The Ordered Latent Transition Analysis Model for the Measurement of Learning

Bright Nsowaa

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
under the Executive Committee
of the Graduate School of Arts and Sciences

Columbia University
2018

ABSTRACT

The Ordered Latent Transition Analysis Model for the Measurement of Learning

Bright Nsowaa

Several statistical models have been developed in educational measurement to determine and track the performance of students in longitudinal studies. An example of a model designed for such purpose is the latent transition analysis (LTA) model. The LTA model (Graham, Collins, Wugalter, Chung, & Hansen 1991) is a type of autoregressive model specifically designed to model transitions between class membership from Time t to Time t+1. The model however makes no assumption of ordering of the latent statuses and the transition probabilities. This project extends the LTA model by using the ordering technique proposed by Croon (1990) to introduce inequality constraints on the response probabilities of the LTA model. This new approach, referred to as the ordered latent transition analysis (OLTA) model, ensures ordering of the students' learning levels (known as statuses under LTA), and the transition probabilities. Simulation study was conducted in order to determine the adequacy of parameter recovery by OLTA as well as to evaluate the performance of the information criterion (AIC and BIC) in selecting the appropriate number of levels in the model. The simulation results showed good parameter recovery overall. Additionally, the AIC and BIC performed comparably well in selecting the correct transition model, but the AIC outperformed the BIC for the selection of optimal number of levels. An example of OLTA analysis of empirical data on reading skill development is presented.

Contents

List of Tables

List of Figures

Acknowledgements

I would like to thank my advisor, Matthew S. Johnson, for his inviolable support, encouragement and the countless opportunities he offered, enabling me to complete this dissertation. It was a great honor to work with him. I would like to thank the members of my dissertation committee: Young - Sun Lee, Bryan Keller, Charles Lang, and Ronald Neath, for their helpful and insightful comments.

I would like to thank my parents, Richard Gyamera and Paulina Owusu for their unflinching love and dedication, thank you for your sacrifices. I would like to thank my siblings, Richmond Gyamera, Sandra Gyamera, and Felix Gyamera.

I would like to thank my boy, Yedidia Nsowaa. A special thank -you to my wife Irene. I have been able to complete this work because of your companionship, understanding, encouragement, and prayers.

To Richard Gyamera and Paulina Owusu

Chapter 1

Introduction

The National Research Council (NRC) defines Learning progressions as" descriptions of the successively more sophisticated ways of thinking about a topic that can follow one another as children learn about and investigate a topic" (NRC, 2007, p.214). At the heart of learning progressions is the description of how students' knowledge and understanding of a topic or concept develop, and become more sophisticated over time. Students understandably have different experiences and as a result have a different understanding with different levels of reasoning (Battista, 2011). The concept of progression provides the ability to track students on the learning path, and to ensure that necessary resources are deployed to help them achieve higher academic goals.

Determining abilities of students at Time t, and or tracking their development at Time t+1 require measurement, and several statistical models have been developed in educational measurement to aid in this process;- specifically, models such as the item response theory (IRT) model (Birnbaum,1968; Rasch,1960), latent class analysis (LCA) model (Lazarsfeld & Henry, 1968), and the ordered latent class analysis (OLCA) model (Croon, 1990) measure students' learning at a single time point. However, the longitudinal IRT model (Fischer, 1989), Bayesian Knowledge Tracing (BKT) model (Corbett &Anderson, 1995), the Latent Transition Analysis (LTA) model (Graham, Collins, Wugalter, Chung, & Hansen, 1991), and other related models measure students' learning at multiple time points.

The LTA model in particular is a type of autoregressive model specifically designed to model transitions between class membership from Time t to Time t+1. The model makes no assumption about ordering of the latent classes (often referred to as statuses). Under the LTA model, the latent statuses are not ordered at Time t, and the transitions at Time t+1 are considered without any ordering. This project extends the LTA model by employing the techniques of ordered latent class analysis model (Croon, 1990) to introduce inequality constraints on the response probabilities of the LTA model.

For ease of demonstration, we will refer to this new technique as the Ordered Latent Transition Analysis (OLTA). The OLTA model ensures ordering of students' learning levels (known as statuses under LTA) at Time t, whilst providing the ability to tracking their development at Time t+1. Again unlike the LTA model, the transition probabilities are also ordered under this new approach. The OLTA model provides a practical solution to common measurement problems: which level do student(s) belong at the initial measurement, and how well have they developed over time. This new technique will enable researchers to test several stage-sequential models concerning human development. Second, the procedure can be used to assess the efficacy of an intervention program, and also estimate the differential effectiveness of such interventions for subjects in different levels. Also, not only is the OLTA model suitable for educational measurement, it can be used to model alcohol cessation or adolescent delinquent behavior, and so on.

The principle behind the ordered learning levels simply stems from the idea that students' categorized as belonging to level k are considered to have higher cognitive skills than those belonging to levels 1,...,(k -1). In order to ensure the potency of the OLTA model, we next present the results of the simulation experiment specifically designed to address issues

concerning the adequacy of parameter recovery. Finally, we present a real data example where several competing models of reading skill acquisition are tested in a sample of Pre-kindergarten children.

1.1 The Concept of Learning Progressions

It is well noted from research that children naturally follow some developmental progressions in terms of learning and development (Clements & Sarama, 2009). From a cognitive standpoint, there is a fundamental difference between learning and development. Pellegrino (2009) suggested that some knowledge is acquired only through purposeful /deliberate teaching, while others are universally acquired through natural development. For example, certain Mathematical concepts like algebra, and mathematical notations are acquired through deliberate teaching whiles fundamentals of ordinality, for instance, seem to develop naturally in children without instruction (Pellegrino, 2009).

The concept of Learning progressions / Learning trajectories in Mathematics and other related disciplines have gained a lot of traction over the years. The growing interest in learning progressions have in some ways provided a shift in emphasis from existing teaching modules to a more coordinated sequential teaching aimed at "developing scientific and mathematical knowledge with accompanying cognitive and metacognitive practices"(Duschl, Maeng, & Sezen, 2011). Pellegrino (2009) defines learning progressions (trajectories) as "empirically grounded and testable hypotheses about how students' understanding of, and ability to use, core concepts and explanations and related disciplinary practices grow and become more sophisticated over time with appropriate instruction".

One of the characteristics of learning progression is the implicit recognition that students follow multiple pathways in gaining mastery of concepts rather than one general sequence (NRC, 2007, p. 220). The hypothesized pathways or the conjectured routes are tested against real evidence of success; if the pathways prove unproductive, they are corrected and retried (Daro, Mosher, & Corcoran, 2011). As a result, the iterative experimental process continues until researchers are able to find an "efficient sequence" (Mosher, 2011).

1.2 Description of Latent Variable Models

There have been several examples in the social, behavioral, and health sciences where researchers have been able to use statistical model to classify individuals into distinct groups or categories. For example, Coffman, Patrick, Palen, Rhodes, and Ventura (2007) were able to identify distinct groups or categories of some high school seniors in the United Sates who were motivated, differently, into drinking. In addition, Bulik, Sullivan, and Kendler (2000) were able to classify a sample of twins into six distinct categories/subgroups of disordered eating.

The latent variable is not measured directly, but since it describes the interdependence of observed variables, two or more observed variables help to measure the latent variable indirectly. The observed variables are measured with error, but the latent variable is error-free. In the field of psychology and related fields, latent variables are often known as constructs (Pedhazur & Schmelkin, 1991). Figure 1.1 describes a hypothetical latent variable. In this figure, latent variables are labeled using the oval symbol, and four observed variables (e.g., X_1, X_2, X_3, and X_4) are labeled with square symbols. The corresponding errors associated with measuring the variables are contained in the circles.

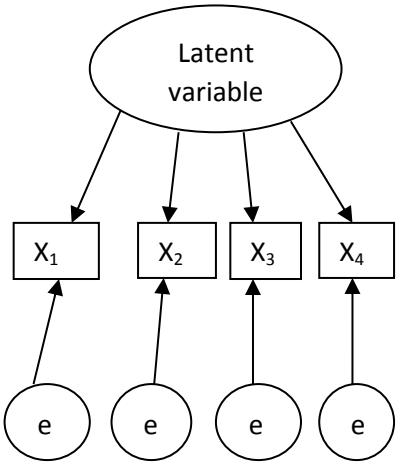


Figure 1.1 latent variable with four observed variables.

Bartholomew (1983) grouped latent variable models into four categories: factor analysis, latent trait analysis, latent profile analysis, and latent class analysis. Table 1.1 depicts how latent variable models are organized. The classifications of the models entirely depend on the nature of the latent and the observed (indicator) variables. When a model has a continuous latent variable, and a continuous indicator variable, it is known as factor analysis. Models with continuous latent variables and categorical indicators are known as latent trait analyses, also known as item response theory (Baker & Kim, 2004; Embretson & Reise, 2000; Lord, 1980; Van der Linden & Hambleton, 1997).

Further, models with categorical latent variables and continuous indicators are known as latent profile analysis (Vermunt & Magidson, 2002; Moustaki, 1996). For latent class analysis, both the latent variable and the indicators are categorical.

	Continuous Latent Variable	Categorical Latent Variable
Continuous Indicators	Factor analysis	Latent profile analysis
Categorical Indicators	Latent trait analysis	Latent class analysis

Table 1.1 Classification of Latent Variable Models

For latent class and latent trait models, the indicators are dichotomous, ordinal or nominal, and the conditional distributions of the models are assumed to be binomial or multinomial distributions (Bartholomew, 1987; Heinen, 1996).Table 1.1clearly depicts the fundamental difference between latent class and latent trait models; the latent trait models have a continuous latent variable but the latent class model contains "classes" which are discrete in nature and aid in categorizing people into homogenous groups (Heinen, 1993). For categorical latent variables, "qualitative differences exist between groups of people or objects"(Ruscio & Ruscio, 2008), and under continuous variables, "people or objects differ quantitatively along one or more continua" (p. 203).

Researchers have paid considerable attention to continuous latent variables over the years, and as a result, a lot of work has been done in this area (Bollen, 1989; Klein, 2004). Categorical latent variable models on the other hand have not been enjoying the same exposure. Recently however, there seems to be a growing interest in this area. Various aspects of the latent class analysis; including ordered latent class analysis model (Croon, 1990), and others, are gradually gaining traction. Croon's (1990) idea is based on classifying individuals into ordinal categories by placing order constraints on the response probabilities. The practical implications of this model make it more attractive for measuring educational data.

Croon's (1990) ordered latent class analysis (OLCA) model has provided inspiration for this project. At the core of this study is the idea that if individuals are ordered on the latent

continuum because of their ability or responses to behavioral issues, we expect changes in their ability to occur over time. For instance, individuals who belong to a lower class, because of their lower abilities, may acquire skills overtime and transition to higher classes. These kinds of changes are most apparent under longitudinal studies and the latent transition analysis (LTA) model is particularly suited to model transitions between class memberships while highlighting the complexities of such progressions overtime.

1.3 Ordering latent classes

Students in general do not have the same abilities, some have high abilities on certain concepts (i.e. a particular topic in mathematics) and others do not. Educational testing provides a platform for measuring students' abilities, and to also distinguish the levels of these abilities. Educational testing could also help to order students on the ability continuum.

For this reason, taking interest in accounting for stochastic ordering of latent classes in educational data analysis enriches interpretation. Two approaches are considered for imposing order to latent classes on a unidimensional scale: a parametric and a less or non-parametric approaches. The Item Response Theory (IRT) is a typical parametric functional form for describing the relationship between the item response probability and the latent classes. Though the IRT approach may well help provide order among classes, the nature of the model is such that the classes are measured in interval-level scale; which of course implies equal distance between ordered classes. The assumptions under this approach are stronger, and the constraints impose on the classes can be too restrictive which may reduce the attractiveness of this approach.

Croon (1990) proposed a less or non-parametric approach for ordering latent classes by imposing inequality restrictions on item response and cumulative item response probabilities.

Unlike the IRT approach, the non-parametric model proposed by Croon is considered ‘weak’ mainly because of its less-restrictive assumptions. However, the nature of the assumptions makes this model more attractive and quite frankly advantageous. Under this approach, the parametric assumptions on item response probability, or the normality assumptions on ability distribution are relaxed. Croon created an algorithm for the maximum likelihood estimation of the parameters. But the maximum likelihood estimation under the proposed order constraints has had a history of resulting to a local maxima (Van Onna, 2002).

In order to avoid the issue of local maxima, and to achieve global maximum, researchers have proposed using many different starting values for the algorithm (Vermunt, 1997), also there are other estimation procedures such as the Bayesian estimation approach which have been proven to be efficient for parameter estimation (Hoijtink & Molenaar, 1997; Hoijtink, 1998; Van Onna, 2002). This project has adopted Croon's non-parametric approach towards ordering latent classes.

Assumptions of ordered Latent Class Analysis (OLCA)

Croon's work on ordered latent classes "is a more broadly applicable way to investigate whether a unidimensional item response model provides a reasonable description and explanation of the subject's response to the different items, without making strong assumptions about the functional relationship between latent and manifest variables" (Croon, 1990, p. 188). The main assumptions of the OLCA model are homogeneity, local or conditional independence, unidimensionality, and monotonicity.

Homogeneity

The assumption of Homogeneity is essential to the success of the OLCA concept. In OLCA, individuals are grouped into homogeneous sub-groups referred to as classes. The fundamental principle of latent classes is that no one belongs to more than one latent class, and the probability for a particular response to a particular item j, solely depends on the class to which the individual belongs.

Local Independence (LI)

For a latent variable model, observed variables, $X = (X_1, \ldots, X_j)$, and latent variable, U, are assumed to be jointly distributed over a population (Holland & Rosenbaum, 1986). The conditional distribution function of X given U is represented as

$$F(x_1, \ldots, x_j|u) = P(X_1 \leq x_1, \ldots, X_j \leq x_j|U = u), \quad (1.1)$$

and the assumption of local independence posits that $X_1, \ldots, X_j$ are conditionally independent given U, otherwise expressed as

$$F(x_1, \ldots, x_j|u) = \prod_{j=1}^{J} F_j(x_j|u) \quad (1.2)$$

for all $x_1, \ldots, x_j$ and u.

$$\text{then,} \quad F_j(x_j|u) = P(X_j \leq x_j|U = u) \quad (1.3)$$

The local independence assumption stipulates that conditional on the latent variable, the observed or indicator variables are independent. This posits that the relations that exist among observed variables are explained by the latent classes, and it is local because the assumption is held within each latent class.

Unidimensionality (U)

The Unidimensionality assumption is fundamental in latent variable models; it states that the observed or indicator variables are assumed to be measuring only one trait, attribute, or ability.

Monotonicity (M)

Croon (1990, 1991) introduced a model to ensure the stochastic ordering of latent classes by imposing inequality constraint directly on the item step response functions (ISRFs). Croon's proposal forms a core assumption of OLCA which makes the ISRFs nondecreasing function of the latent trait. Van Onna (2002) depicted this assumption very well:

Let $\rho_{jrc} = P\ (X_j = r \mid c\ )$ represent the probabilities of individuals choosing category r on item j, given latent class q. Imposing constraints on the cumulative probabilities yield $\rho^*_{jrc} = P\ (X_j \geq r \mid c\ ) = P\big(Y_{jr} = 1 \,\big|c\ )$. All persons responding to item j with response category $r, r+1, \ldots, r_j$ have passed item step j r, for those individuals $Y_{jr} = 1$.

Now, $\rho^*_{jrc} = \sum_{k=r}^{R_j} \rho_{jkc}$

In accordance with the monotonicity assumption, the ISRFs should be nondecreasing with the latent classes to ensure that higher scores are associated with higher latent classes,

$$\rho^*_{jrc} \leq \rho^*_{jr,c+1} \qquad (1.4)$$

Equation (1.4) represents the monotonicity assumption. The NIRT model that satisfies the assumption of monotonicity (M), together with the assumptions of local independence (LI), and

Unidimensionality (U) is referred to as monotone homogeneity (MH) model (Holland & Rosenbaum, 1986; Meredith, 1965; Mokken, 1971; Mokken & Lewis, 1982).

1.4 Markov Chain Models

In recent years, several statistical methodologies have been developed to measure possible change over time, closely as possible. The advancement of these statistical models is essential to testing theories and stage-sequential developments. The growth curve modeling (Rogosa, Brandt, & Zimowski, 1982; Willett, 1988; Meredith & Tisak, 1990; Willett & Sayer, 1994; Singer & Willett, 2003) for instance has enjoyed a lot of success over the years with respect to measuring growth in continuous variables. However, there are instances in the study of human development where the primary interest is the change in qualitative status of individuals over time. The primary statistical model for the measurement of change in qualitative status over time is the manifest Markov chain model (Kaplan, 2008).

Markov chain models are suitable for items with responses categorized as Yes/No; Pass/Fail; Agree/Disagree; Democrat/Republican/independent; Employed/Unemployed; etc., which are repeatedly measured overtime with the same respondents in the sample, and that the dynamics of change are modeled over time; in order to determine change, stability, or both (Langeheine & Van de Pol, 2002). Markov chain models have been applied in many areas including learning; cognitive development; epidemiology; attitudinal change; voting behavior/pattern; and consumer behavior (Langeheine & Van de Pol, 2002). In recent years, several extensions have been added to the manifest Markov chain model, which have undoubtedly enhanced developmental research. These extensions are built to improve on the drawbacks of the manifest Markov chain model, such as accounting for measurement error, and allowing each chain to follow its own dynamics.

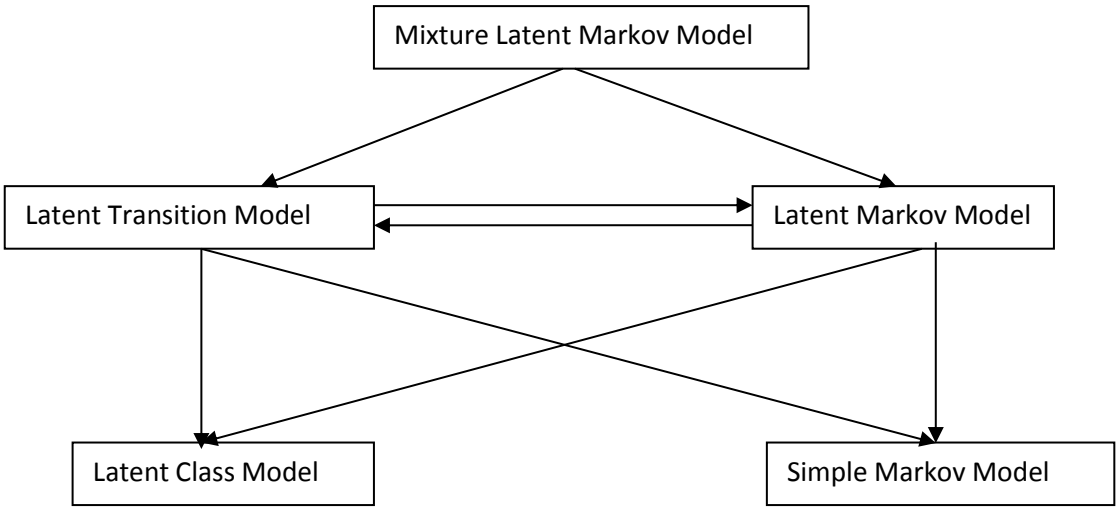


Figure 1.2. A modified hierarchy of Markov models (adapted from Kaplan, 2008). Arrows represent special cases.

The Simple Markov Chain

For the Simple Markov chain model and the remaining of Markov chain models in this section, the data of interest are observed categorical responses. The Simple Markov model consists of a single chain, the model assumes that the probability for a subject to be in a specific state at time point t solely depends on the state the subject was in at time point t – 1. The model ignores the influence of the state in earlier time points, for example t -2. Also, the model assumes Population homogeneity: this implies that the dynamics across time derived by the model hold for all subjects. Furthermore, the model assumes that the data are measured without error.

Researchers including (Anderson, 1954; Goodman, 1962; Wiggins, 1973; Bartholomew, 1981; Van de Pol & de Leeuw, 1986; Langeheine, 1988) have all noted the restrictive nature of these assumptions, and that perhaps explains the difficulty in fitting Simple Markov chain in any given set of data (Langeheine & Van de Pol, 2002).

The Markov chain model with T = 3 points in time, is given by

$$P_{ijk} = \delta_i^1 \tau_{j/i}^{2\ 1} \tau_{k/j}^{3\ 2}, \quad (1.5)$$

where all parameters are manifest quantities: δ^1 represents the observed initial marginal distribution at time1, $\tau_{j/i}^{2\ 1}$ and $\tau_{k/j}^{3\ 2}$ are observed transition probabilities for a transition from time 1 to time 2 and from time 2 to time 3 respectively. Of course, the Simple Markov can be specified in such a way that allows transition probabilities to be constant across time or differ across time (Kaplan, 2008).

The Latent Markov Model

One of the drawbacks of the Simple Markov model described earlier is the assumption that the data are measured without error. This of course implies that the observed responses flawlessly measure an individual's true latent state. Langeheine & Van de Pol (2002) described the error-free assumption as unrealistic for social science researchers. The latent Markov model, developed by Wiggins (1973) addressed the error-free assumption by allowing for correction of errors, thereby obtaining transition probabilities at the latent level.

For T = 3 points in time, the latent Markov model can be written as

$$P_{ijk} = \sum_{a=1}^{A} \sum_{b=1}^{B} \sum_{c=1}^{C} \delta_a^1\ \rho_{i/a}^1 \tau_{b/a}^{2\ 1}\ \rho_{j/b}^2\ \tau_{c/b}^{3\ 2}\ \rho_{k/c}^3. \quad (1.6)$$

Where $\rho^1_{i/a}$ represents the response probability associated with category i given membership in latent status a. $\rho^2_{j/b}$ represents the response probability associated with category j given membership in latent status b. Remaining response probability is interpreted similarly. Transition from time 1 to time 2 in latent status membership is represented by $\tau^{2\ 1}_{b/a}$, and the transition from time 2 to time 3 in latent status membership is captured by $\tau^{3\ 2}_{c/b}$.

It is clear from our discussion of the LTA model and the latent Markov model that there is no fundamental difference between the two models. The difference may be the practicality of the model, since the LTA model is conceivably viewed as being an ideal for studying developmental changes (Kaplan, 2008).

Mixture Latent Markov Model (the Mover-Stayer Model)

Thus far, the Markov Chain models described assume that the sample of observations comes from a single population described by a single Markov chain and one set of parameters. In some cases however, the population comprises a finite mixture of subpopulation, and using the same Markov model approach may lead to biased estimates and misleading conclusions of the developmental processes under consideration.

A better approach for addressing this issue involves the combination of Markov-chain-based models under the assumption of a mixture distribution. This approach is known as the mixture latent Markov model. The Mover-Stayer model (Blumen, Kogan, & McCarthy, 1955) is a special case of the mixture latent Markov model. The Mover-Stayer model primary consists of two Markov chains: Movers follow the usual Markov chain, which is individuals transitioning

across latent statuses over time while Stayers remain in their initial status and do not transition across statuses.

The mixture latent Markov model can be written as

$$P_{ijk} = \sum_{s=1}^{S} \sum_{a=1}^{A} \sum_{b=1}^{B} \sum_{c=1}^{C} \pi_s \, \delta_{a\,/s}^{1} \, \rho_{i/as}^{1} \tau_{b/as}^{2\;1} \, \rho_{j/bs}^{2} \, \tau_{c/bs}^{3\;2} \, \rho_{k/cs}^{3}. \qquad (1.7)$$

where π_s represents the proportion of S latent chains and the remaining parameters are interpreted as in Equation 1.6, except that they are conditioned on membership in Markov chain s. Special cases can be derived from equation 1.7, for example, with s = 1, equation 1.7 reduces to the latent Markov model in Equation 1.6. Also, with s = 1 and without transition probabilities, the model in Equation 1.7 reduces to the latent class model.

1.5 Latent Transition Analysis (LTA)

"Latent transition analysis (LTA) is a variation of latent class model that is designed to model not only the prevalence of latent class membership, but the incidence of transitions over time in latent class membership" (Collins & Lanza, 2010, p.181). The model was introduced by Graham, Collins, Wugalter, Chung, & Hansen (1991), and is considered to be an extension of the latent Markov model by permitting the use of multiple indicator variables to test complex models.

Markov model is a well-known model which has been in existence for relatively longer period; with a long standing contribution to psychology (Anderson, 1954). However, Wiggins (1973) and Lazersfeld & Henry (1968) introduced latent Markov model in the setting of latent

class theory. Until researchers such as Bye & Schechter (1986), Van de Pol & de Leeuw (1986), and Van de Pol & Langeheine (1989) introduced feasible methods for estimating model parameters, there was no meaningful application of the model presented by Wiggins, Lazersfeld and Henry.

The "state" of the latent variable presents a fundamental difference between the latent class theory and LTA. Under latent class theory, the latent variable is primarily static but latent variable under LTA is dynamic (Graham, Collins, Wugalter, Chung, & Hansen, 1991). Dynamic latent variables keep changing over time in a methodical manner. Collins & Cliff (1990) created the contrast between static and dynamic variables, according to the authors (Collins, 1991b, 1991a; Collins, Cliff, & Dent, 1988) the contrast between static and dynamic variables is vital in the sense that the regular measurement theories were not developed for dynamic variables, and that new methods are required to reasonably measure dynamic variables.

LTA has been employed to especially test varying psychological constructs that are based on stage/sequential development, for example, children's drawing development (Humphreys & Janson, 2000), smoking cessation (Velicer, Martin, & Collins, 1996),the progression of health-risk behavior (Reboussin, Reboussin, Liang, & Anthony, 1998), and modeling substance use prevention (Graham, Collins, Wugalter, Chung, & Hansen, 1991). In LTA, individual's class membership at a particular time of measurement is often referred to as the individual's latent status. Three different parameters are of interest when considering LTA: latent status prevalences, item-response probabilities, and transition probabilities.

Latent status prevalences

Latent status prevalences in LTA perform essentially the same functions as their counterparts (latent class prevalences) in LCA. The fundamental difference between the two is

that the latent status prevalences produce vector of latent status prevalences at each occasion. Researchers are able to inspect the latent statuses at each time to determine the most prevalent latent status as well as, the least prevalent one. The latent status prevalences provide useful information, but nothing relating to the extent to which respondents transition from one latent status to the other. The transition probabilities carry such information.

Item- response probabilities

The item-response probabilities in LTA perform the same function as the item-response probabilities in LCA. Just like in LCA, researchers are able to assign labels to latent statuses based on the item-response probabilities in LTA. Just like the latent status prevalences, differences exist between item-response probabilities in LCA and LTA. The difference is that in LTA, each time has its own matching item-response probabilities. Suppose measurements are taken between two times t and t + 1, there will be two sets of item-response probabilities corresponding to time t and t+ 1 respectively.

Transition Probabilities

Unlike the latent status prevalence and the item-response probabilities, the transition probabilities are unique to LTA, and they are of paramount interest in LTA. In many respects, the transition probabilities are considered to be the bed rock of LTA analysis; they show how transitions occur between latent statuses from one time to the next (Graham, Collins, Wugalter, Chung, & Hansen, 1991; Collins & Lanza, 2010, p. 195). The transition probabilities are usually

arranged in a matrix form with the rows corresponding to time 1, and the columns corresponding to time 2. Let us consider the general representation of the transition probability matrix below:

$$\begin{array}{c} \\ \text{Time 1 status} \end{array} \begin{array}{c} \text{Time 2 status} \\ \begin{bmatrix} \tau_{1|1} & \tau_{2|1} & \tau_{3|1} & \dots \\ \tau_{1|2} & \tau_{2|2} & \tau_{3|2} & \dots \\ \tau_{1|3} & \tau_{2|3} & \tau_{3|3} & \dots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix} \end{array}$$

The tau (τ) parameters within the matrix represent the transitioning across latent statuses from one time to the next. From the matrix we can safely say that $\tau_{y|x}$ represents the probability of membership in latent status y at time t+1, conditional on membership of latent status x at time t. The values on the diagonal matrix represent the probability of being in a particular latent status at time t+ 1, conditional on being in that same latent status at time t. The tau parameters above the diagonal elements of the probability transition matrix represent the probability of transitioning to an "advanced" status, and the parameters below the diagonal elements represent the probability of transitioning back to a former status (Velicer, Martin, & Collins, 1996).

Chapter 2

Literature Review

2.1 Introduction

This chapter begins with the benefits of learning progression, and also focuses particularly on the Assessment aspect of learning progression; which is primarily concerned with modeling students' learning. Students' learning can be modeled in two forms: modeling item responses, and modeling growth. Models such as the Item Response Theory (IRT) Model, Latent Class Models, and Ordered Latent Class Models are specially designed to modeling item responses at a single time point. In terms of modeling students' growth, specific models to be considered are the Latent Transition Analysis (LTA) model, which is a special case of latent Markov model, a special case of Hidden Markov Model in Bayesian knowledge tracing, and the longitudinal IRT models. The concept of learning progression and the selected models mentioned are reviewed in this chapter.

2.2 Background of the learning progression framework

2.2.1 Learning Progression and its Benefits

The National Research Council (NRC) defines Learning progressions as" descriptions of the successively more sophisticated ways of thinking about a topic that can follow one another as children learn about and investigate a topic" (NRC, 2007, p.214). Learning Progression and learning trajectories have been used interchangeably, but the intended use of the concepts reveal some subtle differences. For instance, learning trajectory is likely to be developed if the sole aim is designing and testing a curriculum. However, if one is interested in formative assessment

system relevant to many curricula, learning progressions are more likely to be developed (Battista, 2011). In spite of the seeming differences, this study uses learning progression and learning trajectory interchangeably, and also describes students' learning in numerical fashion as a way of measuring their progress.

Generally, students do not develop at the same pace and therefore are not expected to start from, or end at, the same position on any given progression at any given stage (Corcoran, Mosher & Rogat, 2009). The concept of learning progression is vital in the sense that it has the potential to provide instructors with the requisite framework necessary to appropriately respond to the differences in students' progress at any stage, and to adapt instructions that are relevant to the needs of each student in order to help them achieve their learning goals. Progressions in this manner help to improve standards and curriculum.

Curricula improvement is one of the main benefits of learning progressions. Successful progressions are able to focus on students' progress vis-a-vis the instructions they have experienced to create a reliable curriculum framework needed to determine the order and manner in which a particular skill(s) is/are to be taught (Corcoran, Mosher, & Rogat, 2009). In the nutshell, successful learning progressions help provide and also specify what students are expected to know at a particular stage of their academic lives , as well as, what they are expected to do (Pellegrino, 2009). This is huge because it motivates the creation of curricula that are closely linked to students' progress (Duschl, Maeng, & Sezen, 2011).

Also, students' assessments are improved under the learning progression concept. The bedrock of learning progression is the description of how students' performances, reasoning or sophistication develop from Time t to Time t + 1 (Corcoran, Mosher, & Rogat, 2009). This means information gathered from assessing the students are insightful, and the instructors are

likely to have a good understanding on that information and use them to effectively address the needs of the students.

Lastly, learning progressions improve instruction (Clements, & Sarama, 2009; Corcoran, Mosher & Rogat, 2009; Battista, 2011; Duschl, Maeng, & Sezen, 2011). Since learning progressions provide instructors with the tools that aid their understanding on students' progress and their likely levels, they are essentially being guided on their instructional choices. These instructional choices and goals influenced by learning progressions help teachers address difficulties students may be facing in their quest to achieve their academic goals. According to Sztajn,et. al., (2012), teachers previously constructed models on students learning based on their own understanding on how students learn. Learning progressions however, present a paradigm shift in the sense that teachers are making sense of students learning and progression based on scientific research (Sztajn, et.al, 2012; Confrey & Maloney, 2010).

Several studies point to the positive impact learning trajectories have on instruction. For instance, on her attempt to study how teachers learn and use learning trajectories, Mojica (2010) conducted a study on fifty-six teachers for 8 weeks, where the participants were introduced to the concept of learning trajectories. The author claimed that the participating teachers gained insight on the trajectory framework, and subsequently applied the concept in the classroom by making students thinking an integral part of instructional decisions. Also Clements, et al., (2011) conducted a randomized trial study with 42 schools to assess the effectiveness of trajectories as an intervention. The researchers found that students in the experimental group experienced growth in their knowledge of mathematics above and beyond the students in the control group. After examining the classroom practices of the teachers, the researchers found that teachers in

the experimental conditions were more attentive to the needs of students and taught mathematics better than their counterparts in the control group.

2.2.2 Measurement and Assessment

"Assessment includes the processes of gathering evidence about students' knowledge and abilities as related to the tasks to which they respond as well as making inferences from that evidence about what students know or can do more generally" (Gotwals & Songer, 2013). Pellegrino (2009) posited that assessment is useful only when it is linked with curriculum and instruction. According to the Pellegrino, as curriculum provides the general outlines of how certain academic goals could be achieved, and also determining the depth of content in a particular subject area, instruction must also be effective such that students can attain any form of mastery in any subject area based on the goals set forth by curriculum.

Instructions are carried out in a variety of ways, and students engage in diverse activities as part of instruction. Students' activities must be evaluated by educators and instructors at some point to determine how effective the instruction is, and how well the students are progressing and developing mentally. This is one of the reasons why assessment is very critical in our educational system; it plays a distinctive role by providing the means to effectively measure the educational outcome, the capabilities and competencies of students. Assessment takes many forms, there are informal ones that involve instructors organizing class tests, administering pop quizzes, or providing home works to students, and so on. The formal assessments, such as the state assessment are large-scale in nature.

Students' responses to assessment items help reveal how knowledgeable they are and how well they understand a concept, it also helps educators and researchers determine students'

performances and progress (Duschl, Maeng, & Sezen, 2011). Teachers receive item responses from students in the classroom, and they have to interpret the data, and make sense of it. This unique responsibility of teachers have prompted some researchers lately to adapt the learning progression framework to help train teachers to effectively interpret classroom data in order to track students' academic progress (Plummer & Slagle, 2009). Instruction and Assessment are improved (Heritage, 2008), when the outcome of assessment help teachers adapt or adjust instructional style to meet the needs of students.

Statistical models are integral in assessment. Model techniques such as the Rasch model have commonly been used to assess the validity and consistency of students' score. The Rasch model describes the relationship between students' ability level and item difficulty (Embretson & Riese, 2000). Briggs and Alonzo (2009) considered the Rasch model technique ineffective for determining all the attributes of the item. The authors then used the Attributed Hierarchy Method (AHM) instead. Briggs and Alonzo concluded that the AHM provides better understanding of the construct especially for educators or researchers creating Ordered Multiple Choice items. Several models such as the item response theory (IRT) model, Latent class model, etc., described in subsequent sections are suitable for modeling student's outcome.

2.2.3 The Concept of Levels

The Levels Concept plays a critical role in learning progressions. At the heart of learning progressions is how students' understanding or knowledge of a topic becomes more sophisticated over time with appropriate instructions. Students do not all of a sudden gain knowledge or develop full understanding of a topic or concept; they go through a process (levels of understanding) before attaining sophistication on the topic or concept.

Researchers use of the term level is different from stage; a stage is age dependent and is considerable period of time in which a specific cognition occurs in various domains (i.e. Piaget's stages of cognitive development), but a level is independent of age; it is a period of time in which students get prominent cognition for a specific concept (Clement & Battista, 1992). This means that an older person or student can be categorized as belonging to say level 1 on a particular concept due to his/her experiences and interaction to the concept. However, a relatively younger person or student who has a lot of experience and interaction with those same concepts may be categorized as belonging to a higher level; say level 3.

Some of the ingredients that determine the "Level" of the student are the experiences of the students and their interactions with the concept in question, and instructions they may have received. Two types of levels are described in research: A "weak" and a "strong" level. A level is described as "weak" when they are arranged in sequence of sophistication, one above the other, without class inclusion relationship among them, when a set of levels are arranged in sequence of sophistication, one above the other, with the presence of class inclusion relationships among them, it is referred to as a "strong" level (Battista, 2011). This means that for a "strong" level, students who are categorized as reasoning at level k are believed to have progressed through reasoning at the lower levels 1, 2,3,...(k -1). Battista (2007) suggested that being "at" a level means that the student is cognitively developed in a manner that put him in a position to think about a topic or a concept in a particular way.

2.3 Statistical Models of abilities measured at a single time point

2.3.1 Item Response Theory Models

The item response theory (IRT) models have been used widely in educational testing and research. The purpose of these models is to assess item characteristics, and to make inferences on

examinee's abilities in a specific content area. The classic assumptions of IRT models include unidimensionality (UN), local independence (LI), and monotonicity (M). Various models have been proposed for IRT models, and depending on the method of estimation, these models basically fall into two categories: parametric and nonparametric methods. Parametric item response theory (PIRT) models are grouped into two on the basis of how items are scored. For binary or dichotomous IRT models, items are scored into categories, such as pass or fail, true or false, etc. For polytomous IRT models, items are scored in more than two categories, such as Likert items. Three related IRT models under binary IRT models are widely used and are popular in psychometric literature: The Rasch model (or one-parameter model), the two-parameter logistic model, and Birnbaum's three-parameter model. Also, the three related models commonly used under polytomous IRT models are the rating scale model, the graded response model, and the partial credit model. The maximum marginal likelihood estimation (MMLE) procedure is often used to fit the PIRT models (Bock & Aitkin, 1981).

2.3.2 Parametric Item Response Theory (PIRT) Models

Binary IRT Models

For the analysis of binary test items, three IRT models are predominant in psychometric literature. The Rasch (1960) model, often referred to as the one-parameter logistic model (1PL) takes the form

$$P_i(\theta) = \frac{e^{D\alpha_i(\theta - \beta_i)}}{1 + e^{D\alpha_i(\theta - \beta_i)}}, \qquad (2.1)$$

where θ is the ability level of examinees, β_i is an item difficulty parameter, α_i is an item discrimination parameter, and D is a scaling constant (generally D = 1.702). $P_i(\theta)$ represents the probability of an examinee responding correctly to an item. Under the Rasch model, the item discrimination parameter α_i is fixed at unity (Birnbaum, 1968), such that the only item parameter to be estimated is the item difficulty. Constraining the item discrimination index to unity under 1PL ensures that the item response functions (IRFs) do not intersect, and this property is referred to as the item ordering property (Sijtsma & Hemker, 2000; Sijtsma & Junker, 1996).

The two-parameter logistic (2PL) model has a different appreciation for the item discrimination parameter α_i. Unlike the1PL, the 2PL puts no restrictions on the $\alpha_i's$, and items may vary in discrimination. The 2PL takes the form

$$P_i(\theta) = \frac{e^{D\alpha_i(\theta-\beta_i)}}{1+e^{D\alpha_i(\theta-\beta_i)}}\ , \qquad (2.2)$$

The 2PL contains estimates of item difficulty and item discrimination. Birnbaum (1968) introduced a three-parameter model; which adds a guessing parameter c_i to the 2PL. Birnbaum's three-parameter model takes the form

$$P_i(\theta) = c_i + (1-c_i)\frac{e^{D\alpha_i(\theta-\beta_i)}}{1+e^{D\alpha_i(\theta-\beta_i)}}, \qquad (2.3)$$

Equation (2.3) follows the assumption that the student/examinee either knows the item and correctly respond with the probability discussed under equation (2.2) or the examinee does not know the item but guesses with the probability of success equal to the value of the guessing parameter c_i. It is well noted in literature (Baker & Kim, 2004, p. 19; Van der Linden & Hambleton, 1997, p.13) that Birnbaum's three-parameter model does not define a logistic function, in spite of that, equation (2.3) is often referred to as the three-parameter logistic (3 PL) model.

Polytomous Item Response Models

Several models have been proposed for modeling polytomous response data. However, only three of such models are reviewed under this section: the graded response model (Samejima, 1969), the partial credit model (Masters, 1982), and the rating scale model.

The graded response model (GRM) deals with ordered response categories such as letter grading (A, B, C, D, and F) used to assess students' performance or such responses as exist in the case of likert rating scales. The model assumes that the cumulative log odds for scoring $m \epsilon\{1$, 2, . . . M} or higher on item j, is a linear function of latent variable θ:

$$log\left(\frac{Pr(Y_j \geq m|\theta}{\Pr(Y_j < m|\theta}\right) = \alpha_j\left(\theta - \beta_{jm}\right),$$

where m is the response category { $1 \leq m \leq M$ }, j represents an item, θ represents latent trait such as ability. Under the GRM, the item -category step parameters, β_{jm} are ordered by the category index j such that $\beta_{j1} < \beta_{j2} < \ldots < \beta_{jM-1}$, whilst the discrimination index α_j are fixed across item categories.

The partial credit model (PCM) (Masters, 1982) is considered a straightforward model containing only two sets of parameters. Unlike the GRM, the PCM belongs to the Rasch family of models, and for that matter shares the distinguishing characteristics such as sufficient statistics, separable item and person parameters, and so on. Successful application of the PCM includes ratings of infant performance (Wright & Masters, 1982); ratings of writing samples (Pollitt & Hutchinson, 1987; Harris et al., 1988); measures of critical thinking (Masters & Evans, 1986) etc.

The model assumes that the adjacent -categories logit is a linear function of latent variable θ:

$$log\left(\frac{\Pr(Y_j = m|\theta)}{\Pr(Y_j = m-1|\theta)}\right) = \alpha_{jm}(\theta - \beta_{jm}),$$

This leads to the item-category response function below

$$P_{jm}(\theta) = \Pr\{Y_{ij} = m|\theta\}$$

$$= \frac{\exp\{\sum_{l=0}^{m} \alpha_j(\theta - \beta_{jl})\}}{\sum_{r=0}^{M_j} \exp\{\sum_{l=0}^{r} \alpha_j(\theta - \beta_{jl})\}}.$$

Unlike the GRM, β_{jm} under PCM and the generalized partial credit model (GPCM) are not necessarily ordered, also the GPCM generalizes PCM to allow for varying discrimination index, α across items (Muraki, 1992).

The rating scale model (RSM) which was first introduced by Rasch (1961) and restructured by Andrich (1978) is considered to be an extension of the Rasch model. The model

makes an assumption that the category scores are equally spaced and the continuation logit is a linear function of latent variable θ:

$$log\left(\frac{\Pr(Y_j \geq m|\theta)}{\Pr(Y_j=m-1|\theta)}\right) = \alpha_j(\theta - \beta_{jm})$$

2.3.3 Nonparametric Item Response Theory (NIRT) Models

According to literature, interest in nonparametric IRT has been in existence prior to the emergence of parametric IRT (Guttman, 1947, 1950a, 1950b). In spite of this fact, parametric IRT has gained enormous recognition partly due to the success of practical implementation of the logistic models in areas such as test equating, item banking, test bias, and also a successful implementation of the computerized adaptive testing. However, the underlying assumptions of the logistic function on parametric item response probability may be too restrictive, sometimes leading to model misfit.

Researchers such as (Holland, 1981; Holland & Rosenbaum, 1986; Rosenbaum, 1984, 1987a, 1987b) helped brought back the concept of NIRT into psychometric literature as a means of studying the minimal set of assumptions needed to be met by any response model; parametric or nonparametric. For instance, Woods & Thissen (2006) found that their proposed spline-based density estimation procedure provided a flexible alternative to the existing procedures that use normal distribution. The Spline, and Kernel regression techniques have also been used to estimate the non-parametric response function (Johnson, 2007; Ramsay, 1991; Ramsay & Abrahamowicz, 1989; Winsberg, Thissen, & Wainer, 1984).

2.3.4 Latent Class Model

Lazarsfeld & Henry (1968) are credited for proposing the latent class analysis (LCA) model for analysis in the social and behavioral sciences. Since parameter estimation was an integral part of the concept, researchers were unable to expand or implement the ideas because there was no reliable method for estimating the parameters. Even though the proposal was generally accepted in the research community, lack of computational power prevented its implementation until Goodman's findings.

Goodman (1974) helped changed the entire narrative by developing maximum likelihood procedure which was a reliable approach for obtaining estimates of latent class model parameters. Since then, there has been widespread application of the latent class model concept in many areas including medicine (Garrett & Zeger, 2000; Qu & Kutner, 1996; Rabe-Hesketh & Skrondal, 2008; Uebersax & Grove, 1990), and marketing (Dillon & Kumar, 1994; Jain & Chen, 1990; Swait & Adamowicz, 2001). Goodman's procedure for obtaining parameter estimates is found to be related to the notable expectation-maximization (EM) algorithm (Dempster, Laird, & Rubin, 1977).

Latent class model consists of item-response probabilities ($\rho's$) and estimated class prevalence ($\gamma's$).The notations in this writing on LCA are consistent with ones provided by (Collins & Lanza, 2010, p.39). To conduct latent class analysis on empirical data, contingency table is needed. Since most programs needed for conducting LCA do not allow missing values, the contingency table must include all observed variables needed for the analysis. Suppose there

are j = 1,...,J observed variables, with each j having r_j = 1,..., R_j response categories, the contingency table has W=$\prod_{j=1}^{J} R_j$ cells.

For instance, in a hypothetical example of 3 observed variables (or items) with 3, 2, 2 response categories respectively, J = 3. R_1 = 3 means there are 3 response categories for question 1, R_2 = 2 means there are 2 response categories for question 2, and R_3 = 2 means there are 2 response categories for question 3. So in this hypothetical example, W = $\prod_{j=1}^{J} R_j = R_1R_2R_3$ = 3×2×2 = 12 cells. Let y = (r_1,...,r_j) represent a vector response to J observed variables corresponding to each W cells. If Y represents the array of response patterns, then each response pattern y is connected with probability P(Y = y), and $\sum P(Y = y) = 1$. Since latent class model is associated with categorical indicator variables with categorical latent variable, let L represent the categorical latent variable with c = 1,...,C latent classes.

Also let γ_c represent latent class prevalence. In LCA, examinees or individuals cannot belong to more than one class, hence the concept is considered mutually exclusive. Then,

$$\sum_{c=1}^{C} \gamma_c = 1 \qquad (2.4)$$

$\rho_{j,rj|c}$ Indicates the probability of response r_j to variable j, conditional on latent class membership c. Since respondents supply only one response alternative each to indicator variable j,

$$\sum_{r_j=1}^{R_j} \rho_{j,rj|c} = 1 \qquad (2.5)$$

Assuming local independence, joint probabilities within latent class are provided as follows:

$$P(Y = y, L = c) = P(L = c)P(Y = y|L = c) = \gamma_c \prod_{j=1}^{J} \prod_{r_{j=}1}^{R_j} \rho_{j,rj|c}^{I(y_j = r_j)} \qquad (2.6)$$

γ_c = P(L = c), I($y_j = r_j$) = 1 when j = r_j, and 0 otherwise. The marginal distribution of Y is found as:

$$P(Y = y) = \sum_{c=1}^{C} P(Y = y, L = c). \qquad (2.7)$$

Therefore, $$P(Y = y) = \sum_{c=1}^{C} \gamma_c \prod_{j=1}^{J} \prod_{r_{j=1}}^{R_j} \rho_{j,rj|c}^{I(y_j = r_j)} \quad (2.8)$$

Finally, by using Bayes' Theorem, the posterior probability of membership in latent class c, conditional on response pattern y is given by:

$$P(L = c \mid Y = y) = \frac{P(Y=y \mid L=c)P(L=c)}{P(Y=y)} \quad (2.9)$$

Substituting Equations 2.6 and 2.8 into Equation 2.9 yield

$$P(L = c \mid Y = y) = \frac{(\prod_{j=1}^{J} \prod_{r_{j=1}}^{R_j} \rho_{j,rj|c}^{I(y_j=r_j)})\gamma_c}{\sum_{c=1}^{C} \gamma_c \prod_{j=1}^{J} \prod_{r_{j=1}}^{R_j} \rho_{j,rj|c}^{I(y_j=r_j)}} \quad (2.10)$$

2.3.5 Ordered Latent Class Model

In the traditional latent class model proposed by Lazarsfeld & Henry (1968), we are unable to rank order the latent classes because they are measured at the nominal level, and are devoid of assumptions that would permit such ranking. To ensure ranking of the latent classes, Croon (1990) proposed a new model by imposing inequality constraints directly on item response and cumulative response probabilities. However, if the items are nominal with no ordering of the response categories, imposing order constraint on response probabilities will have no meaning. Croon's approach is useful for items with ordered response categories.

The inequality constraints allow latent classes to be ordered along the latent continuum, and it is associated with the core assumption that individuals in a higher latent classes have a higher probability of responding correctly/positively to an item, it also means that the probability of a negative response to an item is a decreasing function of the latent class number. Suppose in a particular latent class analysis, C different latent classes are ordered along an ordinal latent

continuum such that latent class 1 represent the lowest and latent class C the highest level along the latent continuum, two scenarios immediately follow:

Dichotomous response

For dichotomous items with categories (r = 2), with response category 1(representing a 'negative' or incorrect response) and 2 (representing a 'positive' response), let c represent an arbitrary latent class such that latent class c is lower than latent class c + 1. Again let ρ_{j2c} represent the probability that an individual belonging to latent class c responds positively to item j, then from the core assumption of monotonicity, we achieve the following inequalities:

$$\rho_{j21} \leq \cdots \leq \rho_{j2c} \leq \rho_{j2,c+1} \cdots \leq \rho_{j2C}$$

Polytomous response

For polytomous items with $r > 2$ response categories, the entire set of categories $\{1, \ldots, r-1, r, \ldots, s\}$ is dichotomized into two non-overlapping sets $\{1, \ldots, r-1\}$ and $\{r, \ldots, s\}$ for each response category r: $2 \leq r \leq s$. Dichotomizing the response categories should also reflect the original responses to items. For instance, individuals whose original response belongs to category subset $\{1, \ldots, r-1\}$ is viewed as a negative response, and those responses belonging to subset $\{r, \ldots, s\}$ are recoded as a 'positive' response. Just like the case of dichotomous items, all dichotomizations in this case should also satisfy the core assumption of monotonicity.

Let ρ_{jgc} represent the probability that an individual belonging to latent class c chooses category g on item j, then following the monotonicity condition, the system of inequalities are expressed as follows: for $1 \leq c \leq C-1$, and for $r \leq$ s,

$$\sum_{g=r}^{s} \rho_{jg,c} \leq \sum_{g=r}^{s} \rho_{jg,c+1}$$

It is important to note that $\sum_{g=1}^{s} \rho_{jgc} = 1$, for each latent class c and each item j. This means that for r: $1 \leq r \leq s-1$, and c: $1 \leq c \leq C-1$, the equivalent system of inequalities is provided as follows:

$$\sum_{g=1}^{r} \rho_{jg,c} \geq \sum_{g=1}^{r} \rho_{jg,c+1}$$

Even though this study has adopted Croon's non-parametric approach for ordering classes, parametric or non-parametric order constraints could be used for ordered polytomous categories (Vermunt, 2001). Whilst the non- parametric approach is based on imposing inequality constraint on response probabilities, the parametric method imposes linear equality constraints on the response probabilities.

2.4 Statistical models assuming dynamic latent variables

Selected models to be discussed under this section include the latent transition analysis model; which is considered as a special case of the latent Markov model, a special case of Hidden Markov Model in Bayesian knowledge tracing, and the longitudinal IRT models.

2.4.1 Latent Transition Analysis Model

LTA presents an alternative method to modeling longitudinal data. It is viewed as an extension of latent Markov model (Velicer, Martin, & Collins, 1996). Early work in this field include Collins & Wugalter (1992), Graham, Collins, Wugalter, Chung, & Hansen (1991), Langeheine (1988, 1994), Langeheine & Van de Pol (1990), Van de Pol & de Leeuw (1986), and Van de Pol & Langeheine (1990).

The importance of LTA cannot be overemphasized. LTA may be used to provide answers to pertinent research questions including testing for treatment effect by comparing different groups, assessing the impact of varying measures for each latent status, and testing for different theoretical models pertaining to the pattern of change from one time to the next (Velicer, Martin, & Collins, 1996).

The presentation of the LTA model in this study is entirely consistent with Collins & Lanza (2010, pp.196 - 198). For simplicity, this LTA model is assumed to have no missing data on the indicator variables. Suppose there are j = 1, .., J indicator variables measured at t = 1, .., T times, then j has $r_{j,t}$ = 1, ..., $R_{j,t}$ response categories. Let W = $\prod_{t=1}^{T}\prod_{j=1}^{J} R_j$ represent the number of cells in the contingency table obtained by cross-tabulating the j variables at times T. Let y = $(r_{j,1},\ldots r_{J,T})$ represent the response patterns. Each y corresponds to P(Y = y), and $\sum P(Y = y) = 1$, where Y represents the array of response patterns with W rows and T x J columns.

Also, let the general categorical variable be represented by L. Then L will have S latent statuses, so for L_1 representing the categorical latent variable at Time 1, there is $s_1 = 1,\ldots S$; and for L_2 representing the categorical variable at Time 2, there is $s_2 = 1,\ldots S$. Finally, for L_T

representing the categorical latent variable at Time T, there is $s_T = 1, \ldots S$. Despite the technical possibility of obtaining an empty latent status at a particular time, we are assuming for the purpose of easiness that the latent statuses is the same across times; such that $S_1 = S_2 = \ldots = S_T = S$.

As we discussed above, three parameters are of interest when considering LTA model: latent status prevalences, item-response probabilities, and transition probabilities. With respect to the latent status prevalences, the latent statuses are considered to be mutually exclusive and exhaustive at each Time t, this means that each respondent is a member of one and only one latent status at Time t. So at a particular Time t, the latent status prevalences add up to 1.

$$\sum_{s_{t=1}}^{S} \delta_{s_t} = 1 \qquad (2.11)$$

where δ_{s_t} represents the probability of membership in status s at Time t.

For item-response probabilities, each respondent supplies one and only one response alternative to variable j at a certain time t, so the probabilities of each of the response alternatives to variable j add up to 1.

$$\sum_{r_{j,t=1}}^{R_j} \rho_j, r_{j,t|s_t} = 1 \qquad (2.12)$$

for all j , t. Where $\rho_j, r_{j,t|s_t}$ represents the probability of response $r_{j,t}$ to indicator variable j, conditional on latent status membership s_t at Time t. Again, since latent status membership is mutually exclusive and exhaustive, each row of the transition probability matrix adds up to 1.

$$\sum_{s_{t+1=1}}^{S} \tau_{s_{t+1}} | s_t = 1. \qquad (2.13)$$

Where $\tau_{s_{t+1}}|s_t$ represents the probability of transitioning to latent status s at Time t + 1 conditional on latent status memberships at Time t.

Now, putting all the parameters together,

$$P(Y = y) = \sum_{s_1=1}^{S} \ldots \sum_{s_T=1}^{S} \delta_{s_1} \tau_{s_2|s_1} \ldots \tau_{s_T}|s_{T-1} \prod_{t=1}^{T} \prod_{j=1}^{J} \prod_{rj,t=1}^{R_j} \rho_{j,r_{j,t}|s_t}^{I(y_{j,t}=r_{j,t})} \quad (2.14)$$

The indicator function I($y_{j,t} = r_{j,t}$) equals 1 when the response to variable j = r_j at Time t, and equals 0 otherwise.

Considering two time measurements, equation 2.23 reduces to

$$P(Y = y) = \sum_{s_1=1}^{S} \sum_{s_2=1}^{S} \delta_{s_1} \tau_{s_2|s_1} \prod_{t=1}^{2} \prod_{j=1}^{J} \prod_{r_{j,t=1}}^{R_j} \rho_{j,r_{j,t}|s_t}^{I(y_{j,t}=r_{j,t})} . \quad (2.15)$$

2.4.2 Bayesian Knowledge Tracing Model

The Bayesian Knowledge tracing (BKT) model introduced by Corbett & Anderson (1995) is a special case of the Hidden Markov Model, which models students' knowledge as a latent variable. This latent variable is updated constantly depending on the correctness of information provided by students who have the opportunity to apply a specific skill. BKT has been used in intelligent tutoring systems for a long period specifically with reference to mastery acquisition /learning and problem sequencing. Several intelligent systems such as tutors for computer programming, mathematics, and reading skills have employed the BKT technique to predict the performance of students, and also to ascertain when a student has achieved mastery of a specific skill (Beck & Chang, 2007; Corbett & Anderson, 1995; Koedinger, 2002).

Corbett and Anderson's Bayesian Knowledge Tracing model assumes that at any opportunity given to a student to showcase his/her skill, the student either knows the skill or does not, and may consequently provide a correct or incorrect response. Also, when a student requests

for help from the tutor, the model treats it as an incorrect response. The model also assumes that once a student knows a skill he/she does not forget. Furthermore, the model associates each skill with one set of parameters, and these parameters are consistent for every student.

According to Corbett & Anderson (1995), the BKT model assumes four parameters for each skill, two of which are described as knowledge parameters, and the other two as performance parameters. The knowledge parameters are initial or prior knowledge (L_0) and the acquisition (T) parameter. The initial knowledge parameter P(L_0) is the probability that a student knew the skill prior to being given the opportunity to use the tutor. The acquisition parameter P(T) is the probability that a student will transition from an unknown to known state after being given the opportunity to interact with the tutor. The performance parameters are guess (G) and slip (S).

Generally, we expect students who do not know a skill to provide an incorrect response when given the opportunity to apply the skill. But there is a certain probability P(G) that a student will guess right and provide a correct response even if he /she does not know the skill associated with the question. Correspondingly, it is expected that students who know a skill will provide a correct response when given the opportunity to apply it. But there is a probability P(S) that a student will slip and provide an incorrect response even if he/she knows the skill.

The intelligent tutor constantly updates its estimate that a student knows a skill every time the student provides a first response to a problem step. The system achieves this by first re-calculating the probability that the student knew the skill prior to the response using equations 2.16 and 2.17, and then using equation 2.18, the system calculates the probability that the student learned the particular skill during the problem step. From the parameter values, the probability that a student knows a skill P(L_n) after n opportunities to apply the skill is calculated below.

Except for some few modifications, the equations below are entirely consistent with the ones provided by Baker, Corbett, & Aleven (2008). If we let y represent scores such that P(y =1), and P(y = 0) indicate the respective probabilities of correct and incorrect responses, then

$$\mathrm{P}(L_{n-1} \mid y = 1) = \frac{P(L_{n-1}) * (1-P(S))}{P(L_{n-1}) * (1-P(S)) + (1-P(L_{n-1})) * (P(G))} \quad (2.16)$$

$$\mathrm{P}(L_{n-1} \mid y = 0) = \frac{P(L_{n-1}) * P(S)}{P(L_{n-1}) * P(S) + (1-P(L_{n-1})) * (1-P(G))} \quad (2.17)$$

$$P(L_n \mid Action_n) = P(L_{n-1} \mid Action_n) + \left(\left(1 - P(L_{n-1} \mid Action_n)\right) * P(T)\right) \quad (2.18)$$

where $Action_n$ represents correct (y =1) or incorrect (y=0) responses after n opportunities.

Researchers have been working on the BKT concept to improve the predictability and interpretability of the intelligent tutoring system. In their original work, Corbett & Anderson (1995) added individualization to their model in order to improve the predictability of the tutoring system, but they achieved mixed results. Relative to the non-individualized model, their proposed individualized BKT model did not show an improvement on the overall predictive accuracy of the tutoring system. In fact, Beck & Chang (2007) pointed out that Corbett and Anderson's approach had a problem with model identifiability. The reason being, different combinations of the model parameters could fit the data equally well but provided different predictive results.

Beck and Chang then proposed using Dirichlet approach to constrain model parameters but Baker, et al. (2008) showed that Beck and Chang's approach was vulnerable to what they described as model degeneracy, where in some cases the probability that a student knows a skill dropped after the student has actually answered three successive questions. Baker, Corbett, & Aleven (2008) proposed using machine learning to contextualize the guess and slip parameter.

Whist this approach showed much improvement than the existing approaches, the model is yet to be validated externally (Pardos & Heffernan, 2010). Pardos & Heffernan proposed a new individualization approach in a Bayesian network framework that simultaneously fit individualized and skill specific parameters in a single step. Using the individualized prior parameters, the authors have shown big improvement on the predictive nature of their model compared to the standard knowledge tracing model.

2.4.3 Longitudinal IRT Model

Over the years, educational research and assessment have paid particular attention to measuring the performance of students in a longitudinal fashion; by considering the performance of students between testing time points (Fischer, 1989). Some of the researches aimed at measuring students' growth have been conducted on small- scale data, while others have been conducted on a larger -scale educational data. For instance, Davier, Xu, & Carstensen (2011) employed multidimensional item-response-theory (MIRT) models for longitudinal IRT on a larger-scale educational data. Meaningful growth or change in students' performance between testing time points can be measured successfully either by focusing on individuals (Andersen, 1985; Andrade & Tavares, 2005; Embretson, 1991) or focusing on groups (Fischer, 1973, 1976, 1989; Wilson, 1989).

2.4.3a Measuring Growth at the Group level

Fischer (1977a, 1977b, 1983a) developed linear logistic model with relaxed assumptions (LLRA) within the generalized Rasch model framework, for the purpose of determining change in dichotomous item score matrices between two time points. Suppose we have subject S_v, for v = 1, . . . , N subjects, i = 1, . . . , k items, and two time points T_1 and T_2, let θ_{vi} represent latent parameter for S_v on trait dimension D_i with item I_i as an indicator. Also let δ_v represent the

amount of change in S_v within time points T_1 and T_2. Let w_{vh} represent the elements of the matrix w which is deemed given. Again, let β_h represent basic parameters for treatment effects, interactions, experimental conditions, and so on. Fischer then formulated the LLRA equations as follows:

$$P(y = 1|S_{v,}I_{i,}T_1) = \frac{exp(\theta_{vi})}{1+\exp(\theta_{vi})}, \quad (2.19)$$

$$P(y = 1|S_{v,}I_{i,}T_2) = \frac{exp(\theta_{vi}+ \delta_v)}{1+\exp(\theta_{vi}+ \delta_v)}, \quad (2.20)$$

where $\delta_v = \sum_{h=1}^{u} w_{vh}\beta_h$. (2.21)

Among the assumptions of LLRA model were (a) using the same test items at the two time points T_1 and T_2, and (b) local independence for responses. The very idea that the model was based on using same items at the two time points presented a measurement challenge: Unless the well-known issues of Testing, practice, or memory effects were somehow accounted for in the model, it run the risk of yielding misleading results. Another issue was that if it was determined that the amount of change between T_1 and T_2 was large as one would reasonably expect in developmental studies of children, the appropriateness of item difficulty level at T_2 could be questioned especially if those same items were used at an appropriate difficulty level at T_1.

Fischer (1977b) proposed a "hybrid model" to overcome the problems posed by the LLRA model. The hybrid model is based on having pairs of items I_i and I_l, both items have different difficulty levels but are measuring the same latent dimension. The hybrid model is hybrid in the sense that it merged the assumption of multidimensionality of LLRA model (a latent trait per pair of items) with the assumption of unidimensionality of the Rasch model (within pairs of items).

Fischer (1989) extended the hybrid model to designs with multiple time points with different sets of items per time point, so long as" one unidimensional subscale is available per latent trait". Unlike the LLRA model, the hybrid model considers any number of time points, and it is also designed to account for difficulty levels of items used at different time points. Suppose we have G_g treatment groups with g = 1, . . . , n groups , for subject $S_v \in G_g$ (v = 1, . . . , N subjects, i = 1, . . . , k dimensions D_i, t = 1, . . , s time points T_t, and j = 1, . . , m treatments M_j). Let θ_{vi} represent the parameter of subject $S_v \in G_g$ on latent dimension D_i; σ_{it} represents the easiness parameter. Also let $\acute{q}_{g|t}$ represent dose of treatment $M_j$provided to subjects $S_v \in G_g$ up to time point T_t. η_j represents treatment effect ; and the amount of change in subject $S_v \in G_g$ up to time point T_t is represented by δ_{gt}. Fischer (1989) represented the general hybrid model as:

$$P(y = 1 \mid S_v, D_i, T_t) = \frac{\exp(\theta_{vi} + \sigma_{it} + \delta_{gt})}{1+\exp(\theta_{vi} + \sigma_{it} + \delta_{gt})} \quad (2.22)$$

where $\delta_{gt} = \sum_j \acute{q}_{gjt}\eta_j + \sum_{j<l} \acute{q}_{gjt}\acute{q}_{glt}\rho_{jl} + (T_t - T_1)\tau$ (2.23)

ρ_{jl} represents the first order treatment interactions, and τ represents trend. Another model of importance to the hybrid model is the linear logistic test (LLTM) model which was introduced by Fischer (1983a). The LLTM is deemed as a form of Rasch model which has linear constraint imposition on its item parameters. The difference though between the LLTM and the hybrid LLRA model is that the LLTM describe subjects in that model each by a scalar parameter, but in the hybrid model, each subject is described by a vector parameter (Fischer, 1989).

Fischer (1989) successfully interpreted equations (2.22) and (2.23) as LLTM seen in equations (2.24) and (2.25) because of the linear nature of the exponents in the hybrid model. Fischer posited that the advantage of interpreting the hybrid model as LLTM lies in the easiness of deriving estimation equations for the "basic" parameters β_h in the LLTM framework.

$$P(y = 1 \mid S_v \in G_{g,}\, D_i\,, T_t) = \frac{\exp(\theta_{vi} + \alpha_{git})}{1+\exp(\theta_{vi} + \alpha_{git})}, \quad (2.24)$$

with $\alpha_{git} = \sum_h w_{git;h}\beta_h$, (2.25)

where g = 1, . . , n groups $G_{g,}$ i = 1, . . . , k dimensions D_i , and t = 1, . . , s time points T_t.

Wilson (1989) also developed the Saltus model in the dichotomous form which is based on the progression of students through developmental stages. One of the fundamental assumptions of the Saltus model is that developmental stages or levels are represented by class C, and all students or persons belonging to a particular class c respond to all items in the way that are consistent with the class membership. The Saltus model differs from the LLTM in the sense that the Saltus model treats student's current stage as latent, but item difficulties are split into known components under LLTM (Von Davier, Xu, & Carstensen, 2011). The Saltus model is also applied to polytomous items (Draney, 1996; Draney & Wilson, 2007; Wilson & Draney, 1997).

2.4.3b Measuring Growth at the individual level

Anderson (1985) proposed a model to assess individual growth through repeated administration of same items in different time points. Embretson (1991) noted that the abilities of the Anderson's model are specific to the time point; they do not reflect change but rather the ability level at each time point. Based on the Anderson's model, changes across time points are

derived by calculating the differences between time point -specific abilities (Von Davier, Xu, & Carstensen, 2011).

Von Davier, Xu, & Carstensen (2011) expressed Anderson's model as follows:

$$P\left(Y_{ijk} = 1 | \theta_{jk}, \beta_i\right) = \frac{exp(\theta_{jk} - \beta_i)}{1 + exp(\theta_{jk} - \beta_i)}, \quad (2.26)$$

where θ_{jk} represents the ability of an individual j at time point k, and β_i represent the difficulty of item i. The item difficulty remains constant at different time points (since same items are being used), but the abilities corresponding to each time points may of course differ. Once items are repeated on several time points, item responses may be affected by testing or memory effect or both. Embretson (1991) proposed a multidimensional Rasch model for learning and change (MRMLC) to assess individual differences in change.

For MRMLC model, test items do not have to repeat as in the case of Anderson's model. However, the MRMLC assumes that at each time point, the numbers of abilities go up. Von Davier, Xu, & Carstensen (2011) represented the MRMLC as

$$\mathrm{P}\left(Y_{ijk} = 1 | \left(\theta_{j1}, .., \theta_{jk}\right), \beta_i\right) = \frac{exp(\sum_{m=1}^{k} \theta_{jm} - \beta_i)}{1 + exp(\sum_{m=1}^{k} \theta_{jm} - \beta_i)}, \quad (2.27)$$

θ_{jm} and β_i are defined as in (2.18). The MRMLC can also be applied to polytomous items (Fischer, 2001).

Chapter 3

Introduction

Several statistical models have been developed to model the performance or growth of students over time. As discussed in section 2.4.1, the LTA model proposed by Graham and colleagues is a type of autoregressive model specifically designed to model transitions between class memberships from one time point to the next. However, based on literature found on LTA, the latent statuses are unordered. This study focuses on an extension of the LTA model to a concept referred to as ordered latent transition analysis (OLTA) model. The OLTA model espouses the idea that it is reasonable (in educational measurement) to order students on the ability continuum whilst tracking their development overtime. The OLTA principle is in agreement with the tenet of learning progression framework; which is that students are of different levels on the progression path. Under the OLTA model, the latent statuses are referred to as learning levels.

3.1 Objectives of this study

This study extends the LTA model by imposing inequality restrictions on the item response probabilities using Croon's technique (Croon, 1990). In order to ensure ordered learning levels, the inequality restrictions are placed on item-response probabilities at time point 1, whilst tracking the progression of the students at the second time point. The goals of this study are as follows:

- Develop and evaluate an EM algorithm to estimate the OLTA model, and to determine the statistical properties of the estimated model
- Selection Techniques
    - Three simulations are involved, the first simulation is to determine how well the AIC, and BIC select the optimal (learning) levels
    - The second simulation is aimed at determining how good the AIC and BIC are in selecting correct transition model that best fits the data.
    - The third simulation is used to determine
        - Learning level prevalences
        - Item response probabilities
        - Transition Probabilities
        - Bias, Mean Square Error (MSE), and Root Mean Squared Error (RMSE) of parameter estimation
- The final objective is to demonstrate the use of OLTA model in a real data set.

3.2 The LTA Model

For the purpose of simplicity, this model is presented using only two time points (occasions) of measurement. The first measurement is taken at time t, and the second is taken at time t + 1. Let us assume that we have only three items such that item 1, has $i, i', = 1, \dots I$ response categories, item 2 has $j, j', = 1, \dots J$ response categories, and item 3 has $k, k', = 1, \dots K$ response categories, where i, j, k are response categories associated with time t, and the second time point t + 1 is associated with response categories i', j', k'. There is an exogenous static latent variable that divides the student population into levels c = 1,...C, measured by indicator

with m = 1, . . . M response categories. Let us assume there are u , v = 1, . . . , S learning levels with u representing learning level at time t, and v representing learning level at time t + 1.

With a very minor modification to Collins & Wugalter (1992), the estimated proportion of response pattern P(Y) is represented as follows:

$$P(Y = y) = \sum_{c=1}^{C} \sum_{u=1}^{S} \sum_{v=1}^{S} \gamma_c \rho_{m|c} \delta_{u|c} \rho_{i|u,c} \rho_{j|u,c} \ \rho_{k|u,c} \tau_{v|u,c} \rho_{i'|v,c} \rho_{j'|v,c} \rho_{k'|v,c} \qquad (3.1)$$

For the purpose of clarification, γ_c is what is known to be latent class prevalences under latent class models; they are henceforth referred to as level prevalences under OLTA model. With respect to equation 3.1, γ_c represents the proportion in level c; $\delta_{u|c}$ represents the proportion in learning level u at Time t conditional on membership in level c; that is, the proportion of level c members whose learning level is u at Time t. An element of latent transition probability matrix $\tau_{v|u,c}$ represents the probability of membership in learning level v at Time t+1 conditional on membership in learning level u at Time t and membership in level c; that is, the proportion of those level c and learning level u at Time t who are in learning level v at Time t+1. Also, $\rho_{i|u,c}$ represents the probability of response i to item 1 at Time t, conditional on membership in learning level u at Time t and on membership in level c; $\rho_{i'|v,c}$ represents the probability of response i' to item 1 at Time t + 1, conditional on membership in learning level v at Time t + 1 and on membership in level c, and so on. $\rho_{m|c}$ also represents the probability of having a value of m on the indicator of level membership, conditional on membership in level c.

3.3 Item Response Probabilities under the OLTA model

Currently, literature concerning LTA model have unordered levels. This project extends the LTA model by imposing inequality restrictions (Croon, 1990) on the item response probabilities. In order to ensure ordered learning levels, the inequality restrictions are placed on item-response probabilities at time point 1. So, at time point 1, if we let ρ_{jrs} represent the probability that a subject belonging to learning level s answers with category r to item j, then the monotonicity condition lead to the following system of inequalities on these response probabilities: for $1 \leq s \leq S-1$ and for $1 \leq k \leq m$, they should satisfy:

$$\sum_{r=k}^{m} \rho_{jr,s} \leq \sum_{r=k}^{m} \rho_{jr,s+1.}$$

Since for each learning level s and each item j one has

$$\sum_{r=1}^{m} \rho_{jrs} = 1,$$

The first system of inequalities is equivalent to the following one:

$$\sum_{r=1}^{k} \rho_{jr,s} \geq \sum_{r=1}^{k} \rho_{jr,s+1}$$

For k: $1 \leq k \leq m-1$ and s: $1 \leq s \leq S-1$. For the analysis with S learning levels and items with m response categories we have $(m-1) \times (S-1)$ linearly independent inequalities of this kind per item.

The inequalities described above is the heart of the OLTA model, the idea that the learning levels are ordered at time 1; such that subjects belonging to a higher learning level have a higher probability of responding correctly to an item. Generally, some restrictions are considered, and imposed when dealing with LTA models. Collins & Wugalter (1992) have described two general restrictions that are deemed useful: restricting parameter to a specified value, and setting estimated parameters to be equal to each other. The OLTA model will adopt the same principle of parameter restrictions.

In order to enhance the conceptual integrity of the model, the item response probabilities ($\rho's$) will be constrained to be equal across times. The practical reason for constraining item response probabilities to be equal across times is to aid in stabilizing estimation and enhancing identification (Collins & Lanza, 2010, p. 213). Under the LTA model, a lot of item response probabilities are produced; especially when several time points are being considered. Similarly, OLTA models are bound to yield several item-response probabilities, so providing parameter restrictions on the model will help to greatly reduce the number of estimated parameters and increase the degrees of freedom.

Model Identification

The OLTA model faces identification problems similar to LTA models or any other latent class model. There is a necessary but insufficient condition for model identification, that is the number of item parameters to be estimated must not exceed one less than the number of response patterns. Goodman (1974) posited that we could inspect model identification by checking the rank of the matrix of partial derivatives of (all but one) the cell probabilities with respect to the individual parameters to be estimated. If it is established however that the matrices

are of full rank, then the model is identified. Otherwise, we must decrease the number of parameters being estimated.

Two Time Points with Different Items per Time Point

In educational measurement, presenting the same items to students in different time points T_1 and T_2 could create a number of measurement problems. Since students have the capacity to remember previous questions, the well-known practice and / or testing effect could render misleading results. Also, if same questions are given in different time points, the difficulty level of the items obviously remain constant, however, we expect students to mature, so items that may be of an appropriate difficulty level at time T_1 may not be appropriate at time T_2; since students will almost surely provide correct responses to those items.

In order to prevent extreme probability responses, this paper is adopting the concept from Fischer (1989) by assuming pairs of items (I_i, I_l), with both items measuring the same latent dimension but having different difficulty levels.

3.4 Hypothesis Testing about Change between Times

Several research questions concerning change over time may be of interest. For instance, a general question may arise as to whether or not there is a significant amount of change between the two time points 1 and 2. Several models may be used to address this problem, but for the purpose of this study, the generic models used are specified below:

Time 2 level

$$\text{Time 1 level} \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Model A. Fixed transition probabilities

As a baseline, we will test for Model A. This is achieved by fixing all the diagonal transition probability elements to 1 for each learning level, and the rest of the elements in each row are also fixed to 0. In a nutshell, no transition probabilities are estimated for Model A, and the Model espouses the idea that learning level membership at time point 2 is the same as that of time point 1 for all subjects.

Time 2 level

$$\text{Time 1 level} \begin{bmatrix} \tau_{1|1} & \tau_{2|1} & \tau_{3|1} & \tau_{4|1} & \tau_{5|1} \\ \tau_{1|2} & \tau_{2|2} & \tau_{3|2} & \tau_{4|2} & \tau_{5|2} \\ \tau_{1|3} & \tau_{2|3} & \tau_{3|3} & \tau_{4|3} & \tau_{5|3} \\ \tau_{1|4} & \tau_{2|4} & \tau_{3|4} & \tau_{4|4} & \tau_{5|4} \\ \tau_{1|5} & \tau_{2|5} & \tau_{3|5} & \tau_{4|5} & \tau_{5|5} \end{bmatrix}$$

Model B. Unrestricted/Saturated transition probabilities

Model B has the full transition probability matrix. This model allows for transition from a less advanced learning level to a more advanced one, or from a more advanced learning level to a less

advanced one. From a cognitive standpoint, Mosher (2011) argued that students who attain higher levels may fall back to a previous level when the conditions at the higher levels become a little unbearable (i.e. facing more difficult problems).

Time 2 level

$$\text{Time 1 level} \begin{bmatrix} \tau_{1|1} & \tau_{2|1} & \tau_{3|1} & \tau_{4|1} & \tau_{5|1} \\ 0 & \tau_{2|2} & \tau_{3|2} & \tau_{4|2} & \tau_{5|2} \\ 0 & 0 & \tau_{3|3} & \tau_{4|3} & \tau_{5|3} \\ 0 & 0 & 0 & \tau_{4|4} & \tau_{5|4} \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Model C. "Growth only" Model

Model C is particularly useful when investigators have reasons to believe that it is impossible for subjects to transition from an advanced learning level at Time 1 to a less advanced learning level at Time 2. The restrictions shown in Model C are consistent with the hypothesis of the OLTA model. The fundamental idea of the OLTA model is that students who are identified to be associated with a higher learning level at Time t under the learning progression framework, are not expected to transition to a less advanced learning level at Time t + 1. It is entirely possible for students belonging to an advanced learning level at Time t to transition to less advanced learning level at Time t +1 when conditions become hostile (Mosher, 2011). However, this backward transition is not expected under the OLTA model within the learning progression framework.

3.5 Assessing Model Fit

Assessment of model fit is essential to LTA or OLTA modeling. The fit of OLTA model is determined in the same manner as latent class or LTA models. OLTA model treat data like a big contingency table, with each cell matching a response pattern. A model being tested predicts the response pattern frequencies. If the model fits the data well, the predicted frequencies will be close to the observed frequencies, and the goodness-of-fit statistic will be small. On the other hand, if the assumed model fits the data poorly, there will be a large disparity between the predicted frequencies and the observed frequencies, and the goodness-of-fit statistic will be large.

For the purpose of assessing model fit, we employ the widely used likelihood-ratio goodness-of-fit statistic, denoted by G^2(Read & Cressie, 1988). The G^2 is approximately distributed as a chi -square with degrees of freedom K-P-1, where K represents the number of response patterns, and P represents the number of estimated parameters. To adequately assess the fit of the three models in section 3.6, the well-known Akaike's Information Criterion (AIC) and the Bayesian Information Criterion (BIC) (Anderson, 1982) will be employed in addition to G^2.

As a measure of goodnesss-of-fit, the AIC considers the number of model parameters (p) that are being estimated in the model.

$$AIC = -2 * lnL + 2 * p.$$

The individual AIC values are unmeaningful and are much affected by sample size. The AIC is rescaled to a more interpretable form as:

$$\Delta_i = AIC_i - AIC_{min} ,$$

where AIC_{min} is the minimum of AIC_i possible values. The Δ_i have straightforward interpretation and "allow a quick strength-of-evidence comparison and ranking of candidate hypotheses or models" (Burnham & Anderson, 2004). Let us consider some rule of thumbs that are deemed useful in assessing the relative value of models in the set: Models having $\Delta_i \leq 2$ have significant support, those in which $4 \leq \Delta_i \leq 7$ have considerably less support, and models having $\Delta_i \geq 10$ have no support (Anderson, 1982; Burnham & Anderson, 2004).

The BIC on the other hand considers the number of parameters (p) and the number of observations (N) as a measure of goodness-of-fit.

$$BIC = -2 * lnL + (\ln N) * p.$$

Of the two criterion information, the BIC applies larger penalties per parameter of ln(N), and as a result turns to generally select simpler models. As with AIC, the model with the smallest value of BIC among all possible assumed models is selected.

3.6 Parameter Estimation

The parameters $\gamma, \delta, \tau, and\ \rho$ can be estimated using the EM algorithm (Dempter, Laird, & Rubin, 1977). The EM algorithm iterates between the E-step, and M-step. In the E-step, the

expected values of the sufficient statistics; which are the response pattern proportions are calculated based on the observed complete data and parameter estimates. In the M-step, we obtain new parameter estimates given the current estimated sufficient statistics in order to maximize the likelihood function. With these new parameter estimates, another E-step can be executed to obtain new parameter estimates, and so on. The suggested technique for estimating the OLTA model parameters by EM algorithm takes advantage of the fact that model can be treated as a constrained latent class analysis model. For instance, if we have a series of assessments that assume K = 6 learning levels assessed at T= 3, time points then there are a total of $K^J = 6^3 = 216$ possible learning trajectories (e.g., 111, 112, . . . , 666).

The OLTA model then implies both equality and inequality constraints on latent class model parameters. Specifically, because the model is a latent transition model, the K^J class probabilities are constrained by Markov transition model assumptions. Furthermore, the response probabilities are constrained according to which class an individual is in at a given time point, and also restricted by the ordering constraints. The E-step of the suggested technique remains unchanged from the standard latent class analysis model; however, the M-step is modified to handle the various constraints. For dichotomous items, the ordering constraint for the response probabilities will use the pooled-adjacent-violators-algorithm (PAVA; Ayer et al., 1955; Robertson, Wright & Dykstra, 1988; de Leeuw, Hornik & Mair, 2009). Ordering constraints for polytomous items can be handled using the methods described by El Barmi & Johnson (2006), which uses Lagrange multipliers. The detailed estimation of the OLTA model can be found in Appendix A.

Generation of Data

Since there is a finite set of items with finite response categories, there is a finite set of possible responsible patterns. Just like LTA models, every OLTA model yields a corresponding vector of predicted response pattern probabilities. In this simulation study there are such probability vectors corresponding to the combinations of the two proposed models, and the two numbers of items. The probability vectors will be used to generate random data. A random number between 0 and 1 will be generated by means of a uniform random number generator for each simulated subject. This number will then be compared to the cumulative response pattern probability vector in order to place the simulated subject in one of the response pattern cells.

Statistical Properties

The simulation study is designed to study the statistical properties of the proposed EM algorithm. Specifically we will examine the bias and root mean squared error of the model parameters. For example if θ is the true value of some model parameter of interest, and $\hat{\theta}_r$ is the estimate obtained from replicated data set r, then we will approximate the bias of the estimator with the observed average

$$\text{Bias}\left(\hat{\theta}\right) \approx \frac{1}{R}\sum_{r=1}^{R}\hat{\theta}_r - \theta$$

The root mean squared error will be approximated with

$$\text{RMSE}(\hat{\theta}) \approx \frac{1}{R}\sum_{r=1}^{R}(\hat{\theta}_r - \theta)^2.$$

The bias and RMSE will be approximated for each of the experimental conditions and compared.

In addition to using the simulated data sets to examine the statistical properties of the estimators, they will be used to study the ability of the different fit statistics (e.g., AIC, BIC, etc.) described in Section 3.5 to select the correct transition model. For each simulated data set i will estimate the OLTA model with each of the transition models described in Section 3.4, and the proportion of times each model was selected as the best fitting model will be reported. The hope would be that the fit statistics would tend to choose the correct model a large majority of time.

Chapter 4

Simulation

4.1 Introduction

This chapter presents three detailed simulation studies. The first study is aimed at ascertaining the ability of the EM algorithm to recover the OLTA model parameters, and to ensuring the robustness of the estimation procedure under several conditions. The second study is aimed at establishing how well the information criteria (AIC and BIC) select the true transition model. The third simulation study is to establish how well the AIC and BIC select the optimal learning level.

4.2 Study 1

In this study we examined the effects of including four different scenarios in our model estimation: true saturated model estimated as saturated; true saturated model estimated as growth; true growth model estimated as growth, and true growth model estimated as saturated. By examining and contrasting these four estimation scenarios, we are able to determine when it is absolutely necessary to use a particular model estimation, and also determine situations that allow for both growth and saturated models to be used interchangeably under the OLTA model.

Having good items are necessary for the success of any model. For this reason items were generated from three different discrimination indexes. The discrimination index is the log odds ratio for correct responses between two adjacent levels. Chen, Cohen, & Chen (2010) described log odds ratio of 0.2, 0.5, and 0.8 as small, medium, and large, respectively. For the purpose of

this simulation experiment, we selected low discrimination (alpha =0.5), medium discrimination (alpha = 1), and high discrimination (alpha = 2). Adding the discrimination indexes to the simulation design was useful in examining the behavior of the response probabilities, and the transition models under several conditions, and also helped to ascertain the accuracy of the order constraint imposed under the OLTA model.

4.2.1 Method

- Model Type. Two different models were investigated: The Saturated model (Model B) and the 'Growth' model (Model C). The unrestricted/saturated model involves noncumulative development with the possibility of developmental reversals. The 'Growth' model on the other hand involves cumulative development with no developmental reversals. Each model involves two time points and two conditions of learning levels: three and five leaning levels. Except for some minor changes, the true transition matrices for the three learning levels in the simulation for both Growth and Saturated models were adopted from Collins & Wugalter (1992). The true transition matrices for Growth and Saturated models are respectively provided as follows:

$$\begin{bmatrix} .5 & .3 & .2 \\ 0 & .6 & .4 \\ 0 & 0 & 1 \end{bmatrix} \text{ and } \begin{bmatrix} .5 & .3 & .2 \\ .1 & .4 & .5 \\ .1 & .3 & .6 \end{bmatrix}.$$

The initial learning proportions for the two models, also adopted from (Collins & Wugalter, 1992) were the same: 0.5, 0.3, and 0.2 for learning levels 1, 2, and 3 respectively. The true tau

parameters for the five learning levels for the Growth and Saturated models are respectively provided as follows:

$$\begin{bmatrix} .4 & .3 & .2 & .1 & .0 \\ .0 & .4 & .3 & .2 & .1 \\ .0 & .0 & .5 & .3 & .2 \\ .0 & .0 & .0 & .6 & .4 \\ .0 & .0 & .0 & .0 & 1.0 \end{bmatrix} \text{ and } \begin{bmatrix} .4 & .2 & .2 & .1 & .1 \\ .2 & .3 & .3 & .1 & .1 \\ .1 & .1 & .3 & .3 & .2 \\ .0 & .1 & .2 & .4 & .3 \\ .0 & .0 & .1 & .1 & .8 \end{bmatrix}$$

With initial learning proportions 0.3, 0.3, 0.2, 0.1, and 0.1 for learning levels 1, 2, 3, 4, and 5 respectively.

- Number of items. The number of items was manipulated in order to vary the sparseness of the contingency table. We examined two different item sizes of 10 and 20 dichotomous items each for the two time points, and for each learning level. Each item was generated for low, medium, and high discriminations for each of the two time points, and for each learning level.
- Learning Levels: 3 and 5 learning levels were considered.
- Sample Size. The sample size also affects the sparseness of the contingency table. Two different sample sizes: N = 1000 and N=2000 were used for each of the conditions listed above.
- Model Estimation. Four different scenarios were considered: true saturated model estimated as saturated; true saturated model estimated as growth; true growth model estimated as growth, and true growth model estimated as saturated.

4.2.2 Results of Simulation Study 1

As already indicated, ninety-six different conditions were used for this simulation study, but for the purpose of simplicity, few plots will be discussed in this chapter. For all the plots considered in this chapter and beyond, the dash lines "-------" represent the correct model estimation: when true Growth model is estimated as Growth, or when true Saturated model is estimated as Saturated.

The solid lines " ── " on the other hand, represent model misspecification: when true Growth model is estimated as Saturated or when true Saturated model is estimated as Growth.

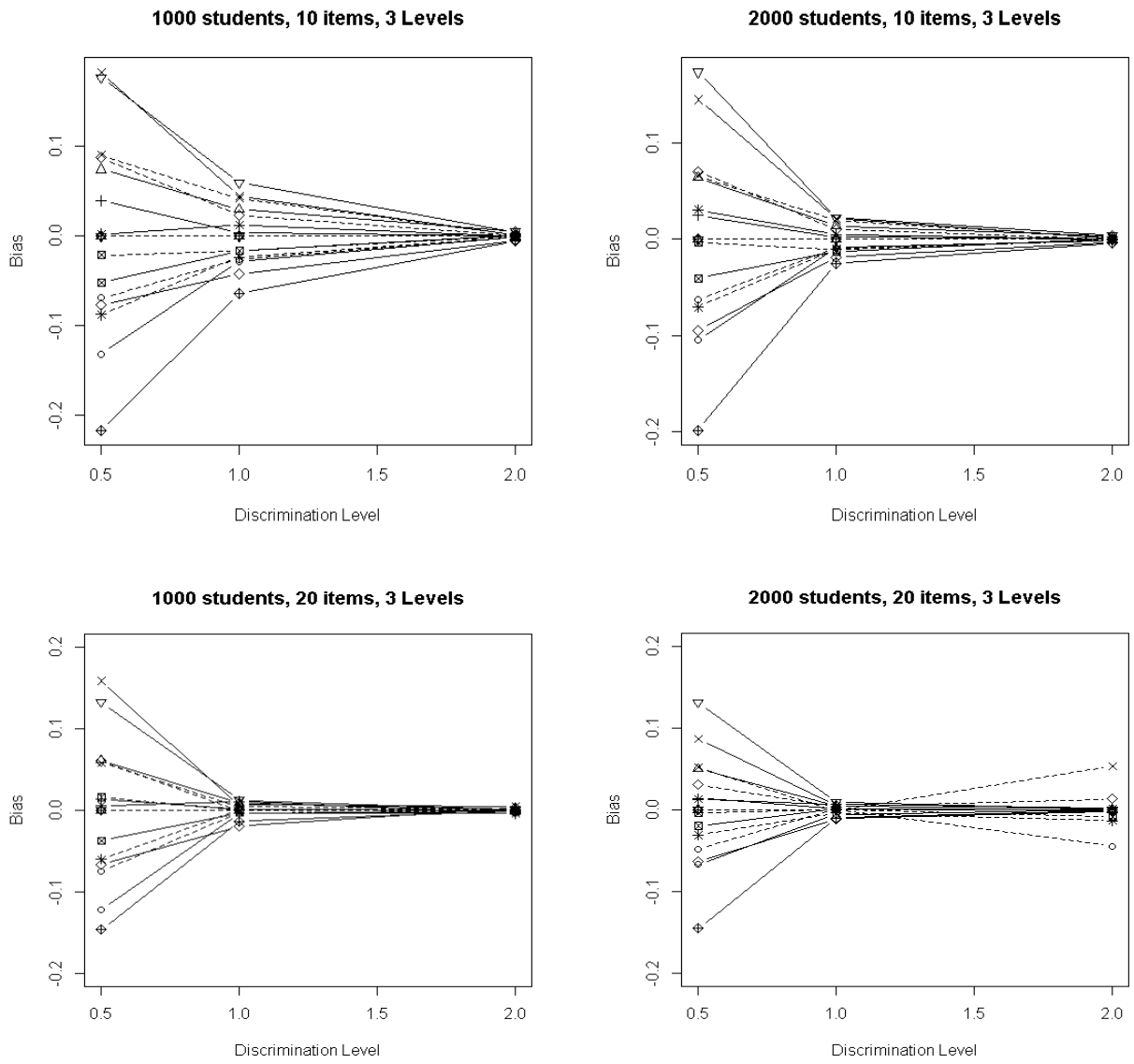


Figure 4.1. Biases for Transition probabilities when the true model is Growth. The dash lines "-------" represent Growth model estimated as Growth, and the solid lines " ___" represent Growth model estimated as Saturated. Each line corresponds to a different transition in the transition matrix.

Figure 4.1 summarizes the (bias) results of the simulation with two sample sizes of 1000, and 2000, two item sizes of 10 and 20 items, with 3 learning levels. If the parameter recovery is good, we expect the average parameter estimate over replications to be equal to the (true) parameter values and the bias to be significantly small. Overall, the mean parameter estimates

are generally quite close to the true parameter values. The biases and the mean square errors (MSE) depicted in figures 4.1, and 4.2 vary, decreasing somewhat in the larger sample size condition.

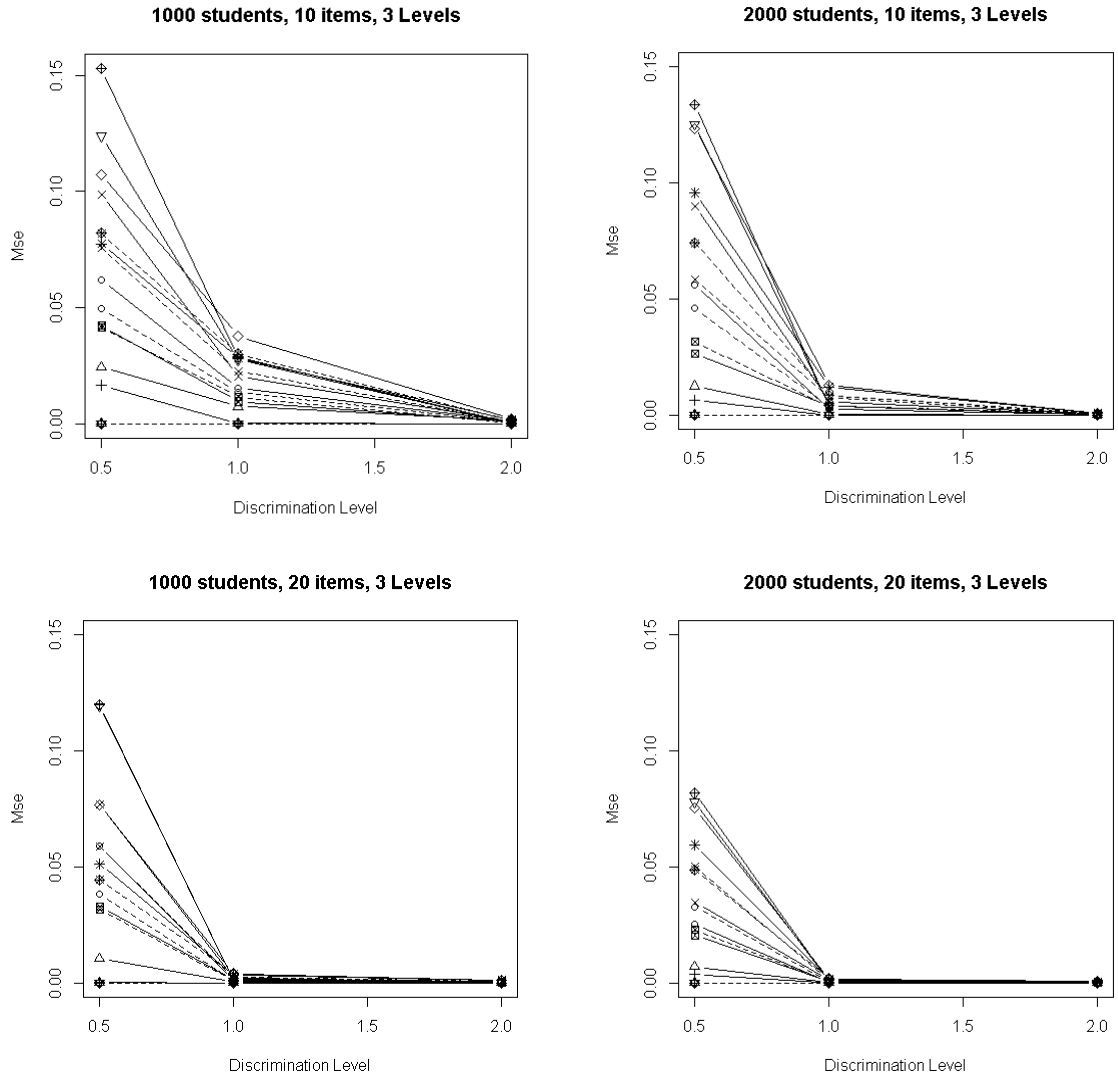


Figure 4.2. MSE for Transition probabilities when the true model is Growth. The dash lines "-------" represent Growth model estimated as Growth, and the solid lines " ___" represent Growth model estimated as Saturated. Each line corresponds to a different transition in the transition matrix.

With respect to the four different scenarios considered for the simulation study, the above plots represent the average biases and MSE's for estimation of transition probabilities when the true parameter is the Growth model. The bias and the MSE seem to depend only on the sample size, item size, and the discrimination index. The type of model estimation seems to be a non-factor in this case. Specifically, the bias and the MSE decreased for items with at least a medium discrimination level for both model specifications: correctly estimating the model as Growth, or "incorrectly" estimating the Growth model as saturated.

As already stated, increasing sample size had a profound effect on the model estimation, as clearly shown in figures 4.1 and 4.2, the bias and the MSE further decreased for items with at least a medium discrimination for larger sample size (i.e., for N = 2000). Also, addition of ten more items dramatically improved parameter recovery. The biases and MSE reduced drastically, and in all cases the mean parameter estimates are substantially closer to the true parameter values. An interesting outcome from this estimation scenario is that even with the true model being a Growth model, correctly estimating the model as growth is as good as, "incorrectly" estimating the model as Saturated especially with items with at least a medium discrimination.

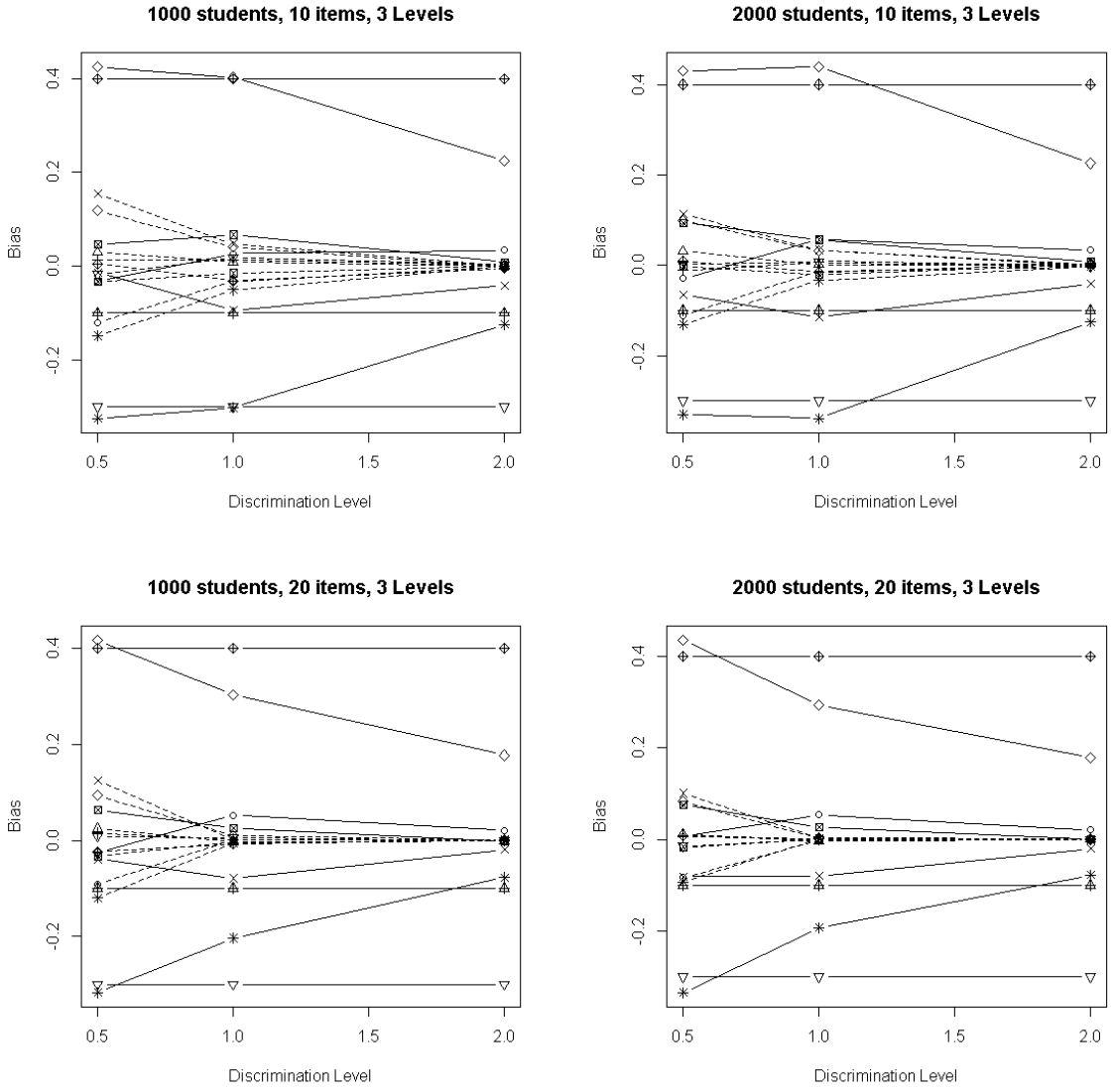


Figure 4.3. Biases for Transition probabilities when the true model is Saturated. The dash lines "-------" represent Saturated model estimated as Saturated, and the solid lines " ___ " represent Saturated model estimated as Growth. Each line corresponds to a different transition in the transition matrix.

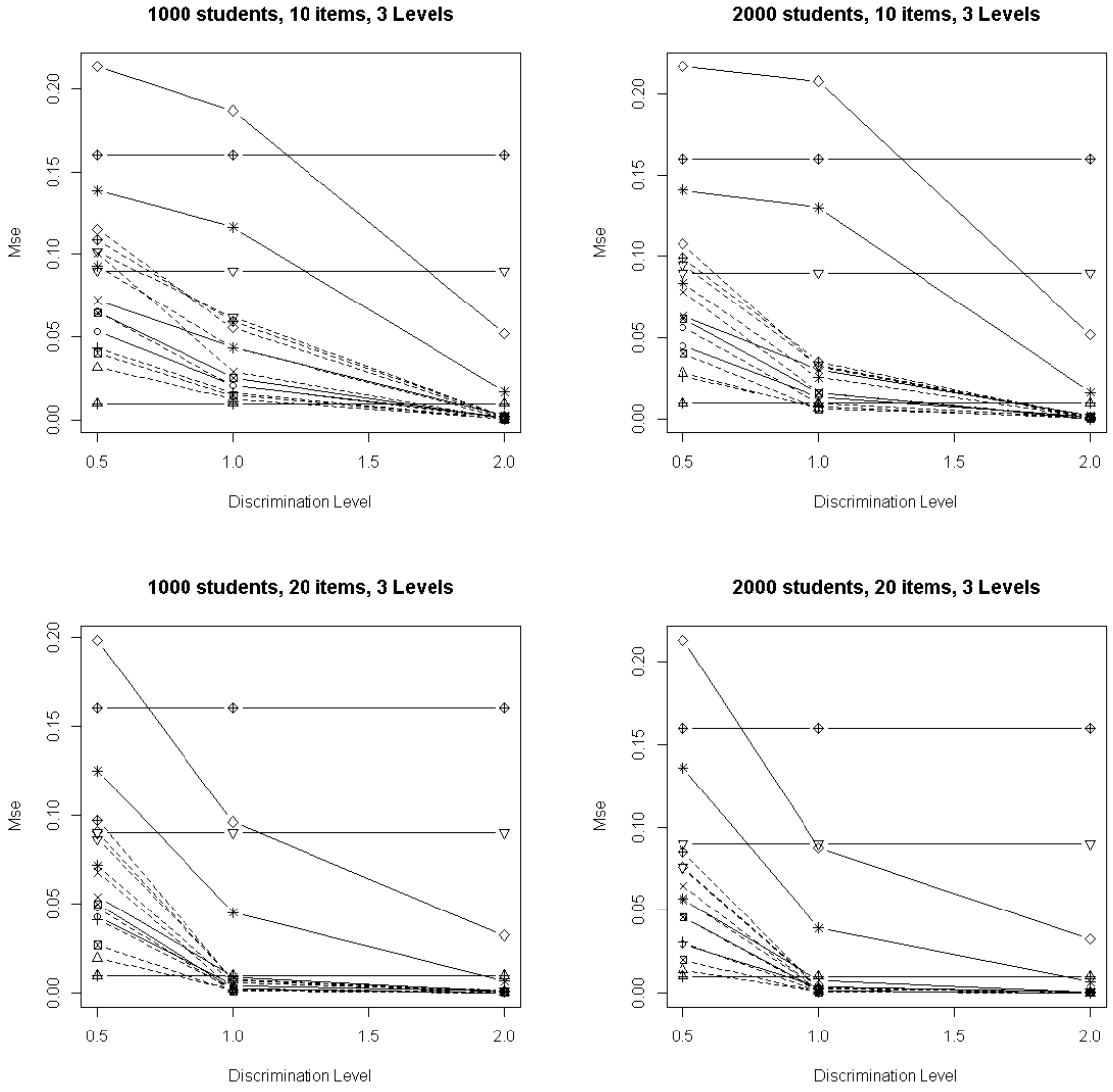
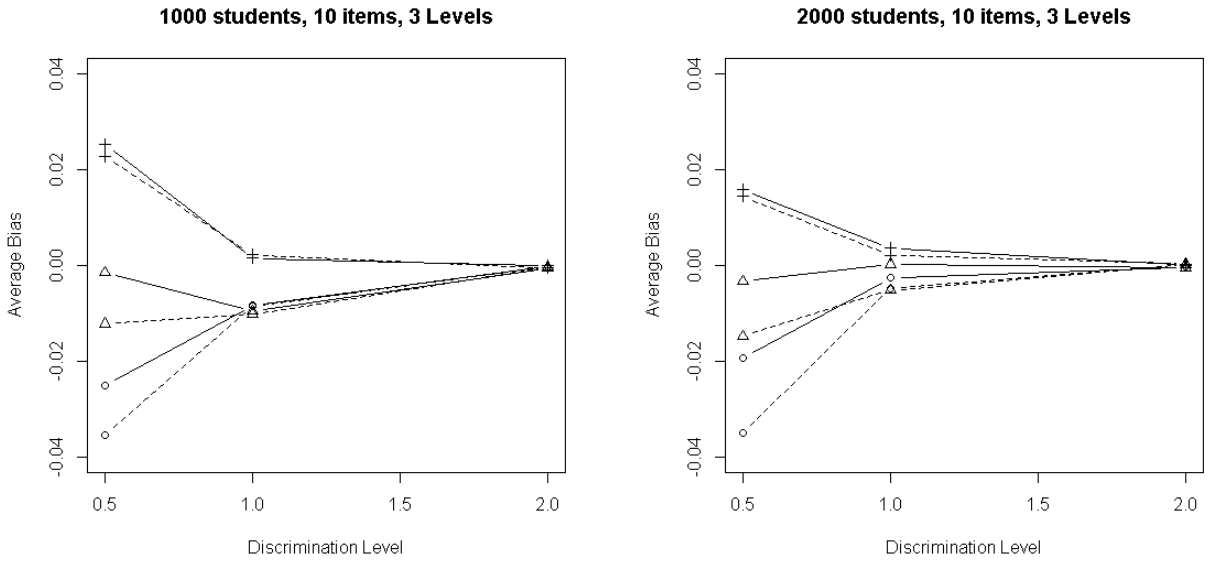


Figure 4.4. MSE for Transition probabilities when the true model is Saturated. The dash lines "-------" represent Saturated model estimated as Saturated, and the solid lines " ▬ " represent Saturated model estimated as Growth. Each line corresponds to a different transition in the transition matrix.

Figures 4.1 and 4.2 show that having items with at least a medium discrimination, the model may be estimated as saturated even if the true model is a Growth model. Unlike figures 4.1 and 4.2, model specification is critical for figures 4.3, and 4.4. Estimating a true saturated model as growth model raises deeper concerns with the parameter recovery. Specifically for this simulation experiment, we notice the mean parameter estimates deviating farther from the true

parameter values resulting in larger biases and MSE's in figures 4.3, and 4.4 respectively. The results clearly show that increasing the sample size and /or the item size only contribute to the sparseness of the data, but do not in any way improve the parameter recovery when a true saturated model is estimated as growth.

Figures 4.3, and 4.4 have also depicted patterns of very good parameter recovery when the saturated model is correctly estimated as saturated. Especially increasing the sample size and/or item size make the mean parameter estimates as closer to the true parameter values as possible, thereby causing a dramatic reductions in the biases and mean square errors.



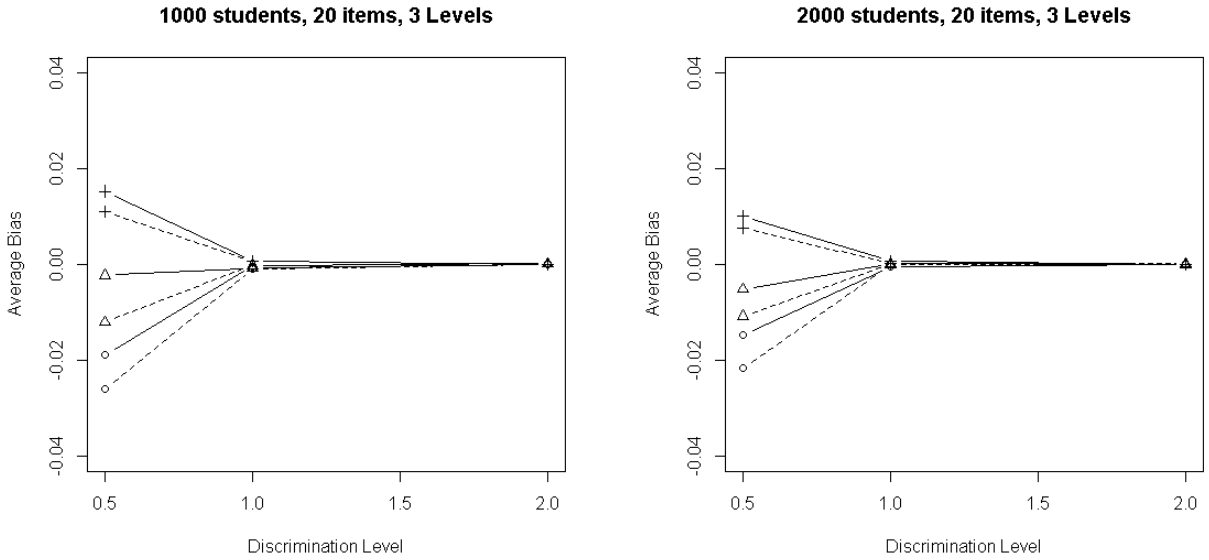


Figure 4.5. Biases for Item Response probabilities when the true model is Growth. The dash lines "-------" represent Growth model estimated as Growth, and the solid lines "___" represent Growth model estimated as Saturated. The lines are averaged over items, and each line corresponds to a different transition in the transition matrix.

Apart from the Transition probabilities which are considered the bed rock of the OLTA model, the item response probabilities are also useful. The OLTA model imposes parameter restrictions on the item response probabilities in order to achieve ordering of the learning levels. Though students sometimes slip and respond negatively to items with low difficulty, the ordering of the learning levels generally ensure that students associated with higher learning levels have higher probability of positively responding to items than those in lower leaning levels. So in the practical sense, the case for having good items cannot be ignored, and Figures 4.5, and 4.6 clearly highlight the importance of having items with at least a medium discrimination level.

Figures 4.5, and 4.6 show a good parameter recovery for good items. The plots also show that increasing the item size, and /or the sample size contributes to a huge reduction of the biases and the MSE's. As a consequence of having items with at least a medium discrimination, the

parameter recovery for the item response probabilities are good whether a true Growth model is being estimated as Growth, or a true Growth model is being estimated as saturated. This same recovery pattern is observed in figures 4.9, and 4.10 for estimation of initial learning proportions.

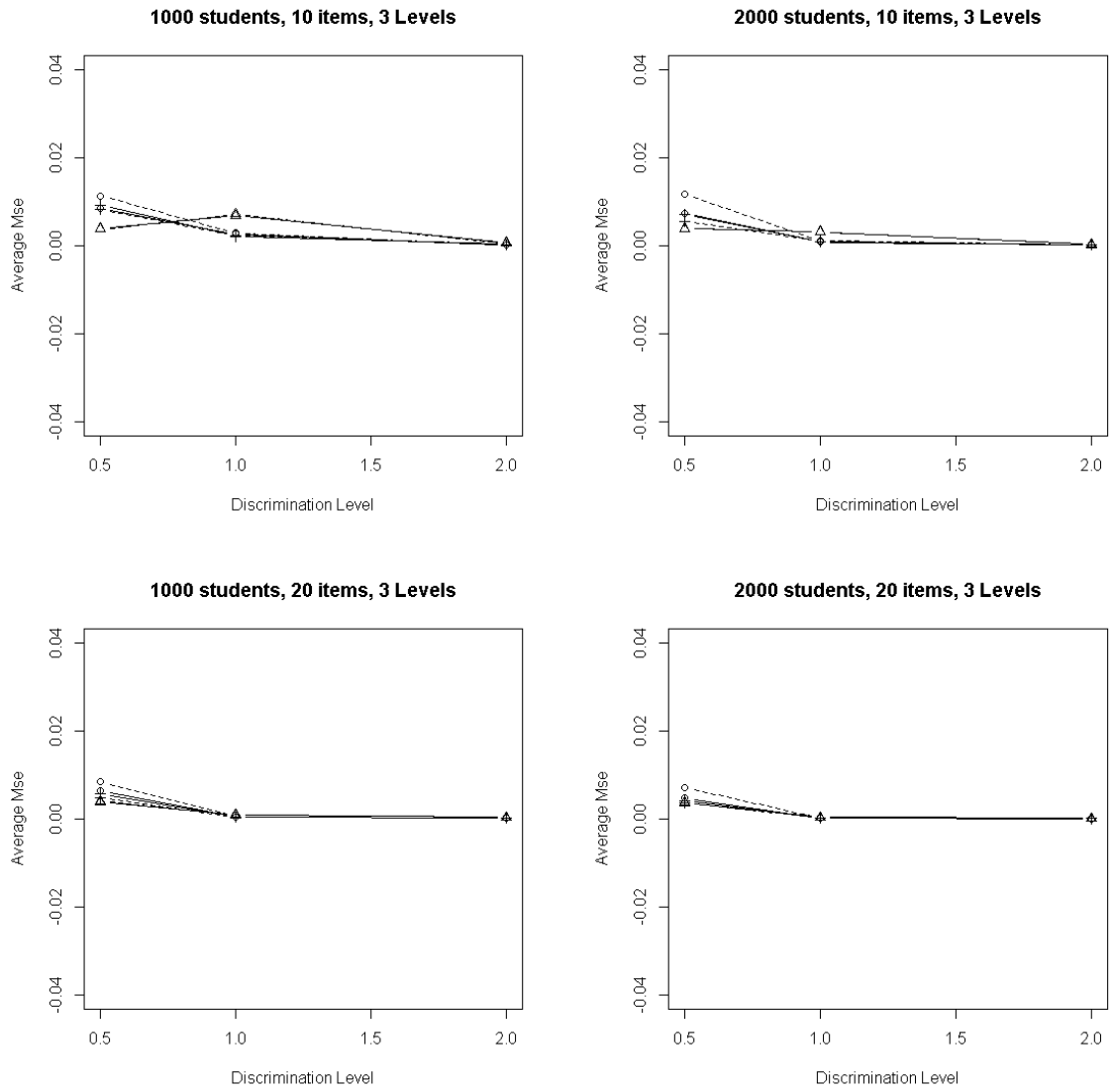


Figure 4.6. MSE for Item Response probabilities when the true model is Growth. The dash lines "-------" represent Growth model estimated as Growth, and the solid lines "___" represent Growth model estimated as Saturated. The lines are averaged over items, and each line corresponds to a different transition in the transition matrix.

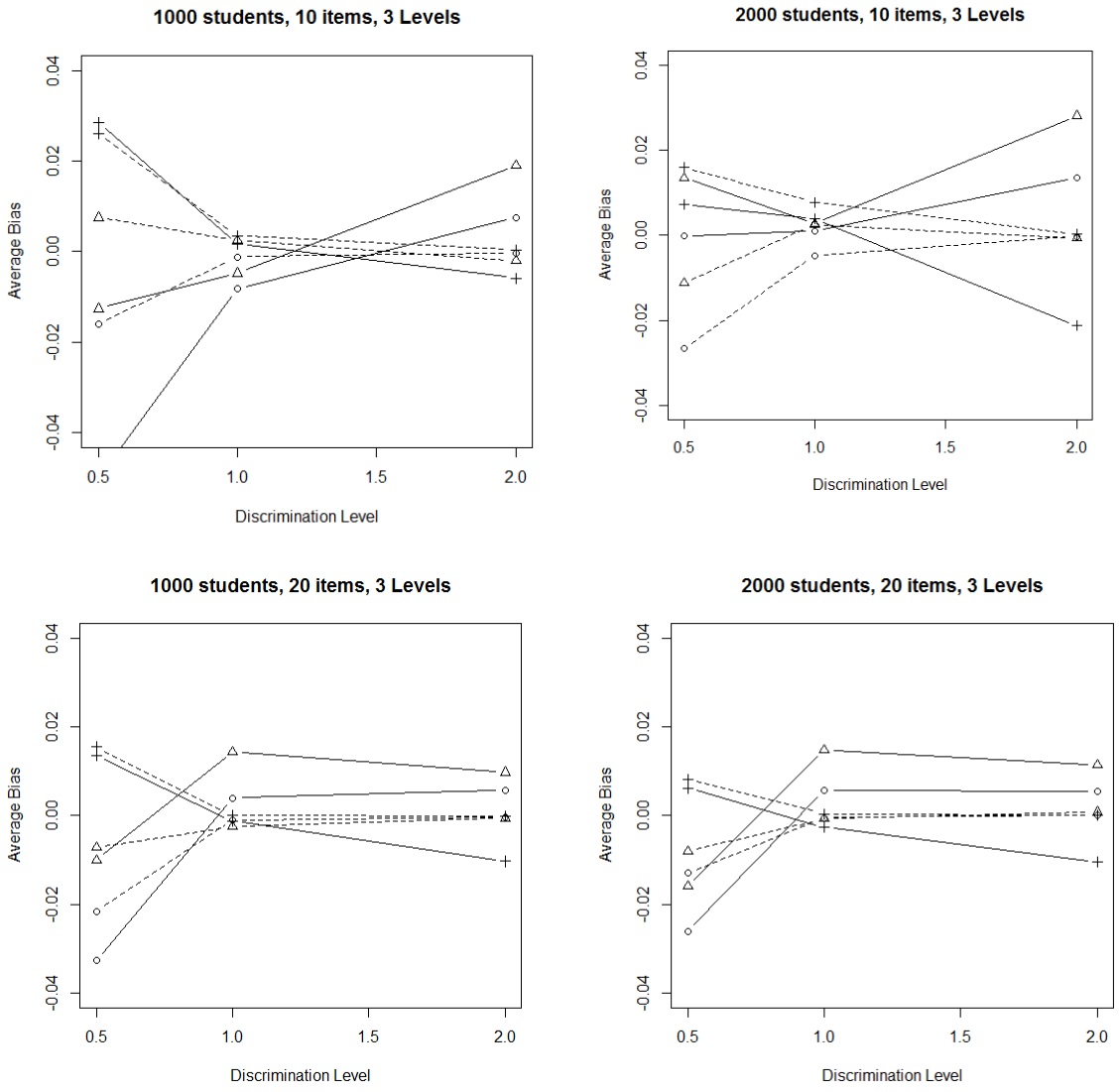


Figure 4.7. Biases for Item Response probabilities when the true model is Saturated. The dash lines "-------" represent Saturated model estimated as Saturated, and the solid lines " ___" represent Saturated model estimated as Growth. The lines are averaged over items, and each line corresponds to a different transition in the transition matrix.

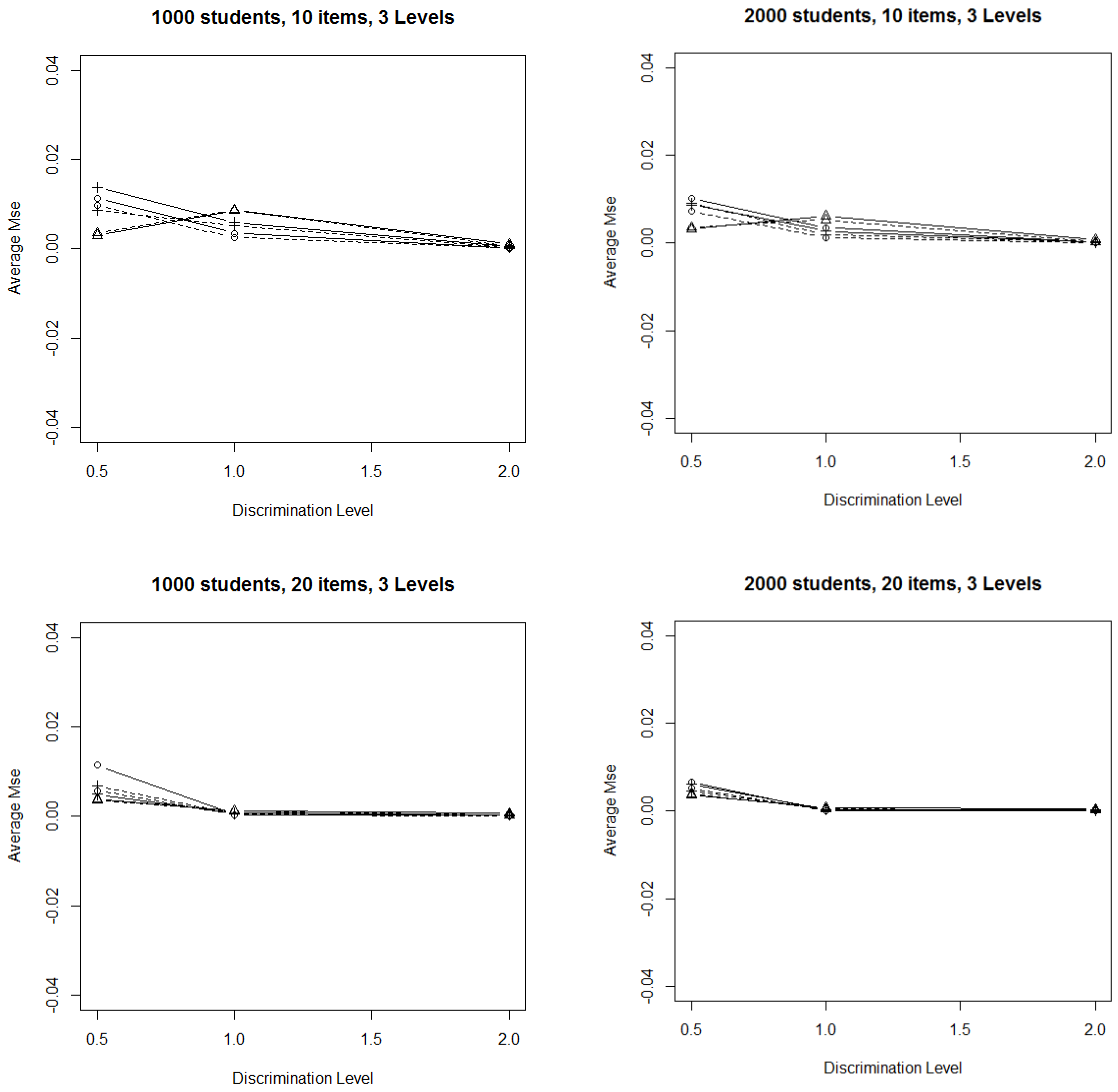


Figure 4.8. MSE for Item Response probabilities when the true model is Saturated. The dash lines "-------" represent Saturated model estimated as Saturated, and the solid lines "——" represent Saturated model estimated as Growth. The lines are averaged over items, and each line corresponds to a different transition in the transition matrix.

So far, we have seen a pattern of parameter recovery when a true saturated model is estimated as Growth. Specifically, as in the case of the transition probabilities, the bias and the means squared errors increase, and the mean of the parameter estimate farther deviates from the true parameter values. In the case of item response probabilities and estimation of initial learning proportions, estimating a true saturated model as Growth also increase the biases as seen in

figures 4.7, and 4.11 respectively. However, the MSE's tell a different story, for items with at least a medium discrimination level, the MSE is almost zero (figure 4.8), and this seems to be the case whether a true saturated model is being estimated as saturated, or a true saturated model is being "incorrectly" estimated as Growth. A similar pattern is also observed in figure 4.12.

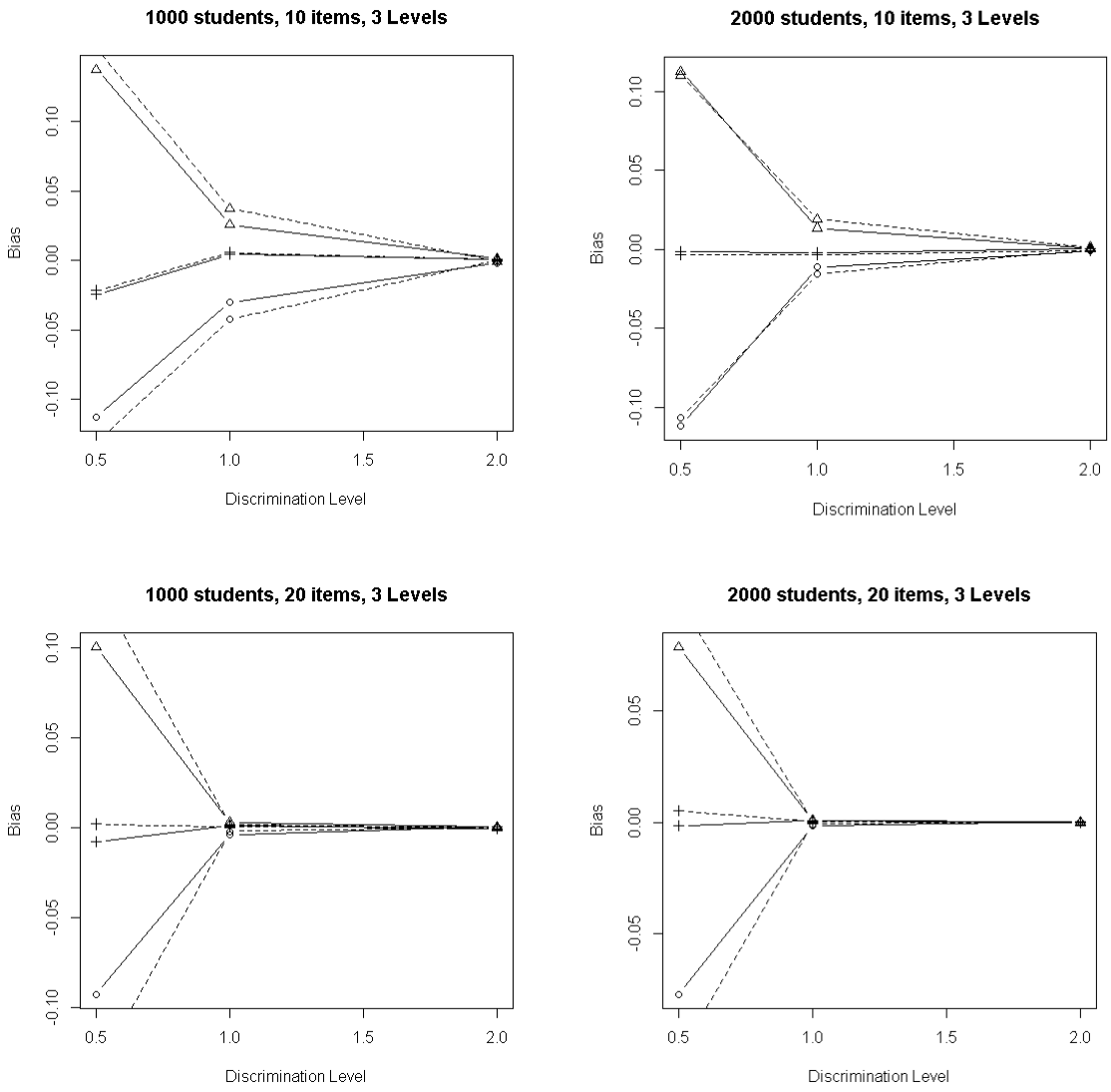


Figure 4.9. Biases for Initial learning levels when the true model is Growth. The dash lines "-------" represent Growth model estimated as Growth, and the solid lines " ___" represent Growth model estimated as Saturated. Each line corresponds to a different transition in the transition matrix.

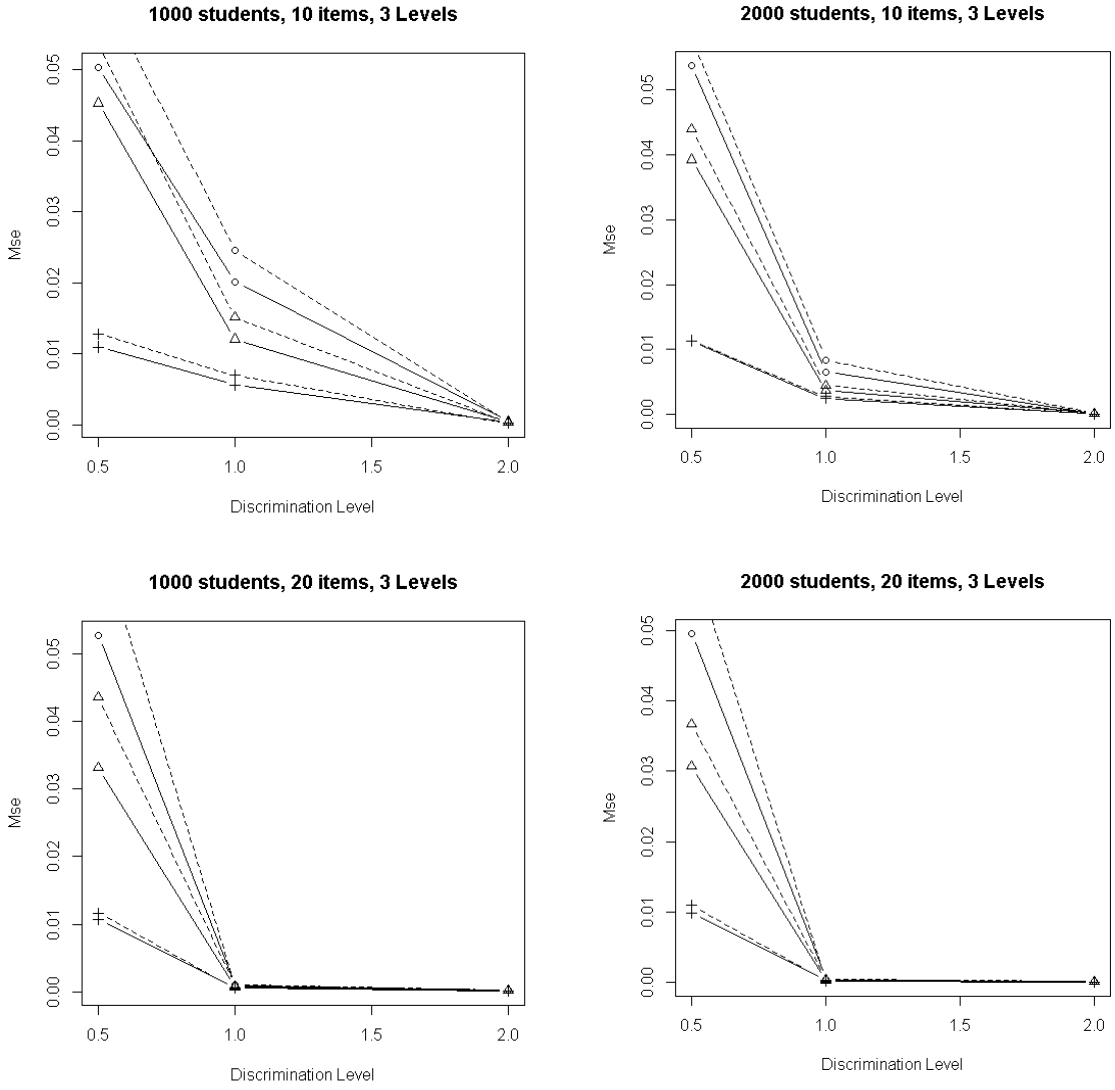


Figure 4.10. MSE for Initial learning levels when the true model is Growth. The dash lines "-------" represent Growth model estimated as Growth, and the solid lines " ___" represent Growth model estimated as Saturated. Each line corresponds to a different transition in the transition matrix.

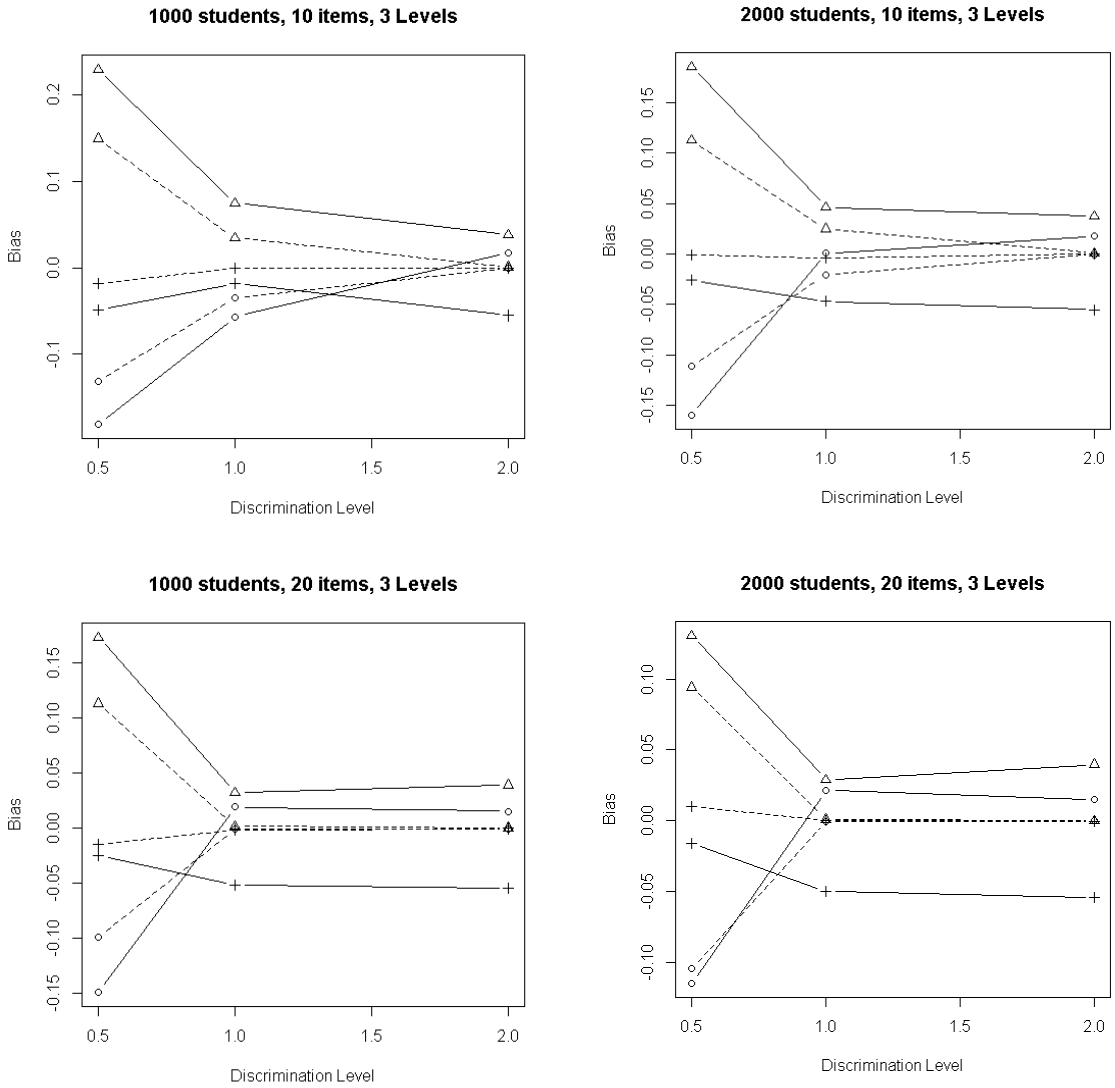


Figure 4.11. Biases for Initial learning levels when the true model is Saturated. The dash lines "-------" represent Saturated model estimated as Saturated, and the solid lines " ___ " represent Saturated model estimated as Growth. Each line corresponds to a different transition in the transition matrix.

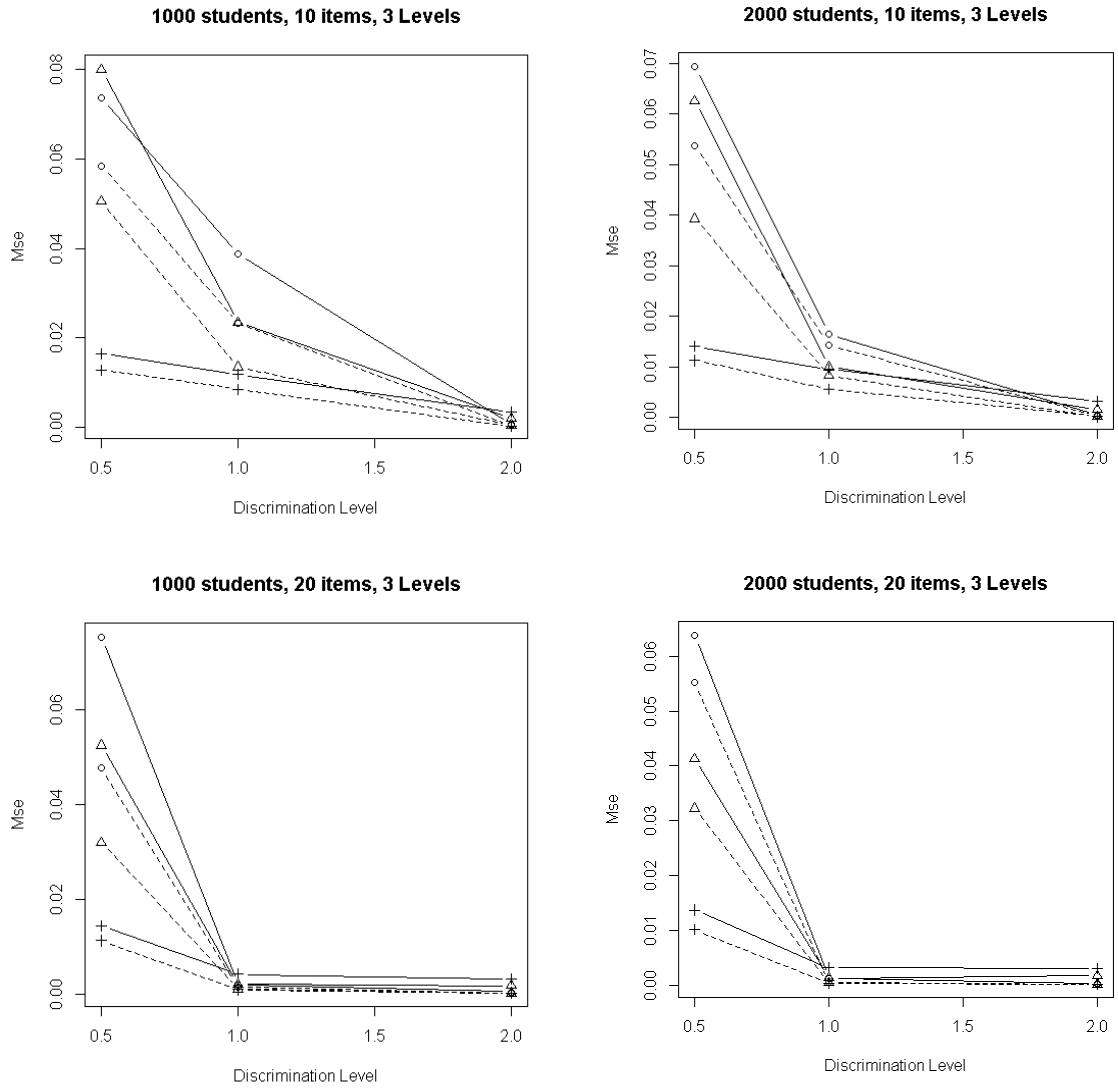


Figure 4.12. MSE for Initial learning levels when the true model is Saturated. The dash lines "-------" represent Saturated model estimated as Saturated, and the solid lines " ___ " represent Saturated model estimated as Growth. Each line corresponds to a different transition in the transition matrix.

4.3 Study 2

This study is aimed at establishing how well the AIC and BIC select the correct transition model under varying conditions. We achieved this goal by determining the proportion of time the fit indices selected the correct transition model; for a true Growth model, and for a true Saturated model. The simulation was conducted for N = 1000 and 2000, item sizes of 10 and 20, and two discrimination indices ($\alpha = 1$, and $= 2$). A discrimination index of 0.5 was later removed from this particular experiment because the low discrimination items always resulted in local maxima, and we could not rely on the results. Also 3 and 5 levels were used as seen in tables 4.1 and 4.2 respectively.

4.3.1 Results of Simulation Study 2

			A I C		B I C	
N	J	α	Growth	Saturated	Growth	Saturated
1000	10	1	0.91	1.00	0.98	1.00
1000	10	2	0.99	1.00	1.00	1.00
1000	20	1	1.00	1.00	1.00	1.00
1000	20	2	0.98	1.00	1.00	1.00
2000	10	1	0.98	1.00	1.00	1.00
2000	10	2	0.97	1.00	1.00	1.00
2000	20	1	1.00	1.00	1.00	1.00
2000	20	2	0.96	1.00	1.00	1.00

Table 4.1 Proportions of 3 level model correctly identified by AIC and BIC

We from table 4.1 that when the true model is saturated, both the AIC and BIC correctly identified the saturated model 100% of the time for all the conditions. This same scenario is observed in table 4.2. In Table 4.2, the AIC and BIC again identified the true saturated model 100% of the time; for the case when the true model is saturated with 5 levels.

			AIC		BIC	
N	J	α	Growth	Saturated	Growth	Saturated
1000	10	1	0.97	1.00	1.00	1.00
1000	10	2	0.94	1.00	0.98	1.00
1000	20	1	0.98	1.00	1.00	1.00
1000	20	2	0.98	1.00	0.98	1.00
2000	10	1	0.96	1.00	1.00	1.00
2000	10	2	0.98	1.00	0.98	1.00
2000	20	1	1.00	1.00	1.00	1.00
2000	20	2	1.00	1.00	1.00	1.00

Table 4.2 Proportions of 5 level model correctly identified by AIC and BIC

For a true Growth model with 3 levels, table 4.1 shows that the BIC correctly identified the Growth model in at least 98% of the time for each condition. Again, for the true Growth model with 5 levels, Table 4.2 shows that the BIC correctly identified the correct model in at least 98% of the time for each of the conditions studied. For the same Growth model with 5 levels, Table 4.2 indicates that the AIC selected the correct model 94% of the time, for 1000 subjects with 10 items, and a discrimination index of 2. The 94% endorsement may not be considered great; compared to the overall performance of the AIC, in several other conditions under the same model.

Furthermore, Table 4.1 shows that except for one condition, the AIC selected the true 3 level Growth model at least 96% of the time in each of the conditions studied. For a sample size of 1000, item size of 10, and a discrimination index of 1, the AIC selected the true 3 level Growth model 91% of the time. The 91% correct model identification seems to be on the low side considering the overall performance of the AIC in correctly identifying the true 3 level Growth model. In the nutshell, the AIC and the BIC did exceptionally well in identifying the true Saturated model for the 3 and 5 levels. Also the performance of the fit indices in selecting the true Growth model for the 3, and 5 levels are generally comparable.

4.4 Study 3

In an attempt to ascertain how well the AIC and BIC identify the correct/optimal number of levels, we conducted a simulation experiment under varying conditions for 3 and 5 levels. We considered the following model specifications: 3 level Growth and saturated models, and 5 level Growth and saturated models. We also conducted additional experiment for opposing models with adjacent levels. Specifically, in the case of a true 3 level growth model, we also included a saturated model for the purpose of comparison, and in each case simulated for 1, 2, 3, 4, and 5 levels. We repeated same concept for a true 3 level saturated model. Furthermore, in the case of a 5 level growth model, we added a simulation for a saturated model, and in each case simulated for 2, 3, 4, 5, 6, and 7 levels. We employed this same idea for the true 5 level saturated model.

The performance of the AIC and the BIC was determined by how often the criterion index opted for the true level model. We considered a satisfactory performance of the fit indices to be an endorsement of at least 90% for the true level model for each of the conditions studied. Eight different conditions were considered for the simulation study. For the subjects, we used N

= 1000 and 2000, item sizes of 10 and 20, and two discrimination indices ($\alpha = 1$, and $\alpha = 2$). A discrimination index of 0.5 was later removed from this particular experiment because low discrimination items always resulted in local maxima; and we could not rely on the results. As already noted, the simulation was conducted for 3 and 5 learning levels as seen in section 4.4.1.

4.4.1 Proportion of True 3 Level model selection by AIC and BIC

			AIC					AIC				
			Saturated					Growth				
			Level					Level				
N	J	α	1	2	3	4	5	1	2	3	4	5
1000	10	1	0	1	3	0	0	0	0	**92**	4	0
1000	10	2	0	0	2	0	0	0	0	**97**	1	0
1000	20	1	0	0	2	0	0	0	0	**96**	2	0
1000	20	2	0	0	1	0	0	0	0	**97**	2	0
2000	10	1	0	0	2	0	0	0	0	**93**	5	0
2000	10	2	0	0	0	0	0	0	0	**99**	1	0
2000	20	1	0	0	4	0	0	0	0	**96**	0	0
2000	20	2	0	0	2	0	0	0	0	**98**	0	0

Table 4.3 Proportion of times AIC selected the true level for 3 level Growth model Correct level specification in bold to enable interpretation.

From Table 4.3, the true model under consideration is a Growth model with 3 levels. The effectiveness of the fit indices clearly depends on the number of times the correct models with the correct levels are chosen out of the hundred simulations. We from Table 4.3 that for the first experimental condition (N = 1000, J = 10, and α =1), the AIC correctly chose the 3 level Growth model 92% of the time, preferred a 4 level Growth model 4% of the time, incorrectly endorsed the 3 level saturated model 3% of the time, and also opted for the 2 level saturated model 1% of

the time. When the sample size was increased to 2000, and the discrimination index was increased to 2, the AIC had its highest endorsement of 99% for choosing the correct 3 level Growth model. In that same condition, the AIC selected the 4 level growth model 1% of the time. For a sample size of 2000, an item size of 20, and a discrimination index of 2, the AIC correctly selected the 3 level Growth model 98% of the time, but incorrectly chose the 3level saturated model 2% of the time. Even when the sample size was decreased to 1000 with 20 items, and a discrimination index of 2, the AIC opted for the correct 3 level model 97% of the time. In general, the performance of the AIC in selecting the correct 3 level growth model for each of the conditions studied was satisfactory.

			BIC					BIC				
			Saturated					Growth				
			Level					Level				
N	J	α	1	2	3	4	5	1	2	3	4	5
1000	10	1	0	0	0	0	0	0	93	**7**	0	0
1000	10	2	0	0	0	0	0	0	0	**100**	0	0
1000	20	1	0	0	0	0	0	0	0	**100**	0	0
1000	20	2	0	0	0	0	0	0	0	**100**	0	0
2000	10	1	0	0	0	0	0	0	47	**53**	0	0
2000	10	2	0	0	0	0	0	0	0	**100**	0	0
2000	20	1	0	0	0	0	0	0	0	**100**	0	0
2000	20	2	0	0	0	0	0	0	0	**100**	0	0

Table 4.4 Proportion of times BIC selected the true level for 3 level Growth model
Correct level specification in bold to enable interpretation.

We a contrast in model identification when the AIC is compared to BIC for same conditions as displayed in Tables 4.3 and 4.4 respectively. It is clear from table 4.4 that for a sample size of

1000, with an item size of 10, and a discrimination index of 1, the BIC performed poorly by incorrectly endorsing the 2 level Growth model 93% of the time , whilst choosing the correct 3 level model only 7% of the time. When the sample size was increased to 2000 for the same condition, the BIC again performed poorly by wrongly endorsing the 2 level Growth model 47% of the time, whilst weakly choosing the correct 3 level model 53% of the time. Interestingly, the poor performance of the BIC is associated with those conditions with item size of 10, and a discrimination index of 1. Apart from the two poor performances by the BIC, the fit index was perfect in endorsing the true 3 level growth model, in the rest of the conditions.

Though the BIC never selected any level under the saturated model in all the conditions studied, its tendency for selecting the simple model, in this case the 2 level growth model, makes it a bit unreliable; especially for items with lower discrimination index, and small item sizes. In the nutshell, the AIC outperformed the BIC in selecting the true 3 level growth model for the conditions studied.

			A I C					A I C				
			Saturated					Growth				
			Level					Level				
N	J	α	1	2	3	4	5	1	2	3	4	5
1000	10	1	0	15	**78**	7	0	0	0	0	0	0
1000	10	2	0	0	**98**	2	0	0	0	0	0	0
1000	20	1	0	0	**95**	5	0	0	0	0	0	0
1000	20	2	0	0	**98**	2	0	0	0	0	0	0
2000	10	1	0	0	**85**	15	0	0	0	0	0	0
2000	10	2	0	0	**98**	2	0	0	0	0	0	0
2000	20	1	0	0	**99**	1	0	0	0	0	0	0
2000	20	2	0	0	**98**	2	0	0	0	0	0	0

Table 4.5 Proportion of times AIC selected the true level for 3 level Saturated model. Correct level specification in bold to enable interpretation.

			BIC					BIC				
			Saturated					Growth				
			Level					Level				
N	J	α	1	2	3	4	5	1	2	3	4	5
1000	10	1	0	100	**0**	0	0	0	0	0	0	0
1000	10	2	0	0	**100**	0	0	0	0	0	0	0
1000	20	1	0	29	**71**	0	0	0	0	0	0	0
1000	20	2	0	0	**100**	0	0	0	0	0	0	0
2000	10	1	0	100	**0**	0	0	0	0	0	0	0
2000	10	2	0	0	**100**	0	0	0	0	0	0	0
2000	20	1	0	0	**100**	0	0	0	0	0	0	0
2000	20	2	0	0	**100**	0	0	0	0	0	0	0

Table 4.6 Proportion of times BIC selected the true level for 3 level Saturated model. Correct level specification in bold to enable interpretation.

Interestingly, the AIC and BIC never preferred the Growth model in the case of the true 3 level saturated model; as displayed in Tables 4.5, and 4.6 respectively. However, the fit indices occasionally preferred different number of levels. Specifically, for N = 1000, J = 10, and $\alpha = 1$, the AIC correctly chose the 3 level saturated model 78% of the time, incorrectly chose the 4 level saturated model 7% of the time, and also chose the 2 level saturated model 15% of the time. In contrast, the BIC performed worse by wrongly endorsing the 2 level saturated model 100% of the time, for those same conditions. Again, for 1000 subjects, 20 items, and a discrimination index of 1, the BIC weakly endorsed the correct 3 level saturated model 71% of the time, whilst opting for the 2 level saturated model 29% of the time. For the same condition, the AIC correctly identified the 3 level saturated model 95% of the time.

For a sample size of 2000, item size of 10, and a discrimination index of 1, the AIC preferred the 4 level saturated model 15% of the time, but it correctly endorsed the 3 level saturated model 85% of the time. For the same condition, the BIC wholly endorsed the simpler model; as the fit index incorrectly chose the 2 level saturated model 100% of the time. The AIC clearly outperformed the BIC for selecting the true 3 level saturated model.

4.4.2 Proportion of True 5 Level model selection by AIC and BIC

			AIC						AIC					
			Saturated						Growth					
			Level						Level					
N	J	α	2	3	4	5	6	7	2	3	4	5	6	7
1000	10	1	0	0	0	0	0	0	0	2	88	**9**	1	0
1000	10	2	0	0	0	1	0	0	0	0	1	**87**	10	1
1000	20	1	0	0	0	0	0	0	0	0	5	**92**	3	0
1000	20	2	0	0	0	2	0	0	0	0	0	**97**	1	0
2000	10	1	0	0	0	0	0	0	0	0	70	**28**	2	0
2000	10	2	0	0	0	0	0	0	0	0	1	**93**	5	1
2000	20	1	0	0	0	3	0	0	0	0	0	**97**	0	0
2000	20	2	0	0	0	1	0	0	0	0	0	**97**	2	0

Table 4.7 Proportion of times AIC selected the true level for 5 level Growth model
Correct level specification in bold to enable interpretation.

We considered 5 level models to investigate the proportion of times the fit indices selected the true 5 level model. Tables 4.7 and 4.8 represent the proportion of times the 5 level growth model was selected by AIC and BIC respectively. Table 4.8 shows that the AIC preferred the 4 level growth model to the 5 level model; in two different conditions. Specifically, with for a 1000 sample size with 10 items; and a discrimination level of 1, the AIC preferred the 4 level

model 88% of the time. Also, Increasing the sample size by a 1000 for the same condition did not help improve the chances of selecting the true level since the AIC still preferred the 4 level model 70% of the time. For a sample size of 1000 with 10 items, and a discrimination level of 2, the AIC correctly endorsed the 5 level model 87% of the time. This endorsement may not necessarily be considered a strong one; due to the fact that the fit index has selected the right model in at least 93% of the time in other conditions .

The BIC is well known for its preference for simpler models, so it is not entirely surprising that the fit index preferred models with lower levels in some of the conditions considered in Table 4.8. To be specific, the BIC wholly preferred a 3 level model to the 5 level model, for a sample size of 1000 with 10 items, and a discrimination level of 1. The BIC opted for a 3 level model 72% of the time, for a sample size of 2000 with 10 items, and a discrimination level of 1. Also there are other conditions that the BIC preferred the 4levelmodel to the 5level model: The BIC selected the 4 level model 58% of the time, for a 1000 sample size with 10 items, and a discrimination level of 2. Also, the fit index selected the 4 level model 98% of the time, for a condition with 2000 sample size, 20 items, and a discrimination index of 1. For situations where the BIC selected the right (5 level) model, the endorsements are at least 97% high, except for one condition (N=2000, J= 10, and α =2), where the BIC weakly selected the correct model 81% of the time. The results in tables 4.8 and 4.9 clearly suggest that the AIC outperformed the BIC.

			BIC						BIC					
			Saturated						Growth					
			Level						Level					
N	J	α	2	3	4	5	6	7	2	3	4	5	6	7
1000	10	1	0	0	0	0	0	0	0	100	0	**0**	0	0
1000	10	2	0	0	1	0	0	0	0	0	58	**41**	0	0
1000	20	1	0	0	0	0	0	0	0	0	0	**100**	0	0
1000	20	2	0	0	0	2	0	0	0	0	1	**97**	0	0
2000	10	1	0	0	0	0	0	0	0	72	28	**0**	0	0
2000	10	2	0	0	0	0	0	0	0	0	19	**81**	0	0
2000	20	1	0	0	0	0	0	0	0	0	98	**2**	0	0
2000	20	2	0	0	0	1	0	0	0	0	0	**99**	0	0

Table 4.8 Proportion of times BIC selected the true level for 5 level Growth model
Correct level specification in bold to enable interpretation.

			AIC						AIC					
			Saturated						Growth					
			Level						Level					
N	J	α	2	3	4	5	6	7	2	3	4	5	6	7
1000	10	1	0	35	57	**8**	0	0	0	0	0	0	0	0
1000	10	2	0	0	27	**67**	6	0	0	0	0	0	0	0
1000	20	1	0	0	57	**42**	1	0	0	0	0	0	0	0
1000	20	2	0	0	0	**98**	2	0	0	0	0	0	0	0
2000	10	1	0	0	90	**9**	1	0	0	0	0	0	0	0
2000	10	2	0	0	4	**91**	5	0	0	0	0	0	0	0
2000	20	1	0	0	10	**86**	4	0	0	0	0	0	0	0
2000	20	2	0	0	0	**100**	0	0	0	0	0	0	0	0

Table 4.9 Proportion of times AIC selected the true level for 5 level Saturated model
Correct level specification in bold to enable interpretation.

				BIC						BIC				
				Saturated						Growth				
				Level						Level				
N	J	α	2	3	4	5	6	7	2	3	4	5	6	7
1000	10	1	0	100	0	**0**	0	0	0	0	0	0	0	0
1000	10	2	0	25	75	**0**	0	0	0	0	0	0	0	0
1000	20	1	0	82	18	**0**	0	0	0	0	0	0	0	0
1000	20	2	0	0	3	**97**	0	0	0	0	0	0	0	0
2000	10	1	0	100	0	**0**	0	0	0	0	0	0	0	0
2000	10	2	0	0	96	**4**	0	0	0	0	0	0	0	0
2000	20	1	0	0	100	**0**	0	0	0	0	0	0	0	0
2000	20	2	0	0	0	**100**	0	0	0	0	0	0	0	0

Table 4.10. Proportion of times BIC selected the true level for 5 level Saturated model Correct level specification in bold to enable interpretation.

Tables 4.9 and 4.10 display the proportion of true level selected by the AIC and BIC respectively for the 5 level saturated model. The performances of the fit indices are not the best, and certainly not what we expected. For instance, in Table 4.11, the BIC selected the correct 5 level model in only two of the eight conditions under consideration. The fit index selected the 5 level model 97% of the time, for a sample size of 1000 with 20 items, and a discrimination index of 2. Also, the BIC opted for the 5 level model 100% of the time, for a sample size of 2000 with 20 items, and a discrimination level of 2. In each of the two conditions that the BIC correctly selected the 5 level model, the item size was 20, and the discrimination index was 2. The BIC opted for a lower level model when the condition under consideration had an item size of 10, and/or a discrimination index of 1.

For a sample size of 10 and a discrimination level of 1, the BIC wholly endorsed the 3 level model for a1000 subjects. Again, for those same conditions, the BIC selected the 3 level model 100% of the time when the subjects were increased by a 1000. Also, maintaining a sample size of 1000 with a discrimination index of 1, the BIC opted for the 3 level model 82% of the time when the item size was 20. Furthermore, for 3 of the 8 conditions, the BIC preferred the 4level model to the 5 level model. For instance, the BIC wholly preferred the 4 level saturated model to the 5level model, for a condition that consisted of 2000 subjects with 20 items, and a discrimination index of 1. But for 2000 subjects with 10 items, and a discrimination index of 2, the fit index selected the 4 level model 96% of the time. However, for 1000 subjects with 10 items, and a discrimination index of 2, the BIC preferred the 4 level model 75% of the time.

We consider the performance of the AIC in Table 4.10 in three forms: enormous endorsement of the correct model, weak endorsement of the correct model, and preference for the wrong model. In two of the conditions the AIC weakly endorsed the correct 5 level saturated model in at most 86% of the time. For conditions that included an item size of 20, and a discrimination level of 2, the AIC enormously, and correctly selected the 5 level model in at least 98% of the time. There are other three conditions that the AIC preferred the 4 level saturated model to the 5 level model. Specifically, the AIC selected the 4 level model 57% of the time, for 1000 subjects, 10 items, and a discrimination level of 1. The AIC expressed same preference (57%) for the 4 level model for 1000 subjects with 20 items, and a discrimination index of 1. But with 2000 subjects with 10 items, and a discrimination index of 1, the AIC had a high endorsement of 90% for the 4 level model. Though the performances of both fit indices did not look good, none of the fit index opted for a growth model when the conditions under consideration were saturated in nature.

Chapter 5

Real Data Analysis

5.1 Introduction

In this chapter, the ordered latent transition analysis model is applied to real data, administered at two time points from the National Center for Early Development and Learning's (NCEDL's) pre-kindergarten study in eleven states. The goal is to investigate the number of learning levels, and also to determine the type of transition matrix that characterizes the progression of student's learning. One of the advantages of using the OLTA model to measure students' learning is that the model recognizes that students' have different ability levels, and the ordered nature of the model ensures that students' are appropriately placed on the learning continuum based on their abilities.

In this analysis, three parameters are of interest: the learning level prevalences, item-response probabilities, and the transition probabilities. The most important parameter to be investigated is the transition probabilities. Specifically, we wanted to know how well the saturated or the growth model fits the data. Model fit is determined by AIC and BIC. Optimal balance of model fit and parsimony are associated with a smaller criterion value. A model with the minimum AIC or BIC might be selected as the optimal model. However, due to the varying penalties associated with the AIC and the BIC, they often disagree on what the optimal model is.

Data Description

The data analyzed here are a combined study obtained by the National Center for Early Development and Learning (NCEDL) from two major studies aimed at understanding the

variations among the state-funded pre-kindergarten (pre-k) programs. The two major studies are the Multi-State Study of Pre-Kindergarten and the State-Wide Early Education Programs (SWEEP) study.

Data collection for the Multi-State Study of Pre-Kindergarten occurred in six states during the 2001 - 2002 school year. The participating states had contributed significant amount of resources to pre-k initiatives. The states were: California, Georgia, Illinois, Kentucky, New York, and Ohio. Forty schools were selected using stratified random sampling in each state. The teachers assisted data collectors in recruiting children for the study, and also helped them determine the eligibility of children. The participating children (a) were old enough for kindergarten in the fall of 2002, (b) did not have an Individualized Education Plan and (c) spoke English or Spanish well enough to understand simple instructions.

Data collection for the SWEEP study occurred in five states during the 2003-2004 school year. The states were Massachusetts, New Jersey, Texas, Washington, and Wisconsin. Hundred randomly selected state-funded pre-kindergarten sites were selected for participation in each of the five states. Initially, 465 sites participated in the fall. Two sites withdrew participation in the spring, resulting in 463 sites participating in the spring. The participating teachers helped the data collectors to recruit children, and also help determine the eligibility of children. The participating children (a) were old enough for kindergarten in the fall of 2002, (b) did not have an Individualized Education Plan and (c) spoke English or Spanish well enough to understand simple instructions.

The combined NCEDL data contains the item scores for 2892 pre- kindergarten children in the eleven states. The data also consist of 9 assessment items administered at two time points. The items assessed nine skills which are provided below:

1. Uses complex sentence structures.
2. Understands and interprets a story or other text read to him/her.
3. Easily and quickly names all upper-and lower-case letters of the alphabet.
4. Produces rhyming words.
5. Predicts what will happen next in stories.
6. Reads simple books independently.
7. Uses different strategies to read unfamiliar words.
8. Demonstrates early writing behaviors.
9. Demonstrates an understanding of some of the conventions of print.

The items were scored on a five-point rating scale:

1. Not yet: Child has not yet demonstrated the skill.
2. Beginning: Child is just beginning to demonstrate the skill.
3. In progress: Child demonstrates skill, knowledge, or behavior with some regularity but varies in level of competence.
4. Intermediate: Child demonstrates skill with increasing regularity and average competence
5. Proficient: Child demonstrates skill competently and consistently.

For the purpose of our study, and to easily analyze the data, the dataset was dichotomized such that rating scores ranging from 1 to 3 are recoded as 0, and scores from 4 to 5 are recoded as 1.

5.2 Item Analysis

	Saturated Model		Growth Model	
Levels	AIC	BIC	AIC	BIC
2	35250.52	35393.79	35368.87	35506.17
3	32759.73	32992.55	32903.69	33118.60
4	32204.45	32538.76	32370.45	32668.93
5	31858.23	**32305.96**	32092.36	32480.39
6	**31777.57**	32350.66	31967.62	32451.16
7	31786.16	32496.56	31920.55	32505.59
8	31790.40	32650.04	31925.34	32617.82
9	31815.96	32836.78	31951.31	32757.22

Table 5.1. Using AIC and BIC for selecting the appropriate model for the dataset. Saturated model with 6 learning levels selected as the most appropriate model for the dataset.

Although the AIC and BIC are fundamentally based on different theoretical motivations, they are generally considered to have the same aim; which is to identify good models, even if there is disagreement on what constitute a "good model" (Burnham & Anderson, 2004; Kuha, 2004). There are instances where the two criteria agree on what the "good model" is, and situations where both criteria clearly disagree on what the "good model" is.

Table 5.1 re-enforces the latter. The BIC prefers a 5 level saturated model, but the AIC prefers a 6 level saturated model. We have decided on the saturated model with 6 learning levels as the most appropriate model that best fit the data.

	Level1	Level2	Level3	Level4	Level5	Level6
Item1	0.07	0.22	**0.82**	**0.88**	**0.88**	**1.00**
Item2	0.01	0.07	**0.94**	**0.94**	**0.94**	**1.00**
Item3	0.00	0.19	0.19	**0.72**	**0.77**	**1.00**
Item4	0.00	0.07	0.07	**0.75**	**0.75**	**1.00**
Item5	0.00	0.20	**0.59**	**0.94**	**0.94**	**1.00**
Item6	0.00	0.07	0.17	0.22	**0.64**	**0.88**
Item7	0.00	0.06	0.06	0.06	**0.79**	**1.00**
Item8	0.00	0.05	0.05	0.08	**0.66**	**0.85**
Item9	0.00	0.21	0.21	0.44	**0.68**	**0.93**

Table 5.2. Item response probabilities for saturated model with 6 learning levels
Conditional probabilities > .5 in bold to facilitate interpretation

The item response probability is the probability that a subject correctly responds to an item conditional on his / her learning level. In table 5.2 we provide for each learning level the probability that respondents belonging to that learning level respond positively to each item. It is clear from the table that the six learning levels can be ordered along the learning continuum , with learning level 1 representing the most 'negative' and learning level 6 representing the most 'positive' level within the entire set of learning levels. It is also clear that differences among the nine items have emerged with respect to the manner their response probabilities vary as a function of the learning levels.

	Level Prevalences	
Level	Time 1	Time 2
1	0.514	0.231
2	0.139	0.175
3	0.216	0.191
4	0.065	0.189
5	0.042	0.117
6	0.023	0.096

Table 5.3. Learning level proportions for Time1 and Time2

Two important factors that may have contributed to this functional relationship are item difficulty and its steepness. The steepness of the functional relationship is connected to the discriminatory power of the item. The discrimination index of an item is the log odds ratio for correct responses between two adjacent levels. Items with good discrimination separate skilled examinees from the less skilled, and the non-skilled ones. Table 5.2 clearly shows that the items under consideration have good discrimination. For instance, if we compare items 5 and 4 in this respect, we clearly that the cumulative response probabilities for item 5 are not only higher than those of item 4, but that they also change more dramatically as we move along the learning level continuum. Same dynamics are observed between items 5 and 6, the cumulative response probabilities for item 5 are higher than those in item 6; and also change more drastically as one moves along the learning continuum.

In a sense item 5 may be considered a better indicator of level membership than items 4 and 6. Also comparing items 8 and 9, the cumulative response probabilities for item 9 are higher than those in item 8 along the 6 learning levels. Item 8 also has the least cumulative response probabilities in almost all the six learning levels. It may be considered the most difficult item among all the nine items, or item with the least discriminatory power. In table 5.3 we provide

estimates of the population proportions of participants belonging to each learning level at time 1, and also the level prevalences at time 2. The graphical representation is depicted in figure 5.1.

Table 5.3 shows more than half of the children (a little over 51%) belonged to level 1 at the initial time point. However, more than half of the children who belonged to level 1 at the initial time, moved to higher levels at time 2, leaving just a little over 23% of children in level 1 at time 2. Generally this kind of movement is expected, and is within reason. Also, with the exception of levels 1 and 3, the rest of the levels saw an increase in membership at time 2.

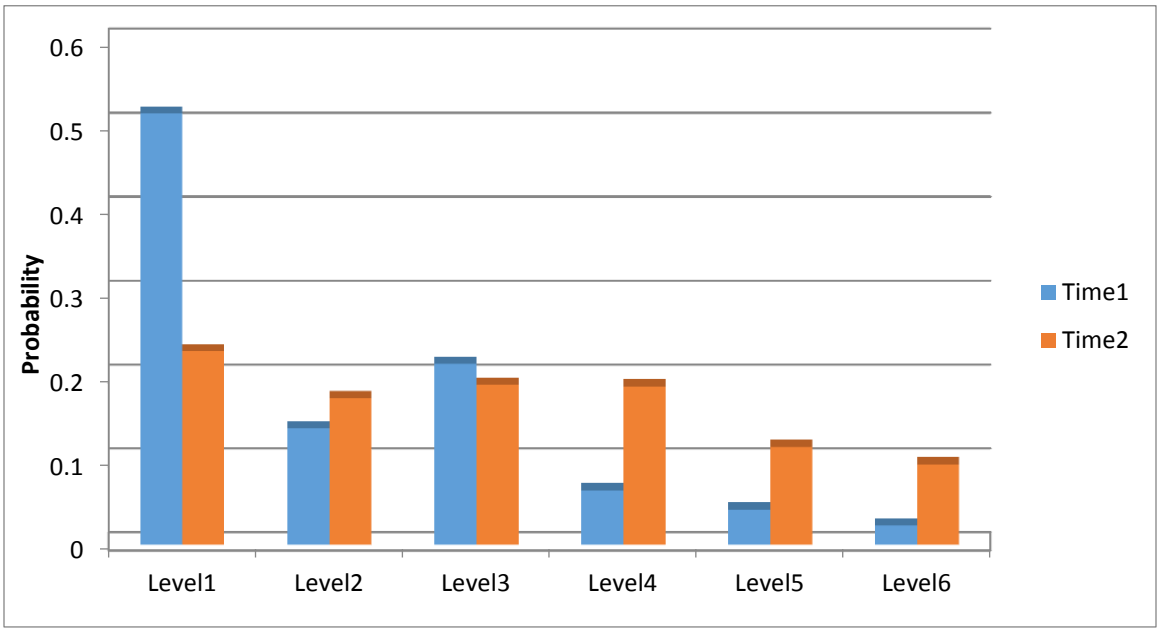


Figure 5.1 Pre-kindergarten Learning level membership probabilities for two time points (NCEDL data, N = 2892).

In table 5.2 the larger conditional probabilities are in bold font to highlight the overall response pattern. These probabilities provide the basis for labeling and interpretation of the learning levels. Learning level 5 is associated with a high probability of individuals responding correctly to all the 9 items. Individuals in this level among other things were likely to have

demonstrated proficiency in reading simple books independently including using different strategies to read unfamiliar words, and also have demonstrated proficiency in early writing behaviors. We labeled this learning level as the "skilled" level. Interestingly, level 6 is associated with children correctly endorsing all the 9 items with even higher probability than those in level 5. The level 6 children have more understanding, and are more skilled than all the other levels including level 5. Because of this, we labeled the level 6 as the "advanced" level. In contrast, individuals in learning level 1 were more likely to have consistently demonstrated to have little to no skill in all the nine items. Even though this level had the highest proportion of individuals (a little over 51%) at the initial time, they simply showed the least skill in the assessment test. More than half of those children gained some skill and consequently moved to upper levels at time 2. Nonetheless, a little over 23% remained at level 1 at time 2. We labeled learning level 1 as the "not ready" level.

There are three other learning levels that reflect different patterns of proficiency. Learning level 2, labeled "inexperienced" level had a somewhat high likelihood of using complex sentence structures(0.22) than understanding and interpreting a story or other text read to them, but the likelihood as noticed is well below 0.5. Individuals in learning level 3 displayed different proficiency pattern, they had high likelihood of using complex sentence structures, and also showed proficiency in understanding and interpreting a story or other text read to them. In addition, they had the ability to predict what will happen next in stories. We labeled this learning level a "developing" level. Apart from demonstrating same proficiency as those in level 3, learning level 4 individuals demonstrated additional ability by being able to easily and quickly naming all upper-and lower-case letters of the alphabet, and producing rhyming words. We labeled learning level 4 the "transitional" level.

			Learning	Levels		
Assigned Label	NotReady	Inexperienced	Developing	Transitional	Skilled	Advanced
Not Ready	**0.42**	0.22	0.20	0.09	0.06	0.02
Inexperienced	0.00	**0.33**	0.11	0.28	0.18	0.10
Developing	0.08	0.06	**0.33**	0.31	0.16	0.06
Transitional	0.00	0.06	0.00	**0.43**	0.17	0.33
Skilled	0.00	0.00	0.10	0.11	**0.40**	0.39
Advanced	0.00	0.00	0.00	0.07	0.00	**0.93**

Table 5.4 Six-Learning -Level Model for Pre-kindergarten assessment test (NCEDL data, N = 2892). Diagonal transition probabilities in bold to enable interpretation.

Table 5.4 depicts the transition probability matrix. This matrix indicates the probability of being in the column learning level at Time 2, conditional on being in the row learning level at Time 1. The diagonal elements of this matrix represent the probability of being in a particular learning level at Time 2 conditional on being in that same learning level at Time 1. Specifically, considering the two time points that the children were assessed, 42% of the children who were considered to be in the "Not Ready" learning level maintained their learning level membership at Time 2 (but 58% moved). Only 6% of those children did well to move to the "skilled" level, whilst 2% moved to the "advanced" level. This is not a surprising outcome, the data clearly showed that children classified to be in the "Not Ready" level initially demonstrated little to no ability with respect to the assessment test, and other factors may have contributed for the 2% to reach the "advanced" level.

In contrast, 93% of children initially in the "advanced" learning level maintained their level membership, but 7% reversed to lower learning levels. The developmental reversal is possible especially when the conditions at Time 2 are unfavorable, or simply some of the children may have slipped. Also 40% of those in the "skilled" level maintained their level

membership, whilst 39% moved to "advanced" level, with 21% of the children reversing to lower levels.

Seventeen percent of the children initially categorized as being in the "Transitional" learning level moved to the "skilled" learning level, and noticeably 33% moved to the "advanced" level, whilst 43% maintained their level membership. Interestingly, none of these children reversed to the "developing" level but 6% actually moved down to the "Inexperienced" level; which is considered lower than the "developing" level. Several factors may have contributed to this development. One of the factors may well be that the 6% guessed and correctly responded to some items at the initial assessment, and their actual ability may not have warranted them being placed at a higher learning level.

Furthermore, children in the "Inexperienced" learning level experienced no developmental reversal. In fact, 67% of them moved to higher levels, with 33% maintaining their level membership. Specifically, 11% moved to the "Developing" level, 28% to the "Transitional" level, 18% to the "skilled" level, and 10% moved all the way to the "advanced" level. Same cannot be said of the children in the "Developing" levels, whilst 33% of them maintained their level membership, 14% reversed to lower learning levels and only 6% transitioned to the "advanced" level.

5.3 The LTA Model

As discussed earlier, the LTA model makes no assumption of the ordering of the levels, and the transition matrices are also considered without order. However, the basis of the OLTA model is the order restriction imposed on the cumulative response probabilities, which facilitates ordering of the levels and the transition probabilities. To provide basis for comparison between

the two models, we shall apply the empirical example provided above to the LTA model. Again, the parameters of interest are the item response probabilities, level prevalences, and the transition matrices. The two competing models to be considered are the saturated and the growth models.

	Saturated Model		Growth Model	
Levels	AIC	BIC	AIC	BIC
2	35250.52	35393.79	35368.87	35506.17
3	32759.73	32992.55	32903.69	33118.60
4	32204.45	32538.76	32369.63	32668.12
5	31640.73	**32088.46**	31822.97	32211.01
6	31520.61	32093.70	31619.64	32103.19
7	31437.51	32147.90	31613.71	32198.74
8	31381.65	32241.29	31467.35	32159.84
9	**31351.64**	32372.46	31384.56	32190.47
10	31379.08	32573.02	31418.98	32344.28

Table 5.5 Summary of information for selecting the appropriate model under LTA
Saturated model with 9 levels selected as the most appropriate model for the data.

The information criteria (table 5.5) provide inconsistent messages about which model best balances parsimony and fit. According to the AIC, the model with nine learning levels is preferred, but according to the BIC, the model with five learning levels is preferred. Due to unordered nature of the LTA model, and also the fact that students generally develop differently, we ultimately chose the nine-level saturated model due to its conceptual appeal.

	Level1	Level2	Level3	Level4	Level5	Level6	Level7	Level8	Level9
Item1	0.02	0.20	0.31	**0.88**	**0.80**	**0.92**	**0.96**	**0.96**	**0.99**
Item2	0.01	0.18	**0.52**	**0.56**	**0.95**	**0.96**	**1.00**	**0.96**	**1.00**
Item3	0.00	0.38	**0.73**	0.03	0.11	0.41	**0.53**	**0.84**	**1.00**
Item4	0.00	0.13	0.46	0.00	0.10	0.36	0.34	**0.97**	**1.00**
Item5	0.01	0.31	**0.58**	0.00	**0.95**	**0.66**	**0.97**	**1.00**	**1.00**
Item6	0.00	0.08	0.41	0.07	0.23	0.10	**0.67**	0.25	**0.84**
Item7	0.00	0.06	**0.79**	0.01	0.01	0.07	**0.58**	0.11	**0.99**
Item8	0.00	0.06	**0.66**	0.00	0.02	0.07	0.49	0.11	**0.83**
Item9	0.01	0.28	**0.56**	0.11	0.15	0.38	**0.69**	0.46	**0.87**

Table 5.6 Item response probabilities for saturated model with 9 learning levels Conditional probabilities > .5 in bold to facilitate interpretation.

Table 5.6 clearly shows different pathways of children reading development. For instance, whilst children in level9 demonstrated skill by highly endorsing each of the nine assessment items, those in level1 showed no skill in all the assessment items. The level2 children showed little skill, they had a somewhat high likelihood of naming all upper-and lower-case letters of the alphabet (0.38), and predicting what will happen next in stories (0.31), but the likelihoods as seen are below 0.5. Their counterparts in level3however showed skill in six out of the nine items, they demonstrated a wide range of skill including naming all upper-and lower-case letters of the alphabet, predicting what will happen next in stories, using different strategies to read unfamiliar words, demonstrating early writing behavior, and understanding of some of the conventions of print. Interestingly, those same children could neither use complex sentence structures nor produce rhyming words. Despite their wide range of skill, level3 children are not considered to be verbal.

Children in level4 are completely opposite to their counterparts in level3 in terms of knowledge, and verbal skills. The level4 children showed no skill in all but two of the nine assessment items. They could not name the letters in the alphabets nor predict what will happen next in stories. Also, they could not use different strategies to read unfamiliar words, and did not demonstrate early writing behaviors. Nonetheless, they are able to use complex sentence structures, and they also understand and interpret a story read to them. Unlike those in level3, children in level4 are considered to be only verbal.

Even though the level4 children could perhaps speak clearly, their skills were limited to the extent that they could not predict what will happen next in stories. Their inability to predict what will happen next in stories separate them from their counterparts in level5. Just like children in level4, those in level5 demonstrated no skill in six of the nine assessment items. They however showed same skill as those in level4, except that they demonstrated an additional skill: they were able to predict what will happen next in stories. The level5 children are considered to be more verbal. Those in Level6 could also understand and interpret a story read to them, use complex sentence structures, and predict what will happen next in stories. However, they also had a somewhat high likelihood of naming all upper-and lower-case letters of the alphabet (0.41). Though the likelihood is below 0.5, notwithstanding, it is one of the contributing factors that separate level6 from level5.

The level7 children demonstrated skill in almost all the nine assessment items. However, they could not produce rhyming words, but also had a high likelihood of demonstrating early writing behaviors (0.49). Children in level8 showed skill in the first five of the nine assessment items, including using complex sentence structures, understanding and interpreting a story or other text read to them, naming all upper-and lower-case letters of the alphabet, etc. However,

they could not do anything else (apart from those five items), for instance, they could not read simple books independently nor use different strategies to read unfamiliar words. Also they did not demonstrate early writing behaviors, but showed a high likelihood of understanding some of the conventions of print (0.46). As already stated, children in level3 and level7 showed a lot of promise, but only those in level9 demonstrated skill in all the nine assessment items.

Table 5.6 shows that children in level1 displayed no skill in all the assessment items whilst their counterparts in level9 showed skill in every single item, this phenomenon by no means reflect ordering of the levels. The item response probabilities reveal some important characteristics of several items which have contributed to the unordered nature of the LTA model. These items were endorsed highly by children in the lower levels than those in the upper level. Specifically, children in level3 displayed higher skill for item3 than their counterparts in level4, level5, and level6. In other words, children in level3 could easily and quickly name all the upper-and lower-case letters of the alphabet, whilst those who are in the upper levels, such as level4, level5, and level6 could not do same. Also, children in level3 could predict what will happen next in stories, but those in level4 could not predict what will happen next in stories.

Furthermore, children in level7 demonstrated higher skill in items 6, 7, 8, and 9 than those in level8. Put differently, the level7 children were able to read simple books independently, used different strategies to read unfamiliar words, demonstrated an understanding of some of the conventions of print, and to some extent demonstrated early writing behaviors, but their counterparts in the upper level (level8) could not do any of these. Interestingly, these same mentioned items were highly endorsed by children in level3 than those in level4, level5, and level6.

One of the parameters of interest in the LTA model is the level prevalences. Table 5.7 show estimates of the population proportions of children belonging to each learning level at time 1, and time 2. The graphical representation is depicted in figure 5.2.

	Level	Prevalences
Level	Time 1	Time 2
1	0.520	0.252
2	0.094	0.135
3	0.012	0.037
4	0.118	0.065
5	0.112	0.098
6	0.037	0.090
7	0.030	0.063
8	0.041	0.129
9	0.033	0.131

Table 5.7 Learning level proportions for Time 1 and Time 2 for LTA model

Table 5.7 shows that at the initial time point, 52% of the children belonged to level1, and a little over 3% belonged to level9. However, more than 50% of the children who initially belonged to level1, moved to upper learning levels at time 2, leaving a little over 25% of the children at level 1 at time 2. In other words, just over 25% of the children were still without skill at the final time point.

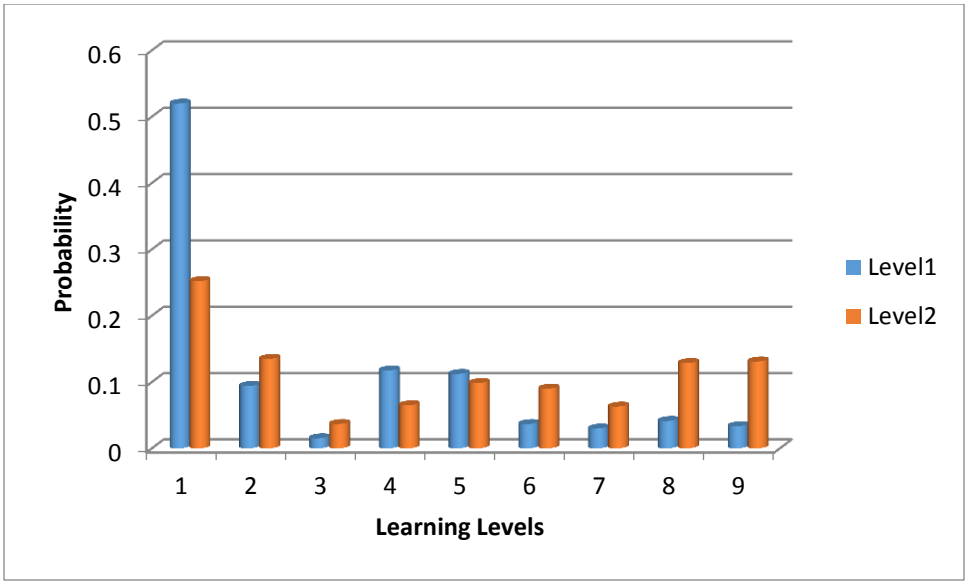


Figure 5.2 Learning level membership probabilities for LTA model.

Although level3 membership saw an increase from 1.2% to 3.7%, it is still the level with the fewest memberships at both time points. This is interesting because the peculiar nature of the level3 membership perhaps contribute to its size. These are the children who displayed skill in almost all the nine items, except that they could not use complex sentence structures. Children in general may not follow that kind of developmental pattern, and that perhaps explain the uniqueness of this level. On the other hand, it is not uncommon at all, to find children who can speak clearly but have no knowledge on the alphabetic system or are unable to read or write. Such children fall within levels 4 and 5, and from table 5.7, level4 and level5 had membership of 11.8% and 11.2% respectively at time1, and 6.5% and 9.8% at time2.

	L e a r n i n g					L e v e l s			
	Level1	Level2	Level3	Level4	Level5	Level6	Level7	Level8	Level9
Level1	**0.44**	0.18	0.02	0.06	0.09	0.05	0.06	0.06	0.03
Level2	0.00	**0.35**	0.16	0.00	0.00	0.17	0.00	0.14	0.18

Level3	0.00	0.00	**0.53**	0.00	0.00	0.00	0.05	0.00	0.41
Level4	0.09	0.03	0.01	**0.20**	0.10	0.35	0.02	0.12	0.09
Level5	0.08	0.03	0.01	0.08	**0.35**	0.00	0.13	0.19	0.14
Level6	0.00	0.00	0.00	0.00	0.00	**0.00**	0.10	0.90	0.00
Level7	0.02	0.02	0.00	0.00	0.00	0.11	**0.34**	0.02	0.48
Level8	0.03	0.03	0.06	0.00	0.00	0.00	0.00	**0.36**	0.52
Level9	0.00	0.00	0.03	0.00	0.00	0.08	0.00	0.07	**0.83**

Table 5.8 Transition probabilities for the 9-level saturated model. Diagonal transition probabilities in bold to enable interpretation.

Table 5.8 depicts the transition probability matrix between time1 and time 2. The diagonal elements of the matrix indicate the probability of being in a particular learning level at Time 2 conditional on being in that same learning level at Time 1. From table 5.8, 44% of the children who were considered to be in level1 maintained their level membership at Time 2 (but 56% moved). Only 3% of those children did well to move to level9, which is considered to be the most skilled level. Eighty three percent of children who were originally associated with the most skilled level (level9) maintained their level membership at Time 2. However, 17% reversed to lower levels.

Thirty five percent of the children who showed very little skill in level2 maintained their level membership (65% moved). Though they displayed little to no skill at Time 1,18% of them moved to level9 at Time 2. For level3, 53% of the members maintained their level membership at Time 2, whilst 47% moved. Forty one percent of the proportion that moved ended up with the most skilled level. This is not a surprising outcome considering how skilled the level3 children were; at the initial time point. It is equally not surprising that only 9% of children considered to be in level4 moved to level 9 at Time 2. We have already established that the level4 children could speak, but could not do anything else, and their lack of skill in other areas was a clear

indication that few would end up in a highly skilled level at Time 2. Also, 20% maintained their level membership but majority of the children (35%) moved to level6, which is considered to be a more verbal level.

Perhaps the most interesting transition happened in level6. Whilst none (0%) of those who were associated with level6 maintained their level membership at Time 2, 10% moved to a skilled level in level7, and 90% moved to level8. One of the factors that could explain this movement is that although the children in level6 were more verbal, they were also characterized with having a high likelihood (0.41) for alphabetic knowledge. Having the skill for alphabetic knowledge may have contributed to the massive movement.

Chapter 6

Discussion

In this project we have discussed and extended the LTA model to OLTA by imposing inequality constraints on the cumulative response probabilities of the LTA model. By employing Croon's idea, we have demonstrated that imposing inequality restriction on the item response and cumulative response probabilities establish an order relation on the set of the learning levels along the level continuum. Most importantly, the OLTA model ensures that we are able to track the upward and downward movements of subjects along the level continuum at several time points.

6.1 Simulation

The simulation study provided interesting results. There is strong evidence that the EM algorithm can recover the OLTA model parameters well under several conditions. The simulation suggests that large sample size improve parameter recovery. We used two sample sizes: N = 1000, and 2000 for the simulation study. Although N = 1000 would not be considered a small sample size, the results showed a decrease in bias and the mean squared error, and a much more improved parameter recovery when the sample size was increased from 1000 to 2000 for the same experimental condition. Collins & Wugalter (1992) found no evidence of bias for N = 300 for LTA parameter recovery, but in this project, we did not establish boundaries on the sample sizes for OLTA parameter recovery. A sample size of 300 may not be considered small but we do not know how that would affect the parameter recovery of the OLTA model, and we

are unsure of how low a sample size should be before the OLTA parameter recovery is negatively affected.

The simulation results also suggest that increasing number of items dramatically improve parameter recovery for conditions with the same sample size. Item sizes of 10 and 20 were used for this study. The results have shown that increasing the item size to 20 even for N = 1000 conditions improved parameter estimation by reducing bias and mean squared errors. Similar to the LTA model, adding more items have the potential to create an impossibly sparse contingency table that could adversely affect parameter estimation (Collins & Wugalter, 1992). But in this study we did not test the limits of the item sizes to ascertain the number of items that can be added before the OLTA model parameter recovery encounter complications. Having said that, the results strongly suggests that adding items with at least a medium discrimination level (i.e. good items) are enormously beneficial to the model. These benefits seem to supersede the risk of increased sparseness. This conclusion on the OLTA model is consistent with the findings of Collins & Wugalter (1992) on LTA parameter estimation.

In order to establish the effectiveness of the OLTA model, and to test the robustness of the parameter recovery, four estimation procedures were employed for each condition in this study: true Growth model estimated as Growth, true Growth model estimated as Saturated, true Saturated model estimated as Saturated, and true Saturated model estimated as Growth. The results showed that even when a true Growth model is estimated as Growth, the bias and mean squared error increase with items of low discrimination. In contrast, correctly estimating a Growth model as Growth for sufficiently large samples sizes, and items with at least a medium discrimination show an excellent parameter recovery; resulting in a drastic decrease of bias and mean squared error. Interestingly, we achieved similar results when a true Growth model was

estimated as Saturated for sufficiently large sample with items of at least a medium index. Thus, with at least a medium discrimination items and large enough sample, estimating a true Growth model as Growth seem to be as good as estimating same model as Saturated.

However, when a true Saturated model is estimated as Growth, the parameter recovery is very poor resulting in a substantial increase of bias and the mean squared error even for sufficiently large sample sizes with items of at least a medium index. But the parameter recovery is excellent for large sample sizes, and items with good discrimination, when a true Saturated model is appropriately estimated as Saturated. The result is evident with a dramatic decrease of bias and mean squared error associated with such estimation procedure. In a nutshell, it is inappropriate to estimate true Saturated model as Growth.

6.1.1 Selection Technique

Through the simulation studies, the performance of the AIC and BIC in selecting the appropriate transition model and also choosing the optimal levels were examined. With respect to the transition model, 3 and 5- level Growth and Saturated models were examined under several conditions. Displayed in tables 4.2 and 4.3, the AIC and BIC correctly identified the true saturated model 100% of the time in all conditions studied for the 3 and 5 level models. The performance of the criterion index in selecting the true Growth model for the 3 and 5 levels were equally impressive. In a nutshell, the AIC and BIC showed comparable abilities in selecting the true transition models.

In addition to determining the performance of the fit indices in selecting the true transition model, we conducted another simulation to ascertain the effectiveness of the AIC and BIC in selecting the optimal levels in a model. For each of the 3 and 5 level models, an opposing model was conducted for the purpose of comparison. Also several levels were included in each study to ensure the credibility of the fit indices in selecting the optimal levels. In the case of the 3 level Growth model, it is clear from tables 4.4 and 4.5 that the AIC did a very good job in identifying and selecting the correct optimal levels in high proportion in a consistent manner. However, the BIC underperformed; it occasionally opted for a different level and sometimes endorsed the correct level in a relatively low proportion. For the 3 level saturated model, the AIC again did well and was consistent in selecting the correct level in a high proportion. Again, the BIC underperformed.

As seen in tables 4.8 and 4.9, the 5 level Growth model presented something different: whilst the AIC was somewhat inconsistent in identifying and selecting the correct level, the BIC performed poorly by displaying consistency in its preference for the 3 and 4 level Growth models to the 5 level Growth model. The AIC displayed same inconsistency in the case of the 5 level saturated model, as the fit index occasionally preferred the 4 level saturated model to the 5 level model, and in some cases weakly endorsing the correct level. The BIC meanwhile; almost wholly endorsed the 3 and the 4 level saturated models instead of the 5 level model. It worth mentioning that despite the seemingly inconsistencies and apparent disagreement and/or agreement of the fit indices, neither the AIC nor the BIC preferred the Growth model in the case where the model under consideration was saturated. However, the fit indices disagreed (occasionally) on the appropriate level.

6.2 Real Data Analysis for the OLTA model

The ability to read in early stages of a child's life is considered as the basis for learning and academic work. It is therefore not surprising that Paris (2005) describes the process of achieving such skill, the greatest childhood achievement. Table 5.4 represents the summary of the data analysis regarding the reading levels of the pre-kindergarteners, and how they transitioned from one level to another with respect to time. The fundamental principle of the OLTA model is that the levels are ordered, and consistent with this principle, table 5.4 shows that children in the "advanced" level demonstrated the highest proficiency on all nine assessment items. The "advanced" level children have the highest ability, followed by those in the "skilled" level, and children associated with the "Not Ready" level are considered the least proficient.

Again, table 5.4 shows an interesting dynamics between the "Developing" and "Transitional" levels. The fundamental difference between children in the "transitional" level and those at the "Developing" level is that the "transitional" level children demonstrated additional proficiency by naming all the upper-and lower case letters of the alphabet easily and quickly, and also were able to produce rhyming words. It is reasonable to suggest that the additional abilities displayed at the "transitional" level were critical in moving 33% of those children to the "advanced" level.

Children's ability to easily and quickly naming all the upper-and lower case letters of the alphabet is important, and played a major role in helping move quite a number of children from the "transitional" level to the "advanced" level. For the purpose of this study, and to fully appreciate children's reading development, we shall attempt to make connections between the results of the OLTA model, and some research findings. For instance, research shows that

children's ability to learn new words improve when they have alphabetic knowledge (Ehri, 2005). It is no secret however that the alphabetic knowledge is recognized as the basis upon which all other words are made. In fact, studies show that letter recognition at kindergarten is usually a powerful predictor of reading a year later (DeHirsch, Jansky & Langford, 1966; Stevenson, Parker, Wilkinson, Hegion, & Fish, 1976; Bruininks & Mayer, 1979).This is an indication that children at the "advanced" level, and to some extent the "skilled" level are on track for a successful academic work at the kindergarten and beyond.

It has been understood that once children become proficient in the alphabetic system, they are able to build their vocabulary easily with sight words (Ehri, 2005). But some children after acquiring the alphabetic knowledge still struggle with printed words, and need a lot attention in order to be comfortable, and become proficient in sight word learning (Ehri & Saltmarsh, 1995; Reitsma, 1983). In particular, Table 5.4 shows that children in the "transitional" level were proficient in the alphabetic system, but those same children could not really demonstrate an understanding of some conventions of print. These children need time, and obviously a lot of practice with sight words in order for them to achieve mastery in printed words. Also, children reading development has been classified into four phases: Pre-alphabetic, Partial alphabetic, Full alphabetic, and consolidated alphabetic phases (Ehri, 1999; Ehri & McCormic, 1998).

According to the authors, children at the Pre-alphabetic phase have little knowledge of the alphabetic system, and are unable to form letter-sound connections in order to read words. These children may be able to guess words from pictures, and pretend to read words they have heard several times, but they are basically nonreaders. Based on the analysis from table 5.4, an argument could be made for the children at the "Developing" level as being on the Pre-alphabetic

phase even as pre- kindergarteners. The reason being, these children have little to no knowledge on the alphabetic system. However, they use complex sentence structures, understand and interpret a story or other texts read to them, and also able to predict what will happen next in stories.

Ehri (2005) also posited that children transition to the partial alphabetic phase after learning the names or sounds of the alphabets that facilitate recollection of how words are read. Children at partial alphabetic phase do not have complete knowledge of the alphabetic system, and as a result; they find difficulty in reading unfamiliar words. It is clear from the cumulative response probabilities in table 5.2 that the "Transitional" level children have some knowledge of the alphabetic system. However, they can neither read independently nor have the ability to decode unfamiliar words. It is therefore reasonable to suggest that the "Transitional" level children are at the partial alphabetic phase with respect to reading as pre-kindergarteners.

"Children become full alphabetic phase readers when they can learn sight words by forming complete connections between letters in spellings and phonemes in pronunciations" (Ehri, 2005). Children at this phase are clearly better readers and spellers than those at the partial alphabetic phase. This is due to the fact that the full alphabetic phase children have an understanding of the grapheme-phoneme connections (Venezky, 1970, 1999). Similar to the "Skilled" level children, those at the full alphabetic phase are able to devise ways to read unfamiliar words. It may be within reason to consider the "Skilled" level children as reaching the full alphabetic phase as pre-kindergarteners. As children are able to commit more sight words into memory, and easily recognize letter patterns that reappear in different words, they are able to rely on letter chunks to read "big" words. Children at this level are classified as being on the consolidated phase (Ehri, 2005). Again, Table 5.2 shows that the "advanced" level children

performed exceedingly well in all the nine assessment items. It may not be a stretch to liken the "advanced" level children to those at the consolidated phase.

6.3 Brief comparison of LTA and OLTA perspectives

In order to provide a context for the OLTA results, we also analyzed the data using the LTA procedure. The LTA procedure selected a 9 level saturated model as the most appropriate to fit the data set. The 9 levels showed several pathways to children reading development. The levels included children without skills to those that are highly skilled. Since the LTA models are considered without ordering of the levels, and the transition probabilities, children that are associated with (say) level 5 are not necessarily considered to possess higher skills than their counterparts in the "lower" levels. For instance, Table 5.6 shows that children at level3 are highly skilled than their counterparts in level4, level5, and level6. Also, children at level7 possess more skills than those in level8.

The characteristics of each of the nine levels portray what one would consider as a natural developmental process. In this case the levels seem to adequately capture children reading development. Certainly, there are children who could neither speak clearly, nor possess alphabetic knowledge. Also, some children could use complex sentence structures, and speak clearly but have no alphabetic or reading skills. Of course, there are others who demonstrate alphabetic, reading, and writing skills but cannot use complex sentence structures. The bottom line is; the LTA procedure seems to have a room for the various combinations of scenarios that may be considered a pathway for children reading development.

Despite the unordered nature of the LTA model, Table 5.6 show what we describe as a natural ordering of levels 4, 5, 6, and 7. Clearly, children at level5 possess more skill than those

at level4, children at level 6 also possess more skill than their counterparts in levels 5, and 4. Furthermore, level7 children have more skills than those in levels 6, 5, and 4. Since the overall levels of the LTA procedure are unordered, we describe the ordering from level 4 to level 7 as partial ordering.

In contrast, the OLTA analyses place restrictions on the cumulative response probabilities to ensure ordering of the levels, and the transition probabilities. Unlike the LTA, children associated with level 5 under OLTA means that those children possess skills that are higher than their counterparts in the lower levels. The OLTA procedure is an effective data reduction tool in the sense that; the ordering nature of the model "cleans up" the data in such a way that clearly depicts the progression of children development. In this case, the progression a child's reading development is clearly specified by the six levels in Table 5.2. For instance, children in level1 showed no skill so we labeled them as "Not Ready", the level2 children showed very little skill so we labeled them as "Inexperienced", those in level3 could speak but had no alphabetic knowledge; neither could they read or write. We labeled the level3 children as "Developing".

Children in level4 could speak just like their counterparts in level3, but they had alphabetical knowledge in addition, so we labeled them "Transitional" and so on. One of the benefits of the OLTA approach is that each of the levels is unambiguously defined, and that could potentially help researchers to easily target the problem level for intervention. Also the clarity of the levels could help researchers to quickly and easily track children who reverse developmentally. With respect to children's reading development, any child/children transitioning from level 5 to level 4 under OLTA approach means that the child has reversed developmentally. But we cannot necessarily draw the same conclusion for the same movement under LTA model.

6.4 Strengths and Limitations of the Ordered Latent Transition Analysis

The discussion above clearly shows important strengths of the OLTA. First, the OLTA technique enables researchers to test several stage-sequential models concerning human development. Second, the procedure can be used to assess the efficacy of an intervention program, and also estimate the differential effectiveness of such interventions for subjects in different levels. Another important feature of the OLTA model is that the ordering natures of the levels help reveal unique characteristics of data. The OLTA model is suitable for educational measurement, modeling alcohol cessation or adolescent delinquent behavior, etc.

However, OLTA also has some limitations. First, the simulation showed that parameter recovery improved dramatically with increasing sample size. As Collins and Wugalter (1992), and Graham et al., (1991) have already established for LTA model, the OLTA model also require relatively large sample size. Second, the over reliance of the AIC and BIC for selecting the appropriate model may not be the best. As the simulation showed, although the criterion index performed very well in selecting the correct transition models, and also selecting the correct 3 level models, the fit indices performed poorly for the selection of the 5 level models. A more reliable and consistent fit indices are badly needed to supplement the AIC and the BIC. Furthermore, future studies are needed to ensure that while ordering is maintained, different learning trajectories are allowed. Also, the LTA procedure showed some partial ordering of the levels. The partial ordering levels could be investigated further in future studies.

References

Andersen, E. B. (1985). Estimating latent correlations between repeated testing. *Psychometrika*, *50*(1), 3-16.

Anderson, E. (1982). Latent structure analysis: A survey. *Scandinavian Journal of Statistics,* 9:1-12.

Anderson, T. W. (1954). Probability models for analyzing time changes in attitudes. Mathematical thinking in the Social Sciences, The Free Press, Glencoe, pp. 17-66.

Andrade, D. F., & Tavares, H. R. (2005). Item response theory for longitudinal data: population parameter estimation. *Journal of Multivariate Analysis*, *95*(1), 1-22.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, *43*(4), 561-573.

Ayer, M., Brunk, H. D., Ewing, G. M., Reid, W. T., & Silverman, E. (1955). An empirical distribution function for sampling with incomplete information. *The annals of mathematical statistics*, *26*(4), 641-647.

Baker, R. S., Corbett, A. T., & Aleven, V. (2008). Improving contextual models of guessing and slipping with a truncated training set. *Human-Computer Interaction Institute*, 17.

Baker, F. B., & Kim, S. H. (Eds.). (2004). *Item response theory: Parameter estimation techniques*. CRC Press.

Bartholomew, D. J. (1987). Latent Variable Models and Factor Analysis. New York, NY: Griffin.

Bartholomew, D. J. (1983). Latent variable models for ordered categorical data. *Journal of Econometrics*, *22*(1-2), 229-243.

Bartholomew, D. J. (1981). *Mathematical methods in social science* (Vol. 1). John Wiley & Sons.

Battista, M. T. (2011). Conceptualizations and issues related to learning progressions, learning trajectories, and levels of sophistication. *The Mathematics Enthusiast*, *8*(3), 507-570.

Battista, M. T. (2007). The development of geometric and spatial thinking. *Second handbook of research on mathematics teaching and learning*, *2*, 843-908.

Beck, J. E., & Chang, K. M. (2007, July). Identifiability: A fundamental problem of student modeling. In *International Conference on User Modeling* (pp. 137-146). Springer Berlin Heidelberg.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In Lord, F. and Novick, M., editors, *Statistical theories of mental test scores*.

Blumen, I. M., Kogan, M & McCarthy, P.J.(1955). *The Industrial Mobility of* labor as a probability process. Ithaca: Cornell University Press.

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

Bollen, K.A.(1989). Structural Equations with Latent Variables, Wiley & Sons, New York, NY.

Briggs, D. C., & Alonzo, A. C. (2012). The psychometric modeling of ordered multiple-choice item responses for diagnostic assessment with a learning progression. In Learning progressions in science (pp. 293-316). Sense Publishers.

Bruininks, A. L. & Mayer, J. H. (1979). Longitudinal study of cognitive abilities and academic achievement. *Perceptual and Motor Skills,* 48,1011-1021.

Bulik, C. M., Sullivan, P. F., & Kendler, K. S. (2000). An empirical study of the classification of eating disorders. *American Journal of Psychiatry*, *157*(6), 886-895.

Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference understanding AIC and BIC in model selection. *Sociological methods & research*, *33*(2), 261-304.

Bye, B. V., & Schechter, E. S. (1986). A latent Markov model approach to the estimation of response errors in multiwave panel data. *Journal of the American Statistical Association*, *81*(394), 375-380.

Chen, H., Cohen, P., & Chen, S. (2010). How big is a big odds ratio? Interpreting the magnitudes of odds ratios in epidemiological studies. *Communications in Statistics—Simulation and Computation*, *39*(4), 860-864.

Chung, H., Lanza, S. T., & Loken, E. (2008). Latent transition analysis: Inference and estimation. *Statistics in medicine*, *27*(11), 1834-1854.

Clements, D., Sarama, J., Spitler, M., Lange, A., and Wolfe, C. B.(2011). Mathematics learned by young children in an intervention based on learning trajectories: A large-scale cluster randomized trial. *Journal for Research in Mathematics Education, 42*, 127–166.

Clements, D. H., & Sarama, J. (2009). Learning trajectories in early mathematics–sequences of acquisition and teaching. *Encyclopedia of language and Literacy Development*, 1-7.

Clements, D. H., & Battista, M. T. (1992). Geometry and spatial reasoning. In D Grous, Handbook of research on mathematics teaching and learning, 420- 464. Reston, VA: National Council of Teachers of Mathematics.

Coffman, D. L., Patrick, M. E., Palen, L. A., Rhoades, B. L., & Ventura, A. K. (2007). Why do high school seniors drink? Implications for a targeted approach to intervention. *Prevention Science*, *8*(4), 241-248.

Collins, L. M., & Lanza, S. T. (2010). Latent class analysis with covariates. *Latent class and latent transition analysis: With applications in the social, behavioral, and health sciences*, 149-177. Wiley; Hoboken, NJ.

Collins, L. M., & Wugalter, S. E. (1992). Latent class models for stage-sequential dynamic latent variables. *Multivariate Behavioral Research*, *27*(1), 131-157.

Collins, L. M. (1991a). The measurement of dynamic latent variables constructs in longitudinal aging research: Quantifying adult development. *Experimental Aging Research, 17*, 13-20.

Collins, L. M. (1991b). Measurement in longitudinal research. In L. M. Collins and J.L. Horn(Eds.), Best methods for analysis of change: Recent advances, unanswered questions, future directions. Washington, DC, US: American Psychological Association.

Collins, L. M., & Cliff, N. (1990). Using the longitudinal Guttman simplex as a basis for measuring growth. *Psychological Bulletin*, *108*(1), 128.

Collins, L. M., Cliff, N., & Dent, C. W. (1988). The longitudinal Guttman simplex: A new methodology for measurement of dynamic constructs in longitudinal panel studies. *Applied Psychological Measurement*, *12*(3), 217-230.

Confrey, J., & Maloney, A. P. (2010). A next generation of mathematics assessments based on learning trajectories. *East Lansing, MI*.

Corbett, A. T., & Anderson, J. R. (1995). Knowledge tracing: Modeling the acquisition of procedural knowledge. *User modeling and user-adapted interaction*, *4*(4), 253-278.

Corcoran, T., Mosher, F. A., & Rogat, A. (2009). Learning Progressions in Science: An Evidence-Based Approach to Reform. CPRE Research Report# RR-63. *Consortium for Policy Research in Education*.

Croon, M. A. (1991). Investigating Mokken scalability of dichotomous items by means of ordinal latent class analysis. *British Journal of Mathematical and Statistical Psychology*, *44*(2), 315-331.

Croon, M. (1990). Latent class analysis with ordered latent classes. *British Journal of Mathematical and Statistical Psychology*, *43*(2), 171-192.

Daro, P., Mosher, F.A., & Corcoran, T. (2011). Learning trajectories in mathematics: A foundation for standards, curriculum, assessment, and instruction (Consortium for Policy Research in Education Report #RR-68). Philadelphia, PA: Consortium for Policy Research in Education.

DeHirsch, K., Jansky, J. J. & Langford, W. S. (1966). *Predicting reading failure.* New York: Harper & Row.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society. Series B (methodological)*, 1-38.

Dillon, W. R., & Kumar, A. (1994). Latent structure and other mixture models in marketing: an integrative survey and overview. *Advanced methods of marketing research*, 295-351.

Draney, K. L. (1996). *The polytomous Saltus model: A mixture model approach to the diagnosis of developmental differences* (Doctoral dissertation, University of California, Berkeley).

Draney, K., & Wilson, M. (2007). Application of the Saltus model to stagelike data: Some applications and current developments. In *Multivariate and mixture distribution Rasch models* (pp. 119-130). Springer New York.

Duschl, R., Maeng, S., & Sezen, A. (2011). Learning progressions and teaching sequences: A review and analysis. *Studies in Science Education*, *47*(2), 123-182.

Ehri, L. C. (2005). Learning to read words: Theory, findings, and issues. *Scientific Studies of reading*, *9*(2), 167-188.

Ehri, L. C. (1999). Phases of development in learning to read words. In J. Oakhill & R. Beard (Eds.), Reading development and the teaching of reading: A psychological perspective (pp. 79–108). Oxford, UK: Blackwell Publishers.

Ehri, L. C., & McCormick, S. (1998). Phases of word learning: Implications for instruction with delayed and disabled readers. *Reading & Writing Quarterly: Overcoming Learning Difficulties*, *14*(2), 135-163.

Ehri, L. C., & Saltmarsh, J. (1995). Beginning readers outperform older disabled readers in learning to read words by sight. *Reading and Writing*, *7*(3), 295-326.

El Barmi, H.,& Johnson, M. (2006). A unified approach to testing for and against a set of linear inequality constraints in the product multinomial setting. *Journal of Multivariate Analysis*, 97:1894-1912.

Embretson, S. E., & Steven, P. Reise.(2000). *Item response theory for psychologists*. Lawrence Erlbaum Associates, Publishers.

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

Fischer, G. H. (2001). Gain scores revisited under an IRT perspective. In *Essays on item response theory* (pp. 43-68). Springer New York.

Fischer, G. H. (1989). An IRT-based model for dichotomous longitudinal data. *Psychometrika, 54*, 599–624.

Fischer, G. H. (1983a). Logistic latent trait models with linear constraints. *Psychometrika*, *48*(1), 3-26.

Fischer, G. H. (1977a). Some probabilistic models for the description of attitudinal and behavioral changes under the influence of mass communication. *Mathematical models for social psychology*, 102-151.

Fischer, G. H. (1977b). Linear logistic trait models: Theory and application. *Structural models of thinking and learning*, 203-225.

Fischer, G. H. (1976). Some probabilistic models for measuring change. *Advances in psychological and educational measurement*, 97-110.

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Actapsychologica*, *37*(6), 359-374.

Garrett, E. S., & Zeger, S. L. (2000). Latent class model diagnosis. *Biometrics*, *56*(4), 1055-1067.

Goodman, L. A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, *61*(2), 215-231.

Goodman, L. A. (1962). Statistical methods for analyzing processes of change. *American Journal of Sociology*, *68*(1), 57-78.

Gotwals, A. W., &Songer, N. B. (2013). Validity Evidence for Learning Progression-Based Assessment Items That Fuse Core Disciplinary Ideas and Science Practices. Journal of Research in Science Teaching, 50(5), 597-626.

Graham, J. W., Collins, L. M., Wugalter, S. E., Chung, N. K., & Hansen, W. B. (1991). Modeling transitions in latent stage-sequential processes: a substance use prevention example. *Journal of consulting and clinical psychology*, *59*(1), 48.

Guttman, L. (1950a). The basis for scalogram analysis. InS. A. Stauffer et al.(eds.), Measurement and Prediction. Princeton, NJ: Princeton University Press.

Guttman, L. (1950b). Relation of scalogram analysis to other techniques. *Measurement and Prediction. Studies in Social Psychology in World War II*, *4*, 172-212.

Guttman, L. (1947). The Cornell technique for scale and intensity analysis. *Educational and Psychological Measurement*, *7*(2), 247-279.

Harris, J., Laan, S., & Mossenson, L. (1988). Applying partial credit analysis to the construction of narrative writing tests. *Applied Measurement in Education*, *1*(4), 335-346.

Heinen, T. (1996). Latent class and discrete latent trait models: Similarities and differences. Thousands Oaks, CA: Sage.

Heinen, T. (1993). Discrete Latent Variable Models. Tilburg University Press.

Heritage, M. (2008). Learning progressions: Supporting instruction and formative assessment. Paper prepared for the Formative Assessment for Teachers and Students, State Collaborative on Assessment and Student Standards of the Council of Chief State School Officers.

Hoijtink, H. (1998). Constrained latent class analysis using the Gibbs sampler and posterior predictive p-values: Applications to educational testing. *Statistica Sinica*, 691-711.

Hoijtink, H. and Molenaar, I. (1997). A multidimensional item response model: Constrained latent class analysis using gibbs sampler and posterior predictive check. *Psychometrika,* 62:171-189.

Holland, P. and Rosenbaum, P. (1986). Conditional association and unidimensionality in monotone latent variable models. *The Annals of Statistics*, 14:1523-1543.

Holland, P. W. (1981). When are item response models consistent with observed data? *Psychometrika*, *46*(1), 79-92.

House, P. A., & Coxford, A. F. (1995). *Connecting Mathematics across the Curriculum. 1995 Yearbook*. National Council of Teachers of Mathematics, 1906 Association Drive, Reston, VA 22091-1593.

Humphreys, K., & Janson, H. (2000). Latent transition analysis with covariates, nonresponse, summary statistics and diagnostics: Modelling children's drawing development. *Multivariate Behavioral Research*, *35*(1), 89-118.

Jain, D., Bass, F. M., & Chen, Y. M. (1990). Estimation of latent class models with heterogeneous choice probabilities: An application to market structuring. *Journal of Marketing Research*, 94-101.

Johnson, M. (2007). Modeling dichotomous item responses with free-knot splines. *Computational Statistics and Data Analysis*, 61:4178-4192.

Kaplan, D. (2008). An overview of Markov chain methods for the study of stage-sequential developmental processes. *Developmental psychology*, *44*(2), 457.

Koedinger, K. R. (2002). Toward Evidence for Instructional Design Principles: Examples from Cognitive Tutor Math 6. Proceedings of PME-NA XXXIII (the North American Chapter of the International Group for the Psychology of Mathematics Education).

Kuha, J. (2004). AIC and BIC: Comparisons of Assumptions and Performance: *Sociological methods & Research, 33*(2)*,* 188- 229.

Langeheine, R. (1994). Latent variables Markov models. In Latent Variables Analysis: Applications for Developmental Research, ed. A von Eye, CC Clogg, pp. 373–95. Thousand Oaks, CA: Sage.

Langeheine, R., & Van de Pol, F. (2002). Latent markov chains. *Applied latent class analysis*, 304-341.

Langeheine, R., & Van de Pol, F. (1990). A unifying framework for Markov modeling in discrete space and discrete time. *Sociological Methods & Research*, *18*(4), 416-441.

Langeheine, R. (1988). New developments in latent class theory. In *Latent trait and latent class models* (pp.77-108). Springer US.

Lazarsfeld, P. and Henry, N. (1968). Latent Structure Analysis. Boston, MA: Houghton Mifflin.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.

Mair, P., Hornik, K., & de Leeuw, J. (2009). Isotone optimization in R: pool-adjacent-violators algorithm (PAVA) and active set methods. *Journal of statistical software*, *32*(5), 1-24.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149-174.

Masters, G. N., & Evans, J. (1986). Banking non-dichotomously scored items. *Applied psychological measurement*, *10*(4), 355-367.

Meredith, W., & Tisak, J. (1990). Latent curve analysis. *Psychometrika*, *55*(1), 107-122.

Meredith, W. (1965). Some results based on a general stochastic model for mental tests. *Psychometrika*, *30*(4), 419-440.

Mojica, G. F. (2010). *Preparing Pre-Service Elementary Teachers to Teach Mathematics with Learning Trajectories*. ProQuest LLC. Ann Arbor, MI 48106.

Mokken, R.J. (1971). *A Theory and Procedure of Scale Analysis with Applications in Political Research*. New York, Berlin: Walter de Gruyter, Mouton.

Mokken, R.J. and Lewis, C. (1982). A nonparametric approach to the analysis of dichotomous item responses. *Applied Psychological Measurement* 6, 417–430.

Mosher, F. (2011). The role of learning progressions in standards-based education reform. Consortium for Policy Research in Education: Policy Briefs (September), 1-16.

Moustaki, I. (1996). A latent trait and a latent class model for mixed observed variables. *British journal of mathematical and statistical psychology*, *49*(2), 313-334.

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

National Council of Teachers of Mathematics (2000). *Principles and standards for school mathematics*. Reston, VA: Author.

Pardos, Z. A., & Heffernan, N. T. (2010). Modeling individualization in a bayesian networks implementation of knowledge tracing. In *International Conference on User Modeling, Adaptation, and Personalization* (pp. 255-266). Springer Berlin Heidelberg.

Paris, S. G. (2005). Reinterpreting the development of reading skills. *Reading research quarterly*, *40*(2), 184-202.

Pedhazur, E. J., & Schmelkin, L. P. (1991). Measurement, design, and analysis: An integrated analysis. *Journal of Econometrics*, 22:229-243.

Pellegrino, J. W. (2009). The design of an assessment system for the Race to the Top: A learning sciences perspective on issues of growth and measurement. *Center for K–12 Assessment & Performance Management, Educational Testing Service. http://www. k12center.org/rsc/pdf/PellegrinoPresenter-Session1. pdf*.

Plummer, J. D., & Slagle, C. (2009). A learning progression approach to teacher professional development in astronomy. In *Learning Progressions in Science (LeaPS) Conference, Iowa City, IA*.

Pollitt, A., & Hutchinson, C. (1987). Calibrating graded assessments: Rasch partial credit analysis of performance in writing. *Language Testing*, *4*(1), 72-92.

Qu, Y., Tan, M., & Kutner, M. H. (1996). Random effects models in latent class analysis for evaluating accuracy of diagnostic tests. *Biometrics*, 797-810

Rabe-Hesketh, S., & Skrondal, A. (2008). Classical latent variable models for medical research. *Statistical methods in medical research*, *17*(1), 5-32.

Ramsay, J. (1991). Kernel smoothing approaches to nonparametric item characteristic curve estimation. *Psychometrika*, 56:611-630.

Ramsay, J.,& Abrahamowicz, M. (1989). Binominal regression with monotone splines: A psychometric application. *Journal of the American Statistical Association*, 84:906-915.

Rasch, G. (1961). On general laws and the meaning of measurement in psychology. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability* (Vol. 4, pp. 321-333). Berkeley: University of California Press.

Rasch, G. (1960). Probabilistic models for some intelligence and achievement tests. *Copenhagen: Danish Institute for Educational Research*.

Read, T., & Cressie, N. (1988). Goodness-Of-Fit Statistics for Discrete Multivariate Analysis. Springer-Verlag, New York.

Reboussin, B. A., Reboussin, D. M., Liang, K. Y., & Anthony, J. C. (1998). Latent transition modeling of progression of health-risk behavior. *Multivariate Behavioral Research*, *33*(4), 457-478.

Reitsma, P. (1983). Printed word learning in beginning readers. *Journal of experimental child psychology*, *36*(2), 321-339.

Robertson, T., Wright, F. T., & Dykstra, R. L.(1988). Order Restricted Statistical Inference. Wiley, New York.

Rogosa, D., Brandt, D., & Zimowski, M. (1982). A growth curve approach to the measurement of change. *Psychological bulletin*, *92*(3), 726.

Rosenbaum, P. R. (1987a). Probability inequalities for latent scales. *British Journal of Mathematical and Statistical Psychology*, *40*(2), 157-168.

Rosenbaum, P. R. (1987b). Comparing item characteristic curves. *Psychometrika*, *52*(2), 217-233.

Rosenbaum, P. R. (1984). Testing the conditional independence and monotonicity assumptions of item response theory. *Psychometrika*, *49*(3), 425-435.

Ruscio, J., & Ruscio, A. M. (2008). Categories and dimensions advancing psychological science through the study of latent structure. *Current Directions in Psychological Science*, *17*(3), 203-207.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika monograph supplement*.

Shouse, A. W., Schweingruber, H. A., & Duschl, R. A. (Eds.). (2007). *Taking science to school: Learning and teaching science in grades K-8*. National Academies Press.

Sijtsma, K. & Hemker, B. (2000). A taxonomy of irt models for ordening persons and items using simple sum scores. *Journal of Educational and Behavioral Statistics*, 49:391- 415.

Sijtsma, K. and Junker, B. (1996). A survey of theory and methods of invariant item ordering. *British Journal of Mathematical and Statistical Psychology*, 49:79-105.

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. Oxford university press.

Stevenson, H. W., Parker, T., Wilkinson, A., Hegion, A. &Fish, E. (1976).Longitudinal study of individual differences in cognitive development and scholastic achievement. *Journal* of *Experimental Psychology, 68,* 377-400.

Swait, J., & Adamowicz, W. (2001). The influence of task complexity on consumer choice: a latent class model of decision strategy switching. *Journal of Consumer Research*, *28*(1), 135-148.

Sztajn, P., Confrey, J., Wilson, P. H., &Edgington, C. (2012). Learning Trajectory Based Instruction Toward a Theory of Teaching. Educational Researcher, 41(5), 147-156.

Uebersax, J. S., & Grove, W. M. (1990). Latent class analysis of diagnostic agreement. *Statistics in medicine*, *9*(5), 559-572.

Van Der Linden, W. J., & Hambleton, R. K. (1997). Item response theory: Brief history, common models, and extensions. In *Handbook of modern item response theory* (pp. 1-28). Springer New York.

Van de Pol, F., & Langeheine, R. (1990). Mixed Markov latent class models. *Sociological methodology*, 213-247.

Van de Pol, F., & Langeheine, R. (1989). Mixed Markov models, mover-stayer models and the EM algorithm. In *Multiway data analysis* (pp. 485-495). North-Holland Publishing Co.

Van de Pol, F., & De Leeuw, J. A. N. (1986). A latent Markov model to correct for measurement error. *Sociological Methods & Research*, *15*(1-2), 118-141.

Van Onna, M. J. H. (2002). Bayesian estimation and model selection in ordered latent class models for polytomous items. *Psychometrika*, *67*(4), 519-538.

Velicer, W. F., Martin, R. A., & Collins, L. M. (1996). Latent transition analysis for longitudinal data. *Addiction*, *91*(12s1), 197-210.

Venezky, R. L. (1999). *The American way of spelling: The structure and origins of American English orthography*. Guilford Press.

Venezky, R. L. (1970). *The structure of English orthography* (Vol. 82). Walter de Gruyter.

Vermunt, J. K., & Magidson, J. (2002). Latent class cluster analysis. *Applied latent class analysis*, *11*, 89-106.

Vermunt, J. K. (2001). The use of restricted latent class models for defining and testing nonparametric and parametric item response theory models. *Applied Psychological Measurement*, *25*(3), 283-294.

Vermunt, J. K. (1997). LEM: A general program for the analysis of categorical data. *Department of Methodology and Statistics, Tilburg University*.

Von Davier, M., Xu, X., & Carstensen, C. H. (2011). Measuring growth in a longitudinal large-scale assessment with a general latent variable model. *Psychometrika*, *76*(2), 318-336.

Wiggins, L. M. (1973). Panel analysis: Latent probability models for attitude and behavior processes. Amsterdam: Elsevier.

Willett, J. B., & Sayer, A. G. (1994). Using covariance structure analysis to detect correlates and predictors of individual change over time. *Psychological bulletin*, *116*(2), 363.

Willett, J. B. (1988). Chapter 9: Questions and answers in the measurement of change. *Review of research in education*, *15*(1), 345-422.

Wilson, M., & Draney, K. (1997). Partial credit in a developmental context: The case for adopting a mixture model approach. *Objective measurement: Theory into practice*, *4*, 333-350.

Wilson, M. (1989). Saltus: A psychometric model of discontinuity in cognitive development. *Psychological Bulletin*, *105*(2), 276.

Winsberg, S., Thissen, D., & Wainer, H. (1984). Fitting item characteristic curves with spline functions. *ETS Research Report Series*, *1984*(2).

Woods, C. M., & Thissen, D. (2006). Item response theory with estimation of the latent population distribution using spline-based densities. *Psychometrika*, 71:281-301.

Wright, B. D., & Masters, G. N. (1982). *Rating Scale Analysis. Rasch Measurement*. MESA Press, 5835 S. Kimbark Avenue, Chicago, IL 60637.

Appendix A

Estimation of the Ordered Latent Transition Analysis Model

Except for some restrictions, parameters of the OLTA model is estimated the same manner as the LTA model. Parameters in the OLTA model include level membership probability at time 1, transition probabilities from time 1 to time 2, time 2 to time 3, time 3 to time 4, and so on, and item response probabilities conditional on levels. As shown by Chung, Lanza & Loken (2008), if we let **S** = (S_1, …, S_T) be the level membership from initial time t = 1 to time T, where S_t = 1, . . . , S. Correspondingly, let $\boldsymbol{Y}_t$ = (Y_{1t}, …, Y_{Jt}) be a vector of J items measuring the level variable S_t, where each variable Y_{jt} takes values 1, . . . , r_m for t = 1, . . , T. The joint probability that the ith individual belongs to I = (s_1, . . . , s_T) and provide item responses y_{is}, …, y_{iT} would be

$$P[\boldsymbol{Y_1} = \boldsymbol{y_{i1}}, \dots, \boldsymbol{Y_T} = \boldsymbol{y_{iT}}, \boldsymbol{S} = \boldsymbol{s}] = \left[\delta_{s_1} \prod_{t=2}^{T} \tau_{s_{t|s_{t-1}}}^{(t)}\right] \times \left[\prod_{t=1}^{T} \prod_{j=1}^{J} \prod_{k=1}^{r_j} \rho_{jkt|s_t}^{I(y_{ijt}=k)}\right] \quad (1)$$

where $\delta_{s_1} = P[S_1 = s_1]$, $\tau_{s_{t|s_{t-1}}}^{(t)} = P[S_t = s_t \mid S_{t-1} = s_{t-1}]$, within each level of s_t for t = 1, . . . , T, we assumed conditionally independence for $Y_{1t}, \dots, Y_{jt}$. This assumption, referred to as local independence, enable us to make inference about the level variable (Lazarsfeld & Henry,1968).We also assume that the sequence S_t forms a first-order Markov chain for t = 2, . . . , T (Chung, Lanza & Loken, 2008).

The marginal prevalence of each level at time t (≥ 2) can be calculated as

$$\delta_{s_t}^{(t)} = P[S_t = s_t] = \sum_{s_1=1}^{S} \dots \sum_{s_{t-1}=1}^{S} \delta_{s_1} \prod_{m=2}^{t} \tau_{s_m|s_{m-1}}^{(m)}$$

From (1), the contribution of the ith individual to the likelihood function of $Y_1, \dots, Y_T$ is given by

$$P[\boldsymbol{Y_1} = \boldsymbol{y_{i1}}, \dots, \boldsymbol{Y_T} = \boldsymbol{y_{iT}}] = \sum_{s_1=1}^{S} \dots \sum_{s_T=1}^{S} P[\boldsymbol{Y_1} = \boldsymbol{y_{i1}}, \dots, \boldsymbol{Y_T} = \boldsymbol{y_{iT}}, \boldsymbol{S} = \boldsymbol{s}]. \quad (2)$$

For the purpose of simplicity, if we consider a sample of n individuals who responded to J binary items measured at two time points, we represent the likelihood function of the constrained LTA model as

$$P[\boldsymbol{Y_1} = \boldsymbol{Y_{i1}}, \boldsymbol{Y_2} = \boldsymbol{y_{i2}}] = \sum_{s_1=1}^{S} \sum_{s_2=2}^{S} \left[\delta_{s_1} \tau_{s_2|s_1} \prod_{t=1}^{2} \prod_{J=1}^{J} \prod_{k=1}^{2} \rho_{jk|s_t}^{I(y_{ijt}=k)}\right], \quad (3)$$

where $\tau_{s_2|s_1} = P[S_2 = s_2 \mid S_1 = s_1]$. In (3), the free parameters are $\theta = (\delta, \tau_1, \dots, \tau_S, \rho_1, \dots, \rho_L)$, where $\delta = (\delta_1, \dots, \delta_{S-1})$, $\tau_s = (\tau_{1|s}, \dots, \tau_{S-1|s})$ and $\rho_s = (\rho_{11|s}, \dots, \rho_{J1|s})$ for s = 1, . . . , S.

Maximum Likelihood estimates for OLTA can be estimated using an EM algorithm.

For the E-Step, we compute the conditional probability that each individual is a member of level s_1 at t = 1 and level s_2 at t = 2 given their responses $y_i = (y_{i1}, y_{i2})$ and current estimates $\hat{\theta}$ for the parameters,

$$\hat{\eta}_i(s_1, s_2) = P[S_1 = s_2,\ S_2 = s_2 \,|\boldsymbol{y_{i1}}, \boldsymbol{y_{i2}}] = \frac{\delta_{s_1}\tau_{s_{2|s_1}}\prod_t\prod_j\prod_k \rho_{jk|s_t}^{I(y_{ijt}=k)}}{\sum_{s_1}\sum_{s_2}\delta_{s_1}\tau_{s_{2|s_1}}\prod_t\prod_j\prod_k \rho_{jk|s_t}^{I(y_{ijt}=k)}}. \quad (4)$$

In the M-step, we update the parameter estimates by

$$\hat{\delta}_{s_1} = \frac{\hat{n}_{s_1}^{(1)}}{n}, \hat{\tau}_{s_{2|s_1}} = \frac{\hat{n}_{(s_1,s_2)}}{\hat{n}_{s_1}^{(1)}}, \hat{\rho}_{jk|s} = \frac{\hat{n}_{jk|s}^{(1)} + \hat{n}_{jk|s}^{(2)}}{\hat{n}_s^{(1)} + \hat{n}_s^{(2)}}, \quad (5)$$

where

$$\hat{n}_{(s_1,s_2)} = \sum_i \hat{\eta}_{i(s_1,s_2),}\ ;\ \hat{n}_{s_1}^{(1)} = \sum_{s_2} \hat{n}_{(s_1,s_2)}, \hat{n}_{s_2}^{(2)} = \sum_{s_1} \hat{n}_{(s_1,s_2)},$$

$\hat{n}_{jk|s}^{(1)} = \sum_{s2}\sum_i I(y_{ij1} = k)\hat{\eta}_{i(s_1,\,s_2)}$, and $\hat{n}_{jk|s}^{(2)} = \sum_{s1}\sum_i I(y_{ij2} = k)\hat{\eta}_{i(s_1,s_2)}$. However, the $\hat{\rho}_{(jk|s)}$ are the unconstrained response probabilities. To enforce the ordering, the constrained estimates are obtained by inputing the unconstrained probabilities, $\hat{\rho}$ into the pooled adjacent violators algorithm (PAVA; Ayer et al.,1955; Robertson, Wright & Dykstra,1988; de Leeuw,Hornik & Mair,2009), which produces the constrained probabilities $\tilde{\rho}$. Iteration occurs between the E-step and the M-step to produce sequence of parameter estimates that converges either to a local or global maximum of the likelihood function (Chung, Lanza & Loken, 2008).

Appendix B

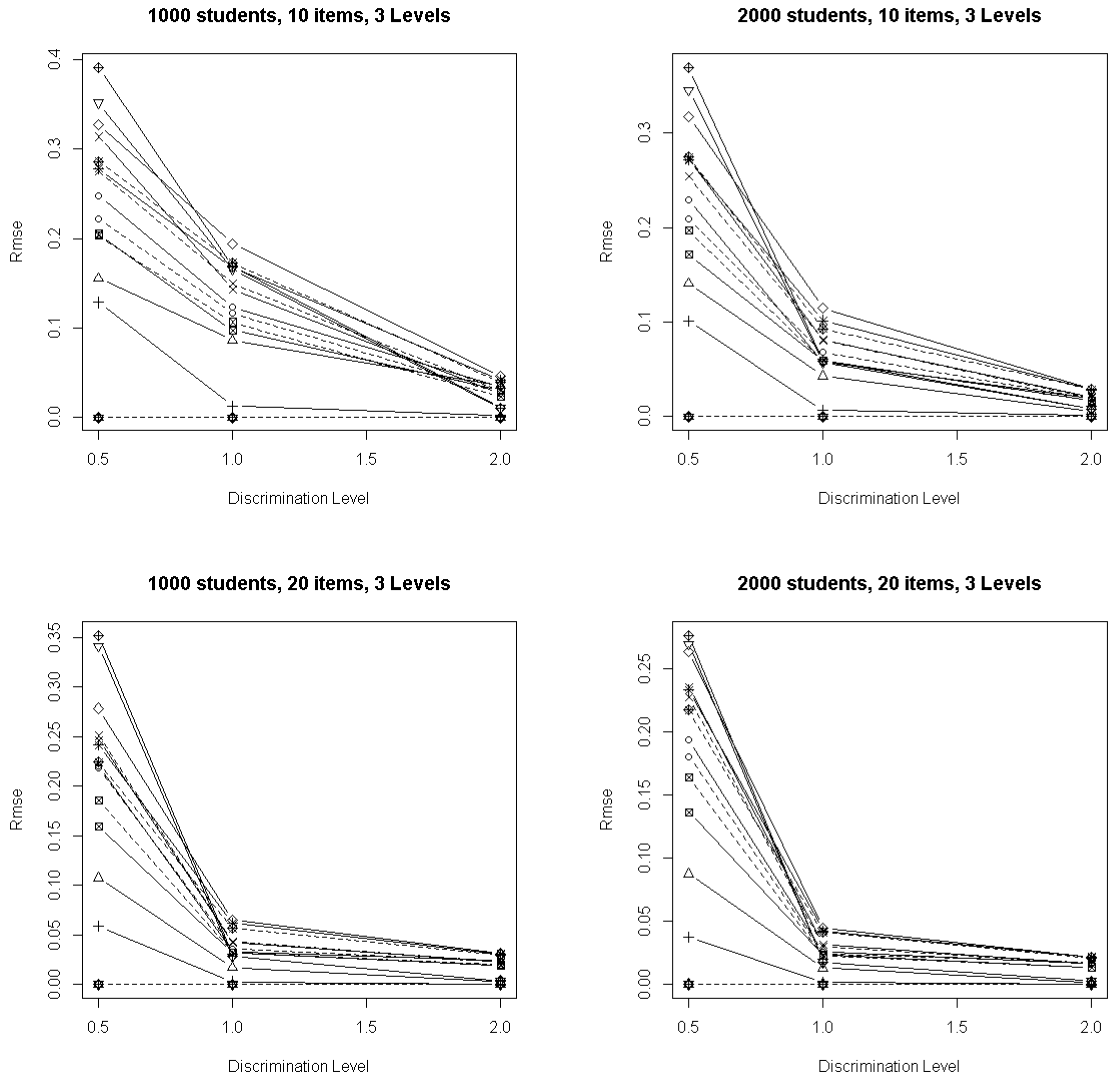


Figure B.1: RMSE for transition probabilities of a 3-level Growth model

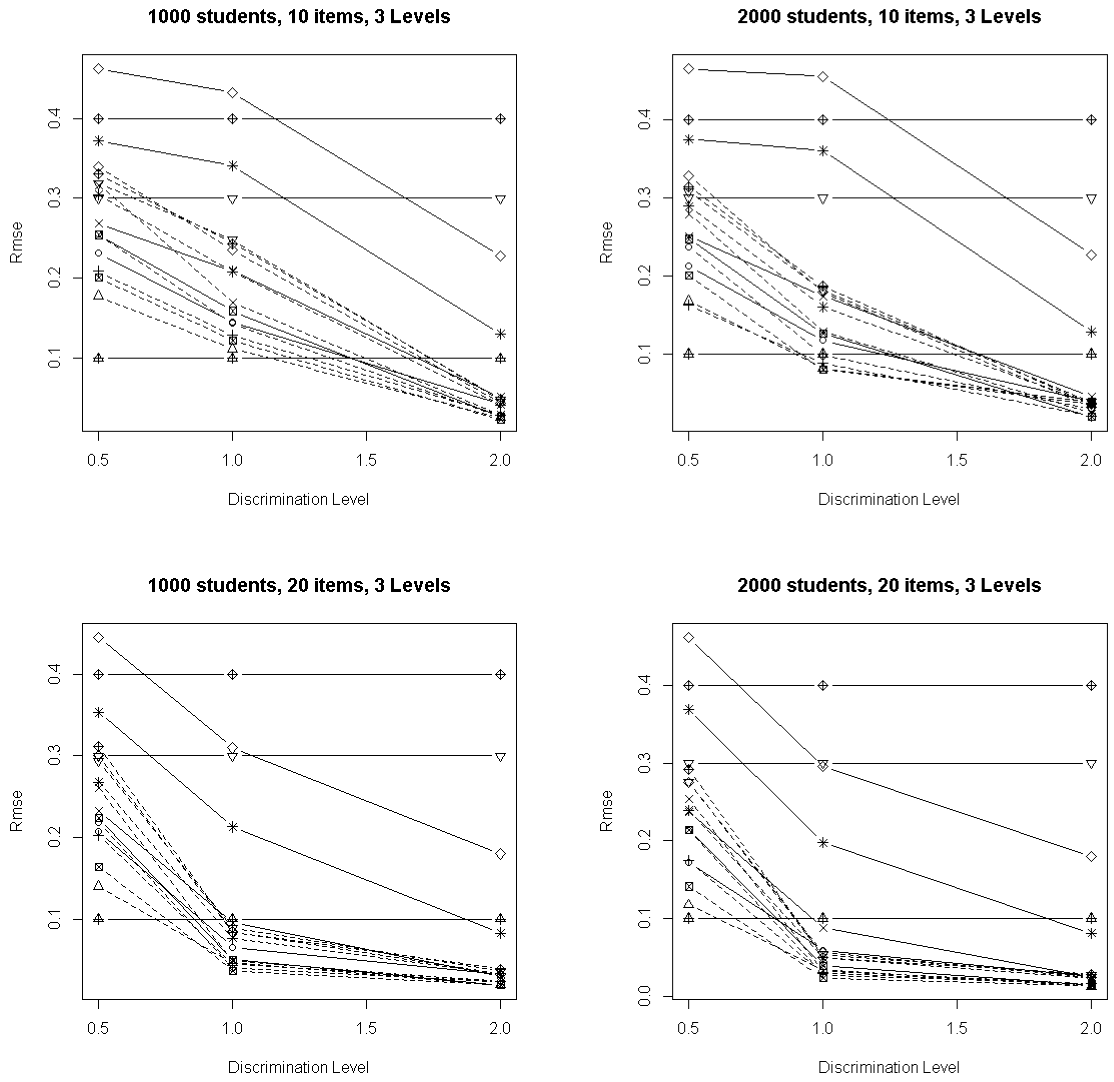


Figure B.2: RMSE for transition probabilities of a 3-level Saturated model

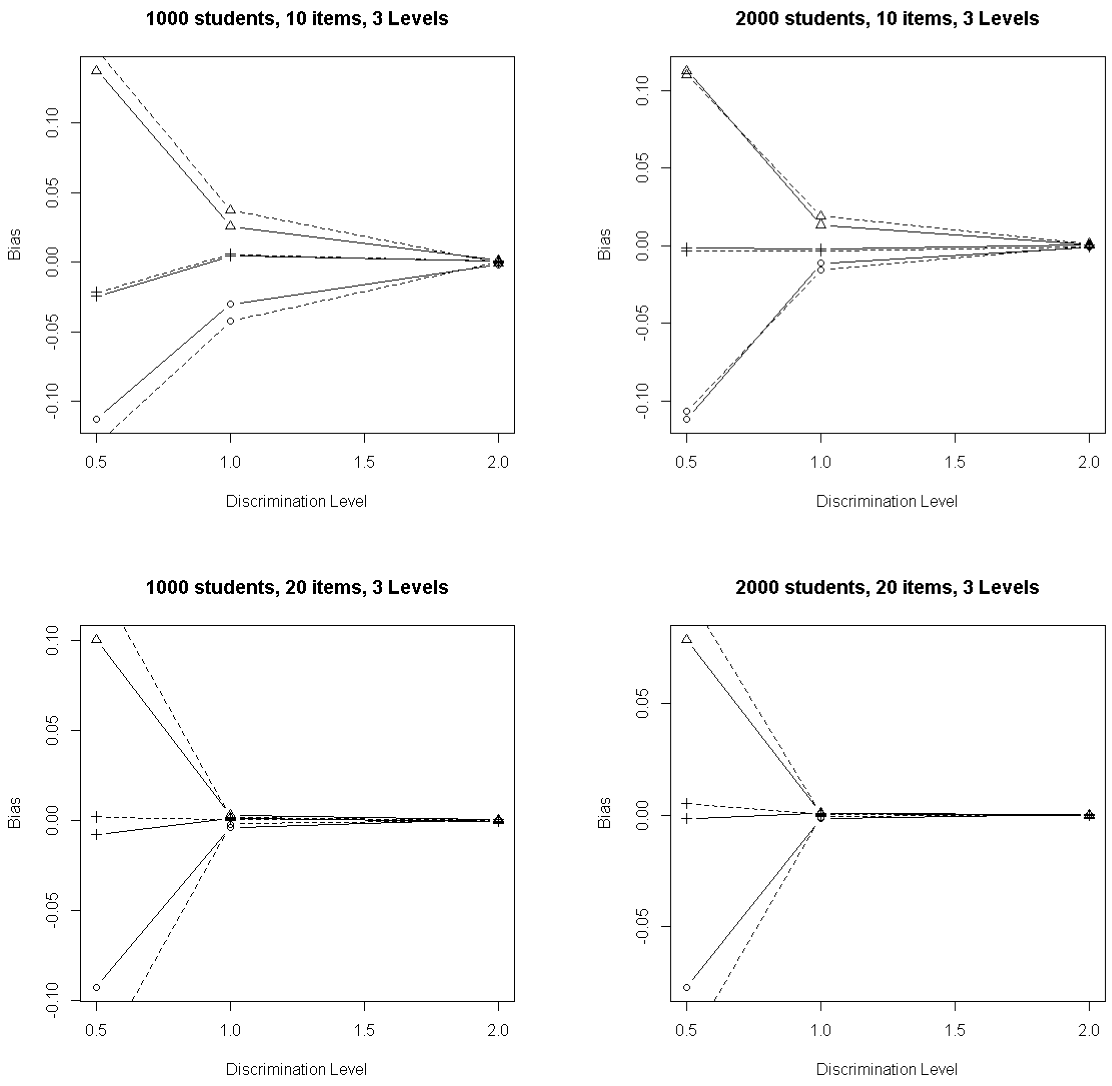


Figure B.3: Biases for learning level prevalences of a 3-level Growth model

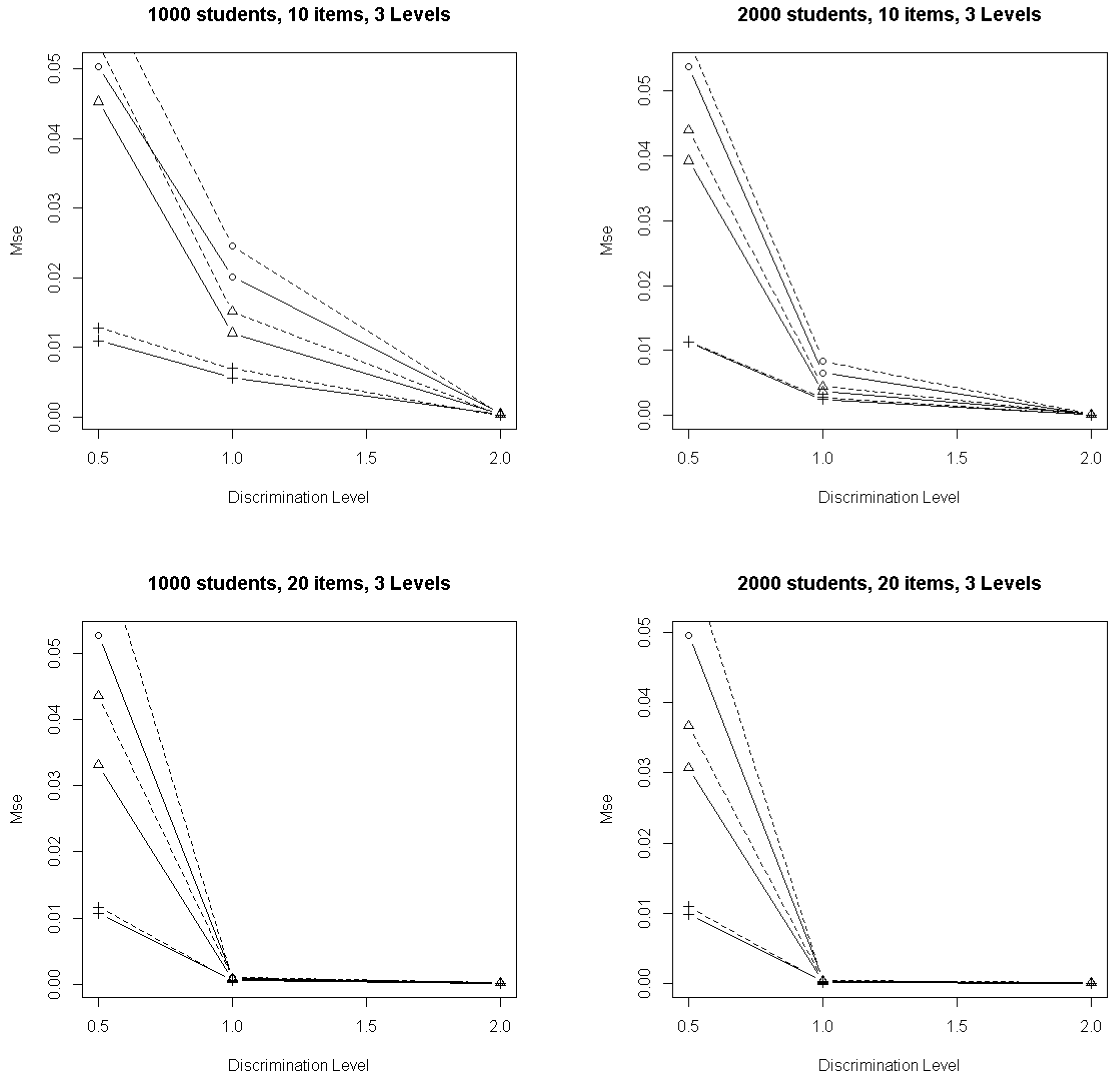


Figure B.4: MSE for learning level prevalences of a 3-level Growth model

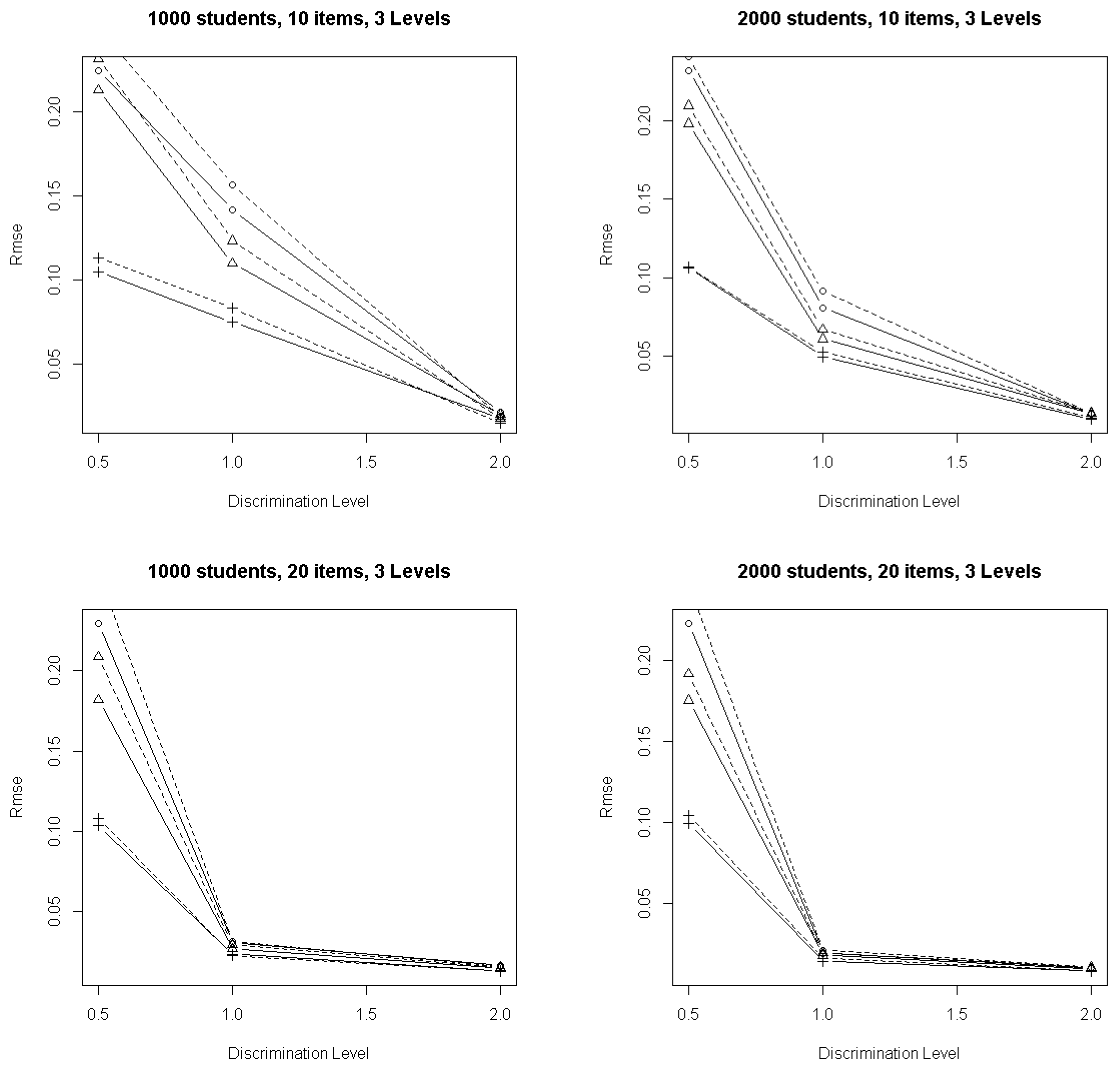


Figure B.5: RMSE for learning level prevalences of a 3-level Growth model

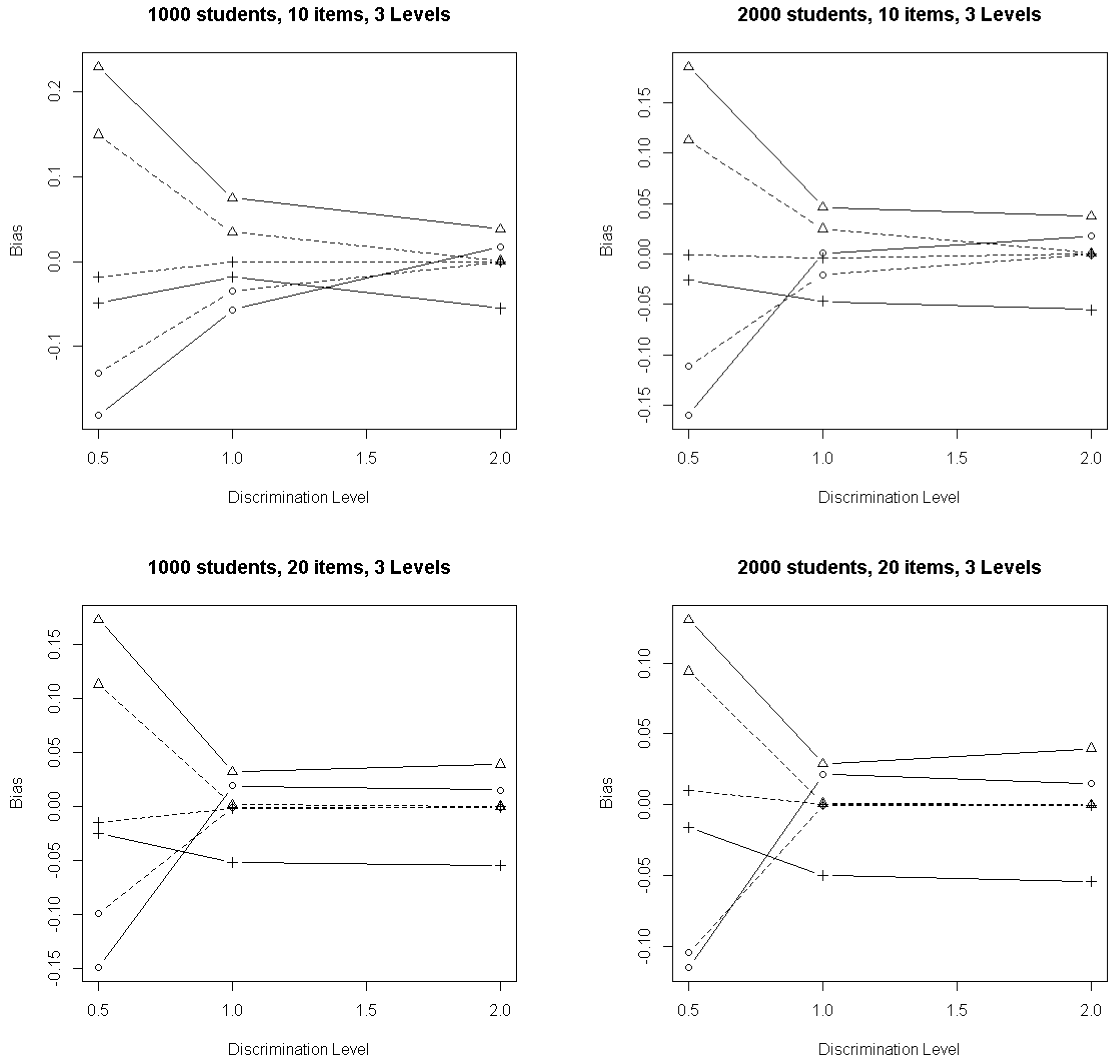


Figure B.6: Biases for learning level prevalences of a 3-level Saturated model

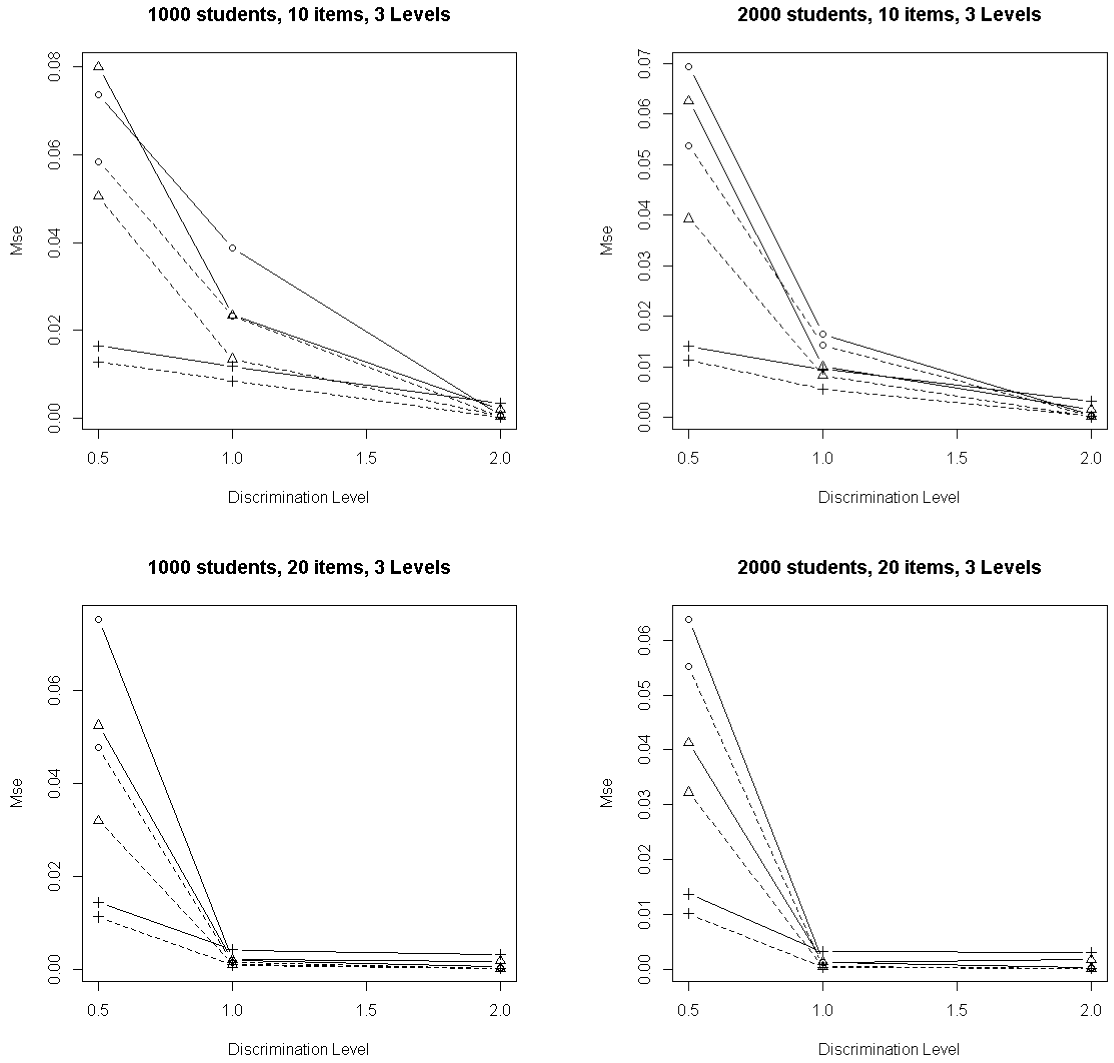


Figure B.7: MSE for learning level prevalences of a 3-level Saturated model

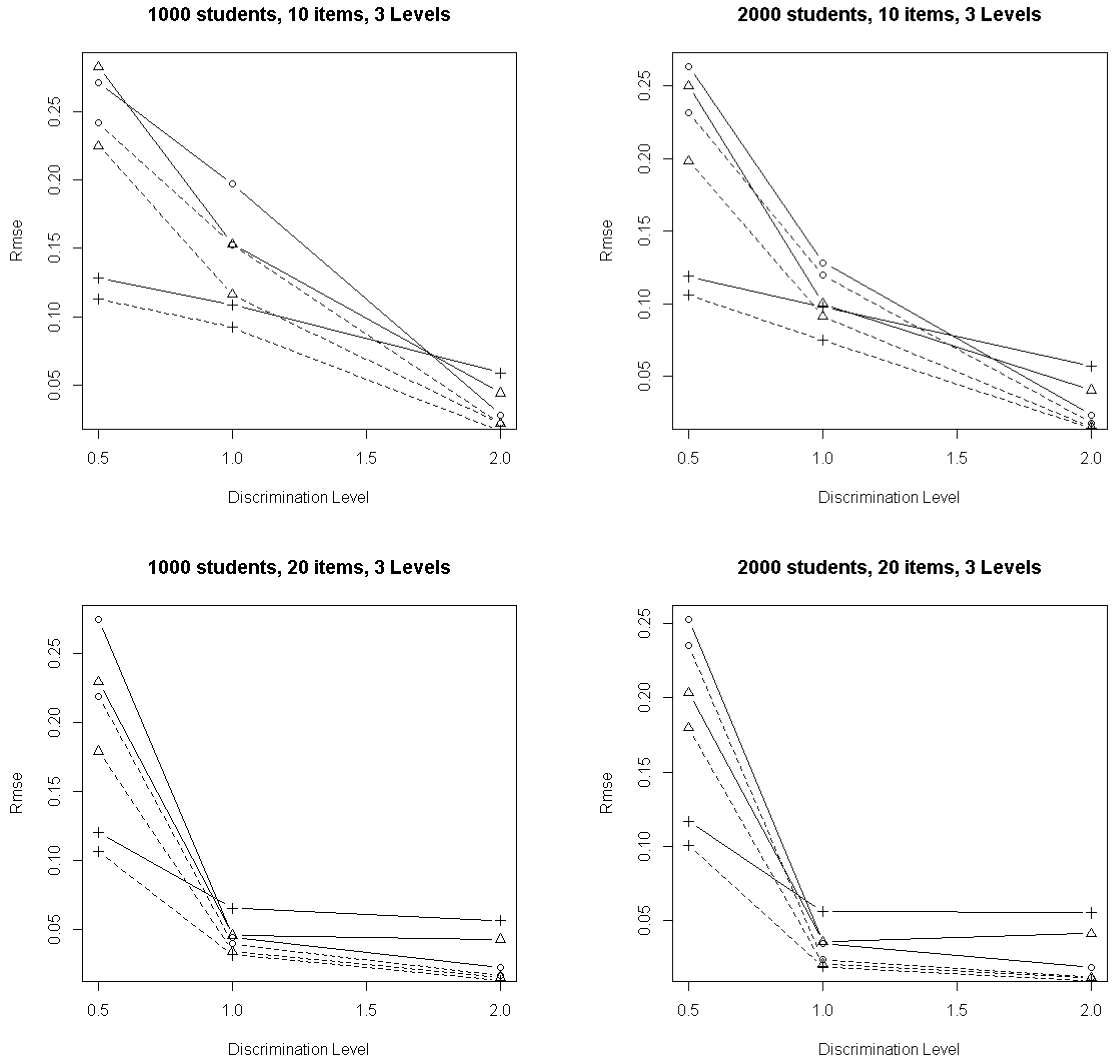
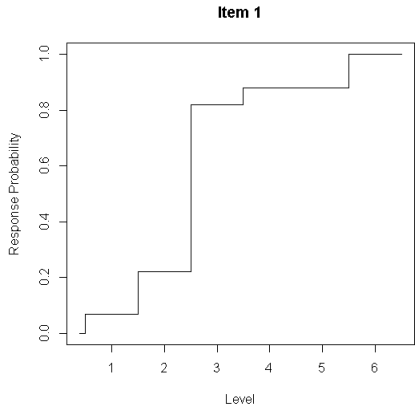
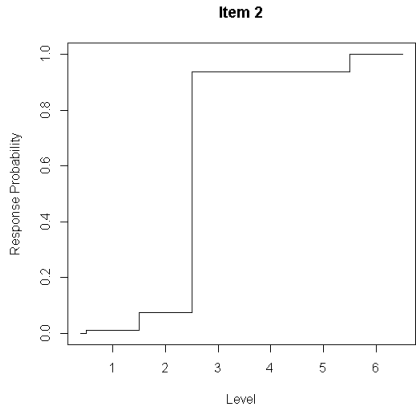
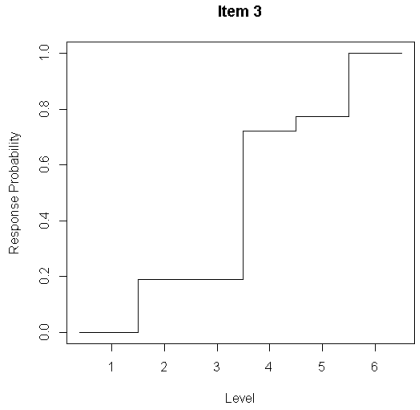
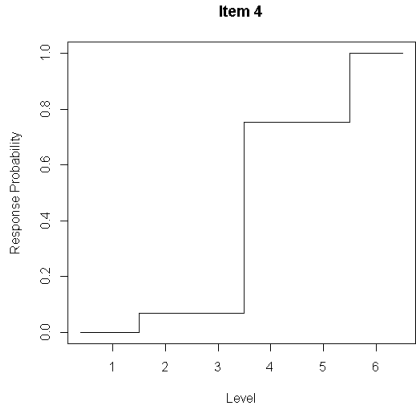


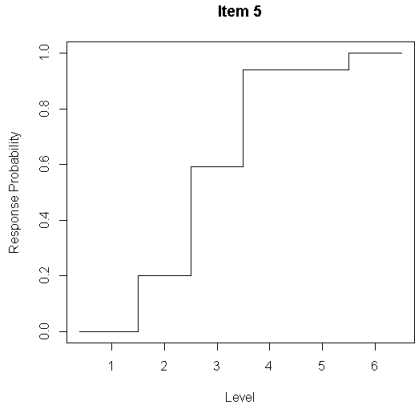
Figure B.8: RMSE for learning level prevalences of a 3-level Saturated model

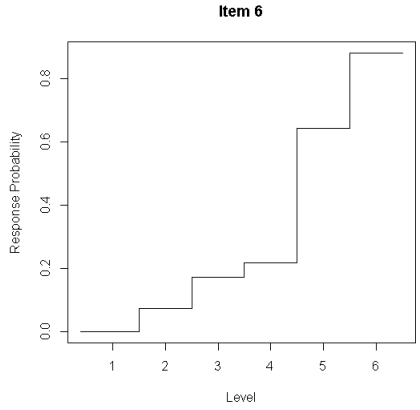

Item 1
Response Probability
0.0
0.2
0.4
0.6
0.8
1.0
1
2
3
4
5
6
Level


Item 2
Response Probability
0.0
0.2
0.4
0.6
0.8
1.0
1
2
3
4
5
6
Level


Item 3
Response Probability
0.0
0.2
0.4
0.6
0.8
1.0
1
2
3
4
5
6
Level


Item 4
Response Probability
0.0
0.2
0.4
0.6
0.8
1.0
1
2
3
4
5
6
Level


Item 5
Response Probability
0.0
0.2
0.4
0.6
0.8
1.0
1
2
3
4
5
6
Level


Item 6
Response Probability
0.0
0.2
0.4
0.6
0.8
1
2
3
4
5
6
Level

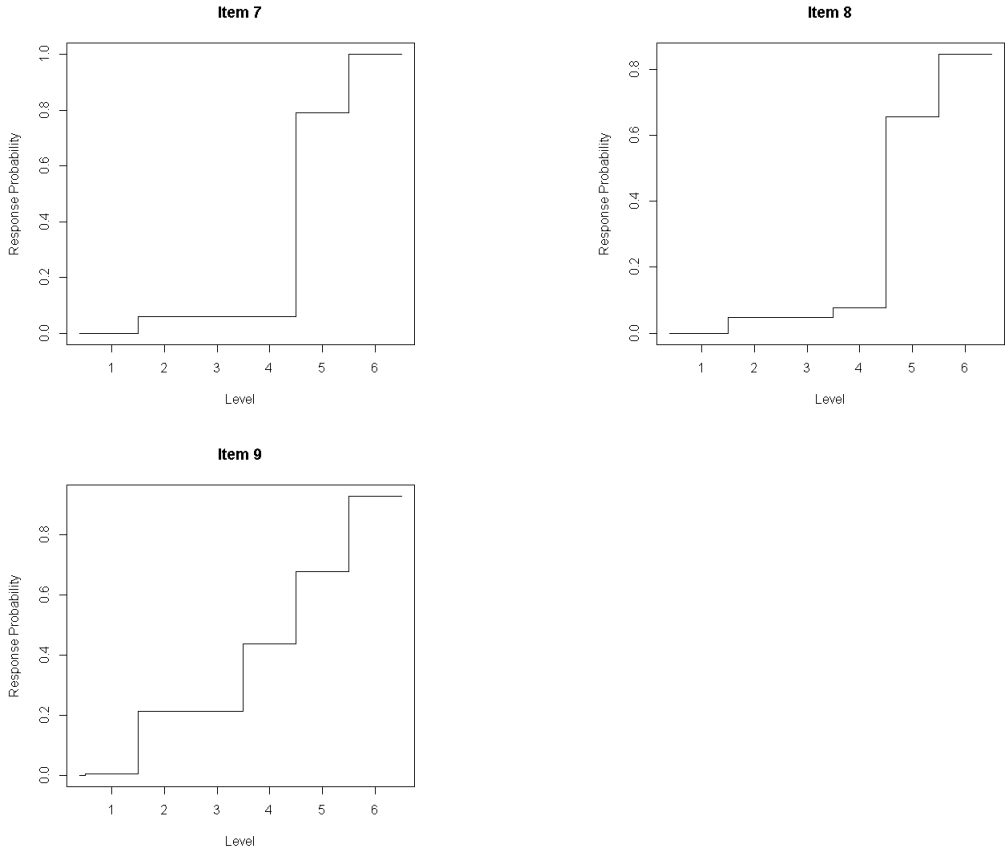
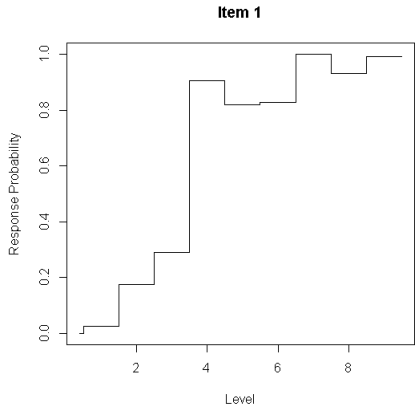
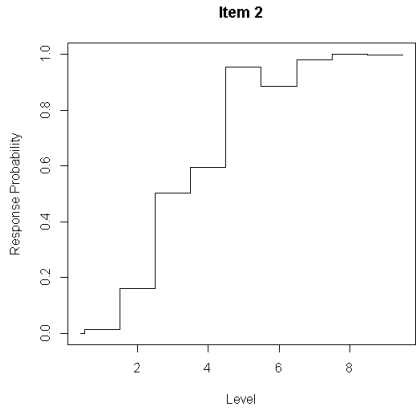
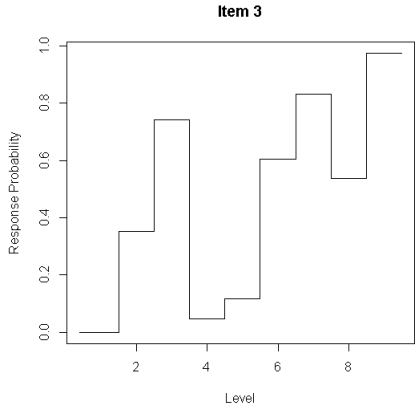
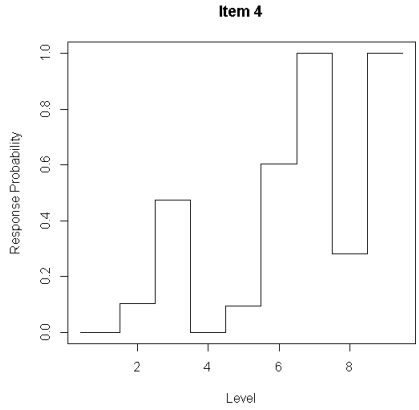


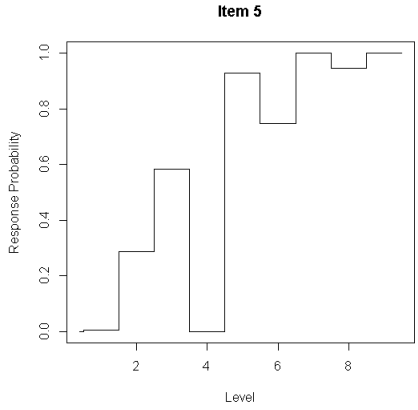
Figure B.9: Item response probabilities for 6 level saturated OLTA model

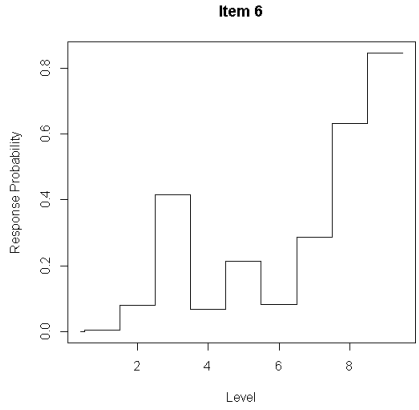

Item 1
Response Probability
0.0
0.2
0.4
0.6
0.8
1.0
2
4
6
8
Level


Item 2
Response Probability
0.0
0.2
0.4
0.6
0.8
1.0
2
4
6
8
Level


Item 3
Response Probability
0.0
0.2
0.4
0.6
0.8
1.0
2
4
6
8
Level


Item 4
Response Probability
0.0
0.2
0.4
0.6
0.8
1.0
2
4
6
8
Level


Item 5
Response Probability
0.0
0.2
0.4
0.6
0.8
1.0
2
4
6
8
Level


Item 6
Response Probability
0.0
0.2
0.4
0.6
0.8
2
4
6
8
Level

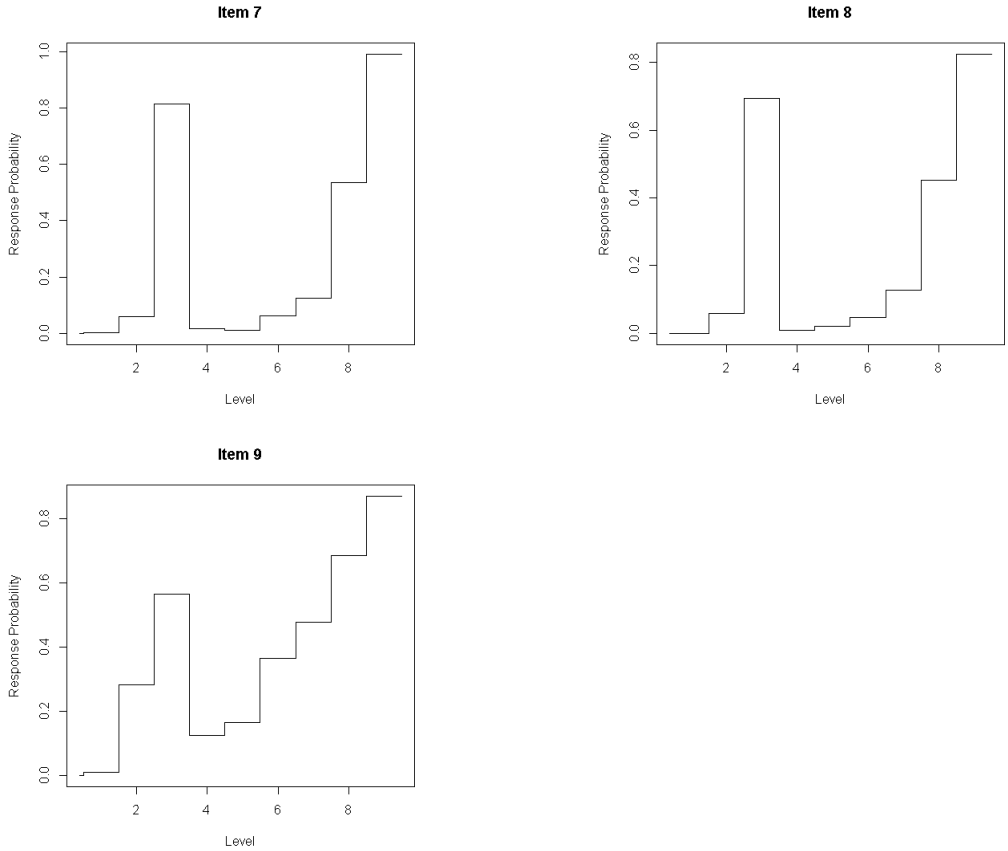


Figure B.10: Item response probabilities for 9 level saturated LTA model