Incorporating Learning Over Time into the Cognitive Assessment Framework

Thesis Proposal
Cassandra Studer

July 29, 2010

Abstract

We propose a variety of models for incorporating learning over time into the cognitive assessment modeling framework. In two of the models, we use Item Response Theory (IRT; VanDerLinden and Hambleton 1997) where we assume that a continuous latent parameter measures a student's general proficiency in the area of interest. In the other two models, we use Cognitive Diagnosis Models (CDMs; Rupp and Templin 2008) where we estimate whether students possess a set of skills as the latent student parameter. In all four models, we assume that students take multiple exams in the same content area over a period of time and that at each time point, we are interested in tracking their learning. Therefore, the models consider what the students knew at the previous time point when estimating their current knowledge. With this information, we believe we can make better predictions about end of year, high-stakes exam scores and inform teachers of areas where students are struggling. We may also be able to compare different methods of teaching to find ones that most promote learning and make some statements about the true rate and variability with which students learn which could help teachers, researchers, and policy makers set more realistic goals for students. Each model is discussed both empirically and mathematically. In a simulation study of one model, the parameters describing student learning were recovered with 94.6% accuracy.

1 Introduction

Schools throughout the country are appraised based on their students' performance on end of year accountability assessments. In fact, the No Child Left Behind Act of 2001 requires schools to meet certain criteria on state exams in order to receive federal funding (Feng et al. 2006). To prepare for these high stakes exams, students are given periodic assessments similar to the one they will see at the end of the year. These benchmark tests are one example of time points over which we may want to measure student learning. Another example is daily interaction with a cognitive tutor to prepare students for an end of unit exam. While serial data points for an individual student naturally arise in a variety of situations, most current psychometric models do not account for the learning that undoubtedly occurs; the ones that do are limited in their use (Anozie and Junker 2007).

In Section 2, we describe some potential benefits of modeling these benchmark test data longitudinally. In Section 3, we propose four models to estimate student learning over multiple testing occasions. These models are a combination of new ideas and extensions and improvements to overcome the limitations of the current techniques of modeling learning. We also present a taxonomical representation of the four proposed learning models to provide the reader with a common framework and a guide for choosing a model. Section 4 presents some preliminary results for a subset of the proposed models. This section is meant to show that our ideas are plausible and can be applied. We also provide a description of the real data set to which we intend to apply the new models. Section 5 describes the methods we intend to employ when assessing and comparing model fits. Section 6 outlines the thesis with details about the timeline that we intend to follow to complete this work. Finally, in the Appendix, we review static cognitive assessment models.

2 Motivation

In this section, we present some potential benefits of modeling student performance on multiple exams as longitudinal data.

We believe that a dynamic, rather than static, approach to tracking individual student learning over the course of a school year will be a better predictor of student performance on end of year state assessments (Anozie and Junker 2007; Ayers and Junker 2008). Because student performance on these exams is a partial determinant for a school's funding, it is imperative that schools know how students are predicted to do. With this information, teachers can devote extra attention to students who are not expected to do well and hopefully avert low scores. Furthermore, we believe our models can be used to discover particular areas upon which teachers should focus more attention. For example, we might determine that certain groups of students need more practice in the area of geometry or even more specifically that a particular student could use more practice or additional instruction in calculating the area of a circle. This type of information could help teachers make more individualized lesson plans (Carver 2001; Bransford et al. 2000).

In addition to giving more information about student learning at the individual level, we believe dynamic models can be used to make statements about student learning in general. More specifically, we may be able to compare interventions, like teaching a new curriculum for a period of time, to discover which most promotes learning (Feng et al. 2009). Instead of assessing pre- to post- test gains, we could actually compare learning rates. Additionally, the estimated learning rates may allow us to make general statements about the true rate and variability with which students learn (Koedinger et al. 2010). This could help teachers, researchers, and policy makers have more realistic expectations about the amount of gain they should expect to see from their students.

3 Dynamic Cognitive Assessment Models

3.1 Static Models

We begin the introduction of dynamic cognitive assessment models with a brief review of static models. In cognitive assessment theory, we assume that student i's performance on item j is due to a combination of his and the item's features. Student features, θ_i, are defined by a student's proficiency in a topic and/or indicators for particular skills. Item features, β_j, are defined by item difficulty, discrimination, and/or guessing and slip probabilities.

Throughout this proposal, we assume that test items are graded dichotomously where $X_{ij} = 1$ if student i answers item j correctly and 0 otherwise. This grading scheme is assumed because it is common to grade assessment items as correct or incorrect, and therefore, the situation often arises naturally. However, the theory can easily be extended to include polytomous responses (Thissen and Steinberg 1986; Hemker et al. 2001). The binary responses follow a Bernoulli distribution where $X_{ij} \sim$ Bernoulli($P(X_{ij} = 1|\theta_i, \beta_j)$). Therefore, assuming local independence (Junker and Sijtsma 2001), we can define the probability of student i's response pattern on J items to be

$$\begin{aligned} &P(X_{i1} = x_{i1}, X_{i2} = x_{i2}, ..., X_{iJ} = x_{iJ}|\theta_i, \beta_1, \beta_2, ..., \beta_J) \\ &\qquad = \prod_{j=1}^{J} P(X_{ij} = 1|\theta_i, \beta_j)^{x_{ij}}(1 - P(X_{ij} = 1|\theta_i, \beta_j))^{1-x_{ij}}. \end{aligned} \tag{1}$$

While all static models fit into this basic framework, the divergence occurs with the choice of $P(X_{ij} = 1|\theta_i, \beta_j)$. Static cognitive assessment models can be split into two categories: Cognitive Diagnosis Models (CDMs; Rupp and Templin 2008) and Item Response Theory (IRT; VanDerLinden and Hambleton 1997) models. CDMs assume the student parameter is a vector of Bernoulli random variables measuring whether a student possesses a set of K skills necessary to do well on an assessment. Therefore, the student parameter, θ_i, is a vector of length K where $\theta_{ik} = 1$ if student i possesses skill k and 0 otherwise. IRT models assume the student parameter, θ_i, is a continuous measure, usually from a normal distribution, of a student's "general propensity to do well" (Junker 1999). The Rasch model (Rasch 1960/1980; Harris 1989), a common IRT model that we will use as an example throughout this proposal, is defined as

$$P(X_{ij} = 1|\theta_i, \beta_j) = \frac{1}{1 + exp(\theta_i - \beta_j)}, \tag{2}$$

where β_j is defined to be the difficulty of item j. For more information on the Rasch and other static models, we direct the reader to the Appendix, Sections A.1 and A.2.

3.2 Common Framework

Just as static cognitive assessment models all derive from Equation 1, the dynamic models that we propose can all be condensed into a common framework. This section presents the notation, assumptions, and equation for that framework.

In order to introduce the dynamic likelihood we refer to throughout this proposal, we define $X_{it} = (X_{it1}, X_{it2}, ..., X_{itJ})$ to be the response vector where X_{itj} is the response of student i on item j at time t. Then $X_i = (X_{i1}, X_{i2}, ..., X_{iT})$ is the complete response pattern for student i. We generally assume that items are graded right or wrong for the reasons noted in Section 3.1 but add that the dynamic theory presented in this proposal could possibly be extended to polytomous response models as in Hemker et al. (2001) and Thissen and Steinberg (1986). We also assume that $\theta_i = (\theta_{i1}, \theta_{i2}, ..., \theta_{iT})$ is the vector of latent student features at each time point.

In addition to θ_i, we introduce $z_i = (z_{i1}, z_{i2}, ..., z_{iT})$, a vector of unobserved states to describe each student's status at each time point. In the four dynamic models that we propose, this latent state can be identical to θ, or an indicator of membership for latent states that describe the distribution of θ. Regardless of the definition, we assume that the relationships between student responses and latent states can be described by the Attributes Assessment Model (Junker 1999), or as it is referred to in the statistical literature, a directed acyclic graph (DAG; Wasserman 2004), which is presented in Figure 1. The conditional independences inherent in a DAG allow us to assume that observations at one time point are independent of the next given a student's latent state, i.e. $X_{it} \perp X_{it-1} | z_{it}$, and the Markov property, i.e. $z_{it+1} \perp z_{it-1}, ..., z_{i1} | z_{it}$.

Using these assumptions, we define a general marginal distribution that is used throughout our description of the dynamic learning models:

$$P(X_i, z_i) \quad = \quad P(z_{i1}) P(X_{i1} | z_{i1}, \beta) \prod_{t=1}^{T-1} P(z_{it+1} | z_{it}) P(X_{it+1} | z_{it+1}, \beta) \tag{3}$$

3.3 Taxonomy

In this section, we discuss four ways to model learning over time within the cognitive assessment framework. These models are called Kalman Filter + IRT, Parameter Driven Process for Change + IRT, Knowledge Tracing + CDM, and Parameter Driven Process for Change + CDM. While the models will be described in subsequent sections, we refer the reader to a taxonomy, Figure 2 in the Appendix, Section B. This taxonomy describes the relationship between the four models in order to construct a framework for the reader.

Succinctly, the two models in the upper half of the figure extend IRT models while the two models in the lower half extend CDMs. The models on the left half track students learning over time by allowing the student parameter, θ, to change at each time point. The models on the right half track learning indirectly by instead tracking student membership in latent states which drive the distributions of the student parameter, θ. We will expand upon these ideas in Sections 3.4 and 3.5 where we also point the reader to the appropriate section of Figure 2.

3.4 Dynamic IRT Models

In this section, we present two ways to include time into a model that assumes a continuous parameter, θ, to describe student proficiency. The first will apply an extended Kalman Filter (KF; Dethlefsen and Lundbye-Christensen 2006) to θ with the aim of tracing the value of this estimate over time to track student learning. The second model is derived from the parameter driven process for change (PDPC; Rijmen et al. 2005; 2008) model. In this model, we group students based on similar response patterns where each group has a different true distribution for θ. Then by tracing the students' path through the latent knowledge state space, we can track learning. In this section, we describe these ideas in more detail.

3.4.1 KF + IRT

In IRT, to evaluate a student's proficiency in a given subject, we simply look at the estimate for his latent variable, θ. Then, one idea for incorporating time into an IRT model is to directly trace a student's value of θ over time. If there were T benchmark tests given over the course of a year, we would find a value for θ, which is identical to z in this case, at each of the T time points. In this case, we could track learning gains and losses through the change in θ.

One approach for estimating a new θ at each time point would be to fit a separate IRT model. However, with this method, we would ignore what we previously knew about the student because we would not be accounting for the value of θ at the previous time points. Therefore, we suggest fitting an IRT model at each time point with an extended Kalman filter (Dethlefsen and Lundbye-Christensen 2006) to account for previous estimates of the students' abilities. We call this model the Kalman Filter + IRT (KF + IRT) model and depict it in the upper left corner of Figure 2.

The Kalman filter is a version of the hidden Markov model where the latent state is considered to be a continuous variable. The latent state of the system, θ, is represented by a real number which is adjusted by a linear operator plus some Gaussian noise at each time point. More specifically, if θ_t is the true state at time t and X_t is an observed response, we have that

$$\begin{aligned} X_t &= C_t\theta_t + \nu_t && (4)\\ \theta_t &= A_t\theta_{t-1} + \delta_t, && (5) \end{aligned}$$

where ν_t and δ_t are process and observation noise, respectively, each from a multivariate normal distribution with mean zero and estimated covariance matrices. Therefore, Equation 4 is static and connects the latent state, θ_t, to the observed responses, X_t. Equation 5 is dynamic and describes the relationship between the latent state and its lagged values (Oud et al. 1999).

With these definitions, we can see that the Kalman filter is a natural way to trace the continuous student ability parameter, θ_i, in a way that is analogous to a hidden Markov model. However, we propose a few changes in order to adapt the Kalman filter to an education context. One change that we propose is to use the extended Kalman filter (Durbin and Koopman 2000) which uses a generalized linear model to describe the relationship between X_{itj} and θ_{it}. This is necessary because in IRT, we are interested in estimating the probability of X_{itj} as a function of θ_{it} and the item features, β_j; therefore, we assume a logit link, which naturally confines the probability to be between 0 and 1. For example, in the Rasch model case, Equation 4 becomes

$$logit(P(X_{itj}|\theta_{it}, \beta_j)) = \theta_{it} - \beta_j. \tag{6}$$

Additionally, we propose changing the relationship between θ_t and θ_{t-1} to be additive at each time increment instead of multiplicative. Then, equation 5 becomes

$$\theta_{it} = \theta_{it-1} + \delta_{it}, \tag{7}$$

where $\delta_{it} \sim N(\alpha, \sigma_\alpha^2)$ is a random effect describing student i's change in proficiency from time $t-1$ to t. Because the relationship between θs at successive time points is additive, the value of δ_{it} is a measure of student i's learning from time $t-1$ to time t. To put the change in perspective, we look at the estimates of α and σ_α which describe the average learning rate and variability of learning for all students in the sample. These estimates are also of interest because in order to compare different methods of teaching and find the one that most promotes learning, we would simply look for the higher average rate of change, α.

The common equation for dynamic learning models, Equation 3, can be rewritten for the KF + IRT model, where $z_{it} = \theta_{it}$, as

$$P(\theta_{i1})P(X_{i1}|\theta_{i1}, \beta) \prod_{t=1}^{T-1} P(\theta_{it+1}|\theta_{it})P(X_{it+1}|\theta_{it+1}, \beta). \tag{8}$$

where we assume $\theta_{i1} \sim N(0,1)$ as is common in the static model. In this way, positive θs are still above average and negative ones are below average. As in Equation 1, $P(X_{it}|\theta_{it}, \beta) = \prod_{j=1}^{J} P(X_{itj}|\theta_{it}, \beta_j)$ using the standard assumption of local independence (e.g. Junker and Sijtsma 2001). Also, while we used the Rasch model to exemplify $P(X_{itj}|\theta_{it}, \beta_j)$, it is also be possible to use any IRT model described in Section A.1. Finally, as described in this section, we assume $\theta_{it+1} = \theta_{it} + \delta_{it}$ where $\delta_{it} \sim N(\alpha, \sigma_\alpha^2)$ describes the extended Kalman Filter that we apply to the latent student cognitive ability parameter.

3.4.2 PDPC + IRT

It is possible that KF + IRT may require estimating too many parameters, be computationally infeasible, or that we would lack the data to support watching individuals move through the latent space. Alternatively, a researcher may not want information about individual students, but only about groups of students for purposes like monitoring teacher performance or developing more focused instruction. In these scenarios, a form of the PDPC model (Rijmen et al. 2005; 2008), called PDPC + IRT and depicted in the upper right corner of Figure 2, may be more appropriate.

In PDPC + IRT, we define Z latent states to describe groups of students with similar response patterns. Then, at each time point, we assume that students transition between these states according to a time homogeneous hidden Markov model and estimate the latent state membership, z_{it}. The distribution of θ_{it} is then dependent on student i's latent knowledge state at time t.

More specifically, we assume that $\theta|z \sim N(\mu_z, \sigma_z^2)$. Then knowing μ_z and σ_z for the Z possible states and a student's trajectory through the latent state space, z_i, as well as the posterior probabilities of being in each of the other latent states, allows us to track that student's learning. By clustering students into groups, we can visualize student learning using the parameters of the θ distribution as opposed to estimating a θ for each individual student as in KF + IRT. This simplifies estimation.

Alternatively, if our goal is to compare learning in different curricula or after an intervention, we would be interested in comparing the transition matrices of the latent knowledge states from the time homogeneous hidden Markov model. The matrix with the higher probability of moving to a more proficient state can be assumed to be the one that most promotes learning.

In PDPC + IRT, we assume that the clusters describing student ability are fixed over the T time points and students transition between them over time. Another method of defining the clusters would be to use trajectory clustering where students who learn similarly over the period of time are grouped together (Stallard 2007; Manrique 2009; Connor 2006; Roeder et al. 1999). While we think this is an important option to explore, it is beyond the scope of this thesis.

The main difference between PDPC + IRT and the PDPC model is that we are interested in watching students move through the latent space while other parameters remain fixed. There is a school of thought, which includes PDPC, that follows student learning by allowing the item features to change over time (Rijmen et al. 2005; 2008; Draney et al. 1995; Pavlik et al. 2009; Cen et al. 2006). We require that β_j remain constant because our main goal is to compare the parameters of the student ability distribution. Therefore, we do not allow any of the variability in latent states to be absorbed by the item parameter assuring that it is contained in the student ability distribution. Another difference is that we do not assume a particular IRT model. Any IRT model can be incorporated into this framework with the constraint that item parameters are constant over time.

The likelihood function for PDPC + IRT is exactly Equation 3 where z_{i1} is the initial state and $P(z_{it+1}|z_{it})$ is the transition probability in a Markov chain. $P(X_{it}|z_{it}, \beta) = \int_\theta P(X_{it}|\theta_{it}, \beta_j)P(\theta_{it}|z_{it})$ where we use the law of total probability to expand the responses to be dependent on θ. By the local independence assumption, $P(X_{it}|\theta_{it}, \beta_j) = \prod_{j=1}^{J} P(X_{itj}|\theta_{it}, \beta_j)$, where the probability of a successful response is any of the static IRT models described in Section A.1. As described in this section, we assume $\theta_{it}|z_{it} \sim N(\mu_z, \sigma_z^2)$.

3.5 Dynamic CDM Models

The dynamic models we have presented thus far assume the latent student parameter is a continuous variable describing a student's overall proficiency or in other words, that they fit into the IRT framework. However, similar approaches can be taken if we are interested in estimating whether students have mastered a set of skills. In static CDMs, presented in detail in the Appendix, Section A.2, we typically want to estimate θ as a vector of K binary latent variables, defined as the presence or absence of individual skills which the exam is covering. We present two ways of incorporating time into such a model. The first is an extension of Knowledge Tracing (KT; Corbett et al. 1995), a method of directly tracing students learning and forgetting skills over time. The second approach uses the PDPC method as its base. This section will describe these ideas in more detail.

3.5.1 KT + CDM

In the KT + CDM model, depicted in the lower left corner of Figure 2, we assume that skills are assigned to items before an exam is given and the goal is to estimate whether students have mastered each skill of interest at each time point based both on how they respond to items and whether they had the skill at the previous time point. Therefore, at each time point, we assume that every skill is allowed to transition between mastered and not mastered independently of the other skills according to a time homogeneous hidden Markov model. We are willing to relax these assumptions as we learn to work with and draw inferences from the complex models.

In this scenario, as introduced in Section 3.2, the latent states, z, are identical to the vector of skills, θ. Then, to define the marginal likelihood, we start with the KF + IRT likelihood, Equation 8 where now $\theta_{it} = (\theta_{it1}, \theta_{it2}, ..., \theta_{itK})$ and let $\theta_{i1} = \prod_{k=1}^{K} P(\theta_{i1k})$ be the product of K initial state probabilities in a Markov chain where we assume each $\theta_{i1k} \sim Bern(p_k)$, just as in a static model. Similarly, $P(\theta_{it+1}|\theta_{it}) = \prod_{k=1}^{K} P(\theta_{it+1k}|\theta_{itk})$ is the product of K transition probabilities from the Markov chain describing the acquisition of skills. Finally, as in Equation 1, $P(X_{it}|\theta_{it}) = \prod_{j=1}^{J} P(X_{itj}|\theta_{it})$, where $P(X_{itj}|\theta_{it})$ is defined to be any of the CDM models described in Section A.2.

With this KT + CDM model, we can follow students learning and forgetting skills over time. Furthermore, we intend to accommodate the likely scenario that skills are not learned independently using a model like that presented by (DeLaTorre and Douglas 2004). In essence, they use a hierarchical IRT model to describe the dependencies between skills. Also, in order to compare different methods of teaching, one could estimate and compare different transition matrices for the acquisition of skills for the different methods.

The KT + CDM model differs from the KT model (Corbett et al. 1995) in that KT was developed to estimate skill mastery in a particular cognitive tutor setting where the items were designed such that each entry the user made was in reference to one particular skill and the probability that the user had that skill was updated immediately. KT + CDM expands this model to incorporate the more common situation of multiple skill items on multiple item exams. Furthermore, KT + CDM allows us to use any CDM whereas current KT is restricted to the NIDA model.

With KT + CDM, we can track learning in a situation that does not depend on an update after every student entry. It allows for paper and pencil benchmark tests which may be more realistic in the classroom but retains the interpretability of the KT model. If a teacher gives her class benchmark exams and wants to see how they are learning, using the KT + CDM algorithm will allow us to tell her specifically which skills each student has learned at each time point.

3.5.2 PDPC + CDM

Just as in IRT, it is possible that the KT + CDM method may become computationally infeasible, or that there are not enough data to support the method because individual skills are not attempted often enough or there are many skills to be estimated. Alternatively, a more parsimonious model with information at a coarser grain size than knowing whether each student has each skill at every time point may be desired or may be a better fit, as discussed in Section 5. In this scenario, we may resort to an adaptation of the PDPC model for CDMs. We call this model PDPC + CDM and depict it in the lower right corner of Figure 2.

Like PDPC + IRT, we cluster students into latent states, z, which can be thought of as groups of students with similar skill or response patterns. Students are then allowed to transition between these latent states according to a time homogeneous hidden Markov model. In order to track students learning over time, we follow the trajectory of their latent states, the posterior probabilities of being in each of the other latent states at each time point, and the parameters of the latent state distributions. In a scenario with multiple skills, the number of latent states would not necessarily be equal to 2^K, the number of skill patterns, but for each latent state, we would be interested in the probability of having each of the K skiills. In this situation, the probability of knowing some skills may increase while others may decrease as students transition between states.

Just as in PDPC + IRT, we add the constraint that any parameter in the CDM that describes the item (or skill as in the NIDA model) as opposed to the student must be independent of both time and the latent state. This allows us to make comparisons of the distributions associated with those latent states. The likelihood for PDPC + CDM is exactly the same as that of PDPC + IRT where we assume $P(X_{itj}|\theta_{it}, \beta_j)$ is a CDM instead of an IRT model. We also assume that

$\theta_{it}|z_{it} \sim \prod_{k=1}^{K} Bern(p_{kz})$, where p_{kz} is the probability of having skill k for students in state z. In this model, we assume that skills are either known or unknown.

4 Preliminary Results

In order to illustrate these methods of incorporating learning into the cognitive assessment framework, we have simulated some data. In this section, we describe one data set and present some preliminary results when applying the PDPC + IRT model as the model of choice.

4.1 Simulated Data

In this section, we describe the data set that we simulated.

1. We chose to simulate data for $1,000$ students answering a 15 item test at 5 time points.

2. We randomly assigned students to the initial states with equal probability.

    - We chose 2 latent knowledge states.

3. We generated the path each student took through the latent knowledge state space.

    - We defined the transition matrix to be $P = \left[\begin{smallmatrix} 0.75 & 0.25 \\ 0.25 & 0.75 \end{smallmatrix}\right]$

4. We generated θs according to the latent class students belonged to at each time point.

    - We set $\mu_1 = -1$, $\mu_2 = 1$, and $\sigma_1 = \sigma_2 = 0.25$.
    - We chose these means for simple interpretation. We think of each state as a group of students who does not do very well (μ = -1) and a group of students who does do relatively well (μ = 1). We call these group 1 and group 2, respectively.
    - While it is not necessary for the standard deviations to be equal, we do feel it is going to be important for them to be small so that the states are separable. If the states are not separable, estimation of state membership will be difficult because in reality, a student could belong to either state. However, more exploration will better inform us as to what constraints need to be placed on these values.

5. We generated student responses.

    - We assumed a Rasch model.
    - The difficulties of the items were taken to be a sequence of 15 numbers evenly distributed between -2 and 2. This ensures that we have items able to differentiate between students of most possible abilities (given our θs come from normal distributions with means -1 and 1 and standard deviations 0.25).

The goal now is to retrace these steps and estimate each student's path over the five time points and the parameters of the two θ distributions. To do this, we will use Markov Chain Monte Carlo (MCMC) methods and the statistical program WinBUGS (Lunn et al. 2000).

4.2 Model

In order to pilot the PDPC + IRT method, we employed the following model:

$$X_{itj} \sim Bernoulli(p_{ijt}),$$

where

$logit(p_{ijt}) = \theta_{it} - \beta_j$
$\theta_{it}|z_{it} = z \sim \text{Normal}(\mu_z, 0.25)$
$\mu_z \sim \text{Normal}(0, 1)$
$z_{i1} \sim \text{Categorical}(K, \pi_0)$
$z_{it} \sim \text{Categorical}(K, P[z_{it-1},])$
$K = 2$
π_0 is a vector of length K where each entry is $\frac{1}{K}$
P is the known transition matrix from the simulation
β is the known vector of item difficulties from the simulation

In this model, we assumed many parameters were known. We did this in order to maintain simplicity in the initial stages of developing the model. Specifically, we allowed the number of latent states, the transition matrix and initial probabilities of being in each state, the standard deviations of the θ distributions, and the item difficulties to be deterministic. As we get more comfortable with the model, we will start to estimate these values as well. We did, however, estimate the path each student took through the latent state space and his proficiency, θ, at each time point. In addition, we estimated the means of the two θ distributions.

4.3 Results

Using WinBUGS, we ran 5 chains with different starting points for $6,000$ iterations, a burn-in of 600, and thinning of 27. This resulted in 200 simulated values which we use to calculate sample statistics. Specifically, we report the median values pooled across the five chains. In this section, we discuss the Gibbs sampler, compare the results to the "true," simulated values, and discuss problems that we encountered.

Because this is a parameter driven process for change model, we are particularly interested in whether we were able to recapture the means of the $\theta|z = 1$ and $\theta|z = 2$ distributions. In fact, we find that the estimated values (with 95% credible intervals) are very close to the true values of $\mu_1 = -1$ and $\mu_2 = 1$ with $\hat{\mu_1} = -1.001$ $(-1.029, -0.971)$ and $\hat{\mu_2} = 1.002$ $(-0.973, 1.029)$. The $\hat{R}$ for these estimates are 1.10 and 1.06, respectively. Figure 3, along with the fact that $\hat{R} \leq 1.1$

(Gelman et al. 2004) for each value, show that the five simulated sequences adequately converged. Therefore, we were able to recapture these parameter values within rounding to the one-hundredths place. The narrow credible intervals, approximately 0.06 in width, also indicate that the recovery was accurate and successful.

Next, we consider the recovery of the latent states, z_{it}. Because there are only two latent states, the median value that we use as the result corresponds to the latent state that is frequented most by the MCMC chains. If we ignore time and look at the $N \cdot T = 5,000$ latent states that we are trying to estimate, we find that $4,729/5,000$ (94.6%) estimated latent states matched the students' states in the simulated data set. Therefore, students were generally classified to their true state. Random chance would have classified approximately $2,500$ states correctly so we are doing 44.6% better.

We also found that only $228/1,000$ (22.8%) students were assigned to the wrong state at an average of 1.19 (sd = 0.43) time points. Therefore, $772/1,000$ (77.2%) students were placed in the correct state at all 5 time points. Even if a student is placed in an incorrect state at one time point, his states at other time points are still likely to be correctly determined. With no estimation and given our transition matrix, the best we could do is use the maximum likelihood trajectory where we assume that a student stays in the same state at all time points. This would correspond to putting approximately 16% of students into the correct latent state at all five time points. We achieve 77.2%. Therefore, we are estimating correct trajectories for 61.2% more students.

Figure 4 shows the simulated chains for 3 randomly selected students and time points with added noise so that the chains were not completely overlapping. We see that for these three students, the five chains agreed on latent state placement at each iteration. While this is the case for the majority of student time points $(3,224/5,000)$, there are some cases where the chain switches from one state to the other at least once.

Figure 5 shows an example, for student 13 at time 2, where the five chains do not agree. We claim that this discrepancy is a good result because if we look up this student's response pattern at time 2, we find he successfully answered the eight items with difficulty less than or equal to 0 and incorrectly answered the seven items with difficulty greater than 0. This would lead us to believe that his θ value is approximately 0 instead of 0.69 as we know from the simulated data set. Therefore, because we do not have a latent state corresponding to a θ of 0, it makes sense that the five chains disagree. In fact, if we average the θ estimates across chains, we indeed get a mean of 0.08 which more closely matches that student's data. Therefore, our estimation process is suffering because we do not have a latent state in which he would belong. This seems to be the case in many of the cases where there are switches from one state to another. It is possible that once we relax the assumption that we know the number of latent states, this issue will automatically alleviate itself. Otherwise, we will need to find a way to better define the number of latent states for WinBUGS and be amenable to changing that after some exploratory work.

Another possible explanation for the chains to be switching state assignments is the label switching phenomenon (Jasra et al. 2006). While we do not believe that possibility to be likely at this time since our assumptions are so strict, we do believe it is going to be an issue in the future.

Finally, we do not present many results about the latent student parameter, θ. This is done purposely for two reasons. First, the focus of the PDPC + IRT model is not on the individual students' θs but on their trajectories through the latent state space and the distributions describing those states. Second, we do not present those results because, as we just described with the student who has a "true" $\theta = 0.69$ but behavior of a student with $\theta = 0$, there is enough discrepancy between the "true" θ and the corresponding response patterns that comparing the θs is not particularly interesting. In the future, we may try to simulate a data set where the response patterns match more closely the true θ. More items would be a good method of more accurately estimating θ.

We take these preliminary results about the recovery of the distributional means and latent trajectories as good evidence that our methods are feasible. Obviously, there is still a lot that we must do including trying the other three models, making the PDPC + IRT model less strict (because we assumed so many parameters were known), making comparisons to the baseline model, and applying the models to real data sets. We look forward to continuing to work with these methods.

4.4 Real Data

To pilot our models in a real scenario, we may use a subset of data from the Assistment System on line mathematics tutor (Heffernan et al. 2001). The goal of the Assistment System is to prepare students for the mathematics portion of the Massachusetts Comprehensive Assessment System (MCAS), Massachusetts' end of year exam meant to satisfy the requirements of NCLB (Ayers and Junker 2008). In the Assistment System, students are given items in a cognitive tutor setting over the course of the academic year. We will be interested in looking at a subset of these items that focus on a particular area such as geometry. Then we can apply our methods and track student learning in that subject over the course of the year. We also have access to the student's MCAS scores and can compare predictions using our methods to predictions using a static model. This is important because it is one of the main methods of model fit that we intend to employ as described in Section 5. Another nice aspect of the Assistment data is that it is coded as individual skills so we can not only try the IRT models but the CDM ones as well.

Other options for real data are cognitive tutor data sets found in the PSLC DataShop (Koedinger et al. 2008). These are desirable because there are often more items covering a more narrow range of topics. Therefore, we would expect better estimates from the dense plethora of data than we would find in the Assistment system. Unfortunately, we may not be able to find a data set that covers more than a few days' time or that has a posttest as natural as the MCAS exam. Therefore, we may use one or both types of data sets throughout our analysis of the learning models.

5 Model Comparison

In this section, we describe the methods we will employ to assess the fit of our models on real data.

Ayers and Junker (2008) found they could better predict students' end of year MCAS exam scores using a student's cognitive ability estimate, θ, as opposed to the number of correct items a student completed in the Assistment system. Following their work and that of Anozie and Junker (2007); Schofield (2007), we intend to use an errors-in-variables regression to predict exam scores. We would like to find that a dynamic estimate of θ is a better predictor than a static one. Therefore, we will compare predictive results like the cross-validation average difference between true and predicted MCAS scores (mean absolute deviation (MAD) scores), assuming we use the Assistment data set, and the cross-validation mean square error (MSE). We may also add a middle level of learning where we assume a dynamic model with a common rate of change for all students. Ideally, our models with the individual changes in cognitive ability for each student will best predict end of year state exam scores. This not only would lend credibility to the models but would be good evidence for teachers, researchers, and policy makers that the individual learning estimates are more valuable than static ones and they could then use the estimates to better inform instruction, experiments, and policy.

6 Proposed and Future Work

In this section, we outline the chapters we foresee going into the thesis along with proposed draft completion dates.

1. Introduction: May, 2011

2. Static Model Review: completed
   We will present the common static framework from the proposal as well as the review of static models.

3. Common Framework for Dynamic Models: completed
   We will present the common dynamic framework from the proposal.

4. Individual Models (4 chapters): September, 2010 - February, 2011
   We will present both an empirical and mathematical description of each of the four models we propose. In each case we will discuss the results of simulations on a variety of data sets including ones built specifically for that model, for one of the other models and possibly for other cognitive theory models altogether. We can then discuss the estimation and parameter recovery of those simulations and hopefully present some insight into the advantages and disadvantages of each model. Where appropriate, we will discuss the difficulties we encounter along with any computational tricks we discover for overcoming these difficulties. We may also develop some guidelines for when each model is appropriate, e.g. perhaps a certain sample size or number of skills is needed before considering a particular model.

5. Application: February - April, 2011
   We will apply at least one model to a real data set and present substantive conclusions relating to predictive power when comparing to predictions from a static model. We may also be able to give insight into the true rate and variability with which students learn.

6. Summary, Conclusions and Future Work: May, 2011

7 Conclusion

We propose four ideas for incorporating learning over time into the cognitive assessment framework. These models extend both IRT and CDM models so the user has the choice of whether he wants to define a student's latent cognitive ability as a continuous measure of overall ability or a discrete vector of skills. With the extensions to these models that we describe, we hope to develop a solid theoretical basis for which to include time in cognitive assessment models.

We expect that applying these models to real data sets will be informative to teachers, researchers, and policy makers for many reasons. On one hand, it may allow us to discover more about the rate at which students learn. We can then compare the actual learning rates with expectations which could be useful for helping teachers become better instructors and everyone have more realistic expectations about what to expect in terms of student learning over relatively short periods of time. We can also expect detailed information about individual students which has the potential to inform teachers so that they can better prepare for end of year accountability exams. Finally, the models presented can be used to make comparisons of learning between different types of experimental interventions and curricula.

Obviously, there is still a lot of work that must be done before these models are ready to be used mainstream. However, with the timeline presented in Section 6, we believe we can do accomplish this feat in the next year.

Appendices

A Static Models

In the main part of this proposal, we extend the theory behind cognitive assessment models to account for the dynamic case. In this section, we present some of the most commonly used static models that can be used in the new framework. We also note that while we assume dichotomous responses, i.e. $X_{ij} = 1$ if student i correctly answers item j and 0 otherwise, polytomous models are also possible (Studer 2009; Hemker et al. 2001; Thissen and Steinberg 1986).

A.1 Item Response Theory Models

In Item Response Theory (IRT; VanDerLinden and Hambleton 1997) models, the student parameter is usually defined to be normally distributed and is essentially measuring a student's "general propensity to do well" (Junker 1999). It's often a univariate parameter called θ but could also be multivariate. The item parameter, typically called β, can also be univariate or multivariate and describes the difficulty and other attributes of the item. In this section, we will present a short description of many of the static IRT models that are commonly used by defining the different enumerations of the student and item parameters.

All of the IRT models presented use a logistic link to define $P(X_{ij} = 1|\theta_i, \beta_j)$, the probability that student i correctly answers item j. Furthermore, we assume that the student parameter is a random effect (Holland 1990) and the item parameter is a fixed effect. Therefore, IRT models are a variation on mixed effects logistic regression models (Raudenbush and Bryk 2002; Hedeker 2005; McCullagh and Nelder 1989).

- **3PL Model** In the three parameter logistic (3PL) model [1],

  $$P(X_{ij} = 1|\theta_i, \beta_j) = g_j + \frac{1 - g_j}{1 + exp(a_j(\theta_i - b_j))}, \tag{9}$$

  where

  - β_j is assumed to be multidimensional with $\beta_j = (a_j, b_j, g_j)$.
  - θ_i is assumed to be unidimensional.

  When $P(X_{ij} = 1|\theta_i, \beta_j)$ is plotted against different values of θ, we call it an Item Characteristic Curve (ICC). Each of the parameters in β have both a mathematical and psychometric definition. (Harris 1989)

[1] Some authors add a constant to the 3PL, 2PL and 1PL models such that, for example with the 3PL model,$P(X_{ij} = 1|\theta_i, \beta_j) = g_j + \frac{1-g_j}{1+exp(1.7a_j(\theta_i - b_j))}$. They do this so that the approximation to the Normal Ogive model, which we do not discuss in this paper, is computationally simpler. We do not make this distinction as the notationally simpler model without 1.7 is sufficient for the logistic model discussed in this proposal.

  - Mathematically, a_j is the slope of the ICC. Psychometrically, a_j is the discrimination parameter. Higher values of a_j indicate that the item better differentiates between students at b_j, the point of inflection. Easy items (items with low b_j's) are used to discriminate between students with low abilities where as hard problems are used to discriminate between students with high abilities.
  - Mathematically, b_j is the point of inflection on the ICC. Psychometrically, b_j is the item difficulty parameter. At the pivot point, where proficiency equals difficulty or in other words, where $\theta_i = b_j$, student i has a 50% chance of correctly solving question j. Harder problems have higher values of b_j while easier problems have lower values of b_j.
  - Mathematically, g_j is the limit of the curve as θ goes to $-\infty$. Psychometrically, g_j is a guessing parameter which allows students, even those with low abilities, to answer questions correctly by chance.

- **2PL Model** The two parameter logistic (2PL) model is a more specific model than the 3PL model that assumes no guessing, $g_j = 0$. In other words, the 2PL model assumes that students have no chance of correctly guessing the answer. They must have some proficiency to get an answer right. (Harris 1989)

$$P(X_{ij} = 1|\theta_i, \beta_j) = \frac{1}{1 + exp(a_j(\theta_i - b_j))}, \tag{10}$$

  where

  - β_j is assumed to be multidimensional with $\beta_j = (a_j, b_j)$.
  - θ_i is assumed to be unidimensional.

  The mathematical and psychometric definitions of these parameters remain unchanged from the 3PL model.

- **1PL or Rasch Model** The 1PL model is an even less general model than the 2PL model and is most commonly known as the Rasch model (Rasch 1960/1980). This model sets $g_j = 0$ and $a_j = 1$. Therefore, students have no chance of guessing the correct answer and furthermore, all questions on the exam are presumed to be equally discriminating. (Harris 1989) Then

$$P(X_{ij} = 1|\theta_i, \beta_j) = \frac{1}{1 + exp(\theta_i - \beta_j)}, \tag{11}$$

  where

  - β_j is assumed to be unidimensional; hence, we substitute β_j for b_j.
  - θ_i is assumed to be unidimensional.

The mathematical and psychometric definition of β_j remains unchanged from b_j in the 3PL model.

- **Multidimensional IRT** In multidimensional IRT (MIRT; VanDerLinden and Hambleton 1997) models, we are interested in changing the dimension of the student parameter. Many tests are constructed to measure more than one type of ability. For example, with an exam that covers both reading and math material or less disparate, two math subjects like probability and integration, we would most likely be more interested in separate estimates of the proficiencies for each subject. In this case, θ, for each student, would be two dimensional. (DeBoeck and Wilson 2004) In this case, the ICC remains the same as the 1PL, 2PL or 3PL models but
  - β_j can be from the 3PL model: (a_j, b_j, g_j), the 2PL model: (a_j, b_j) or the 1PL model: (b_j).
  - θ_i is a linear combination of $(\theta_{1i}, \theta_{2i}, ..., \theta_{Ki})$ where K is the desired dimension of θ.

- **Hierarchical Models** Another way to expand the student parameter is hierarchically (Raudenbush and Bryk 2002). We may expect all students with similar characteristics to have the same ability. For example all students in the same class were taught by the same teacher and so we may expect them to have the same ability. In this case, we would actually be estimating fewer than N student parameters (the number of teachers in this scenario). Alternatively, we may think that students from the same class are similar but not exactly the same. In this case we build in a class hierarchy. The hierarchy can be extended to any parameter that describes students such as the school they attend, the district they are in, the teacher they have, their gender, etc. These models are an extension to the logistic models because we still use a 1-, 2- or 3- PL model. (DeBoeck and Wilson 2004) In this case, the ICC remains the same as the 1PL, 2PL or 3PL models but
  - β_j can be from the 3PL model: (a_j, b_j, g_j), the 2PL model: (a_j, b_j) or the 1PL model: (b_j).
  - θ_i is a linear combination of θs describing the student's properties. For example, there could be a separate θ for the teacher, school, district and then one that directly corresponds to the additional variation in student i.

- **LLTM** In the Linear Logistic Test Model (LLTM) we are interested in reducing the dimension of the item parameter. Instead of assuming that each question has its own difficulty, we might be interested in grouping similar questions together. For example if an exam tests multiple skills, we may classify individual items by the skills necessary to complete them and then be interested in the difficulty of the skill as opposed to the item. These are a subclass of the logistic models where we still use a 1-, 2- or 3- PL model. (DeBoeck and Wilson 2004) Therefore, the ICC remains the same as the particular model chosen but
  - Whichever IRT model is chosen, each element in β_j is a linear combination of D item features. For example, in the Rasch model, $\beta_j = (\beta_{1j}, \beta_{2j}, ..., \beta_{Dj})$.

  - θ_i is assumed to be unidimensional.

- **Multidimensional LLTM** Finally, we could adjust the dimensionality of both the student and item parameters. This is known as the multidimensional LLTM (DeBoeck and Wilson 2004). Again, the ICC corresponds to the IRT model chosen but

  - Whichever IRT model is chosen, each element in β_j is a linear combination of D item features. For example, in the Rasch model, $\beta_j = (\beta_{1j}, \beta_{2j}, ..., \beta_{Dj})$.
  - θ_i is a linear combination of $(\theta_{1i}, \theta_{2i}, ..., \theta_{Ki})$ where K is the desired dimension of θ.

A.2 Cognitive Diagnosis Models

In Cognitive Diagnosis Models (CDMs; Rupp and Templin 2008), the student parameter is defined to be a vector of Bernoulli's and is essentially measuring whether a student possesses a set of skills necessary to do well on the assessment. Therefore, the student parameter, θ_i is a vector of length K where $\theta_{ik} = 1$ if student i possesses skill k and 0 otherwise for each of K skills. Because we are interested in knowing whether a student has a specific set of skills, it is necessary to know which skills are required to answer each question correctly. We need an expert defined design matrix, often called Q, where $q_{jk} = 1$ indicates that a correct answer to item j require a student to possess trait k and is 0 otherwise (Barnes 2005). In CDMs, the item parameter is typically multidimensional and describes different attributes of the item. In this section, we present a short description of some common static CDMs that are commonly used in cognitive assessment.

- **DINA Model** In the Deterministic Input; Noisy "And" gate (DINA) model,

  $$P(X_{ij} = 1|\theta_i, \beta_j) = (1 - s_j)^{\xi_{ij}} g_j^{1-\xi_{ij}}. \tag{12}$$

  In this model, a new indicator variable ξ_{ij} is presented, which equals 1 if student i has mastered all skills necessary to correctly answer item j and 0 if he has failed to master at least one skill. In other words, it is the ideal response in that if there were no guessing and slipping allowed, ξ would be equal to the student's response. Mathematically,

  $$\xi_{ij} = \prod_{k=1}^{K} \theta_{ik}^{q_{jk}}. \tag{13}$$

  The item parameter, β_j, is two dimensional with slip and guess parameters. The slip parameter, $s_j = P(X_{ij} = 0|\xi_{ij} = 1)$, is the probability that a student gets question j incorrect even though he has mastered all of the necessary skills. The guess parameter, $g_j = P(X_{ij} = 1|\xi_{ij} = 0)$, is similar to the guess parameter in the 3PL model. It is the probability a student correctly answers question j even though he doesn't have all of the required skills. (Macready and Dayton 1977; Haertel 1989; Junker and Sijtsma 2001)

  Explicitly,

- β_j is assumed to be multidimensional with $\beta_j = (s_j, g_j)$.
- θ_i is assumed to be multidimensional with $\theta_i = (\theta_{i1}, \theta_{i2}, ..., \theta_{iK})$.

- **NIDA Model** In the Noisy Input, Deterministic "And" gate (NIDA) model,

$$P(X_{ij} = 1|\theta_i, \beta_j) = \prod_{k=1}^{K}((1 - s_k)^{\theta_{ik}} g_k^{1-\theta_{ik}})^{q_{jk}}. \tag{14}$$

The NIDA model is similar to the DINA model except we assume that each *skill*, as opposed to each *item*, has an associated guess and slip parameter. (Maris 1999; Junker and Sijtsma 2001)

Explicitly,

- β is a skill parameter as opposed to an item parameter and is therefore subscripted by k. It is assumed to be multidimensional with $\beta_k = (s_k, g_k)$.
- θ_i is multidimensional with $\theta_i = (\theta_{i1}, \theta_{i2}, ..., \theta_{iK})$.

- **RedRUM** In the Reduced Reparametrized Unified Model (RedRUM),

$$P(X_{ij} = 1|\theta_i, \beta_j) = \pi_j \prod_{k=1}^{K}(r_{jk}^{1-\theta_{ik}})^{q_{jk}} \tag{15}$$

The RUM model is a generalization of the NIDA and DINA models where β_j is defined as π_j, the maximal probability of success on item j, and r_{jk}, the penalty for each skill, k, not possessed (DiBello et al. 1995; Junker 2007; Hartz 2002).

If we constrain this model, it becomes exactly the NIDA or DINA model depending on the parametrization. If we let $\pi_j = \prod_{k=1}^{K}(1 - s_{jk})^{q_{jk}}$, $r_{jk} = \frac{g_{jk}}{1-s_{jk}}$, and rearrange terms we will discover the DINA and NIDA models (assuming for the NIDA model: $s_{jk} = s_k$ and $g_{jk} = g_k$ and for the DINA model: $s_{jk} = s_j$ and $g_{jk} = g_j$ Hartz 2002).

Explicitly,

- β_j is assumed to be multidimensional with $\beta_j = (\pi_j, r_{jk})$.
- θ_i is assumed to be multidimensional with $\theta_i = (\theta_{i1}, \theta_{i2}, ..., \theta_{iK})$.

B Figures

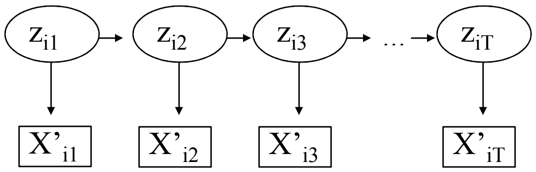
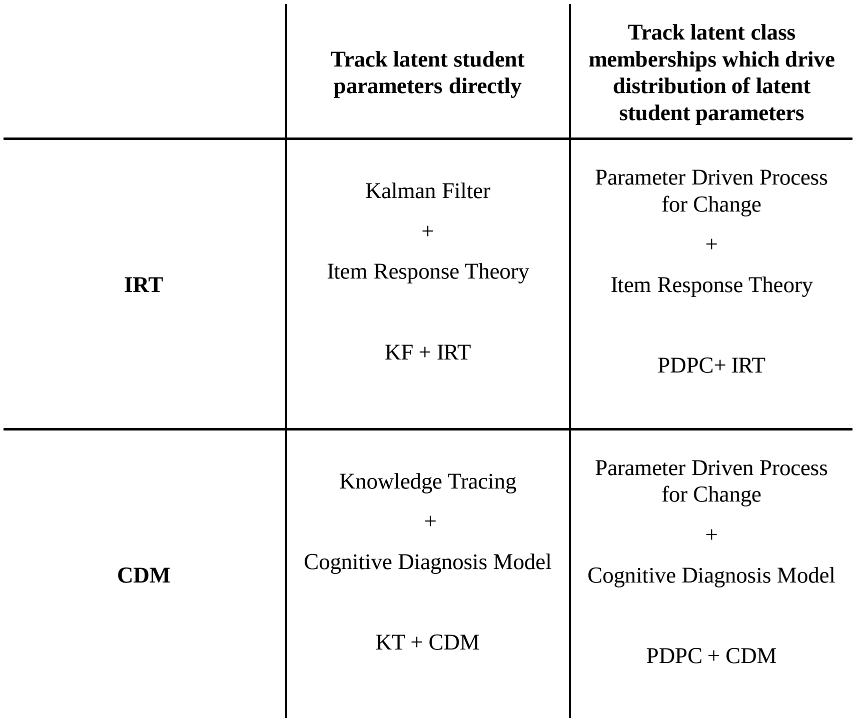


Figure 1: A directed acyclic graph that depicts the relationship between student responses and latent knowledge states over time. The conditional independences inherent in directed acyclic graphs help us define the common framework for dynamic models.



	Track latent student parameters directly	**Track latent class memberships which drive distribution of latent student parameters**
IRT	Kalman Filter + Item Response Theory KF + IRT	Parameter Driven Process for Change + Item Response Theory PDPC+ IRT
CDM	Knowledge Tracing + Cognitive Diagnosis Model KT + CDM	Parameter Driven Process for Change + Cognitive Diagnosis Model PDPC + CDM

Figure 2: A graphical depiction of the four methods we present to incorporate time into cognitive assessment models.

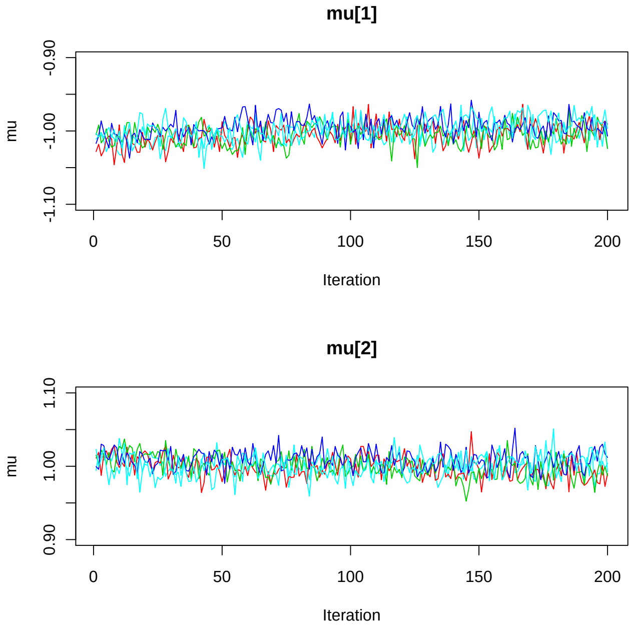


Figure 3: Each plot shows the five MCMC chains for estimating μ_1 and μ_2. There are 200 iterations from the original 2000 due to a burn-in of 600 and thinning of 27. All five chains overlap and span a small range with $\hat{R} \leq 1.1$ indicating convergence.

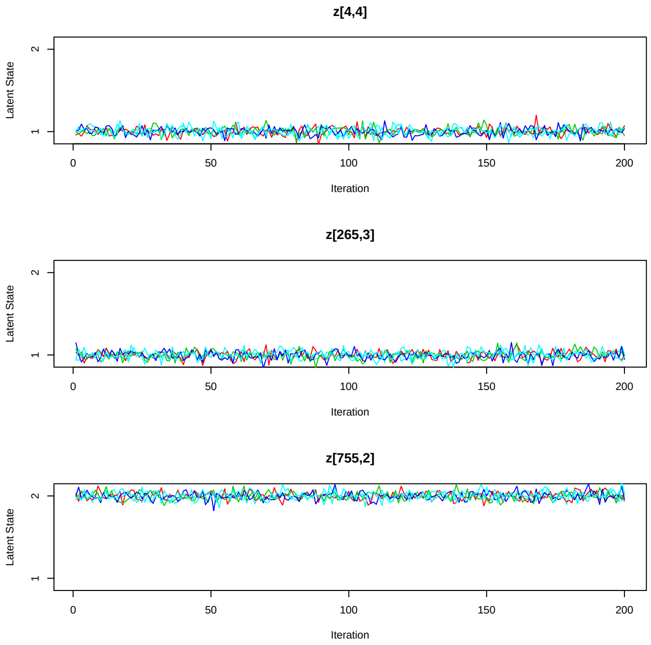


Figure 4: Each plot shows the five MCMC chains for estimating 3 random students' latent states at 3 random time points (where $z_{it} = z[i,t]$) from top to bottom. There are 200 iterations from the original $6,000$ due to a burn-in of 600 and thinning of 27. We added noise to the binary results so that the chains would not be completely overlapping. In these cases, all five of the chains agreed at all simulation points about which state the students belonged.

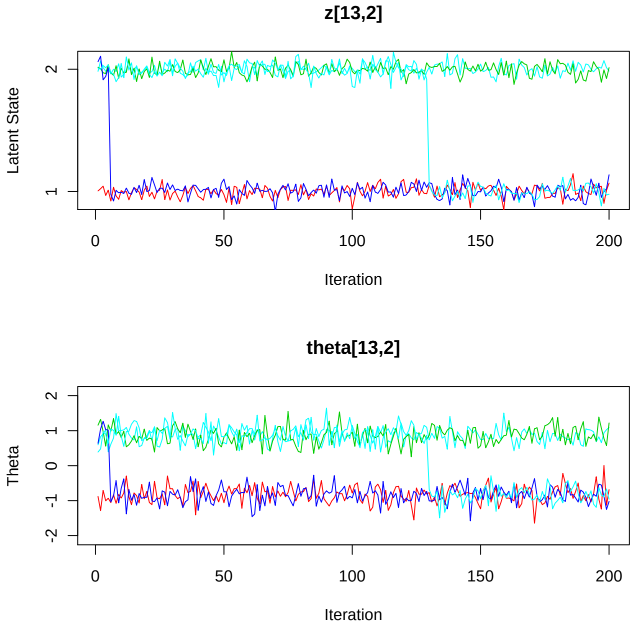


Figure 5: The two plots show the five MCMC chains for estimating student 13's latent state and θ at time $t = 2$. There are 200 iterations from the original $6,000$ due to a burn-in of 600 and thinning of 27. We added noise to the binary results of the latent state so that the chains would not be completely overlapping. We see that there was little agreement about which state this student should be classified to. However, looking at the student's response pattern indicates that he may have a θ value close to 0 which means he doesn't fit into either latent state and so it's not surprising that the estimation procedure has trouble placing him.

References

Anozie, N. and Junker, B. (2007), "Investigating the utility of a conjunctive model in Q-matrix assessment using monthly student records in an online tutoring system," *Proposal submitted to the National Council on Measurement in Education 2007 meeting*.

Ayers, E. and Junker, B. (2008), " IRT Modeling of Tutor Performance To Predict End-of-year Exam Scores," *Educational and Psychological Measurement*, 68, 972–987.

Barnes, T. (2005), "Q-matrix Method: Mining Student Response Data for Knowledge," *Proceedings of the AAAI Workshop on Educational Data Mining Pittsburgh (AAAI Technical Report)*.

Bransford, J., Brown, A., and Cocking, R. (2000), *How people learn: Brain, Mind, Experience, and School*, Washington DC: National Academy Press.

Carver, S. (2001), *Cognition and instruction: Enriching the laboratory school experience of children, teachers, parents, and undergraduates* , Mahwah, NJ: Lawrence Erlbaum Associates, Inc., In Carver & Klahr (Eds.) Cognition and instruction: Twenty-five years of progress.

Cen, H., Koedinger, K., and Junker, B. (2006), *Lecture Notes in Computer Science: Learning Factors Analysis - A General Method for Cognitive Model Evaluation and Improvement*, Berlin: Springer.

Connor, J. (2006), "Multivariate Mixture Models to Describe Longitudinal Patterns of Frailty in American Seniors," *Thesis, Department of Statistics, Carnegie Mellon University*.

Corbett, A., Anderson, J., and O'Brien, A. (1995), "Student Modeling in the ACT Programming Tutor," *Chapter 2 in P. Nichols, S. Chipman & R. Brennan, Cognitively Diagnostic Assessment. Hillsdale, NJ: Erlbaum*.

DeBoeck, P. and Wilson, M. (2004), *Explanatory Item Response Model: A Generalized Linear and Nonlinear Approach*, New York: Springer.

DeLaTorre, J. and Douglas, J. (2004), "Higher-order Latent Trait Models for Cognitive Diagnosis," *Psychometrika*, 69, 333–353.

Dethlefsen, C. and Lundbye-Christensen, S. (2006), "Formulating State Space Models in R with Focus on Longitudinal Regression Models," *Journal of Statistical Software*, 16.

DiBello, L., Stout, W., and Roussos, L. (1995), "Unified Cognitive/Psychometric Diagnostic Assessment Likelihood-Based Classification Techniques," *Chapter 15 in P. Nichols, S. Chipman & R. Brennan, Cognitively Diagnostic Assessment. Hillsdale, NJ: Erlbaum*.

Draney, K., Pirolli, P., and Wilson, M. (1995), "A measurement model for complex cognitive skills," *Chapter 15 in P. Nichols, S. Chipman & R. Brennan, Cognitively Diagnostic Assessment. Hillsdale, NJ: Erlbaum*.

Durbin, J. and Koopman, S. (2000), "Time Series Analysis of Non-Gaussian Observations Based on State Space Models from both Classical and Bayesian Perspectives," *Journal of the Royal Statistical Society B*, 62, 3–56, with discussion.

Feng, M., Heffernan, N., and Beck, J. (2009), "Using learning decomposition to analyze instructional effectiveness in the ASSISTment system ," *Proceeding of the 14th International Conference on Artificial Intelligence in Education*, 523–530.

Feng, M., Heffernan, N., and Koedinger, K. (2006), "Addressing the Testing Challenge with a Web-Based E-Assessment System that Tutors as it Assesses," *Association for Computing Machinery*, 307–316.

Gelman, A., Carlin, J., Stern, H., and Rubin, D. (2004), *Bayesian Data Analysis*, Boca Raton: Chapman and Hall/CRC, 2nd ed.

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

Harris, D. (1989), "Comparison of 1-, 2-, and 3-Parameter IRT Models," *Items: Instructional Topics in Educational Measurement*, 157–163.

Hartz, S. (2002), "A Bayesian framework for the Unified Model for assessing cognitive abilities: Blending theory with practice," *Unpublished doctoral dissertation, University of Illinois at Urbana-Champaign.*

Hedeker, D. (2005), *Generalized Linear Mixed Models. In B. Everitt & D. Howell (Eds.)*, Wiley, New York: Encyclopedia of Statistics in Behavioral Science.

Heffernan, N., Koedinger, K., Junker, B., and Ritter, S. (2001), "Using Web-Based Cognitive Assessment Systems for Predicting Student Performance on State Exams," *Technical Report, Institute of Educational Statistics: US Dept. of Education, & Dept. of Computer Science Worcester Polytechnic Institute Univ.*

Hemker, B., VanDerArk, L., and Sijtsma, K. (2001), "On Measurement Properties of Continuation Ratio Models," *Psychometrika*, 66, 487–506.

Holland, P. (1990), "On The Sampling Theory Foundations of Item Response Theory Models," *Psychometrika*, 55, 577–601.

Jasra, A., Holmes, C., and Stephens, D. (2006), "Markov Chain Monte Carlo Methods and the Label Switching Problem in Bayesian Mixture Modeling," *Statistical Science*, 20, 50–67.

Junker, B. (1999), "Some statistical models and computational methods that may be useful for cognitively-relevant assessment," *Prepared for the Committee on the Foundations of Assessment, National Research Council.*

— (2007), “Some Issues And Applications In Cognitive Diagnosis And Educational Data Mining,” *New Trends in Psychometrics*.

Junker, B. and Sijtsma, K. (2001), “Cognitive Assessment Models with Few Assumptions, and Connections with Nonparametric Item Response Theory,” *Applied Psychological Measurement*, 25, 258–272.

Koedinger, K., Cunningham, K., Skogsholm, A., and Leber, B. (2008), “An open repository and analysis tools for fine-grained, longitudinal learner data. In Baker, R.S.J.d., Barnes, T., Beck, J.E. (Eds.),” *1st International Conference on Educational Data Mining, Proceedings. Montreal, Quebec, Canada*, 157–166.

Koedinger, K., McLaughlin, E., and Heffernan, N. (2010), “A Quasi-Experimental Evaluation of an On-line Formative Assessment and Tutoring System ,” *Journal of Educational Computing Research*, 4.

Lunn, D., Thomas, A., Best, N., and Spiegelhalter, D. (2000), “WinBUGS - a Bayesian modelling framework: concepts, structure, and extensibility ,” *Statistics and Computing*, 10, 325–337.

Macready, G. and Dayton, C. (1977), “The use of probabilistic models in the assessment of mastery,” *Journal of Educational Statistics*, 2, 99–120.

Manrique, D. (2009), “Mixed Membership Multivariate Longitudinal Models with Applications,” *Thesis proposal, Department of Statistics, Carnegie Mellon University*.

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

McCullagh, P. and Nelder, J. (1989), *Generalized Linear Models*, Boca Raton: Chapman and Hall/CRC, 2nd ed.

Oud, J., Jansen, R., VanLeeuwe, J., Aarnoutse, C., and Voeten, M. (1999), “Monitoring Pupil Development by Means of the Kalman Filter and Smoother Based Upon SEM State Space Modeling,” *Learning and Individual Differences*, 11, 121–136.

Pavlik, P., Cen, H., and Koedinger, K. (2009), “Performance Factors Analysis - A New Alternative to Knowledge Tracing ,” *Proceeding of the 14th International Conference on Artificial Intelligence in Education (AIED09)*, 531–538.

Rasch, G. (1960/1980), *Probabilistic models for some intelligence and attainment tests*, Chicago: The University of Chicago Press, 2nd ed., copenhagen, Danish Institute for Educational Research), expanded edition (1980) with foreword and afterword by B.D. Wright.

Raudenbush, S. and Bryk, A. (2002), *Hierarchical Linear Models: Applications and Data Analysis Methods*, California: Sage Publications, 2nd ed.

Rijmen, F., DeBoeck, P., and VanderMaas, H. (2005), "An IRT Model with a Parameter-Driven Process for Change," *Psychometrika*, 70, 651–669.

Rijmen, F., Vansteelandt, K., and DeBoeck, P. (2008), "Latent Class Models for Diary Method Data: Parameter Estimation by Local Computations," *Psychometrika*, 73, 167–182.

Roeder, K., Lynch, K., and Nagin, S. (1999), "Modeling Uncertainty in Latent Class Membership: A Case Study in Criminology," *Journal of the American Statistical Association*, 94, 766–776.

Rupp, A. and Templin, J. (2008), "Unique characteristics of diagnostic models: a review of the current state-of-the-art," *Measurement*, 6, 219–262.

Schofield, L. (2007), "Using cognitive test scores in social science research," *Thesis proposal, Department of Statistics, Carnegie Mellon University*.

Stallard, E. (2007), "Trajectories of Disability and Mortality Among the U.S. Elderly Population: Evidence from the 1984-1999 NLTCS," *North American Actuarial Journal*.

Studer, C. (2009), "A Unifying Framework for Cognitive Assessment Models," *Tech Report?*

Thissen, D. and Steinberg, L. (1986), "A Taxonomy of Item Response Models," *Psychometrika*, 51, 567–577.

VanDerLinden, W. and Hambleton, R. (1997), *Handbook of modern item response theory*, New York: Springer-Verlag.

Wasserman, L. (2004), *All of Statistics: A Concise Course in Statistical Inference*, New York: Springer Texts in Statistics.