Article

Applied Psychological Measurement
2018, Vol. 42(1) 5–23

DOI: 10.1177/0146621617721250
journals.sagepub.com/home/apm
SAGE

A Hidden Markov Model for Learning Trajectories in Cognitive Diagnosis With Application to Spatial Rotation Skills

Yinghan Chen[1], Steven Andrew Culpepper[2], Shiyu Wang[3], and Jeffrey Douglas[2]

Abstract

The increasing presence of electronic and online learning resources presents challenges and opportunities for psychometric techniques that can assist in the measurement of abilities and even hasten their mastery. Cognitive diagnosis models (CDMs) are ideal for tracking many fine-grained skills that comprise a domain, and can assist in carefully navigating through the training and assessment of these skills in e-learning applications. A class of CDMs for modeling changes in attributes is proposed, which is referred to as learning trajectories. The authors focus on the development of Bayesian procedures for estimating parameters of a first-order hidden Markov model. An application of the developed model to a spatial rotation experimental intervention is presented.

Keywords

cognitive diagnosis, Latent class models, learning

Introduction

The increasing presence of electronic and online learning resources presents challenges and opportunities for psychometric techniques that can assist in the measurement of abilities and even hasten their mastery. Cognitive diagnosis models (CDMs) are ideal for tracking many fine-grained skills that comprise a domain, and can assist in carefully navigating through the training and assessment of these skills in e-learning applications. The coupling of models for skill acquisition with item response models in cognitive diagnosis is considered. Modeling

[1]University of Nevada–Reno, USA
[2]University of Illinois at Urbana–Champaign, USA
[3]University of Georgia, Athens, USA

Corresponding Author:
Jeffrey Douglas, Department of Statistics, University of Illinois at Urbana–Champaign, 725 South Wright Street, Champaign, IL 61820, USA.
Email: jeffdoug@illinois.edu

learning is a natural direction for skills diagnosis with the increased presence of online training and interventions designed to promote skill acquisition.

Traditional research in CDMs has assumed that the vector of latent skills under diagnosis is static throughout an assessment. However, e-learning and intelligent tutoring systems provide opportunities to provide training interspersed with assessment, so it is more accurate to think of the latent attribute vector dynamically. Only recently have CDMs been considered for changing attribute vectors, and this has mostly been in the longitudinal data setting (Kaya & Leite, 2016; Li, Cohen, Bottge, & Templin, 2015), in which considerable time could expire between two assessments within which the latent trait is assumed static. Li et al. (2015) used the DINA (Junker & Sijtsma, 2001) model together with a transition model as a means of measuring the effects of an educational intervention. Through simulation studies, Kaya and Leite (2016) studied such transition models for longitudinal applications using both the Deterministic inputs, noisy and gate model (DINA) and the Deterministic inputs, noisy or gate model (DINO) (Templin & Henson, 2006) models.

In e-learning settings, one might reasonably expect that learning could take place between items or between brief learning modules with little separation in time. Studer (2012) proposed the Parameter Driven Process for Change method which indirectly tracks learning by assessing a student's membership in a small number of possible latent states. The number of latent states would be much fewer than the 2^K attribute profiles that exist when K binary skills are being assessed, and the K components of the attribute profile are modeled as independent given the latent state. These latent states essentially serve as higher order latent variables. Some states have higher prevalence for each component and learning is tracked less specifically by considering membership in the latent states rather than charting the change in the attribute profile itself. In this way, fewer transition probabilities are needed to model transitions, but membership in the different latent states has a bearing on changing posterior probabilities for the full attribute pattern as it changes over time. Studer (2012) also considered the case in which there are 2^K latent states in the transition model, one for each attribute pattern, and refers to this as Knowledge Tracing plus CDM, though no specific formulation or examples are provided. Recently, Wang, Yang, Culpepper, and Douglas (2016) considered a full transition model for the entire attribute pattern to address the individualized transition patterns by using a set of latent and observed covariates, such as a general learning ability and practice effects, and applied it to a study of learning spatial reasoning skills in which an intervention was administered between short blocks of items. Recent learning research in CDMs has also considered sequential statistical methods for detection of learning (Ye, Fellouris, Culpepper, & Douglas, 2016), in which methods for detection of learning with minimal delay were introduced. More accurate models for learning could enhance detection, enabling one to more efficiently navigate through a list of skills to be learned while avoiding false detections.

In the data mining field, the method of Knowledge Tracing (Corbett & Anderson, 1994) has emerged as a popular technique for modeling learning, usually in the setting of intelligent tutoring systems. Knowledge Tracing bears much resemblance to learning models for CDM, but traditionally focuses on one attribute at a time. In fact, Studer (2012) showed that Knowledge Tracing is mathematically equivalent to an extension of the Noisy inputs, deterministic and gate model (NIDA) (Junker & Sijtsma, 2001; Maris, 1999) model for multiple time points, with the restriction that each item depends on a single skill. However, recent extensions to Knowledge Tracing have been introduced that incorporate multiple skills at once and allow for different item parameters (González-Brenes, Huang, & Brusilovsky, 2014; González-Brenes & Mostow, 2013; Pardos & Heffernan, 2010; Y. Xu & Mostow, 2012). As the body of research on dynamic latent class modeling grows, the parallel tracks of CDM and Knowledge Tracing can be expected to produce ideas that can benefit both methodologies at once.

The purpose of this article is to propose a class of dynamic models for tracing changes in attributes, which is referred to as learning trajectories. There are two main contributions compared with the existing literature. The first is to introduce the notion of learning trajectories within the CDM framework. Different types of learning trajectories, from the most general case, unstructured trajectories, to nondecreasing patterns, are defined and illustrated with examples. A Bayesian Modeling framework to capture each type of learning trajectory is also established. Especially, the authors discuss and compare their model complexities, the number of parameters, and the prior distributions for each proposed model. The second contribution is that one of the proposed model through a Monte Carlo simulation and a real data application are evaluated. In practical settings, researchers may have a pool of items to administer over time to evaluate learning trajectories. There is no prior research on how to design such tests to ensure accurate model parameter recovery. Accordingly, the Monte Carlo simulation includes conditions to evaluate different test designs for studies of learning trajectories, and the results can provide some guidelines on how to design an efficient test to guarantee the accuracy of the estimation of item parameters and learning transition matrix. The real data application demonstrates how to use the proposed model to capture students' learning trajectories in a spatial rotation experiment with learning interventions. The estimated transition matrix has implications on how students learn those skills, and this information can be used to refine learning interventions to improve the speed at which skills are mastered.

This article includes five sections. The "Enumerating Learning Trajectories" section introduces readers to the notion of learning trajectories within the CDM framework. One consideration of this section is to enumerate the complexity of different learning trajectories with an aim of identifying the level of generality that may be possible to consider in real problems. The "Modeling Learning Trajectories" section presents a Bayesian model formulation for estimating CDMs. In particular, this section provides an overview of several prior distributions for modeling learning trajectories over time. The "Monte Carlo Simulation Study" section reports results from a Monte Carlo study that is designed to assess parameter recovery of a first-order hidden Markov model (FOHM). The "Application to Spatial Rotation Skill Acquisition" section presents an application of a FOHM model to a spatial rotation experimental intervention. The "Discussion" section provides concluding remarks and recommendations for future research.

Enumerating Learning Trajectories

This section considers the circumstance where the latent attributes may change over time. Let α_{ikt} be one if individual i $(i = 1, \ldots, N)$ possesses attribute k $(k = 1, 1 \ldots, K)$ at time t $(t = 1, \ldots, T)$. Define a trajectory of attribute profiles for individual i as $\alpha_i = (\alpha_{i1}, \ldots, \alpha_{iT})'$ where $\alpha_{it} = (\alpha_{i1t}, \ldots, \alpha_{iKt})'$ represents student i's attribute profile at time t. Let $j = 1, \ldots, J_t$ index the number of items administered at time t and define the total number of items as $J = \sum_{t=1}^{T} J_t$. Let $\mathbf{Q}_t$ be the $J_t \times K$ Q-matrix at time t with zeros and ones that denote the skills needed to correctly answer the J_t items at time t.

Let the total set of attribute vectors be $\boldsymbol{\mathcal{A}} = \left\{\boldsymbol{\alpha} : \boldsymbol{\alpha} \in \{0, 1\}^K\right\}$, so that $\boldsymbol{\alpha}_c \in \boldsymbol{\mathcal{A}}$ denotes one of the possible 2^K attribute vectors. Define the set of attribute profiles with k mastered skills as $\boldsymbol{\mathcal{A}}_k = \left\{\boldsymbol{\alpha} : \boldsymbol{\alpha} \in \{0, 1\}^K, \|\boldsymbol{\alpha}\|_1 = k\right\}$ and note that $\boldsymbol{\mathcal{A}} = \cup_{k=1}^{K} \boldsymbol{\mathcal{A}}_k$. Recall that the cardinality of $\boldsymbol{\mathcal{A}}_k$ is the number of attribute profiles in $\boldsymbol{\mathcal{A}}_k$ and is denoted by $|\boldsymbol{\mathcal{A}}_k| = \binom{K}{k}$ and the total number of attribute profiles is $|\boldsymbol{\mathcal{A}}| = \sum_{k=0}^{K} \binom{K}{k} = 2^K$.

Unrestricted Learning Trajectories

Definition 1: In the absence of restrictions on the discrete latent class space, the set of learning trajectories at time t is recursively defined as

$$\boldsymbol{\mathcal{A}}^{(t)} = \boldsymbol{\mathcal{A}}^{t-1} \times \boldsymbol{\mathcal{A}} = \boldsymbol{\mathcal{A}}^{t}. \tag{1}$$

Note the parentheses in the superscript are dropped and $\boldsymbol{\mathcal{A}}^t$ is used to denote the set of learning trajectories at time t. The definition implies the set of attribute trajectories for time $t=1$ is $\boldsymbol{\mathcal{A}}^{(1)} = \boldsymbol{\mathcal{A}} = \boldsymbol{\mathcal{A}}^1$. Similarly, the set of attribute trajectory classes for time $t=2$ is $\boldsymbol{\mathcal{A}}^{(2)} = \boldsymbol{\mathcal{A}}^1 \times \boldsymbol{\mathcal{A}} = \boldsymbol{\mathcal{A}}^2$. That is, students can theoretically switch from the initial states in $\boldsymbol{\mathcal{A}}^1$ to any attribute class in $\boldsymbol{\mathcal{A}}$ after receiving feedback and/or instruction. Rules concerning Cartesian products imply the cardinality of $\boldsymbol{\mathcal{A}}^T$ is the product of sets is $|\boldsymbol{\mathcal{A}}^T| = 2^{KT}$; that is, learning trajectories grow exponentially with time and the number of attributes.

Example 1: Consider $\boldsymbol{\mathcal{A}}^T$ for $K=2$ and $T=2$. If $K=T=2$, $\boldsymbol{\mathcal{A}}^1 = \{00, 10, 01, 11\}$ and $\boldsymbol{\mathcal{A}}^2$ includes 16 learning trajectories:

$$\boldsymbol{\mathcal{A}}^2 = \boldsymbol{\mathcal{A}}^1 \times \boldsymbol{\mathcal{A}} = \left\{ \begin{array}{l} (00,00), (00,10), (00,01), (00,11), (10,00), (10,10), (10,01), (10,11), \\ (01,00), (01,10), (01,01), (01,11), (11,00), (11,10), (11,01), (11,11) \end{array} \right\}.$$

■

Nondecreasing Learning Trajectories

Clearly, the discrete latent space grows exponentially as t increases, and the set $\boldsymbol{\mathcal{A}}^t$ can be restricted to study a subset of the 2^{KT} learning trajectories. For instance, in education, a reasonable assumption is that students acquire skills over time in a nondecreasing fashion. Stated differently, learning can be assumed to be an absorbing state where mastered attributes cannot be unlearned. This subsection compares the set of nondecreasing learning trajectories with $\boldsymbol{\mathcal{A}}^t$.

Definition 2: Nondecreasing learning and skill acquisition trajectories are defined as

$$\boldsymbol{\mathcal{A}}^t_+ = \boldsymbol{\mathcal{A}}^t \cap \{\boldsymbol{\alpha} : \boldsymbol{\alpha}_t \boldsymbol{\alpha}_{t-1},\ \forall\, t>1\}. \tag{2}$$

That is, $\boldsymbol{\mathcal{A}}^t_+$ is the set of learning trajectories that are nondecreasing. It is straightforward to find the cardinality of $\boldsymbol{\mathcal{A}}^t_+$ by counting the number of ways of picking the change point (i.e., the point where α_{ikt} changes from zero to one). For a given attribute, there are T possibilities for the change point if learning occurs and one possibility if the attribute remains a zero over time for a total of $T+1$ outcomes. Consequently, the number of nondecreasing change points for K attributes is $|\boldsymbol{\mathcal{A}}^T_+| = (T+1)^K$.

Example 2: Revisiting Example 1 with $K=2$ and $T=2$, for $K=2$ and $T=2$, there are a total of nine nondecreasing learning trajectories within $\boldsymbol{\mathcal{A}}^2_+$. The set of nondecreasing learning trajectories, $\boldsymbol{\mathcal{A}}^2_+$, includes the following elements:

$$\mathcal{A}_+^t = \{(00,00),(00,10),(00,01),(00,11),(10,10),(10,11),(01,01),(01,11),(11,11)\}.$$

■

Any models of student learning trajectories necessarily must propose a manner in which students transition between attribute classes over time. One limitation for practice is that the number of learning trajectories grows exponentially with t, which implies that larger datasets would be needed to accurately model the share of the population in any given learning trajectory. The learning assumption significantly reduces the number of possible learning trajectories. For purpose of demonstration, Table 1 reports the cardinality of $\mathcal{A}_+^T$ and $\mathcal{A}^T$ for different values of T and K. The values in Table 1 show that even for the case where $K = 2$ and $T = 6$, there are 4,096 unrestricted learning trajectories, which could pose challenges in real applications. In contrast, the learning assumption reduces the number of learning trajectories by a factor of nearly 100 from 4,096 to 49. An important observation from Table 1 is that the number of unrestricted learning trajectories grows quickly even for a modest $K = 4$ with a total of 2^{24} possible trajectories. Although there are relatively fewer nondecreasing learning trajectories, for $K = 4$ and $T = 6$, researchers would need substantial data to estimate probabilities of membership in one of the possible 2,401 classes.

Modeling Learning Trajectories

This section discusses strategies for modeling unrestricted and nondecreasing learning trajectories. Let $\boldsymbol{\alpha}_l' = (\boldsymbol{\alpha}_{l1}, \ldots, \boldsymbol{\alpha}_{lT})$ be a learning trajectory such that $\boldsymbol{\alpha}_l \in \mathcal{A}^T$. Let Y_{ijt} denote the random variable for person i to item j at time t and let y_{ijt} represent the observed value. Let $\boldsymbol{y}_i' = (\boldsymbol{y}_{i1}, \ldots, \boldsymbol{y}_{iT})$ denote vectors of observations over time of the random variables $\boldsymbol{Y}_i' = (\boldsymbol{Y}_{i1}, \ldots, \boldsymbol{Y}_{iT})$ with $\boldsymbol{y}_{it}' = (y_{i1t}, \ldots, y_{iJ_t t})$ and $\boldsymbol{Y}_{it}' = (Y_{i1t}, \ldots, Y_{iJ_t t})$. Let the collection of item parameters over time be denoted by $\zeta' = (\zeta_1, \ldots \zeta_T)$ with $\zeta_t' = (\zeta_{1t}, \ldots, \zeta_{J_t t})$, and ζ_{jt} denotes the item parameters for item j at time t.

The probability of a correct response on item j at time t for an individual with learning trajectory l is $P(Y_{ijt} = 1 | \alpha_{it} = \alpha_{lt}, \zeta_{jt})$. Assuming that responses are independent given α_i, the probability of observing subject i's responses to items $j = 1, \ldots, J_t$ at time t conditioned upon $\boldsymbol{\alpha}_{it} = \boldsymbol{\alpha}_{lt}$ and the item parameters ζ_t is

$$p(\boldsymbol{Y}_{it} | \alpha_{lt}, \zeta_t) = \prod_{j=1}^{J_t} \left[P\left(Y_{ijt} = 1 | \alpha_{it} = \alpha_{lt}, \zeta_{jt}\right)\right]^{y_{ijt}} \left[1 - P\left(Y_{ijt} = 1 | \alpha_{it} = \alpha_{lt}, \zeta_{jt}\right)\right]^{1-y_{ijt}}. \tag{3}$$

Assuming item responses are independent over time given the learning trajectory of attribute profiles and item parameters implies that the likelihood for individual i is

$$p(\boldsymbol{Y}_i | \alpha_l, z) = \prod_{t=1}^{T} p(\boldsymbol{Y}_{it} | \alpha_{lt}, \zeta_t). \tag{4}$$

The posterior probability that $\boldsymbol{\alpha}_i = \boldsymbol{\alpha}_l$ given the observed responses and item parameters is proportional to the product of $p(\boldsymbol{Y}_i | \boldsymbol{\alpha}_l, \zeta)$ and $p(\boldsymbol{\alpha}_l)$, where $p(\boldsymbol{\alpha}_l)$ is the prior probability that individual i has learning trajectory l. Certainly, there are many choices for $p(\boldsymbol{\alpha}_l)$. The following discussion describes several models for $\boldsymbol{\alpha}_i$ that differ in terms of complexity and the number of parameters.

Table 1. Cardinality of Unrestricted and Nondecreasing Learning Trajectories for Values of K and T.

	$\lvert\mathcal{A}_+^T\rvert=(T+1)^K$				$\lvert\mathcal{A}^T\rvert=2^{KT}$			
T	$K=1$	$K=2$	$K=3$	$K=4$	$K=1$	$K=2$	$K=3$	$K=4$
1	2	4	8	16	2	4	8	16
2	3	9	27	81	4	16	64	256
3	4	16	64	256	8	64	512	4,096
4	5	25	125	625	16	256	4,096	65,536
5	6	36	216	1,296	32	1,024	32,768	1,048,576
6	7	49	343	2,401	64	4,096	262,144	16,777,216

Unstructured Trajectories

Consider the case where individuals can freely move through attribute classes over time (i.e., losing an attribute is possible). In this case, as noted above, there are 2^{KT} possible learning trajectories $\boldsymbol{\alpha}_l$. Let $\pi_l = P(\boldsymbol{\alpha}_i = \boldsymbol{\alpha}_l \in \mathcal{A}^T)$ denote the structural probability of having learning trajectory l characterized by starting at α_{l1}, proceeding through $\boldsymbol{\alpha}_{lt}$ for $1<t<T$, and terminating at $\boldsymbol{\alpha}_{lT}$. Let $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_{2^{KT}})'$ be the 2^{KT} vector of unstructured learning trajectory profiles, so the categorical prior for α_i is

$$p(\boldsymbol{\alpha}_i|\pi) = \prod_{l=1}^{2^{KT}} \pi_l^{\mathcal{I}(\boldsymbol{\alpha}_i=\boldsymbol{\alpha}_l)}. \tag{5}$$

The posterior probability that individual i's followed learning trajectory α_l conditioned upon the data, item parameters, and structural parameters in the unstructured case is

$$p(\boldsymbol{\alpha}_i|\boldsymbol{Y}_i, \boldsymbol{\zeta}, \pi) \propto p(\boldsymbol{Y}_i|\boldsymbol{\alpha}_l, \boldsymbol{\zeta})\left(\prod_{l=1}^{2^{KT}} \pi_l^{\mathcal{I}(\boldsymbol{\alpha}_i=\boldsymbol{\alpha}_l)}\right). \tag{6}$$

DINA Bayesian model formulation. The prior for the general unstructured model for learning trajectories can be included as a prior for any CDM Bayesian formulation. Consider the case of the DINA model for the item response function, $P(Y_{ijt}=1|\boldsymbol{\alpha}_{it}=\boldsymbol{\alpha}_{lt}, \boldsymbol{\zeta}_{jt})$. Also, let item parameters be invariant over time (i.e., $s_{jt}=s_j$ and $g_{jt}=g_j$), and suppose Q is known. A Bayesian model for estimating the general unstructured model for the DINA CDM is

$$Y_{ijt}\big|\boldsymbol{\alpha}_{it}, s_j, g_j \sim \text{Bernoulli}\left(\left(1-s_j\right)^{\eta_{ijt}} g_j^{1-\eta_{ijt}}\right),\ \eta_{ijt} = \prod_{k=1}^{K} \alpha_{ikt}^{q_{jk}}, \tag{7}$$

$$p(\alpha_i|\boldsymbol{\pi}) \propto \prod_{l=1}^{2^{KT}} \boldsymbol{\pi}_l^{\mathcal{I}(\alpha_i=\alpha_l)},\ 0<\pi_l<1,\ \sum_{l=1}^{2^{KT}} \pi_l = 1, \tag{8}$$

$$\boldsymbol{\pi} \sim \text{Dirichlet}(\boldsymbol{\delta}_0),\ \boldsymbol{\delta}_0 = \left(\delta_{01}, \ldots, \delta_{0,2^{KT}}\right), \tag{9}$$

$$p\left(s_j, g_j\right) \propto s_j^{\alpha_s-1}\left(1-s_j\right)^{\beta_s-1} g_j^{\alpha_g-1}\left(1-g_j\right)^{\beta_g-1}\mathcal{I}\left(0 \le g_j < 1-s_j \le 1\right). \tag{10}$$

Equation 9 is a conjugate Dirichlet prior for the unrestricted learning trajectory probabilities.

Nondecreasing Learning Trajectories

As noted above in Table 1, using the prior in Equation 6 would require substantial observations for even modest T and K to accurately recover parameters. One way to reduce the computational burden in educational contexts is to assume learning trajectories are nondecreasing. As noted above, there are $(T+1)^K$ possible nondecreasing attribute trajectories. One strategy is to specify the prior for α_i as categorical over the $(T+1)^K$ nondecreasing trajectories,

$$p(\boldsymbol{\alpha}_i|\boldsymbol{\pi}_+)=\prod_{l=1}^{(T+1)^K}\pi_{+l}^{\mathcal{I}(\alpha_i=\alpha_l)}\propto\mathcal{I}(\mathbf{0}_K\boldsymbol{\alpha}_{i1}\boldsymbol{\alpha}_{i2}\cdots\boldsymbol{\alpha}_{iT}\mathbf{1}_K)\prod_{l=1}^{2^{KT}}\boldsymbol{\pi}_l^{\mathcal{I}(\alpha_i=\alpha_l)}, \tag{11}$$

where $\pi_{+l}=P(\boldsymbol{\alpha}_i=\boldsymbol{\alpha}_l\in\mathcal{A}_+^T)$ is the probability of being in the nondecreasing attribute profile l and $\boldsymbol{\alpha}_{i,t-1}\boldsymbol{\alpha}_{it}$ if $\alpha_{ik,t-1}\le\alpha_{ikt}$ for $k=1,\ldots,K$. The posterior probability that individual i follows nondecreasing attribute trajectory α_l is

$$p(\boldsymbol{\alpha}_i|\boldsymbol{Y}_i,\boldsymbol{\zeta},\boldsymbol{\pi}_+)\propto p(\boldsymbol{Y}_i|\boldsymbol{\alpha}_l,\boldsymbol{\zeta})\left(\prod_{l=1}^{(T+1)^K}\pi_{+l}^{\mathcal{I}(\alpha_i=\alpha_l)}\right). \tag{12}$$

DINA Bayesian model formulation. Consider the DINA model item response function for the nondecreasing attribute trajectories with parameters for item j invariant over time (i.e., $s_{jt}=s_j$ and $g_{jt}=g_j$) and Q is known. A Bayesian model for estimating the nondecreasing model is obtained by updating the prior for $\boldsymbol{\alpha}_i$ by replacing Equation 6 with 11 as,

$$p(\boldsymbol{\alpha}_i|\boldsymbol{\pi})\propto\prod_{l=1}^{|\mathcal{A}_+^T|}\pi_{+l}^{\mathcal{I}(\alpha_i=\alpha_l)},\ 0<\pi_{+l}<1,\ \sum_{l=1}^{|\mathcal{A}_+^T|}\pi_{+l}=1, \tag{13}$$

$$p\sim\text{Dirichlet}(\boldsymbol{\delta}_0),\ \boldsymbol{\delta}_0=\left(\boldsymbol{\delta}_{01},\ldots,\boldsymbol{\delta}_{0,|\mathcal{A}_+^T|}\right). \tag{14}$$

FOHM Model

Table 1 shows that the cardinality for the set of unrestricted and nondecreasing learning trajectory sets may be too large for datasets observed in typical educational research studies. An alternative is to use a more parsimonious approximation to larger sets of the unrestricted and nondecreasing learning trajectories.

One approximation as employed in prior research (e.g., see Kaya & Leite, 2016, for an example with $t=2$) is to consider a hidden Markov model with first-order transition probabilities, $\boldsymbol{\Omega}$. That is, let π_1 denote initial class membership probabilities at time $t=1$ and let $\boldsymbol{\Omega}$ be a $2^K\times2^K$ matrix of first-order transition probabilities between classes. Specifically, the elements of $\boldsymbol{\Omega}$ denote the chance of transitioning from class c to c' between any two time periods. Let elements of $\boldsymbol{\Omega}$ be denoted as $\omega_{c'|c}=p(\boldsymbol{\alpha}_{it}=\boldsymbol{\alpha}_{c'}|\boldsymbol{\alpha}_{i,t-1}=\boldsymbol{\alpha}_c)$ for all t. The prior for α_i under an unrestricted FOHM model for $t\ge2$ is

$$\begin{aligned}p(\boldsymbol{\alpha}_i|\boldsymbol{\pi}_1,\boldsymbol{\Omega})&=p(\boldsymbol{\alpha}_{i1}|\boldsymbol{\pi}_1)\prod_{t=2}^{T}p(\boldsymbol{\alpha}_{it}|\boldsymbol{\alpha}_{i,t-1},\boldsymbol{\Omega})\\&=\left(\prod_{\alpha_c\in\mathcal{A}}\pi_{1c}^{\mathcal{I}(\alpha_{i1}=\alpha_c)}\right)\prod_{t=2}^{T}\left(\prod_{\alpha_{c'}\in\mathcal{A}}\omega_{c'|c}^{\mathcal{I}(\alpha_{it}=\alpha_{c'})}\right)^{\mathcal{I}(\alpha_{i,t-1}=\alpha_c)}.\end{aligned} \tag{15}$$

The unrestricted FOHM model requires significantly fewer parameters to be estimated than both $\mathcal{A}^T$ and $\mathcal{A}_+^T$. In fact, there are 2^K+4^K total parameters (i.e., 2^K elements in $\boldsymbol{\pi}_1$ and 4^K elements in $\boldsymbol{\Omega}$). A clear advantage of the FOHM model is that the model imposes some structure to reduce the parameters to a set that can be more easily recovered in applied educational settings. Furthermore, the number of parameters for the FOHM model does not grow with T. However, it should be emphasized that the FOHM model is an approximation and may not capture the true underlying nuance found in the unstructured and nondecreasing learning trajectories $\mathcal{A}^T$ and $\mathcal{A}_+^T$.

The full conditional distribution for $\boldsymbol{\alpha}_{it}$ depends upon t for the FOHM model. For attribute profiles at time t such that $1<t<T$, $\boldsymbol{\alpha}_{it}$ is conditioned upon the adjacent attributes $\boldsymbol{\alpha}_{i,t-1}$ and $\boldsymbol{\alpha}_{i,t+1}$. Accordingly, the full conditional probability that $\boldsymbol{\alpha}_{it}=\boldsymbol{\alpha}_{lt}$ given that $\boldsymbol{\alpha}_{i,t-1}=\boldsymbol{\alpha}_{c,t-1}$ and $\boldsymbol{\alpha}_{i,t+1}=\boldsymbol{\alpha}_{c',t+1}$ is

$$\begin{aligned} p(\boldsymbol{\alpha}_{lt}|\boldsymbol{Y}_i,\boldsymbol{\zeta},\boldsymbol{\Omega},\boldsymbol{\alpha}_{c,t-1},\boldsymbol{\alpha}_{c',t+1}) &\propto p(\boldsymbol{Y}_i|\boldsymbol{\alpha}_l,\boldsymbol{\zeta})p(\boldsymbol{\alpha}_{lt}|\boldsymbol{\alpha}_{c,t-1},\boldsymbol{\Omega})p(\boldsymbol{\alpha}_{c',t+1}|\boldsymbol{\alpha}_{lt},\boldsymbol{\Omega}) \\ &\propto p(\boldsymbol{Y}_{it}|\boldsymbol{\alpha}_{lt},\boldsymbol{\zeta}_t)\omega_{l|c}\ \omega_{c'|l} \end{aligned}, \quad (16)$$

where $\omega_{l|c}=p(\boldsymbol{\alpha}_{it}=\boldsymbol{\alpha}_l|\boldsymbol{\alpha}_{i,t-1}=\boldsymbol{\alpha}_c)$ and $\omega_{c'|l}=p(\boldsymbol{\alpha}_{i,t+1}=\boldsymbol{\alpha}_{c'}|\boldsymbol{\alpha}_{it}=\boldsymbol{\alpha}_l)$ are first-order transition probabilities. At time , the conditional probability that $\boldsymbol{\alpha}_{i1}=\boldsymbol{\alpha}_{l1}$ given the item responses and parameters and $\boldsymbol{\alpha}_{i2}=\boldsymbol{\alpha}_{c2}$ is

$$p(\boldsymbol{\alpha}_{l1}|\boldsymbol{Y}_i,\boldsymbol{\zeta},\boldsymbol{\Omega},\boldsymbol{\alpha}_{c2})\propto p(\boldsymbol{Y}_{i1}|\boldsymbol{\alpha}_{l1},\boldsymbol{\zeta}_1)\pi_l\ \omega_{c|l}, \quad (17)$$

where $\pi_l=p(\boldsymbol{\alpha}_{i1}=\boldsymbol{\alpha}_{l1})$ is the baseline prior probability. Similarly, for time $t=T$, the conditional probability that $\boldsymbol{\alpha}_{iT}=\boldsymbol{\alpha}_{lT}$ given $\boldsymbol{Y}_i$, $\boldsymbol{\zeta}$, and $\boldsymbol{\alpha}_{i,T-1}=\boldsymbol{\alpha}_{c,T-1}$ is

$$p(\boldsymbol{\alpha}_{lT}|\boldsymbol{Y}_i,\boldsymbol{\zeta},\boldsymbol{\Omega},\boldsymbol{\alpha}_{c,T-1})\propto p(\boldsymbol{Y}_{iT}|\boldsymbol{\alpha}_{lT},\boldsymbol{\zeta}_T)\omega_{l|c}. \quad (18)$$

Given the priors $p(\boldsymbol{\omega}_c)$ and $p(\boldsymbol{\pi}_1)$, the full conditional distribution for $\boldsymbol{\Omega}$ and $\boldsymbol{\pi}$ can be derived as follows:

$$\begin{aligned} p(\boldsymbol{\Omega},\boldsymbol{\pi}_1|\boldsymbol{\alpha},\boldsymbol{\zeta}) &\propto \prod_{l=1}^{2^K}\prod_{c=1}^{2^K}\left(p(\boldsymbol{\alpha}_t=\boldsymbol{\alpha}_c|\boldsymbol{\alpha}_{t-1}=\boldsymbol{\alpha}_l,\boldsymbol{\omega}_l)p\left(\omega_{c|l}\right)\right)^{\mathcal{I}(t>1)}\left(\pi_{1,l}p(\pi_{i,l})\right)^{\mathcal{I}(t=1)} \\ &\propto \prod_{l=1}^{2^K}\prod_{c=1}^{2^K}\omega_{c|l}^{\tilde{N}_{c|l}}\pi_{1,l}^{\tilde{N}_{0,l}}p(\boldsymbol{\omega}_l)p(\boldsymbol{\pi}_1) \\ &\propto \prod_{l=1}^{2^K}\prod_{c=1}^{2^K}\omega_{c|l}^{\tilde{N}_{c|l}+\delta_{l,c}}\pi_{1,l}^{\tilde{N}_{0,l}+\delta_{0,l}} \end{aligned}. \quad (19)$$

where $\tilde{N}_{c|l}$ is the number of subjects that have attribute profile $\boldsymbol{\alpha}_l$ at time t and change into $\boldsymbol{\alpha}_c$ at time $t+1$, $t=1,\ldots,T-1$, and $\tilde{N}_{0,l}$ is the number of subjects that initially have attribute profile $\boldsymbol{\alpha}_l$. Let Dirichlet distributions be the conjugate priors for $\boldsymbol{\pi}_1$ and each $\boldsymbol{\omega}_c$, $c=1,\ldots,2^K$, that is,

$$\boldsymbol{\pi}_1=\left(\pi_{1,1},\ldots,\pi_{1,2^K}\right)\sim\text{Dirichlet}(\boldsymbol{\delta}_0),\ \boldsymbol{\delta}_0=\left(\delta_{0,1},\ldots,\delta_{0,2^K}\right), \quad (20)$$

$$\boldsymbol{\omega}_c=\left(\omega_{1|c},\ldots,\omega_{2^K|c}\right)\sim\text{Dirichlet}(\boldsymbol{\delta}_c),\ \boldsymbol{\delta}_c=\left(\delta_{1|c},\ldots,\delta_{2^K|c}\right), \quad (21)$$

then as shown in Equation 19, the posterior for $\boldsymbol{\pi}_1$ will be Dirichlet($\delta_0+\tilde{\boldsymbol{N}}_0$), where $\tilde{\boldsymbol{N}}_0=(\tilde{N}_{0,1},\ldots,\tilde{N}_{0,2^K})$, and posteriors for $\boldsymbol{\omega}_c$ will be Dirichlet($\delta_c+\tilde{\boldsymbol{N}}_c$), where $\tilde{\boldsymbol{N}}_c=(\tilde{N}_{1|c},\ldots,\tilde{N}_{2^K|c})$.

FOHM Model, Nondecreasing Attributes

The previous subsection introduced the unrestricted FOHM model, and there may be substantive areas where it is reasonable to assume attributes are nondecreasing over time. A nondecreasing FOHM model can also be formulated that restricts transitions to nondecreasing states. Specifically, let $\boldsymbol{\Omega}_+$ include zeros for elements of $\boldsymbol{\Omega}$ that correspond with losing an attribute. In this case, the prior for $\boldsymbol{\alpha}_i$ becomes,

$$\begin{aligned} p(\boldsymbol{\alpha}_i|\boldsymbol{\pi}_1,\boldsymbol{\Omega}_+) &= p(\boldsymbol{\alpha}_{i1}|\boldsymbol{\pi}_1)\prod_{t=2}^{T} p(\boldsymbol{\alpha}_{it}|\boldsymbol{\alpha}_{i,t-1},\boldsymbol{\Omega}_+) \\ &\propto \prod_{\boldsymbol{\alpha}_c\in\mathcal{A}} \pi_{1c}^{\mathcal{I}(\boldsymbol{\alpha}_{i1}=\boldsymbol{\alpha}_c)} \prod_{t=2}^{T}\left(\prod_{\boldsymbol{\alpha}_{c'}\boldsymbol{\alpha}_{i,t-1}} \omega_{c'|c}^{\mathcal{I}(\boldsymbol{\alpha}_{it}=\boldsymbol{\alpha}_{c'})}\right)^{\mathcal{I}(\boldsymbol{\alpha}_{i,t-1}=\boldsymbol{\alpha}_c)}. \end{aligned} \tag{22}$$

The nondecreasing FOHM includes fewer parameters to estimate than the unrestricted FOHM model. In cases where attributes cannot be unlearned, it is possible to show that $\boldsymbol{\Omega}_+$ includes a total of 3^K transition probabilities. The reduction in parameters between the unrestricted and nondecreasing FOHMs is nontrivial. For $K=1,2,3,4$, $\boldsymbol{\Omega}$ for the unrestricted FOHM includes 4, 16, 64, and 256 elements, respectively, in comparison with 3, 9, 27, and 81 for $\boldsymbol{\Omega}_+$ of the nondecreasing FOHM.

The full conditional distributions for $\boldsymbol{\alpha}_{it}$ under the nondecreasing FOHM model have the same form as Equation 16 through 18 for the unrestricted FOHM model. The only modification is to recognize that some of the transition probabilities are zero.

Higher Order FOHM Model

A parsimonious alternative is to model transition probabilities for each attribute and introduce dependence in transition by conditioning on a higher order factor, θ. Another version of a higher order FOHM model specifies a prior for each α_{ikt} conditional on a higher order learning trait θ for the probability of transitioning and retaining an attribute,

$$\omega_{1|0,ik} = p(\alpha_{ikt}=1|\alpha_{ik,t-1}=0,\theta_i,\gamma_0), \tag{23}$$

$$\omega_{1|1,ik} = p(\alpha_{ikt}=1|\alpha_{ik,t-1}=1,\theta_i,\gamma_1), \tag{24}$$

where $\boldsymbol{\gamma}_0$ and $\boldsymbol{\gamma}_1$ are possibly vectors of parameters used to model the transition probabilities as a function of θ. It is important to recognize the role of θ in the higher order FOHM model. The higher order factor captures individual differences in learning rates and is included to model dependence among transitions in attributes over time.

One option for modeling the transition probabilities for attribute k is to, for example, set $\omega_{1|0,ik}=\Psi(\gamma_{00k}+\gamma_{01k}\theta_i)$, where $\Psi(\cdot)$ is a cumulative distribution function. In fact, Wang et al. (2016) provided an example of setting Ψ as a logistic function.

The distribution of α_{ikt}, given $\alpha_{ikt,t-1}$, θ, and the transition model parameters, is

$$p(\alpha_{ikt}|\alpha_{ik,t-1},\theta_i,\gamma_0,\gamma_1) = \left(\omega_{1|1,ik}^{\alpha_{ikt}}\left(1-\omega_{1|1,ik}\right)^{1-\alpha_{ikt}}\right)^{\alpha_{ik,t-1}}\left(\omega_{1|0,ik}^{\alpha_{ikt}}\left(1-\omega_{1|0,ik}\right)^{1-\alpha_{ikt}}\right)^{1-\alpha_{ik,t-1}}. \tag{25}$$

If attribute transitions are independent given θ, the prior probability for $\boldsymbol{\alpha}_i$ conditioned upon initial class membership $\boldsymbol{\pi}_1$ and transition parameters $\boldsymbol{\gamma}_0$ and $\boldsymbol{\gamma}_1$ is

$$p(\boldsymbol{\alpha}_i|\boldsymbol{\pi}_1, \boldsymbol{\gamma}_0, \boldsymbol{\gamma}_1) = \prod_{\boldsymbol{\alpha}_c \in \mathcal{A}} \pi_{1c}^{\mathcal{I}(\boldsymbol{\alpha}_{i1} = \boldsymbol{\alpha}_c)} \int \left(\prod_{t=2}^{T} \prod_{k=1}^{K} p(\alpha_{ikt}|\alpha_{ik,t-1}, \theta_i) \right) p(\theta_i) d\theta_i, \tag{26}$$

where $p(\theta_i)$ is a prior distribution for θ.

The aforementioned discussion of the higher order FOHM model allows for individuals to lose attributes when transition from on time to the next. The learning assumption can easily be enforced by setting $\omega_{1|1,ik} = 1$ and $\alpha_{ikt}\alpha_{ik,t-1}$ for all k and t.

Monte Carlo Simulation Study

Overview

This section reports Monte Carlo simulation studies evaluating the recovery of the transition matrix and item parameters based on the FOHM model for nondecreasing learning trajectories. The model by considering $K=4$ attributes, $T=5$ time points, and $J=50$ items with slipping parameters equal to 0.2 and guessing parameters equal to 0.3 is illustated. The transition matrix is set to be the same as the estimated transition probabilities from the real data of the next section (see Table 6). In this simulated test, 50 items are assigned into five blocks with 10 items in each block, and the Q-matrix in each block is complete (Chiu, Douglas, & Li, 2009) for model identification purposes. Wang et al. (2016) conjectured that the item positions may cause biased item parameter estimation because the students' latent attribute profiles are assumed to change with time while the DINA model parameters are static. In this case, the items in the later stage of the test may not receive the sufficient exposure to different latent classes, causing an inaccurate estimation for the corresponding model parameters. To investigate whether there is an order effect of the item positions to model parameter estimation, two test designs are considered. The first mimicked the experimental design in Wang et al. (2016) where the blocks of items are counterbalanced allowing each item to be positioned throughout all test stages to guarantee items will be exposed evenly at different time points, and it is referred as a ''balanced'' design from now on. The second was an unbalanced design where each examinee receives the same order of blocks during the test. For each test design, three levels of sample sizes $N=500$; 2,500; 5,000 were considered, and for each level, 50 independent datasets were generated to assess the performance. A Markov chain of 20,000 iterations with a burn-in of 10,000 iterations is used and the convergence of Gibbs samplers using the multivariate potential scale reduction factor (MPSRF) $\hat{R}$ (Brooks & Gelman, 1998) for chain length of 10,000 is evaluated. Examinees' initial attribute profiles were simulated from the multivariate normal threshold model (e.g., see Chiu et al., 2009), and both independent structure and dependent structure with $\rho=0.5$ of the initial attributes are considered in the simulation. The true distribution of initial attribute profiles is shown in Table 2.

Results

Parameter convergence was evaluated using the MPSRF based on five independent Markov chains of length 10,000 starting from random initial values. The results suggest the Markov chain converges to the posterior distribution after 4,000 iterations, which indicates that a chain length of 20,000 iterations with a burn-in of 10,000 is reasonable for this simulation study.

Figures 1 and 2 show that this method can accurately estimate the slipping and guessing parameters. When the sample size is small, results using the balanced design outperform the results under the unbalanced design, and the difference between the two test designs diminishes as sample size grows. Figure 3 presents the deviance of estimated probabilities for initial

Table 2. True Distribution of Initial Attribute Profiles.

α_c	0000	0001	0010	0011	0100	0101	0110	0111
$\rho=0$	0.038	0.010	0.026	0.006	0.058	0.014	0.038	0.010
$\rho=0.5$	0.122	0.002	0.012	0.001	0.044	0.003	0.013	0.002
α_c	1000	1001	1010	1011	1100	1101	1110	1111
$\rho=0$	0.154	0.038	0.102	0.026	0.230	0.058	0.154	0.038
$\rho=0.5$	0.178	0.013	0.058	0.012	0.194	0.044	0.179	0.122

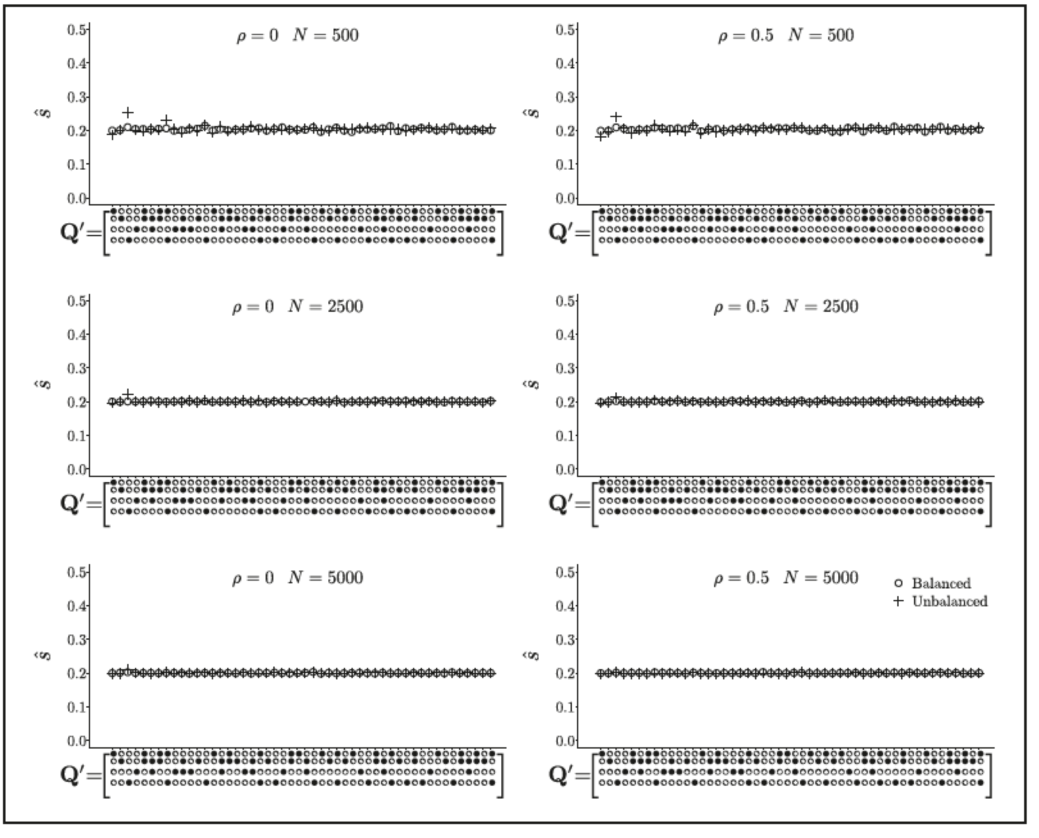

Figure 1. Estimated slipping parameters for different settings.
Note. Results are based on 50 replications.

attribute classes. The bias becomes smaller as the sample size grows, and results from the balanced design tend to have smaller bias than the unbalanced design. Figure 4 presents the deviance of estimated transition probabilities. There are 81 transition probabilities and the last one is always 1 because of the nondecreasing constraint. Again, the accuracy of the estimated transition probabilities improves with the increase of the sample size. In summary, the simulation results indicate it is better to use the counterbalanced design for relatively small sample sizes to ensure accurate recovery of the model parameters.

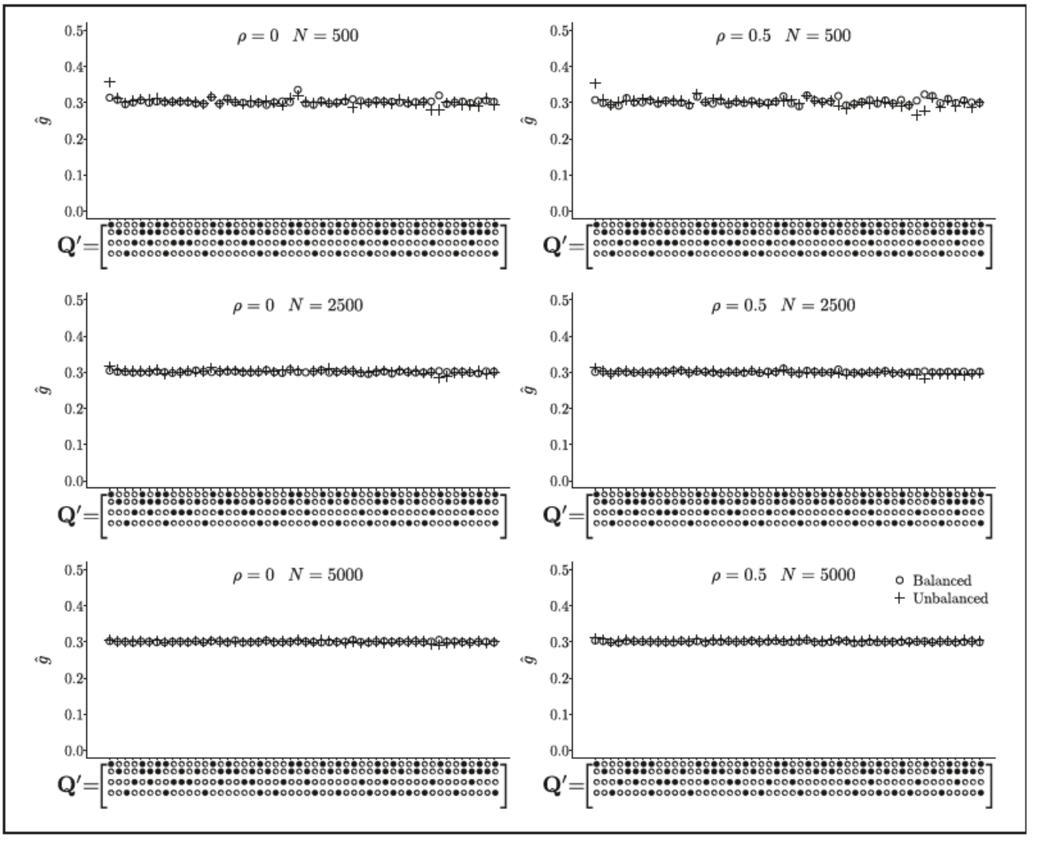

Figure 2. Estimated guessing parameters for different settings.
Note. Results are based on 50 replications.

Application to Spatial Rotation Skill Acquisition

In this section, the FOHM model with nondecreasing attributes assumption is applied to model students' learning of spatial skills. The spatial reasoning questions were developed based on the Revised Purdue Spatial Visualization Test (PSVT-R; Yoon, 2011). Wang et al. (2016) extended this test to incorporate a training tool which is designed to cause learning of rotation tasks. Specifically, participants first answered the questions in a test block then proceeded to a learning block in which they can receive feedback on their responses in the previous test block and learn the corresponding spatial rotation tasks through a learning intervention. The whole test consisted of five test blocks each containing 10 questions, followed by a learning intervention block. The items in the test block included a rotated object, and subjects were presented a new object and must determine which of the five options corresponded to the rotated version. Four mental rotation skills were identified: (a) 90° x axis, (b) 90° y axis, (c) 180° x axis, and (d) 180° y axis. All items included x and y axis rotations with objects of varying complexity. The 10 questions were assembled with similar structures but were different for each block. To eliminate the order effect on the estimation of the item parameters, a counterbalanced test condition was applied. Five versions of test were developed by rotating the five test blocks as the first block in the whole test process. During the experiment, those five tests were randomly assigned to participants to guarantee that different test blocks can have an equal chance to be the first

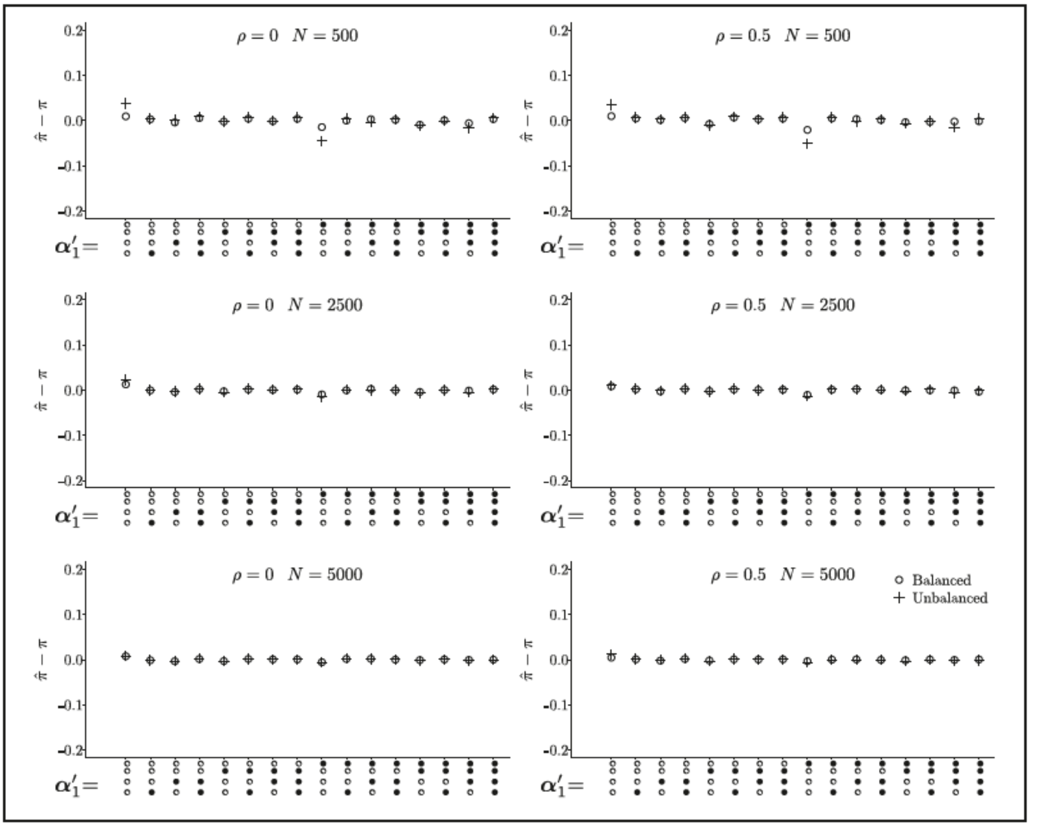

Figure 3. Deviance of estimated prior for initial attribute profile class probabilities.

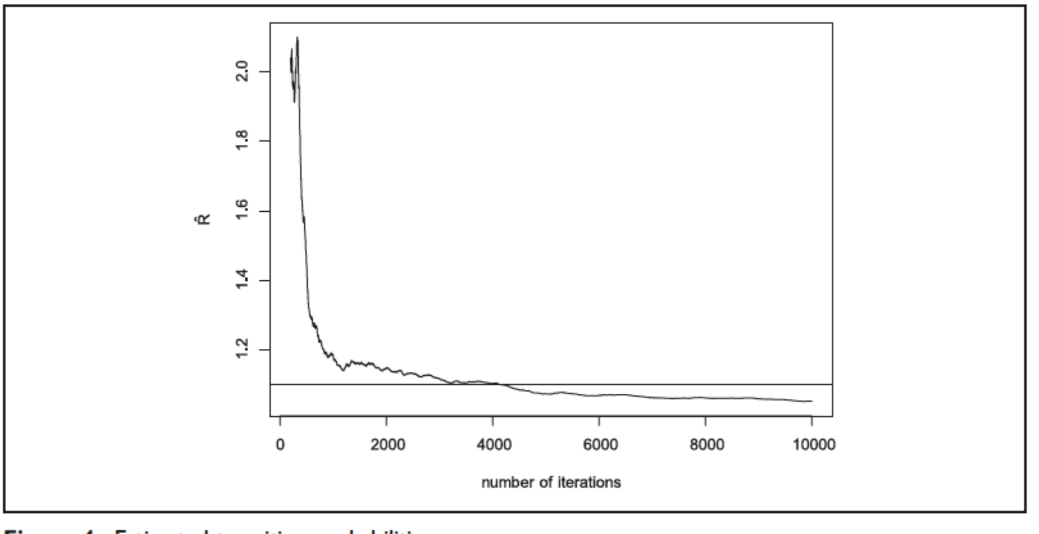

Figure 4. Estimated transition probabilities.

block among the participants. The detailed experiment design can be referred to Wang et al. (2016).

Responses from 351 individuals to those 50 questions were collected from this experiment, and the FOHM model with nondecreasing attributes assumption and DINA Bayesian Model formulation was applied to estimate students' learning trajectories. The five test blocks in the test represented five different time points. Based on previous simulation results, the chain of 30,000 iterations with a burn-in of 15,000 is used to ensure the convergence of the Markov chain.

Table 3 documents the estimates of the DINA item parameters. The results are quite consistent with the estimates from Wang et al. (2016). That is, most of the items have relatively large guessing parameters. This result might be due to the simple shapes for some items and the many distractors that can be easily eliminated from consideration. Students' learning trajectories in terms of the estimated mastery rate for each of the four skills over time, the frequencies of the number of mastered skills at each time point, and the transition matrix are summarized, which are presented in Tables 4 to 6, respectively. From those results, an increase of mastery rates for each skill with time and the transition pattern of different latent classes at each time point are observed.

The transition probability matrix explicitly gives the probabilities of remaining in the same stage or learning more skills from time t to $t+1$, which reveals the learning pattern of current mastered skills. The learning process can be inferred by interpreting the entries in the estimated transition matrix. For example, in Table 6, the ninth row of the estimated transition matrix presents the transition probability for current state of attribute profile $(1, 0, 0, 0)$ to another latent class. It indicates that if a student has mastered a single skill on $90°$ x axis, he or she tends to master more skills next time with probability around 0.87, and the most likely skill to be mastered next is $90°$ y axis or $180°$ x axis. On the contrary, the fifth row shows the transition probability for current state of acquisition on the single skill of $90°$ y axis. The chance for mastering more skills next time is around 0.67, which is lower compared with mastery of $90°$ x axis, and the next mastered skill with this current latent state will most likely to be $180°$ y axis. The second and third rows also imply that students who have mastered skills on $180°$ x axis or y axis will most likely master $90°$ x axis or y axis next time. This in fact also indicates certain hierarchical structure among those four latent skills, mastering the $90°$ might be prerequisite for mastering $180°$.

The distribution of the initial latent classes at Time Point 1 in the transition matrix is slightly different from those estimated from Wang et al. (2016). This might be due to the different assumptions applied to the transition matrix. The FOHM model assumes the time-homogeneity of the transition matrix, that is, the transition probabilities of different latent classes are the same at different points of time, and are the same for each individual. However, the transition probabilities in Wang et al. (2016) depend on different covariates, which could make it different for different individuals and at different points of time. In this sense, FOHM model with nondecreasing learning trajectories assumption is a reduced model from Wang et al. (2016), and it is worth investigating the model comparison in the future to select the most appropriate model for a given data set.

Discussion

Online education is becoming ubiquitous, and when coupled with intelligent tutoring systems and sequenced training modules, it provides an opportunity to systematically train subjects on a long list of prespecified attributes while keeping careful track of progress. This is an ideal setting for utilizing latent class models for cognitive diagnosis. However, the static models that have typically been considered are not adequate when training and learning are the goals, which

Table 3. Item Parameter Estimates for Spatial Rotation Data.

Item	Parameters	
	$\hat{s}_j$	$\hat{g}_j$
1	0.023	0.811
2	0.037	0.749
3	0.109	0.562
4	0.027	0.618
5	0.136	0.487
6	0.519	0.229
7	0.058	0.601
8	0.159	0.425
9	0.017	0.848
10	0.140	0.578
11	0.054	0.620
12	0.056	0.754
13	0.027	0.527
14	0.040	0.819
15	0.181	0.315
16	0.122	0.547
17	0.112	0.324
18	0.056	0.639
19	0.155	0.325
20	0.132	0.370
21	0.059	0.645
22	0.063	0.533
23	0.115	0.446
24	0.117	0.433
25	0.018	0.805
26	0.174	0.323
27	0.176	0.553
28	0.018	0.883
29	0.041	0.794
30	0.259	0.357
31	0.098	0.461
32	0.031	0.522
33	0.027	0.723
34	0.151	0.613
35	0.123	0.384
36	0.141	0.320
37	0.379	0.327
38	0.056	0.751
39	0.165	0.345
40	0.135	0.500
41	0.055	0.569
42	0.110	0.468
43	0.056	0.785
44	0.027	0.659
45	0.123	0.461
46	0.059	0.468
47	0.230	0.287
48	0.447	0.301
49	0.018	0.727
50	0.069	0.479

Note. Results are based on the chain of length 30,000 and burn-in 15,000.

Table 4. Skill Mastery Rate Over Time.

Skills	Time Point 1	Time Point 2	Time Point 3	Time Point 4	Time Point 5
90° x axis	0.530	0.607	0.632	0.652	0.670
90° y axis	0.618	0.678	0.712	0.749	0.781
180° x axis	0.473	0.536	0.575	0.621	0.670
180° y axis	0.459	0.521	0.547	0.601	0.661

Note. Results are based on the chain of length 30,000 and burn-in 15,000.

Table 5. Number of Skill Mastery Over Time.

Number of skills	Time Point 1	Time Point 2	Time Point 3	Time Point 4	Time Point 5
0	121	99	93	81	70
1	37	28	23	23	23
2	33	41	36	32	26
3	13	20	25	26	27
4	147	163	174	189	205

Note. Results are based on the chain of length 30,000 and burn-in 15,000.

suggests that latent class models for learning in cognitive diagnosis will be a fruitful and fertile area for research and applications. Here, some of the possibilities for modeling learning trajectories in cognitive diagnosis, starting with the most general model with 2^{KT} parameters, and then a more parsimonious model for nondecreasing trajectories with $(T+1)^K$ parameters, are laid out. In most applications, this will be too many to calibrate, and the authors further restrict the model to arrive at the FOHM model, with 3^K parameters to estimate in the case of nondecreasing trajectories. This model is quite flexible, and is practical when large samples are available, when K is relatively small, or when a long list of attributes can be partitioned into small subgroups of attributes that are studied together and can be modeled separately with distinct FOHMs. To further reduce the parameters of transition models for learning, covariates may be used, and this is a wide-open area for future research. Covariates may involve demographic variables, measures of the amount of practice one has done, dummy variables for interventions, and many other predictors of learning. To address the heterogeneity of learning rates across subjects in a population, continuous random effects may be used, which can be interpreted as latent variables representing the construct of general learning ability as studied by Wang et al. (2016). Measuring such a construct may be of intrinsic interest, beyond simply serving to address the heterogeneity.

The directions for extending static CDMs to include the dynamics of learning are numerous, but should not outrace the empirical data on learning by much. There is a need for more applied research to gather data applicable to learning research, which will afford the opportunity to investigate the particular models and covariates for learning that provide the best fits and yield the most accurate and useful predictions. The spatial reasoning example in the real data study was gathered carefully in a laboratory according to a designed study, and illustrates how a model such as the FOHM model can be used in practice. The accuracy of learning transition probability estimates was limited by the modest sample size. However, the Markov Chain Monte Carlo (MCMC) algorithm ran efficiently with these data and indicated convergence to the posterior distribution of the parameters. The model complexity of the FOHM model is

Table 6. Estimated Prior for Starting Attributes and Transition Matrix.

	α_c	$\hat{\pi}_l$	$\hat{\omega}_{1\|i}$	$\hat{\omega}_{2\|i}$	$\hat{\omega}_{3\|i}$	$\hat{\omega}_{4\|i}$	$\hat{\omega}_{5\|i}$	$\hat{\omega}_{6\|i}$	$\hat{\omega}_{7\|i}$	$\hat{\omega}_{8\|i}$	$\hat{\omega}_{9\|i}$	$\hat{\omega}_{10\|i}$	$\hat{\omega}_{11\|i}$	$\hat{\omega}_{12\|i}$	$\hat{\omega}_{13\|i}$	$\hat{\omega}_{14\|i}$	$\hat{\omega}_{15\|i}$	$\hat{\omega}_{16\|i}$
1	0000	0.31	0.84	0.01	0.01	0.01	0.02	0.02	0.01	0.01	0.01	0.01	0.02	0.01	0.02	0.01	0.01	0.01
2	0001	0.02		0.13		0.11		0.18		0.10		0.11		0.12		0.12		0.13
3	0010	0.01			0.12	0.11			0.13	0.13			0.16	0.11			0.12	0.11
4	0011	0.01				0.31				0.25				0.21				0.23
5	0100	0.05					0.37	0.15	0.11	0.08					0.09	0.07	0.07	0.07
6	0101	0.02						0.72		0.12						0.09		0.06
7	0110	0.02							0.50	0.17							0.18	0.15
8	0111	0.01								0.78								0.22
9	1000	0.03									0.13	0.09	0.17	0.10	0.19	0.10	0.12	0.10
10	1001	0.01										0.22		0.25		0.29		0.24
11	1010	0.02											0.73	0.08			0.12	0.07
12	1011	0.02												0.43				0.57
13	1100	0.05													0.72	0.08	0.09	0.12
14	1101	0.02														0.47		0.53
15	1110	0.03															0.55	0.45
16	1111	0.38																1.00

Note. Results are based on the chain of length 30,000 with a burn-in of 15,000. The skills represented by α_c are labeled in the order of 90^0 *x* axis, 90^0 *y* axis, 180^0 *x* axis and 180^0 *y* axis. $\omega_{j|i}$ refers to the transition probability from the *i*th attribute profile to the *j*th attribute profile.

difficult to compare with the higher order model with covariates of Wang et al. (2016). The number of structural parameters is much higher in the FOHM model, but the higher order model includes a latent learning ability parameter for each individual, resulting in more actual transition probabilities. Nevertheless, these two models yielded results that were relatively consistent.

A common aspect of the FOHM model and the models of Wang et al. (2016) is that of non-decreasing trajectories. Yet another direction to consider in latent class models for learning is how to build relatively parsimonious models that allow for traits to be learned and unlearned. This modeling of retention may prove especially useful in long range longitudinal studies more than in the relatively short time duration of the spatial reasoning training and assessment.

In this study, it is assumed that the item parameters were invariant over time. It is unclear whether fixing item parameters over time is a necessary condition for identifying the model parameters. Additional theoretical research is needed to establish general model identifiability conditions.

Models for learning will necessarily be more complicated than static models, because they combine the measurement model with a transition model. Computational aspects of fitting these models will be a critical area of research. In the simulation study and real data analysis, the DINA model was used. When attributes are so clearly defined, such as the particular rotation operations in the data example that must be applied conjunctively, the DINA model can fit quite adequately as shown by Wang et al. (2016). However, in applications with less clear attributes, perhaps leading to some Q-matrix misspecification, more general CDMs will prove useful. Coupling the FOHM model or another learning trajectory model with a more general measurement model (e.g., see Henson, Templin, & Willse, 2009; G. Xu, 2016) can present computational challenges and the need for research and software development. Another area of research that will play a vital role is assessment of model fit. The dynamic aspect of this model presents another dimension of fit assessment than is required by static models, and techniques that target particular aspects of misfit will be more useful than global measures of fit. The general method of posterior predictive checking (Sinharay, 2006; Sinharay & Almond, 2007; Sinharay, Johnson, & Stern, 2006) allows one to study residuals that pertain to a particular aspect of fit, such as the measurement model or the transition model, and there is a need for research into the residuals that best evaluate the fit of each. As more data are collected through online learning, new questions and applications will arrive that will require additional methodological research in this emerging area of psychometrics.

Authors' Note

Opinions reflect those of the author(s) and do not necessarily reflect those of the granting agency.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported by a grant from the National Science Foundation Methodology, Measurement, and Statistics program grant #1632023.

References

Brooks, S. P., & Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, *7*, 434-455.

Chiu, C.-Y., Douglas, J. A., & Li, X. (2009). Cluster analysis for cognitive diagnosis: Theory and applications. *Psychometrika*, *74*, 633-665.

Corbett, A. T., & Anderson, J. R. (1994). Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction*, *4*, 253-278.

González-Brenes, J. P., Huang, Y., & Brusilovsky, P. (2014). General features in knowledge tracing to model multiple subskills, temporal item response theory, and expert knowledge. In The 7th International Conference on Educational Data Mining (pp. 84-91). Pittsburgh, PA: University of Pittsburgh.

González-Brenes, J. P., & Mostow, J. (2013). What and when do students learn? Fully data-driven joint estimation of cognitive and student models. In *Proceedings of the 6th International Conference on Educational Data Mining* (pp. 236-240). Pittsburgh, PA: University of Pittsburgh.

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

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

Kaya, Y., & Leite, W. L. (2016). Assessing change in latent skills across time with longitudinal cognitive diagnosis modeling an evaluation of model performance. *Educational and Psychological Measurement*, *77*, 369-388.

Li, F., Cohen, A., Bottge, B., & Templin, J. (2015). A latent transition analysis model for assessing change in cognitive skills. *Educational and Psychological Measurement*, *76*, 181-204.

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

Pardos, Z. A., & Heffernan, N. T. (2010). Modeling individualization in a Bayesian networks implementation of knowledge tracing. In Proceedings of the 18th International Conference on User Modeling, Adaptation and Personalization (pp. 255-266). Big Island, Hawaii.

Sinharay, S. (2006). Model diagnostics for Bayesian networks. *Journal of Educational and Behavioral Statistics*, *31*, 1-33.

Sinharay, S., & Almond, R. G. (2007). Assessing fit of cognitive diagnostic models a case study. *Educational and Psychological Measurement*, *67*, 239-257.

Sinharay, S., Johnson, M. S., & Stern, H. S. (2006). Posterior predictive assessment of item response theory models. *Applied Psychological Measurement*, *30*, 298-321.

Studer, C. (2012). *Incorporating learning over time into the cognitive assessment framework* (Unpublished doctoral dissertation). Carnegie Mellon University, Pittsburgh, PA.

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

Wang, S., Yang, Y., Culpepper, S., & Douglas, J. (2016). *Tracking skill acquisition with cognitive diagnosis models: Application to spatial rotation skills*. Manuscript submitted for publication.

Xu, G. (2016). Identifiability of restricted latent class models with binary response. *Annals of Statistics*, *45*, 675-707.

Xu, Y., & Mostow, J. (2012). Comparison of methods to trace multiple subskills: Is LR-DBN best? In *Proceedings of the 5th International Conference on Educational Data Mining*. (pp. 41-48)

Ye, S., Fellouris, G., Culpepper, S., & Douglas, J. (2016). Sequential detection of learning in cognitive diagnosis. *British Journal of Mathematical and Statistical Psychology*, *69*, 139-158.

Yoon, S. Y. (2011). *Psychometric properties of the revised Purdue Spatial Visualization Tests: Visualization of rotations (the revised PSVT-R)* (Unpublished doctoral dissertation). Purdue University, West Lafayette, IN.