PA was minimal. It's speculated that the reason the results between low and high item quality under the P-DINA-PA were very similar was because estimating the intercept in the model is eqivialent to estimating the guess parameter in the DINA. The guess parameter in the DINA tends to estimate well because typically, it's more likely that the respodents randomly sampled from the population have not mastered all required attributes for a given item. Thus, more data is contributed to estimating the intercept parameters compared to the effect-level parameter. Again, it's speculated that the reason the results between low and high item quality under the P-C-RUM-PA were very similar was because of how effect-level parameters were generated. For example, results may vary more if there was a larger distinction between low and high performing items where $\lambda \sim U(.1, .5)$ for low item quality and $\lambda \sim U(1,1.5)$ for high item quality. As the number of measured attributes increases, the quality of parameter recovery of the λ_0 decreased. Results from the simulation study showed that when dichotomous attribute levels were used, recovery of the λ_0 parameters were better compared using polytomous attribute

levels. This result overall indicates that as the number of latent classes increases, recovery of the λ_0 parameters is poorer.

Table 13. Average MAD with SD between True and Estimated Intercept Parameters

Model	S	K	Low Item Quality			High Item Quality		
			EM	MH-RM	SEM	EM	MH-RM	SEM
P-C-RUM-PA	2	4	.08 (.01)	.11 (.02)	.11 (.02)	.08 (.01)	.10 (.02)	.10 (.03)
		6	.16 (.02)	.17 (.03)	.17 (.03)	.15 (.03)	.17 (.05)	.17 (.05)
	3	4	.10 (.01)	.13 (.02)	.13 (.02)	.09 (.01)	.12 (.03)	.12 (.03)
		6	.18 (.02)	.19 (.03)	.19 (.03)	.17 (.03)	.20 (.04)	.20 (.04)
P-DINA-PA	2	4	.05 (.00)	.05 (.01)	.05 (.01)	.05 (.00)	.05 (.01)	.05 (.01)
		6	.06 (.01)	.07 (.01)	.07 (.01)	.06 (.01)	.06 (.01)	.06 (.01)
	3	4	.05 (.01)	.09 (.02)	.09 (.02)	.06 (.01)	.08 (.01)	.08 (.01)
		6	.06 (.01)	.09 (.02)	.09 (.02)	.07 (.01)	.09 (.02)	.09 (.02)
P-DINO-PA	2	4	.09 (.01)	.10 (.02)	.10 (.02)	.07 (.01)	.08 (.01)	.08 (.01)
		6	.17 (.02)	.19 (.03)	.19 (.03)	.13 (.01)	.14 (.02)	.14 (.02)
	3	4	.11 (.01)	.15 (.02)	.15 (.02)	.09 (.01)	.12 (.03)	.12 (.03)
		6	.21 (.02)	.25 (.04)	.25 (.04)	.19 (.02)	.21 (.03)	.21 (.03)

Note. EM = expectation maximization algorithm; SEM = stochastic expectation maximization algorithm; MH-RM = Metropolis-Hastings Robbins-Monro algorithm; P-C-RUM-PA = polytomous compensatory reparameterized unified model for polytomous attributes; P-DINA-PA = polytomous deterministic input noisy "and" gate model; P-DINO-PA = polytomous deterministic input noisy "or" gate model; S = number of attribute levels; K = number of attributes

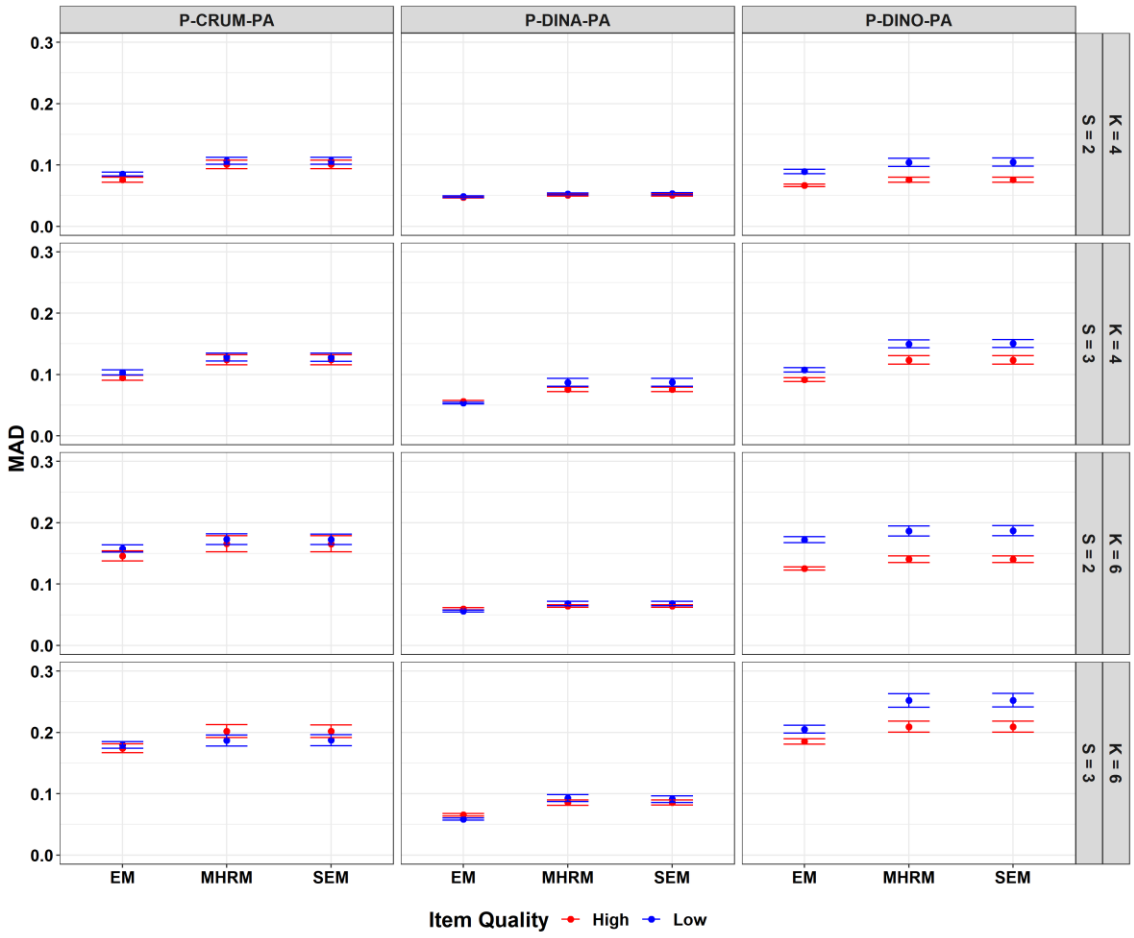

Figure 2. Average MAD with 95% Confidence Intervals between True and Estimated Intercept Parameters

Correct Classification Rates

The results in Table 14 show the average CCR (i.e., $p(CC)$) along with the corresponding standard deviations between true and estimated $\boldsymbol{\alpha}$ for each joint condition. Figure 3 show the average CCR along with the corresponding 95% confidence intervals between true and estimated $\boldsymbol{\alpha}$ for each joint condition. Results from the simulation experiment showed that the $p(CC)$ is similar across the EM, SEM, and MH-RM algorithms for each joint condition. The results indicated that the $p(CC)$ is higher when

item quality is high compared to low item quality. The $p(CC)$ decreased as the number of latent classes increased. In addition, the $p(CC)$ was higher for dichotomous attribute levels compared to polytomous attribute levels. As the number of measured attributes increased, $p(CC)$ was negatively impacted. The $p(CC)$ was the highest when item quality was high and there were the fewest, 2^4, latent classes. The $p(CC)$ was the lowest when item quality was low and there were the maximum, 3^6, latent classes.

Table 14. Average CCR with SD between True and Estimated Attribute Profiles

			Low Item Quality			High Item Quality		
Model	S	K	EM	MH-RM	SEM	EM	MH-RM	SEM
P-C-RUM-PA	2	4	.81 (.01)	.81 (.01)	.81 (.01)	.91 (.01)	.91 (.01)	.91 (.01)
		6	.79 (.01)	.79 (.01)	.79 (.01)	.88 (.01)	.87 (.01)	.87 (.01)
	3	4	.62 (.01)	.61 (.01)	.61 (.01)	.75 (.01)	.75 (.01)	.75 (.01)
		6	.60 (.01)	.60 (.01)	.59 (.01)	.71 (.01)	.70 (.01)	.70 (.01)
P-DINA-PA	2	4	.75 (.01)	.75 (.01)	.75 (.01)	.87 (.01)	.87 (.01)	.87 (.01)
		6	.68 (.01)	.68 (.01)	.68 (.01)	.78 (.01)	.78 (.01)	.78 (.01)
	3	4	.54 (.01)	.54 (.01)	.54 (.01)	.68 (.01)	.68 (.01)	.68 (.01)
		6	.48 (.01)	.47 (.01)	.47 (.01)	.58 (.01)	.58 (.01)	.58 (.01)
P-DINO-PA	2	4	.75 (.01)	.74 (.01)	.74 (.01)	.87 (.01)	.87 (.01)	.87 (.01)
		6	.68 (.01)	.67 (.01)	.67 (.01)	.79 (.01)	.79 (.01)	.79 (.01)
	3	4	.55 (.01)	.54 (.01)	.54 (.01)	.68 (.01)	.68 (.01)	.68 (.01)
		6	.48 (.01)	.47 (.01)	.47 (.01)	.59 (.01)	.58 (.01)	.58 (.01)

Note. EM = expectation maximization algorithm; SEM = stochastic expectation maximization algorithm; MH-RM = Metropolis-Hastings Robbins-Monro algorithm; P-C-RUM-PA = polytomous compensatory reparameterized unified model for polytomous attributes; P-DINA-PA = polytomous deterministic input noisy "and" gate model; P-DINO-PA = polytomous deterministic input noisy "or" gate model; S = number of attribute levels; K = number of attributes

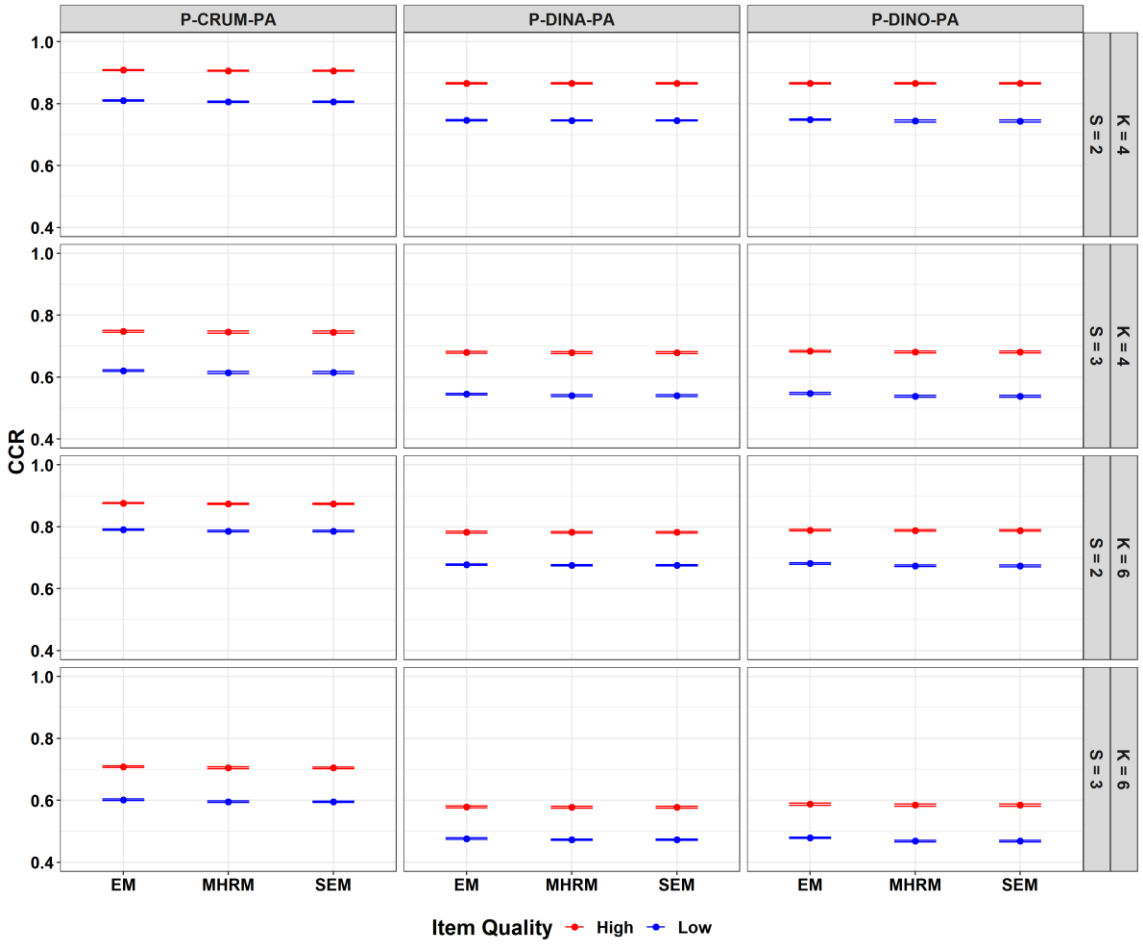

Figure 3. Average CCR with 95% Confidence Intervals between True and Estimated Attribute Profiles

Computational Time

The results shown in Table 15 provide a summary of average computational time and standard deviations between the EM, SEM, and MH-RM algorithms across all replications within a joint condition. Figures 4, 5, and 6 show the computational time results under the P-C-RUM-PA, P-DINA-PA, and P-DINO-PA, respectively. Because it was nearly impossible to obtain convergence when the SEM and MH-RM was used to estimate the P-DINA-PA when 3^8 latent classes were present, only one replication was

done. The results from the simulation experiment showed to be consistent across the three submodels. Computational time for the SEM and MH-RM increased slightly as the number of latent classes increases. The maximum computational time for the SEM and MH-RM was approximately one hour when there were 3^8 latent classes. For the EM, computational time increased exponentially compared to the SEM, and MH-RM when the number of latent classes increased to either 3^7 or 3^8. When the number of latent classes was 3^8, the EM required approximately six hours to complete. The EM also showed to be computationally slower when estimating the P-DINA-PA when there were 3^6 latent classes compared the SEM and MH-RM. In addition, low item quality showed to impact convergence rates (i.e., the algorithms took more iterations to converge) in the EM, SEM, and MH-RM compared to high item quality. For the EM, the differences in the results between low and high quality test items showed to increase as the number of latent classes increased. For the SEM and MH-RM, the differences in the results between low and high quality test items showed to be minimally impacted as the number of latent classes increased. For the P-DINO-PA, the EM showed to be computationally slower under the high item quality condition compared to the low item quality condition when there were 3^8 latent classes present. When estimating the P-DINA-PA and P-C-RUM-PA, computational time was slower under the low item quality condition compared to the high item quality condition when 3^7 possible latent classes were present.

Table 15. Average Computational Time (in minutes) with SD

Model	S	K	Low Item Quality			High Item Quality		
			EM	MH-RM	SEM	EM	MH-RM	SEM
P-C-RUM-PA	2	4	.37 (.05)	10.22 (.11)	9.78 (.22)	.37 (.08)	10.06 (.40)	9.73 (.50)
		5	.66 (.04)	15.20 (.31)	14.44 (.22)	.71 (.09)	15.46 (.23)	14.50 (.16)
		6	1.36 (.26)	25.28 (.85)	23.77 (.45)	1.32 (.16)	24.89 (.51)	23.41 (.49)
		7	4.89 (3.15)	41.68 (.80)	39.10 (.95)	4.48 (1.20)	40.82 (1.28)	38.36 (.53)
		8	15.50 (3.09)	75.66 (1.71)	72.28 (4.43)	13.24 (3.46)	81.07 (12.20)	77.82 (14.96)
	3	4	1.28 (.24)	10.02 (.12)	9.66 (.19)	1.12 (.18)	9.71 (.12)	9.29 (.10)
		5	4.92 (2.12)	14.73 (.31)	13.98 (.29)	3.19 (.28)	14.90 (.24)	13.93 (.21)
		6	15.45 (1.24)	24.80 (.67)	23.27 (.67)	13.59 (1.13)	25.10 (.88)	23.17 (.77)
		7	84.06 (25.73)	39.01 (.63)	36.80 (.70)	75.35 (22.62)	41.61 (.85)	38.30 (.93)
		8	339.55 (62.41)	74.23 (3.74)	69.18 (3.83)	295.82 (43.95)	76.19 (2.39)	68.11 (1.09)
P-DINA-PA	2	4	.30 (.01)	10.05 (.44)	9.26 (.35)	.22 (.01)	9.19 (.30)	8.75 (.12)
		5	.54 (.03)	13.70 (.18)	13.06 (.17)	.39 (.05)	13.73 (.11)	13.24 (.09)
		6	1.34 (.08)	23.20 (.19)	21.89 (.16)	.85 (.09)	22.21 (.34)	21.25 (.14)
		7	3.35 (.31)	37.96 (.80)	35.55 (.84)	2.17 (.27)	36.74 (.70)	35.38 (.39)
		8	12.07 (1.38)	70.16 (1.11)	66.13 (1.59)	8.67 (.85)	74.48 (8.23)	72.55 (7.99)
	3	4	1.51 (.28)	9.15 (.13)	8.73 (.05)	.92 (.07)	9.11 (.15)	8.67 (.09)
		5	5.67 (1.55)	13.83 (.17)	12.98 (.15)	3.00 (.42)	13.76 (.28)	12.97 (.22)
		6	21.90 (3.19)	22.24 (.81)	20.52 (.60)	12.28 (1.22)	22.41 (.38)	20.74 (.28)
		7	102.21 (29.79)	36.33 (.78)	33.52 (.10)	63.35 (15.55)	38.74 (2.16)	35.75 (1.63)
		8	435.80 (-)	71.15 (-)	67.04 (-)	258.89 (-)	70.39 (-)	65.18 (-)
P-DINO-PA	2	4	.35 (.05)	9.35 (.21)	9.20 (.44)	.26 (.03)	9.01 (.11)	8.87 (.21)
		5	.68 (.12)	13.76 (.47)	13.56 (.53)	.49 (.08)	13.28 (.20)	12.96 (.17)
		6	1.33 (.14)	22.88 (2.31)	21.27 (.14)	.95 (.17)	21.66 (.44)	21.10 (.33)

	7	3.73 (.47)	35.77 (.25)	35.13 (.26)	2.55 (.38)	35.74 (.92)	35.10 (.34)
	8	13.86 (1.12)	70.33 (8.82)	62.78 (1.49)	11.48 (1.78)	72.65 (10.23)	65.23 (3.02)
	4	1.40 (.13)	9.54 (1.15)	8.65 (.11)	.94 (.07)	8.81 (.07)	8.56 (.13)
	5	4.64 (1.35)	13.10 (.12)	12.86 (.10)	3.45 (.56)	13.15 (.20)	12.91 (.10)
3	6	14.77 (1.37)	22.71 (2.00)	20.79 (.33)	13.03 (.94)	21.12 (.60)	20.40 (.58)
	7	87.88 (17.20)	37.21 (3.12)	34.00 (.66)	59.46 (8.35)	35.50 (1.08)	34.03 (.93)
	8	278.36 (15.75)	68.88 (9.36)	63.60 (1.02)	301.14 (28.04)	69.30 (10.22)	65.35 (7.26)

Note. EM = expectation maximization algorithm; SEM = stochastic expectation maximization algorithm; MH-RM = Metropolis-Hastings Robbins-Monro algorithm; P-C-RUM-PA = polytomous compensatory reparameterized unified model for polytomous attributes; P-DINA-PA = polytomous deterministic input noisy "and" gate model; P-DINO-PA = polytomous deterministic input noisy "or" gate model; S = number of attribute levels; K = number of attributes

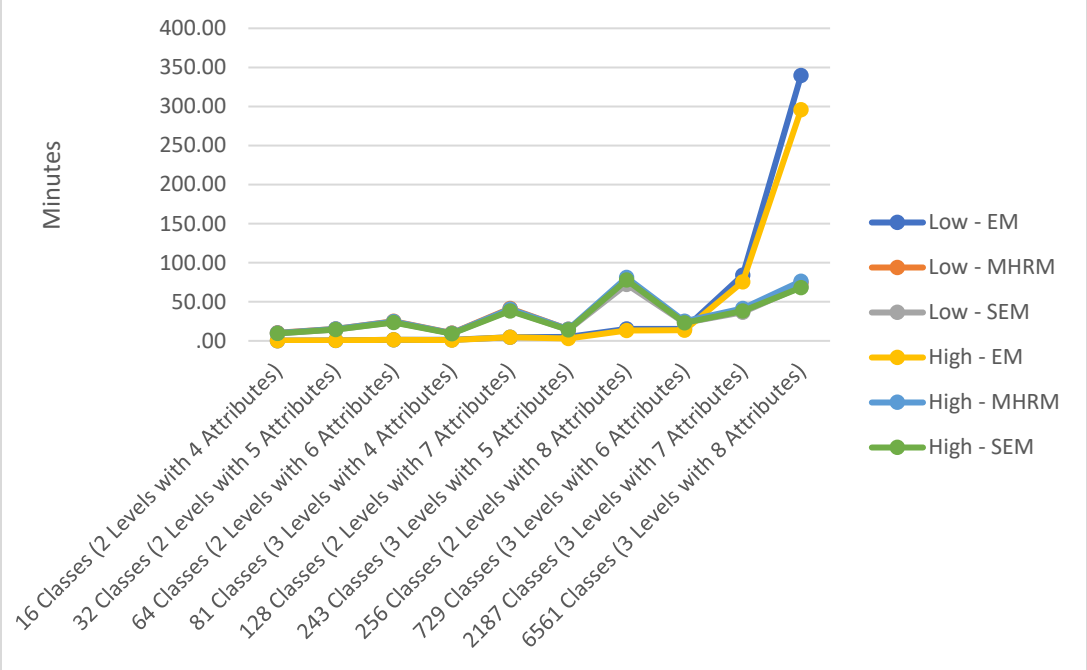

Figure 4. Average Computational Times (in minutes) for the P-C-RUM-PA

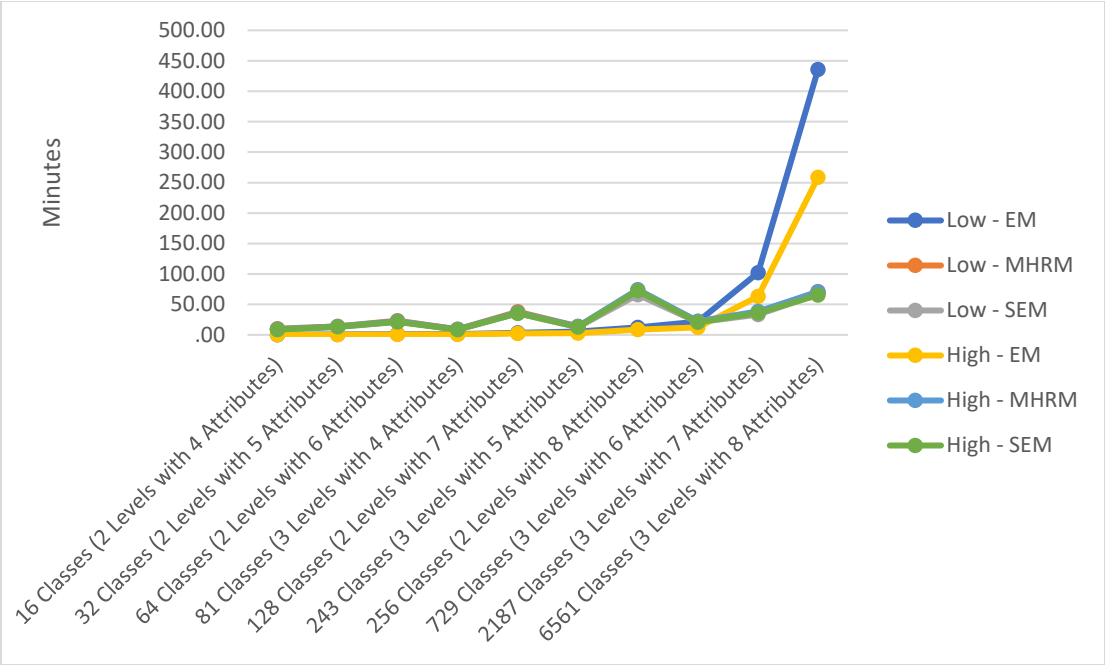

Figure 5. Average Computational Times (in minutes) for the P-DINA-PA

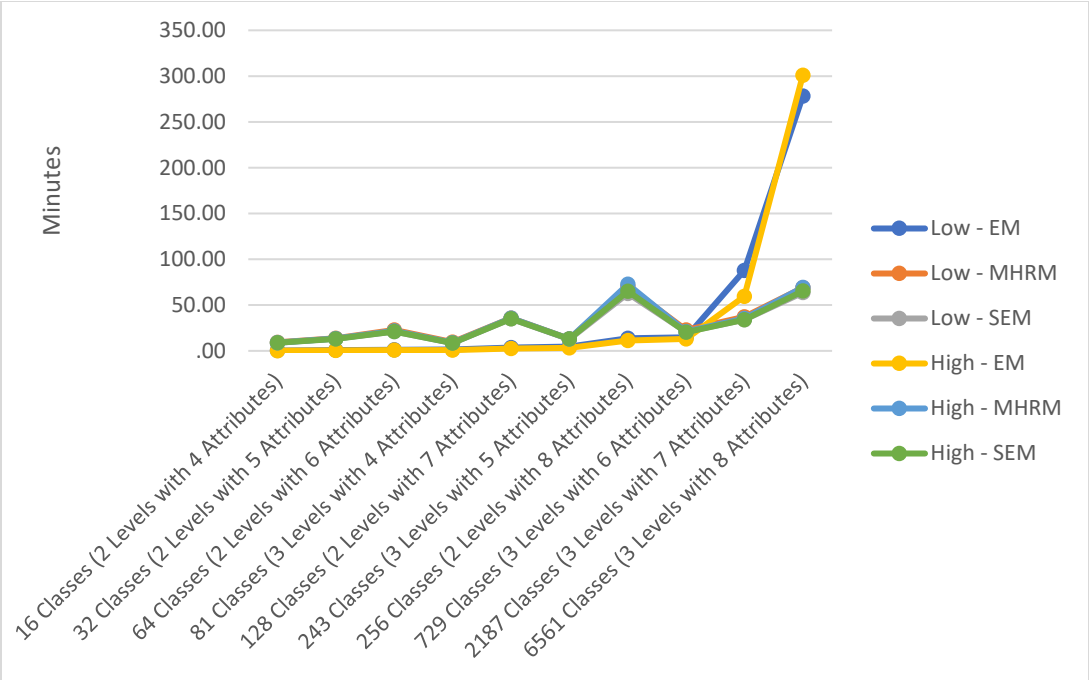

Figure 6. Average Computational Times (in minutes) for the P-DINO-PA

CHAPTER V

DISCUSSION

Conclusions

The primary goal of this study was to utilize polytomous attributes in the polytomous log-linear cognitive diagnosis model (P-LCDM-PA), which is a special case of the general polytomous diagnostic model (GPDM) for polytomous attributes. Then, due to exponentially increasing the number of latent classes, explore the feasibility and efficiency in addition to the quality of parameter estimation of the stochastic expectation-maximization (SEM) and Metropolis-Hastings Robbins-Monro (MH-RM) algorithms relative to the EM algorithm. Recall, with respect to the comparison of SEM and MH-RM to the EM algorithm, the three research questions motivating this study are:

1) To what extent does the SEM and MH-RM algorithms show to be computationally faster over the EM algorithm as the number of latent classes increases?

2) How accurately are the item parameters of the P-LCDM-PA submodels estimated when comparing the SEM, MH-RM, and EM algorithms for estimation?

3) How accurately are examinees attributes (and attribute patterns) estimated when using the SEM, MH-RM, and EM algorithms to estimate the P-LCDM-PA submodels?

Results from the simulation experiment overall showed that the EM recovered the item parameters λ and λ_0 of the P-LCDM-PA submodels better than the SEM and MH-RM. Item parameter recovery between the SEM and MH-RM was very similar. Recall that both the SEM and MH-RM stages I and II used $b = 800$ and $c = 200$ cycles. It's speculated that convergance was obtained in the SEM cycles of the MH-RM, so the additional stage was unnecessary, thus why item parameter recovery results were not better in the MH-RM. The EM and MH-RM have been compared in previous studies by Cai (2010a, 2010b) and Chalmers and Flora (2014) when estimating MIRT models. The results from those research studies have shown that the EM and MH-RM produce very similar item parameter estimates. It's a possiblilty that the MH-RM and SEM has a harder time estimating models with discete latent variables compared to models with continuous latent variables. For the current study, the possible difference in item parameter recovery results between the EM and MH-RM/SEM could be due to convergence criteria. For example, running a burnin of greater than 1,000 could of improved the results in the SEM and MH-RM. Another possibility for differences in item parameter recovery is the fact that the SEM and MH-RM are stochastic algorithms, whereas the EM is a deterministic algorithm. Different convergence criteria were used for the EM and MH-RM. The reason for this was from results obtained during preliminary analysis of the algorithms. Initially, a convergence criteria of .0001 was used for both the EM and MH-RM. The EM showed to have major issues reaching that convergence criteria of .0001 when the number of latent classes became relatively large (e.g., 3^6 latent classes). The convergence criteria for the EM was then increased to .0005. The MH-RM

showed to have major issues reaching that convergence criteria of .0001 under most conditions. This could possibly be due to the stochastic nature of the algorithm. Because of the stochastic nature of the algorithm and convergence issues with such a small criteria, the convergence criteria was then set to .001. This criteria is default in the mirt package (Chalmers, 2012) when using the MH-RM and is also used in the simulation experiment done by Chalmers and Flora (2014). The preliminary analysis also compared item parameter recovery results from the MH-RM using a convergance criteria of .001 and .0005. Results showed that the smaller convergance criteria of .0005 did not improve item parameter recovery compared to using a convergance criteria of .001. Again, it's speculated that convergance was obtained during the SEM cycles of the algorithm, so the additional stage with a set convergance criteria was unnecessary. The probability of correct classification when using all three estimatioin methods, the EM, SEM, and MH-RM, was very similar. As might be expected, high item quality improved item parameter recovery and classification rates when compared to low item quality. As the number of latent classes increased, item parameter recovery and classification rates were negatively impacted across all submodels of the P-LCDM-PA. Specifically, recovery of the item parameters and classification rates were improved when dichotomous attribute levels were used compared to using polytomous attribute levels. As the number of measured attributes increased so does the number of classes and so the model becomes more complex without any addition to the number of items. These results were consistent across all three estimation algorithms.

Results examining computational time showed that the EM was computationally sufficient when the number of possible latent classes was relatively low. However, as the number of latent classes increased to either 3^7 or 3^8, the SEM and MH-RM showed to be computationally faster than the EM. Even though the SEM and MH-RM showed to be computational faster when 3^7 or 3^8 latent classes were present, this did not deteriate from parameter recovery performace of these algorithms compared to the EM. Again, the EM becomes computationally slow because it always has to directly compute the probability of class membership for each respondent. There were some cases where the SEM and MH-RM were slightly computationally faster when polytomous attributes were used compared to dichotomous attributes, e.g., 2^4 compared to 3^4. It's speculated that this is due to the low number of replications used for the compulational time experiment. If the number of replications were to increased to 50 or 100, this behavior would possibly disappear. Generally, lower item quality overall negatively impacted convergence (i.e., the algorithms took more iterations to reach convergence) rates for the EM, SEM, and MH-RM compared to high item quality. The negative impact of low item quality was more profound in the EM compared the SEM and MH-RM. When the EM was used to estimate the P-DINO-PA under the 3^8 latent classes, the low item quality condition showed to be computationally faster than the high item quality condition. It's speculated that this is due to the low number of replications used for the computational time experiment. Increasing the number of replications may dissolve this behavior. To better interpret what an MAD result of .2, .21, and .21 for λ_0 in the EM, SEM, and MH-RM, respectively, let's using an example where the estimated model was the P-DINO-PA.

Recall that the λ_0 is used to define the conditional probability of a response to the c^{th} category or higher for the reference group, which is defined by the respondents who have mastered none of the measured attributes. This conditional probability from the model is equivalent to the guess parameter in the prototypical formulation of the DINO (or DINA). The equivalency of an MAD result of .2, .21, and .21 average difference between true and estimated λ_0 would result in .04, .04, and .04 average difference between the true and estimated guess parameters, respectively.

Implications

The results of this study will have implications towards researchers and practitioners who are interested in developing diagnostic assessments that may contain many attributes and/or dichotomous/polytomous attribute levels and are need more computationally efficient estimation algorithms. Recall the study completed by Sessoms & Henson (2017) examined the number of measured attributes for real application of diagnostic assessments. Results from their study showed that there was wide range in the number of attributes measured, with as few as four attributes (e.g., Li & Suen, 2013a, 2013b) and as many as 23 attributes (e.g., Chen, 2012). The average number of attributes measured was eight ($M = 8.19$, median = 6.5, $SD = 4.95$). The most frequent number of attributes measured was four and eight. Specifically, 25% of the studies measured four attributes, while 19% measured eight attributes. In reality, dichotomous attributes may not be what reasearchers and practiconers are wanting in their diagnostic measures. For example, the science field tends to focus on learning progressions that could be modelled as polytomous attributes, which in turn, can increase the complexity of the models used.

As modeling situations like this comes up, there is a push for fast computers, where the current study brings to the forefront that alternative parameter estimation methods could be used to circumvent these problems. There also can be motivation to using polytomous attributes over continuous latent variable models such as MIRT. A motivation for this could be the case where a researcher or practicioner wants to use CDM but dichotomous attribute definitions of mastery/non-mastery are too limited. Polytomous attributes are a way to stay within specific stages e.g., non-mastery, partial mastery, and mastery, while not adapting fully continuous latent variables into the model.

Recommendations

Typically, larger samples sizes are needed to estimate complex models such as the one studied in this dissertation, which is especially true when estimating higher-order interaction parameters in the P-DINA-PA. The SEM and MH-RM show evidence of being computationally efficient as the number of latent classes is large compared to the EM. Generally speaking, the EM is a sufficient algorithm to use when the number of latent classes is relatively low, whereas the SEM and MH-RM serve no advantage in these scenarios. In fact, when the number of latent classes is low, SEM and MH-RM resulted in high MADs with respect to item parameters estimation. There is also the issue with classification rates becoming poorer when increasing the number of possible latent classes. So, this is something researchers and practitioners need to keep in mind when developing diagnostic assessments that measure many attributes and/or incorporate polytomous attribute levels. A way of circumventing this issue would be to increase the number of items and/or response categories in a diagnostic assessment. This provides

more statistical information about the items and respondents, thus reducing the amount of statistical error associated with item parameter estimates and classifications of individuals. As a result, it is recommend for researchers and practitioners to construct some items in a diagnostic assessment that are open ended responses (i.e., allows for partial credit) instead of all multiple-choice responses. There are cases were the CCR were around .5 when the estimated model was either the P-DINA-PA or P-DINO-PA under 3^6 latent classes and low performing items. A CCR around .5 indicates that there's weak statistical information about a respondents true classification. It would be advised that cases when polytomous attributes are used in combination with an assessment measuring many attributes, results from estimating the P-DINA-PA and P-DINO-PA should be used with caution. Another important results is that even through the item parameters recovered differently between the EM compared to the SEM and MH-RM, the CCR results were very similar between the three algorithms. Thus, if a researcher or practitioners is more concerned with classifications of individuals rather than the item parameter results, any of the three methods would be equally sufficient to use when estimating the P-LCDM-PA. Another important recommendation for researchers and practitioners is to use a MAP approach where the $f(\boldsymbol{\alpha})$ is estimated instead of treated as uniform. The reason for this is that treating the probability of class membership, $f(\boldsymbol{\alpha})$ as being estimated instead of uniform could improve classification rates in respondents i.e., reducing the amount of statistical error associated around the respondents estimated $\boldsymbol{\alpha}$ parameters. Note that when $f(\boldsymbol{\alpha})$ is estimated in calibration procedure, we have statistical information about the population, thus improving the amount of statistic information we

have about the respondents. When estimation is completed using the SEM and MH-RM, $f(\boldsymbol{\alpha})$ can be estimated using the accept/reject sampling method of a MH algorithm, or via Gibbs sampler using a combination of a multinomial and Dirichlet distributional priors. For the EM, obtaining an estimate of $f(\boldsymbol{\alpha})$ is simply done by averaging the posterior probabilities across the sample for each possible latent class.

Limitations

While the results of the current study provide important implications for resesearchers and practicioners, there are a few limitations that were presented in this research study. The number of replications per joint condition could be higher (e.g., 100 replications per condition). However, for the current study, increasing the number of replications would have been very difficult for conditions with 3^7 or 3^8 possible latent classes present due to computational time associated with the EM. The SEM and MH-RM showed to be less problematic because it took an average of one hour or less to complete. Also, the standard deviation of the MAD results for item parameters and classification rates across the 50 replications with a condition was small across all simulation conditions. This result indicated that the results were consist across the 50 replications for each joint condition. Thus, increasing the number of replications per joint condition is not expected to change the results and may not be necessary. Using more informative priors could have improved item parameter recovery, classification rates, and computational time. However, the results from the simulation study (e.g., item parameter recovery) would have been more influenced by the priors during estimation procedures rather than the response data. Another limitation of the current study is that sample size

and test length was not varied as a factor of the simulation design. Item parameter recovery, classification rates, and computational time was not analyzed when smaller sample sizes and test lengths are present. Even with a sample size of $N = 5{,}000$, item parameter recovery was poor under certain conditions. Sample sizes up to $N = 10{,}000$ may be needed for adequate parameter recovery when there are many attributes measured and/or polytomous attribute levels are present in a diagnostic assessment. The challenge of such a sample size requirement is that it narrows the range of applications of such an approach to Larger testing programs such as American College Testing and Educational Testing Services. Whereas many smaller scale applications could really benefit from such a modeling approach. Item parameter recovery was overall poorer when the SEM and MH-RM were used. This difference could have potentially been caused in part, by the method used to generate item intercepts in the simulation study. For example, in some preliminary analysis, simulated intercepts that resulted in $\lambda_0 = (.5, -1, -2.5)$ compared to $\lambda_0 = (1{,}0, -1)$ showed to negatively impact item parameter recovery in the SEM and MH-RM Possible reasons for this could be an issue with how the item parameter priors was specified or problems with the complete-data gradients. The EM consistently used all possible latent classes in the estimation of the models. Another limitation is that the respondent's estimated latent classes were obtained using a uniform prior, $f(\boldsymbol{\alpha})$ in the MAP scoring procedure. Note that a Bayesian approach to classification (i.e., estimation of attribute profiles) can result in higher correct classification rates when the prior is a reasonable approximation of the true distribution of the attribute profile. As a result, the $p(CC)$ is lower when $f(\boldsymbol{\alpha})$ is assumed to be uniform compared to being estimated and

imputed in the scoring procedure. Thus, it is likely $p(CC)$ could have been higher if $f(\boldsymbol{\alpha})$ was estimated in the calibration procedure. Another possible limitation is that the author coded the program and did the best he could to make sure things were running accurately, but there may be some issues and efficiency with the program.

Directions for Future Research

Several future research directions can be explored in relation to this research study. Future research could explore parameter recovery and computational time when smaller sample sizes and survey lengths are used. Typically, a sample size of $N = 5{,}000$ is not realistic in many small-scale testing situations. For the current study, a partially informative prior was used for the item parameters. Future research could explore how influential informative vs. noninformative priors impact item parameter recovery, classification rates, and computational time in relation to the P-LCDM-PA. An informative prior may improve item parameter recovery, classification rates, and computational time, however, the drawback to this is that the data will have less influence on the results obtained from the estimation procedure. This can be more problematic when real data study is implemented because we don't know true values and thus a reasonable prior cannot easily be determined. A direction for future research could include evaluating how random starting values compared to "good" starting values impact convergence rates in the algorithms. Future researchers could explore how the general difficulty of a test impacts item parameter recovery, CCR, and computational time, especially in the SEM and MH-RM. Future research can use the SEM and MH-RM to estimate other CDMs including GDM, G-DINA, unconstrained P-LCDM-PA,

dichotomous LCDM and its submodels. The increase in the number of latent classes as a function of measured attributes in a diagnostic assessment is also applicable in other CDM as well. Another important direction for future research could be comparing other methods to the SEM and MH-RM for reducing computational time such as the OCAC framework presented in Karelitz (2004), defining a higher-order continuous factor, or forming hierarchical structures. These methods could possibly be combined with the SEM and MH-RM to further improve computational time.

Summary

In summary, the primary goal of this study was to utilize polytomous attributes in the P-LCDM-PA, which is a special case of the GPDM for polytomous attributes. Then, due to exponentially increasing the number of latent classes, explore the feasibility and efficiency in addition to the quality of parameter estimation of the SEM and MH-RM algorithms relative to the EM algorithm. The SEM and MH-RM algorithms may be more computationally advantageous over an EM algorithm when there exist many latent classes. As the number of measured attributes increases in a diagnostic assessment, the number of latent classes increases exponentially. The large number of classes is even more problematic when polytomous attribute levels are introduced in the diagnostic assessment. This study will provide researchers and practitioners interested in developing diagnostic assessments that may contain many attributes and/or dichotomous/polytomous attribute levels and are need more computationally efficient estimation algorithms.

REFERENCES

Baker. F. B., & Kim, S-H. (2004). *Item Response Theory: Parameter Estimation Techniques.* Boca Raton, FL: Taylor and Francis Group.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46*(4), 443-459.

Cai, L. (2010a). High-dimensional exploratory item factor analysis by a Metropolis-Hastings Robbins-Monro algorithm, *Psychometrika, 75*, 33-57.

Cai, L. (2010b). Metropolis-Hastings Robbins-Monro algorithm for comfirmatory item factor analysis. *Journal of Educational and Behavioral Statistics, 25*, 207-335.

Cai, L. (2017). flexMIRT®: Flexible multilevel multidimensional item analysis and scoring (version 3.51) [Computer software]. Chapel Hill, NC: Vector Psychometric Group.

Chalmers, P. R. (2012), mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software, 48*, 1-29.

Chalmers, P. R., & Flora, D. B. (2014). Maximum-likelihood estimation of noncompensatory IRT models with the MH-RM algorithm. *Applied Psychological Measurement, 38*, 339-358.

Chen, J., & de la Torre, J. (2013). A general cognitive diagnostic model for expert-defined polytomous attributes. *Applied Psychological Measurement, 37*, 419-437.

Chen, J., & de la Torre, J. (2018). Introducing the general polytomous diagnosis modeling framework. *Frontiers in Psychology*, 9:1474.

Chen, Y. (2012). Cognitive diagnosis of mathematics performance between rural and urban students in Taiwan. *Assessment in Education: Principles, Policy, & Practice, 19,* 193-209.

Chen, Y., Ferron, J. M., Thompson, M. S., Gorin, J. S., & Tatsuoka, K. K. (2010). Group comparisons of mathematics performance from a cognitive diagnostic perspective. *Educational Research and Evaluation*, *16,* 325-343.

Choi, H. J., Templin, J. L., Cohen, A. S., & Atwood, C. H. (2010, April). *The impact of model misspecification on estimation accuracy in diagnostic classification models*. Paper presented at the meeting of the National Council on Measurement in Education (NCME), Denver, CO.

Choi, K. M., Lee, Y., Park, Y. S. (2015). What CDM can tell about what students have learned: An analysis of TIMSS eighth grade mathematics. *Eurasia Journal of Mathematics, Science, & Technology Education*, *11*, 1563-1577.

von Davier, M. (2005). A general diagnostic model applied to language testing data. ETS Research Report No. RR-05-16.

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

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

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

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika, 76*, 179-199.

de la Torre, J., Lam, D., Rhoads, K., & Tjoe, H. (2010). *Measuring grade 8 proportional reasoning: The process of attribute identification and task development and validation*. Paper presented at the annual meeting of the American Educational Research Association, Denver, CO.

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

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood estimation from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society—Series B, 39*, 1–38.

DiBello, L. V., Henson, R. A., & Stout, W. F. (2015). A family of generalized diagnostic classification models for multiple-choice option-based scoring. *Applied Psychological Measurement, 39*, 62-79.

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

Diebolt, J., & Ip. E. H. (1994a). *A stochastic EM algorithm for approximating the maximum likelihood estimate* (Technical Report No. 301). Stanford, CA: Stanford University.

Diebolt, J., & Ip. E. H. (1994b). A stochastic EM: Method and application. In W. Gilks, S. Richardson, & D. Spiegel-halter (Eds.), *Markov chain Monte Carlo in practice* (pp. 259-273). London, England: Chapman & Hall.

Embretson, S. E., & Reise, S. P. (2000). Item Response Theory for Psychologists. Erlbaum.

Gibbons, R. D., Bock, R. D., Hedeker, D., Weiss, D. J., Segawa, E., Bhaumik, D. K., . . . Grochocinski, V. J. (2007). Full-information item bifactor analysis of graded response data. *Applied Psychological Measurement*, *31*, 4-19.

Grünewald, M., Humphreys, K., & Hössjer, O. (2010). A stochastic EM type algorithm for parameter estimation in models with continuous outcomes, under complex ascertainment. *International Journal of Biostatistics, 6,* Article 23.

Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement*, *26*, 333–352.

Hagenaars, J. A. (1990). *Categorical longitudinal data: loglinear panel, trend, and cohort analysis*. Thousand Oaks: Sage.

Hagenaars, J. A. (1993). *Loglinear models with latent variables*. Thousand Oaks: Sage.

Hansen, M. P. (2013). *Hierarchical item response model for cognitive diagnosis*. Unpublished doctoral dissertation, University of California Los Angeles.

Hastings, W. K. (1970). Monte Carlo simulation methods using Markov chains and their applications, *Biometrika*, *57*, 97-109.

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

Henson, R. A., & Templin, J. L. (2009). Implications of *Q*-matrix misspecification in cognitive diagnosis. Unpublished paper.

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

Im, S., & Park, H. J. (2010). A comparison of US and Korean students' mathematics skills using a cognitive diagnostic testing method: Linkage to instruction. *Educational Research and Evaluation*, *16*, 287-301.

Ip, E. H. (1994). A stochastic EM estimator in the presence of missing data: Theory and applications. Technical report, Department of Statistics, Stanford University.

Jaeger, J., Tatsuoka, C., & Berns, S. (2003). Innovative methods for extracting valid cognitive deficit profiles for NP test data in schizophrenia. *Schizophrenia Research*, *60*, 140-140.

Junker, B. W., Patz, R. J., & VanHoudnos, N. M. (2016). Markov Chain Monte Carlo for item response theory. In *Handbook of Item Response Theory, Volume Two: Statistical Tools* (pp. 272-312). Boca Raton, FL: Taylor & Francis Group.

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

Karelitz, T. (2004). *Ordered category attribute coding framework for cognitive assessments*. Unpublished doctoral dissertation, University of Illinois at Urbana-Champaign

Kim, A. (2015). Exploring ways to provide diagnostic feedback with an ESL placement test: cognitive diagnostic assessment of L2 reading ability. *Language Testing*, *32*, 227-258.

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*, 59-81.

Li, H., & Suen, H. K. (2013a). Constructing and validating a q-matrix for cognitive diagnostic analyses of reading test. *Educational Assessment, 18*, 1-25.

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

Louis, T.A. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society—Series B, 44*, 226–233.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading: Addison-Wesley.

Ma, W., & de la Torre, J. (2016). A sequential cognitive diagnosis model for polytomous responses. *British Journal of Mathematical and Statistical Psychology*, *69*, 253-275.

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

Metropolis, N., Rosenbluth, A. W., Teller, A. H., and Teller, E. (1953). Equations of state space calculations by fast computing machines. *Journal of Chemical Physics, 21*, 1087-1091.

Millon, T., Millon, C., Davis, R., & Grossman, S. (2006). *MCMI-III manual* (3rd ed.). Minneapolis: Pearson Assessments.

Monroe, S., & Cai, L. (2014). Estimation of a Ramsey-curve item response theory model by the Metropolis-Hastings Robbins-Monro algorithm. *Educational and Psychological Measurement, 74*(2), 343-369.

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

Muraki, E., & Carlson, J. E. (1995). Full-information factor analysis for polytomous item responses. *Applied Psychological Measurement*, *19*, 73-90.

Nering, M. L., & Ostini, R. (2010). *Handbook of Polytomous Item Response Theory Models*. New York, NY: Taylor and Francis Group.

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

R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL.

Reckase, M. D. (1997). A linear logistic multidimensional model for dichotomous item response data. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 271-286). New York, NY: Springer.

Robbins, H., & Monro, S. (1951). A stochastic approximation method. *The Annals of Mathematical Statistics, 22,* 400-407.

Rupp, A. A., Templin, J., and Henson, R. A. (2010). *Diagnostic Measurement: Theory, Methods, and Applications*. The Guilford Press: New York, NY.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometric Monographs*, 17.

Samejima, F. (1995). Acceleration model in the heterogeneous case of the general graded response model. *Psychometrika, 60*, 549–572.

Sen, S., & Arican, M. (2015). A diagnostic comparison of Turkish and Korean students' mathematics performance on the TIMSS 2011 assessment. *Journal of Measurement and Evaluation in Education and Psychology*, *6*, 238-253.

Sessoms, J., & Henson, R. A. (2017). *Applications of diagnostic classifications models: A literature review and critical commentary.* Unpublished research.

Shu, Z., Henson, R., & Willse, J. (2013). Using neural network analysis to define methods of DINA model estimation for small sample sizes. *Journal of Classification, 30,* 173-194.

Sinharay, S. (2004). Experiences with Markov Chain Monte Carlo convergence assessment in two psychometric examples. *Journal of Educational and Behavioral Statistics, 29,* 461-488.

Sinharay, S., Puhan, G., & Haberman, S. J. (2011). An NCME instructional module on subscores. *Educational Measurement: Issues and Practice*, *30*(3), 29–40.

Svetina, D., Gorin, J. S., & Tatsuoka, K. K. (2011). Defining and comparing the reading comprehension construct: A cognitive psychometric modeling approach. *International Journal of Testing*, *11*, 1-23.

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

Templin, J. (2004). *Generalized linear mixed proficiency models for cognitive diagnosis*. Unpublished doctoral dissertation, University of Illinois at Urbana-Champaign.

Templin, J. (2006). *CDM user's guide*. Unpublished manuscript.

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

Wu, C. F. J. (1983). On the convergance of the EM algorithm. *Annals of Statistics*, *11*, 95-103.

Xu, X., & von Davier, M. (2006). *Cognitive diagnosis of NAEP proficiency data*. Research Report RR-06-08. Princeton, NJ: Educational Testing Service.

APPENDIX A

LOG-POSTERIOR AND DERIVATIVES

Because prior distributions are being imposed on λ and λ_0 in the P-LCDM-PA, the function that will be maximized is the log-posterior, denoted here as f, rather than the log-likelihood. The log-posterior function used to estimate model parameters of the P-LCDM-PA can be presented in the following form

$$logf = \left[\sum_{i=1}^{L}\sum_{j=1}^{J}\sum_{c=0}^{C_j-1} r_{ijc}\log(P_{ijc} - P_{ijc+1})\right] + \sum_{j=1}^{J}\sum_{k=1}^{2^{K_j}} logp(\lambda_{jk}) + \sum_{j=1}^{J}\sum_{c=1}^{C_j-1} logp(\lambda_{jc,0}). \quad (A.1)$$

Here L is the total number of latent classes, $L = \prod_{k=1}^{K} S_k$, J is the total number of items in the diagnostic assessment, and C_j is the total number of response categories for the j^{th} item. In the SEM and MH-RM, the summation of L can be taken over the total sample size N rather than over all latent classes, and r_{ijc} can be set to $I[X_{ij} = c]$. The the prior distribution $p(\lambda_{jk})$ is defined as the log-normal density function with hyperparameters $\mu_{\lambda_{jk}}$ and $\sigma_{\lambda_{jk}}$,

$$p(\lambda_{jk}) = \frac{1}{\lambda_{jk}} \cdot \frac{1}{\sigma_{\lambda_{jk}}\sqrt{2\pi}} exp\left[-\frac{1}{2}\left(\frac{log\lambda_{jk} - \mu_{\lambda_{jk}}}{\sigma_{\lambda_{jk}}}\right)^2\right], \quad (A.2)$$

and taking the log of $p(\lambda_{jk})$ results in

$$logp(\lambda_{jk}) = log\frac{1}{\lambda_{jk}} + logC - \frac{1}{2}\left(\frac{log\lambda_{jk} - \mu_{\lambda_{jk}}}{\sigma_{\lambda_{jk}}}\right)^2. \quad (A.3)$$

where $logC$ is a constant that can be dropped out of the equation. The prior distribution $p(\lambda_{jc,0})$ is defined as the normal density function with hyperparameters $\mu_{\lambda_{jc,0}}$ and $\sigma_{\lambda_{jc,0}}$,

$$p(\lambda_{jc,0}) = \frac{1}{\sigma_{\lambda_{jc,0}}\sqrt{2\pi}} exp\left[-\frac{1}{2}\left(\frac{\lambda_{jc,0} - \mu_{\lambda_{jc,0}}}{\sigma_{\lambda_{jc,0}}}\right)^2\right] \quad (A.4)$$

and taking the log of $p(\lambda_{jc,0})$ results in

$$logp(\lambda_{jc,0}) = logC - \frac{1}{2}\left(\frac{\lambda_{jc,0} - \mu_{\lambda_{jc,0}}}{\sigma_{\lambda_{jc,0}}}\right)^2. \quad (A.5)$$

where $logC$ is a constant. The r_{ijc} in Equation (A.1) is defined as the expected frequency of respondents responding at c^{th} category for the j^{th} item, given the i^{th} latent class, P_{ijc} and P_{ijc+1} are the boundary response probabilities for the j^{th} item at c^{th} category, given the i^{th} latent class. For notational clarity, the indices of *i*, *j* and *k* will be suppressed in the log-posterior function. Following Baker and Kim (2004) and Cai (2010a), differentiating the log-posterior function with respect to model parameters $\boldsymbol{\lambda}$ and $\lambda_{c,0}$ the first-order partial derivative values are defined as

$$\frac{\partial logf}{\partial \boldsymbol{\lambda}} = \sum_{c=0}^{C-1} \frac{r_c}{P_c - P_{c+1}}\left(\frac{\partial P_c}{\partial \boldsymbol{\lambda}} - \frac{\partial P_{c+1}}{\partial \boldsymbol{\lambda}}\right) + \left(-\frac{1}{\boldsymbol{\lambda}} - \frac{log\boldsymbol{\lambda} - \mu_{\boldsymbol{\lambda}}}{\sigma_{\boldsymbol{\lambda}}^2 \boldsymbol{\lambda}}\right), \quad (A.6)$$

$$\frac{\partial logf}{\partial \lambda_{c,0}} = -\left(\frac{r_{c-1}}{P_{c-1} - P_c} - \frac{r_c}{P_c - P_{c+1}}\right)\frac{\partial P_c}{\partial \lambda_{c,0}} + \left(-\frac{\lambda_{c,0} - \mu_{\lambda_{c,0}}}{\sigma^2_{\lambda_{c,0}}}\right), \tag{A.7}$$

where

$$\frac{\partial P_c}{\partial \boldsymbol{\lambda}} = P_c(1 - P_c)\boldsymbol{\alpha}^{**}, \tag{A.8}$$

$$\frac{\partial P_c}{\partial \lambda_{c,0}} = P_c(1 - P_c), \tag{A.9}$$

which are then collected into a $(2^{K_j} - 1) + \left(C_j - 1\right) \times 1$ gradient vector:

$$\nabla logf^T = \left(\frac{\partial logf}{\partial \boldsymbol{\lambda}}, \frac{\partial logf}{\partial \lambda_{c,0}}, \cdots, \frac{\partial logf}{\partial \lambda_{C-1,0}}\right). \tag{A.10}$$

Further differentiating the log-posterior function with respect to model parameters $\boldsymbol{\lambda}$ and $\lambda_{c,0}$ the second-order partial derivative are defined as

$$\begin{aligned}\frac{\partial^2 logf}{\partial \boldsymbol{\lambda} \partial \boldsymbol{\lambda}'} = &\sum_{c=0}^{C-1} -\frac{r_c}{(P_c - P_{c+1})^2}\left(\frac{\partial P_c}{\partial \boldsymbol{\lambda}} - \frac{\partial P_{c+1}}{\partial \boldsymbol{\lambda}}\right)\left(\frac{\partial P_c}{\partial \boldsymbol{\lambda}'} - \frac{\partial P_{c+1}}{\partial \boldsymbol{\lambda}'}\right) \\ &+ \frac{r_c}{P_c - P_{c+1}}\left(\frac{\partial^2 P_c}{\partial \boldsymbol{\lambda} \partial \boldsymbol{\lambda}'} - \frac{\partial^2 P_{c+1}}{\partial \boldsymbol{\lambda} \partial \boldsymbol{\lambda}'}\right) + \left(\frac{1}{\boldsymbol{\lambda}^2} - \frac{1}{\sigma^2_{\boldsymbol{\lambda}} \boldsymbol{\lambda}^2} + \frac{log\boldsymbol{\lambda} - \mu_{\boldsymbol{\lambda}}}{\sigma^2_{\boldsymbol{\lambda}} \boldsymbol{\lambda}^2}\right),\end{aligned} \tag{A.11}$$

$$\begin{aligned}\frac{\partial^2 logf}{\partial \lambda^2_{c,0}} = &-\left(\frac{r_{c-1}}{(P_{c-1} - P_c)^2} + \frac{r_c}{(P_c - P_{c+1})^2}\right)\left(\frac{\partial P_c}{\partial \lambda_{c,0}}\right)^2 \\ &-\left(\frac{r_{c-1}}{P_{c-1} - P_c} - \frac{r_c}{P_c - P_{c+1}}\right)\left(\frac{\partial^2 P_c}{\partial \lambda^2_{c,0}}\right) + \left(-\frac{1}{\sigma^2_{\lambda_{c,0}}}\right),\end{aligned} \tag{A.12}$$

$$\frac{\partial^2 logf}{\partial\lambda_{c-1,0}\partial\lambda_{c,0}} = \frac{r_{c-1}}{(P_{c-1}-P_k)^2}\left(\frac{\partial P_{c-1}}{\partial\lambda_{c-1,0}}\right)\left(\frac{\partial P_c}{\partial\lambda_{c,0}}\right), \tag{A.13}$$

$$\frac{\partial^2 logf}{\partial\lambda_{c+1,0}\partial\lambda_{c,0}} = \frac{r_c}{(P_c-P_{c+1})^2}\left(\frac{\partial P_{c+1}}{\partial\lambda_{c+1,0}}\right)\left(\frac{\partial P_c}{\partial\lambda_{c,0}}\right), \tag{A.14}$$

$$\frac{\partial^2 logf}{\partial\boldsymbol{\lambda}\partial\lambda_{c,0}} = -\frac{r_c}{(P_c-P_{c+1})^2}\left(\frac{\partial P_c}{\partial\lambda_{c,0}}\right)\left(\frac{\partial P_c}{\partial\boldsymbol{\lambda}}-\frac{\partial P_{c+1}}{\partial\boldsymbol{\lambda}}\right)+\frac{r_{c-1}}{(P_{c-1}-P_c)^2}\left(\frac{\partial P_c}{\partial\lambda_{c,0}}\right)$$
$$\left(\frac{\partial P_{c-1}}{\partial\boldsymbol{\lambda}}-\frac{\partial P_c}{\partial\boldsymbol{\lambda}}\right)-\left(\frac{r_{c-1}}{P_{c-1}-P_c}-\frac{r_c}{P_c-P_{c+1}}\right)\left(\frac{\partial^2 P_c}{\partial\boldsymbol{\lambda}\partial\lambda_{c,0}}\right), \tag{A.15}$$

where

$$\frac{\partial^2 P_c}{\partial\boldsymbol{\lambda}\partial\boldsymbol{\lambda}'} = P_c(1-P_c)(1-2P_c)\boldsymbol{\alpha}^{**}\boldsymbol{\alpha}^{**(T)}, \tag{A.16}$$

$$\frac{\partial^2 P_c}{\partial\lambda_{c,0}^2} = P_c(1-P_c)(1-2P_c), \tag{A.17}$$

$$\frac{\partial^2 P_c}{\partial\boldsymbol{\lambda}\partial\lambda_{c,0}} = P_c(1-P_c)(1-2P_c)\boldsymbol{\alpha}^{**}. \tag{A.18}$$

which are then collected into a $(2^{K_j}-1)+\left(C_j-1\right)\times(2^{K_j}-1)+\left(C_j-1\right)$ Hessian matrix:

$$\nabla^2 logf = \begin{pmatrix} \frac{\partial^2 logf}{\partial \boldsymbol{\lambda} \partial \boldsymbol{\lambda}'} & \cdots & \frac{\partial^2 logf}{\partial \boldsymbol{\lambda} \partial \lambda_{c,0}} \\ \vdots & \ddots & \frac{\partial^2 logf}{\partial \lambda_{c-1,0} \partial \lambda_{c,}} \\ \frac{\partial^2 logf}{\partial \boldsymbol{\lambda} \partial \lambda_{c,0}} & \frac{\partial^2 logf}{\partial \lambda_{c+1,0} \partial \lambda_{c,0}} & \frac{\partial^2 logf}{\partial \lambda_{c,0}^2} \end{pmatrix}. \quad \text{(A. 19)}$$

Note in Equation (A.11) the second-order partial derivative with respect to $logp(\boldsymbol{\lambda})$ only applies when $\boldsymbol{\lambda} = \boldsymbol{\lambda}'$ otherwise, 0.