Behavior Research Methods
https://doi.org/10.3758/s13428-022-02041-w

Flow with an intelligent tutor: A latent variable modeling approach to tracking flow during artificial tutoring

Hyeon-Ah Kang[1] · **Adam Sales**[2] · **Tiffany A. Whittaker**[1]

Accepted: 30 November 2022

Abstract
Increasing use of intelligent tutoring systems in education calls for analytic methods that can unravel students' learning behaviors. In this study, we explore a latent variable modeling approach for tracking learning flow during computer-interactive artificial tutoring. The study considers three models that give discrete profiles of a latent process: the (i) latent class model, (ii) latent transition model, and (iii) hidden Markov model. We illustrate application of each model using example log data from Cognitive Tutor Algebra I and suggest analytic procedures of drawing learning flow. Through experimental application, we show that the models can reveal substantive information about students' learning behaviors and have potential utility for describing the learning flow. The models differed in the assumptions and data constraints but yielded consistent findings on the flow states and interaction modalities. Based on our experiential analyses, we discuss strengths and limitations of the models and illuminate areas of future development.

Keywords Intelligent tutoring · Flow · Latent transition analysis · Cognitive Tutor Algebra

Introduction

An intelligent tutoring system (ITS; e.g., ALEKS, ASSISTments, AutoTutor, Easy with Eve, MATHia, MetaTutor, SQL-Tutor) is educational software that provides computerized tutoring. The system employs an artificial intelligent tutor to guide a student through problem sets and provide customized feedback. In ITS, the tutor directly interacts with a student to perform tutoring activities. Since the instruction is mostly achieved by direction interaction between a tutee and a tutor, minimal help is needed from human teachers and it enables cost-effective large-scale individualized learning.

One of the important considerations in implementing ITS is whether a student adequately follows through learning activities with continued attention and engagement. As ITS is typically administered in a self-regulated environment, students can divert from active learning when they encounter challenges. For example, if a student is presented with tasks that exceedingly challenge his skill levels, the student can become frustrated and demotivated to learn new skills. If assigned tasks are too easy and require little effort, a student can also diverge from genuine learning and show effortless behaviors. Examining student's interaction behaviors in these settings can help understand student's learning process and the functioning of ITS. Since many ITS programs are designed to keep students engaged and flow in learning, any distinctive behaviors that deviate from the normal operation would indicate emergence of nonoptimal learning and ill-functioning of ITS.

The purpose of this study is to explore statistical models that can describe students' learning behaviors during artificial tutoring and draw information that can help future ITS refinement and intervention planning. We in particular examine behaviors that illuminate students' learning flow. Flow (Csíkszentmihályi, 1990) is a mental state learners experience when immersed in deep learning. A student in a flow state shows high engagement with the learning activities and tends to gain positive learning outcomes. In ITS, the inbuilt design makes the flow a highly achievable state. Many ITS programs customize tutoring activities to students' skill levels and learning progression, and they generally expect students to flow while learning with the tutor. Modeling students' flow states in this setting can help

✉ Hyeon-Ah Kang
hkang@austin.utexas.edu

[1] University of Texas at Austin, Austin, TX, USA

[2] Worcester Polytechnic Institute, Worcester, MA, USA

Published online: 07 February 2023

understand the students' learning process and when students become subject to suboptimal learning.

For modeling learning flow, we apply latent variable models that give discrete profiles of a latent process. Three models are considered for application: the (i) latent class model (LCM), (ii) latent transition model (LTM), and (iii) hidden Markov model (HMM). These models allow analysis of large-scale multivariate time-series data and can describe systematic effects of contextual variables (e.g., problem effects, student covariate effects). The models differ in the specific ways of characterizing the variables (e.g., permissible indicators, latent state transition, covariate effects) and constraints of estimation software. In this study, we suggest practical strategies of applying the models, addressing the related assumptions and constraints. We show how each model can be applied to accommodate distinct characteristics of the ITS data and draw information relevant to learning flow.

To demonstrate the application, we employ example data from Cognitive Tutor Algebra (CTA).[1] Using log data from a particular time period, we show an analytic process of drawing learning flow. We give a didactic demonstration of data preparation, model formulation, and estimation, and show that the outcomes of the models reveal substantive information about students' learning process. Based on our experimental analysis, we discuss strengths and limitations of the models in describing the ITS data and illuminate areas of future consideration.

The rest of this article is organized as follows. In "Cognitive Tutor Algebra", we introduce CTA and present basic information about the evaluation data. We discuss characteristics of the raw CTA data, arrangements needed for analysis, and analysis steps that apply the latent variable models in phases. Sections "Profiling flow"–"Flow across problems" present specific analyses performed under each model. We discuss model formulation, data preparation, model fitting, and corresponding results. Section "Conclusion" concludes with a summary of the findings and future considerations.

Cognitive Tutor Algebra

Data

The study used Cognitive Tutor Algebra (CTA) I to illustrate the application of latent variable models in the ITS data. The example evaluation data were collected in 2007-08 as a part of an effectiveness study (Pane et al., 2014). The raw data contained observations from N = 2860 students that received tutoring between July 2007 and May 2008. The tutoring was offered during regular curricula under the supervision of teachers. The system contained a total of 637 problems across 106 sections that are nested within 27 units (e.g., algebra level 1, level 2; equation solver level 1, level 2). Across the study period, students received on average 276.826 problems (SD = 193.642), 43.900 sections (SD = 32.523), and 8.927 units (SD = 6.746). The specific problem sets and the order of problems differed by students, teachers, and school districts. Most of the problems were prompted by an artificial tutor following the student's skill mastery, but teachers could reassign students to different sections and the system could also promote students to a new section if a student reaches a maximum number of problems.

Preparation

The raw interaction data bear a number of complications for applying the latent variable models. Since the system administered problems differently according to the students' skill levels, the assigned problems will induce between-subject variance in the evaluation data. In addition, since the tutoring was offered in multiple sessions over a year, the interaction data will exhibit large temporal variance in the students' flow progression. For examining learning flow, it is necessary to reform the raw data and regulate undesired variance.

Our strategy for regulating the variance in this study was to choose one problem unit and examine students' workings on single days. Fixating on one problem unit helps regulate excessive measurement noninvariance. Limiting tutoring times to single days helps reduce temporal variance and dimensionality of latent states. Specifically, we chose an elementary problem unit, *equation solver level* 1(*es*1 hereinafter), and examined the interaction data that were collected on the single days. The *es*1 problems showed the most homogeneous measurement properties (see Appendix A) and it was reckoned that they would induce minimal variance in the indicator variables.

The data extraction was achieved as follows. If a student worked on *es*1 on multiple days, we picked the day the student attempted most problems and examined the student's flow development during the day. Similarly, if a student worked on multiple units on the same day, only the observations from *es*1 were examined to regulate the variance from the other units and problems. We note that, although we carefully prepared the data to exhibit homogeneous measurement properties, we also additionally addressed the measurement noninvariance when models allow modification (e.g., random effect).

[1] CTA is a predecessor of MATHia. We retain the name of CTA since the data were obtained from the 2007–2008 CTA administration (Pane et al., 2014).

Variables

Applying the above strategy led to subset data of N = 2219 students. The students in the final data attempted 50.236 problems on average (SD = 16.102) with a minimum of four and maximum of 151. For evaluating flow, we examined three indicator variables: the interaction time, the number of erroneous attempts, and the number of hints requested. Each indicator variable was transformed to meet the constraints of calibration programs. For example, the timing variable was placed on the log metric to approximate normality. The count variables were used as observed or categorized into three ordinal categories (i.e., none, one, and more than one) and modeled by Poisson or proportional odds models. Along with the interaction indicators, we also made use of student-level covariates when inferring the state membership and transition behaviors. The covariates used include: Pre- and Gain test scores on the standardized test, Race (0 = White and Asian, 1 = Black and multiethnic, 2 = Hispanic and native America), Sex (0 = Female, 1 = Male), and whether a student was enrolled in a free lunch program (0 = No, 1 = Yes).

Analysis

As the cleaned data were obtained as above, we performed analysis in three stages as follows. We first conducted latent class analysis to examine heterogeneity in the interaction data and investigated if the identified heterogeneity can be characterized as distinct latent classes of in- and out-of-flow. Based on the findings from the latent class analysis, we then performed latent transition analysis to track progression of latent states across different tutoring stages. In both analyses, we applied sample-level data (i.e., data for 2219 students) to account for effects of contextual variables (e.g., populational characteristics, problem effects). The last stage analysis was performed on the individual student-level data (i.e., each student's interaction data) using hidden Markov models. Unlike LCM and LTM, which require modification of data to reduce the event times, HMM can model intensive time series and requires minimal data transformation. The third-stage analysis drew on this flexibility and applied HMMs to examine students' learning progression over individual problem-solvings. Since CTA customized problem assignments to each student's skill levels, we surmised that the problems would exhibit weak measurement invariance if conditioned at the student level. We exploited this assumption to track student's learning progression across individual problems.

Table 1 summarizes the analyses performed in each stage. Each analysis was carefully designed to address the assumptions of a model, constraints of a calibration program, and the characteristics of the CTA data. It is important to mention that, across the analyses, we applied the models assuming a small number of latent states. Since our study was mainly interested in modeling discrete flow states, we focused more on the stability of the extracted states, indicator modalities under each state, and the evolution of latent states over time. Our supplementary analysis suggests that allowing more states tends to result in the disintegration of the normal flow state, characterizing different problem-solving strategies. Although unraveling the flow state can help learn different working processes, the identified features generally require subjective interpretation and are difficult to validate beyond face validity. We therefore limit our attention to clear

Table 1 Analysis settings

Analysis	Model	Data	Variable	Measurement invariance
Flow profiles	Random-effect LCM	Sample-level data, aggregated by knowledge components	log time, nhint and nerr on the ordinal scale (0, 1, 2)	modeled through random-effect terms in the structural and measurement models
Flow transition over tutoring stages	Random-intercept LTM	Sample-level data, aggregated for tutoring stages	log time, nhint and nerr on the ordinal scale (0, 1, 2)	modeled through random-intercept latent factors in the measurement model
Flow transition across problems	HMM	Student-level data with problem-level observations	log time, nhint, nerr	assumed weak measurement invariance and addressed possible impact of measurement noninvariance

Note. time = Task interaction time. nhint = Number of hints requested. nerr = Number of erroneous attempts. Throughout the analyses, the same set of covariates were used: Pre/Gain scores on a standardized test, Race, Sex, Enrollment on the free lunch program

bimodality of flow—in- and out-of-flow. In the following sections, we discuss specific analysis conducted under each model, including the model formulation, analytic strategies for CTA, and corresponding results.

Profiling flow

Throughout the article, we apply following notations to describe models. Let $\boldsymbol{S}_i = (S_{it} : t = 0, \ldots, T)$ denote a sequence of flow states that student i went through over T measurement events.[2] The flow state, S_{it}, takes a nominal value from a finite latent space, $S_{it} \in \mathcal{S} \equiv \{1, \ldots, M\}$, and is manifested by a set of indicator variables, $\boldsymbol{Y}_{it} = (Y_{ijt} : j = 1, \ldots, J)$, where j indexes different kinds of indicators (e.g., interaction time, behavioral frequency). A collection of indicator outcomes over time, $\boldsymbol{Y}_i = (\boldsymbol{Y}_{it} : t = 1, \ldots, T)$, then forms multivariate cross-sectional time series and mirrors trajectory of flow states over time. In the event that student's background variables are available (e.g., gender, ethnicity), the covariates, $\boldsymbol{Z}_i = (Z_{ik} : k = 1, \ldots, K)$, can be used when drawing the state profiles. For simplicity, the study assumes time-invariant covariates that remain constant over time; the analyses suggested below however can be easily extended to accommodate the time-variant covariates.

Random-effect latent class model

The first stage analysis applied latent class models to examine heterogeneity in the interaction data. We obtained evaluation data as cross-sectional observations at each time point, $\boldsymbol{Y}_t = (Y_{ijt} : i = 1, \ldots, N;\ j = 1, \ldots, J)$, and examined students' workings at each measurement time. Let $\boldsymbol{Y}_t = (Y_{ijt} : i = 1, \ldots, N;\ j = 1, \ldots, J)$ denote the sample cross-sectional data observed at measurement time t. The LCM then profiles students' latent states based on the homogeneity of the observed interaction patterns. A sequence of states identified over time, $\boldsymbol{S}_i = (S_{it} : t = 0, \ldots, T)$, gives a latent profile of a student i and defines a unique latent class.

The formulation of LCM involves two sub-models: (i) a structural model that describes the probability of a latent state, $P(S_{it} = m)$ $(m = 1, \ldots, M)$, and (ii) a measurement model that describes the probability of indicators given a latent state, $P(\boldsymbol{Y}_{it} \mid S_{it} = m)$. Unlike the other two models discussed below, LCM does not model transition of latent states. It instead assumes that latent states evolve independently over time and it models the sequence of interaction patterns through a distinct latent class profile.

The LCM for CTA was formulated as follows. The structural model that describes the probability of a latent state was formulated on the multinomial logistic regression. In regular settings, the model is parameterized assuming within-class homogeneity. That is, students within the same class are expected to give homogeneous performance when solving problems. In CTA (and in many cases of ITS), this assumption is generally not tenable because students receive different problems according to their skill levels and progression. The difference in the problem assignments induces extra variance in the outcome data and the observed data become no longer independent when conditioned on the latent states. A common approach to addressing this extra variance is to allow random variation when modeling the outcome probabilities (e.g., Qu et al. 1996). That is, a random-effect term is added to the structural model so that the heterogeneity in the outcome probabilities can be explained by both the latent state and the random effect.

In the present setting, the random-effect LCM can be formulated as follows. Let η_i model the idiosyncratic effect of student i on the state probability. The probability of a latent state m can be then modeled as

$$P(S_{it} = m \mid \eta_i,\ \boldsymbol{Z}_i = \boldsymbol{z}_i) = \frac{\exp\left(\beta_{m0} + \boldsymbol{\beta}_m^\top \boldsymbol{z}_i + \eta_i\right)}{\sum_{l=1}^{M} \exp\left(\beta_{l0} + \boldsymbol{\beta}_l^\top \boldsymbol{z}_i + \eta_i\right)} \tag{1}$$

for all t $(= 1, \ldots, T)$. The β_{m0} is an intercept parameter that determines the conditional probability of the latent state m when $\boldsymbol{Z}_i = \boldsymbol{0}$ and $\eta_i = 0$. The slope parameter β_m models the effects of the covariates, $\boldsymbol{Z}_i$, on the logit. To ensure identifiability of the model parameters, we assume that $\eta_i \sim \mathcal{N}(0,\ \sigma_\eta^2)$ with $\sigma_\eta^2 = \mathrm{Var}(\eta_i;\ i = 1, \ldots, N)$ modeling the magnitude of extra variance from the individuals (i.e., beyond covariate effects).

The measurement model can be formulated in a similar way. Since students in ITS typically receive different problem sets that differ in the measurement properties, we again introduce problem-level random effects when modeling the indicator variables. A measurement model for an indicator j is parameterized as

$$P_j(Y_{ijt} \mid S_{it} = m,\ \delta_{jt}) = f_j(\boldsymbol{\psi}_{jm},\ \delta_{jt}), \tag{2}$$

with f_j specifying the functional form of the probability measure of the indicator j and $\boldsymbol{\psi}_{jm}$ giving the parameters of f_j (e.g., location, scale). The δ_{jt} models the idiosyncratic effect of problem t on the indicator j. The functional form

[2] In ITS the event time corresponds to the problems or problem sets. Also note that students can receive different problems and T can vary by students. The study defines T as $\max(T_i : i = 1, \ldots, N)$ and uses as a generic notation for the event length.

of f_j is defined according to the type of a variable. For binary variables, common practice is to use a Bernoulli distribution with a probit or logit link. The ordinal variables are typically modeled by cumulative probability functions such as proportional-odds models, adjacent-categories, or continuation-ratio logit models (Agresti, 2012). The count and continuous variables can be modeled by Poisson regression and a Gaussian model, respectively. The random-effect term, δ_{jt}, in Eq. 2 is parameterized for each j such that $\delta_{jt} \sim \mathcal{N}(0, \sigma^2_{\delta_j})$ and $\sigma^2_{\delta_j} = \text{Var}(\delta_{jt};\ t = 1, \ldots, T)$ models the variance across the problems in indicator j. For graphical illustration of the suggested model, see Fig. 1a.

Integrating the two constituting models, the random-effect LCM is formulated as a finite mixture model:

$$P(\boldsymbol{Y}|\boldsymbol{Z},\ \boldsymbol{\eta},\ \boldsymbol{\delta}) = \prod_{i=1}^{N}\prod_{t=1}^{T}\sum_{S_t\in\mathcal{S}} P(S_t|\eta_i, \boldsymbol{Z}_i)P(\boldsymbol{Y}_{it}|S_t,\ \boldsymbol{\delta}_t), \quad (3)$$

where $\boldsymbol{Y}$, $\boldsymbol{Z}$, and $\boldsymbol{\eta}$ each denote an array of variables for all students, and $\boldsymbol{\delta} = (\boldsymbol{\delta}_t : t = 1, \ldots, T)$ where $\boldsymbol{\delta}_t = (\delta_{jt} : j = 1, \ldots, J)$.

The marginal model (3) can be estimated using a maximum likelihood (ML) or Bayesian estimators. The ML estimation is computationally efficient but can lead to zero variance estimates (Gelman et al., 2013, p. 313).[3] The Bayesian estimation requires more computation time, but it gives more stable estimates once converged. In this study, we apply the ML estimator when exploring the candidate models and apply Bayesian estimation for deriving the parameters of a final model. The ML estimation was performed using an L-BFGS algorithm (Nocedal & Wright, 2006) in Stan, called from `R` (R Core Team, 2020). The Bayesian estimation was implemented using a "No U-Turn" Markov Chain Monte Carlo (MCMC) sampler (Hoffman & Gelman, 2014) in Stan. Details of the estimation routine, including the discussion on the label switching, are presented in Appendix B.

Profiling flow in CTA

The analysis model for CTA was formulated following Eqs. 1 and 2 but some adjustment was made in the measurement model to accommodate the CTA design. In CTA, problems were administered adaptively to students' skill levels, and the interaction data consisted of sparse time series with large measurement points (max $T = 151$). Introducing problem-level random effects in this case will induce computational overhead and challenge convergence of model estimation. To unify the measurement size and reduce the dimensionality to an estimable degree, we defined the measurement unit as a batch of problems that measure similar skills and applied the random-effect LCMs to the agglomerated data. In CTA, problems were arranged according to *knowledge components* (e.g. Ritter et al. 2007) that measure homogeneous skill sets and were administered successively until a student masters each knowledge component. We exploited this arrangement and examined the latent process underlying each knowledge mastery.

Defining a measurement unit as a knowledge component resulted in a total of 18 problem batches with 2219×18 ($N \times T$) interaction data. We applied the random-effect LCMs to the reshaped data, allowing measurement noninvariance in the indicator variables (i.e., δ_{jt} ($j = 1, \ldots, 3, t = 1, \ldots, 18$)). The random-effect terms from the three indicators were jointly modeled by a trivariate normal distribution: $\boldsymbol{\delta} = (\delta_{1t}, \delta_{2t}, \delta_{3t} : t = 1, \ldots, T)^\top \sim \mathcal{N}_3(\mathbf{0},\ \Sigma_{\boldsymbol{\delta}})$. The measurement models for the indicator variables were formulated as follows. The interaction time, Y_{i1t}, was modeled by a Gaussian model on the log metric:

$$\log Y_{i1t}|S_{it} = m,\ \delta_{1t} \sim \mathcal{N}\left(\delta_{1t} + \mu_{1m},\ \sigma^2_{1m}\right) \quad (4)$$

with the unique location and scale parameters for each state. The count variables—the number of errors, Y_{i2t}, and the number of hints requested, Y_{i3t}—were modeled by ordinal logistic regression after being categorized into zero, one, and more than one groups. For example, the probability of the number of errors was modeled as

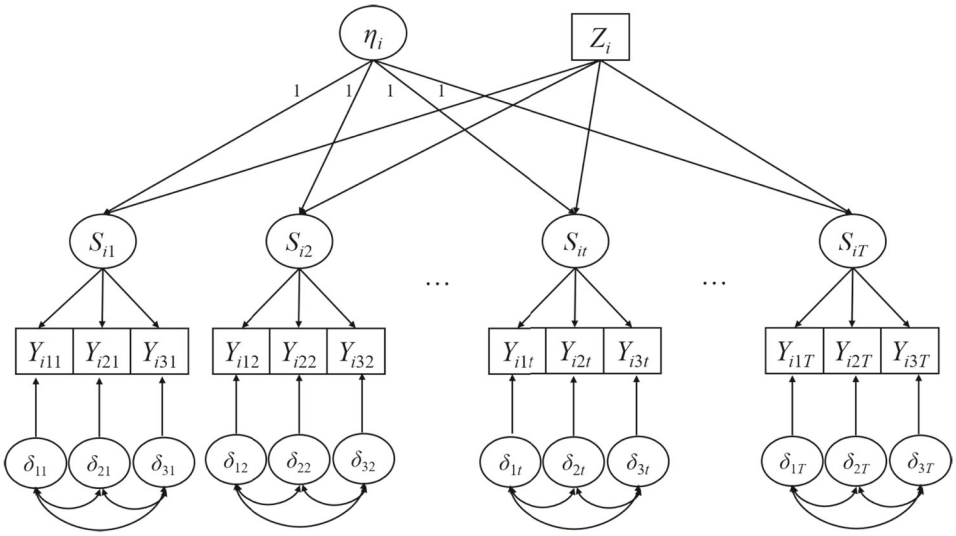
$$P(Y_{i2t} = h|S_{it} = m,\ \delta_{2t}) = \begin{cases} 1 - \text{logit}^{-1}\left(\mu_{2m} + \delta_{2t} - c_{21}\right) & h = 0 \\ \text{logit}^{-1}\left(\mu_{2m} + \delta_{2t} - c_{21}\right) - \text{logit}^{-1}\left(\mu_{2m} + \delta_{2t} - c_{22}\right) & h = 1 \\ \text{logit}^{-1}\left(\mu_{2m} + \delta_{2t} - c_{22}\right) & h > 1 \end{cases} \quad (5)$$

with the location parameter μ_{2m} varying across the latent states, and the step parameters satisfying $c_{22} > c_{21} > 0$. As above, δ_{2t} is the problem-level random effect for errors ($j = 2$). The number of hints was modeled analogously with μ_{3m}, δ_{3t}, c_{31}, and c_{32} each replacing μ_{2m}, δ_{2t}, c_{21} and c_{22}.

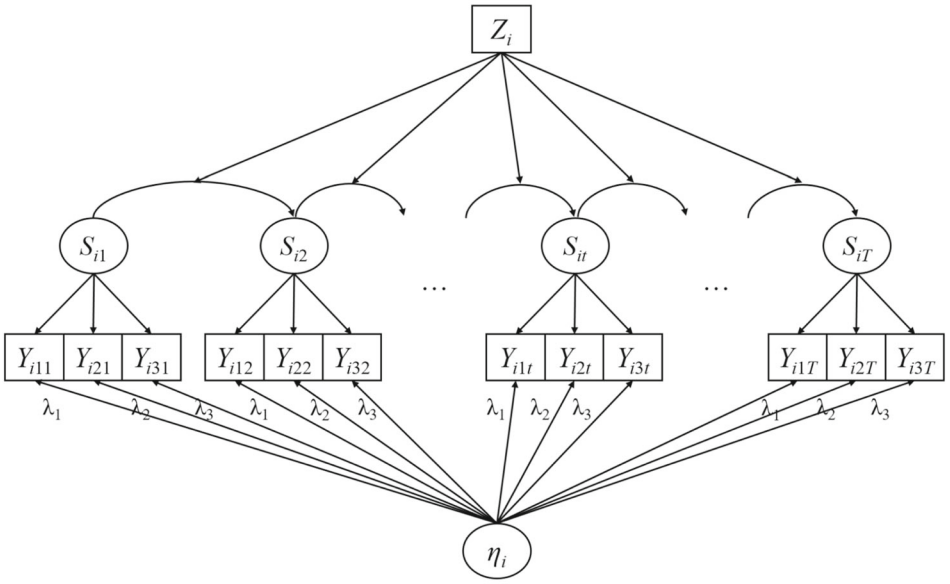
Results

Model comparison Table 2 reports fit statistics of four LCMs that were considered for flow evaluation. The models were fit assuming different structural effects while keeping

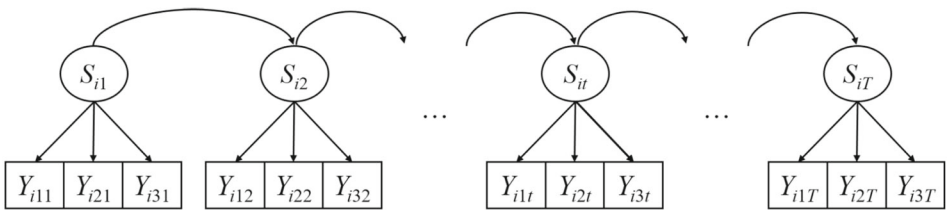
[3] The random-effect LCM can be seen as a hierarchical model with the random-effect terms being modeled at the higher-order level.



(a) Random-Effect Latent Class Model



(b) Random-Intercept Latent Transition Model



(c) Hidden Markov Model

Fig. 1 Latent variable models surveyed. *Note*. S_{it}: Latent state of a student i at measurement time t (= 1, ... , T). Y_{ijt}: Student i's observation data on an indicator j (= 1 (log interaction time); 2 (number of hints asked); 3 (number of errors)) at time t. η_i: Student i's random effect or random intercept. λ_j: Loading of an indicator j on the random effect. Z_i: Student i's background covariates

Table 2 Fit statistics of LCMs

M	Condition	df	−2LL	AIC	CAIC	BIC	ABIC
1		62	−241475	483074	483672	483672	483475
2	HomStud	67	−219935	440005	440650	440650	440437
2	StudEff	2285	−216607	437783	459802	459802	452540
2	StudCov	2291	−216576	437735	459811	459811	452530

Note. M = Number of states at each time point. df = Number of free parameters. LL = -2 log likelihood. AIC = Akaike information criterion (Akaike, 1973). CAIC = Corrected AIC (Burnham & Anderson, 2002; Sugiura, 1978). BIC = Bayesian information criterion (Schwartz, 1978). ABIC = Adjusted BIC (Sclove, 1987). The results were obtained from the ML estimator

the measurement model the same.[4] The one-state model in the first line assumes that all students exhibited the same state with no individual difference in the performance outcome. Since it assumes a single homogeneous latent state, it has no structural component. The following lines give results for the two-state models that assume different student effects. The *HomStud* model assumes that students in the same state had the same state probability, $P(S_{it} = 1) = \beta_0$. The *StudEff* model allows individual differences in the state probability, $P(S_{it} = 1|\eta_i) = \beta_0 + \eta_i$. The *StudCov* models the state probability as a function of both the student random effects and student-level covariates, $P(S_{it} = 1|\eta_i, \boldsymbol{z}_i) = \beta_0 + \boldsymbol{\beta}^\top \boldsymbol{z}_i + \eta_i$.

The results from Table 2 suggest that students indeed showed heterogeneity when working with the CTA. All information criteria preferred the two-state models over the one-state model. Among the two-state models, the criteria that heavily penalize the complexity (i.e., CAIC, BIC, ABIC) recommended the simplest model, *HomeStud*, whereas AIC and the likelihood statistic preferred *StudCov*. Our additional analysis with a likelihood ratio test suggested that *StudCov* achieves significantly better fit than *StudEff* ($\chi^2 = 62$, $df = 6$, $p < .001$). The posterior state probability estimates also indicated that there exists clear heterogeneity across the student demographic profiles. The observations from these analyses point *StudCov* as the most sensible model. We therefore choose *StudCov* as a final model and examine outcomes of *StudCov* more carefully to understand the students' learning modalities.

Measurement model As we identified the final model, we re-estimated the model parameters using MCMC to obtain stable parameter values. The ML estimation, though computationally affordable, yielded zero variance estimates for the student random effects (i.e., $\hat{\sigma}^2_\eta = 0$) despite the strong indication of heterogeneity in the observed data. The Bayesian estimation, although required a longer computation time, converged properly, yielding nonzero variance estimates and other point estimates essentially identical to the ML estimates.[5] Below we examine the outcomes of the Bayesian estimation to infer students' learning flow.

Table 3 reports measurement model parameter estimates. Each row gives distributional parameters of the indicator variables defined in Eqs. 4 and 5. The last line reports marginal probabilities of the latent states, $E[P(S = m)]$ ($m = 1, 2$). The state probability estimates suggest that State 1 was the major latent state underlying the interaction behaviors. Across the evaluated tutoring sessions, students worked under State 1 about 81.02% of times and the remaining 18.97% under State 2. The difference statistics in the last column suggest that students working in the second state requested more hints, made more errors, and spent more time than the students working in the first state. Note that in CTA students received problems adapted to their skill levels and they are generally expected to flow in learning. Considering this inbuilt design, we conclude that State 1 represents the state of flow and State 2 represents deviation from the flow. In State 1, students tended to progress in a timely manner, exerting adequate effort. Students in State 2 tended to exhibit prolonged learning behaviors. We mention that the present results do not inform the cause of deviation; we only surmise that multiple factors could have attributed to the deviating state, for example, fatigue, excessive difficulty, disengagement, gaming, or frustration with the problems.

To understand the degree of heterogeneity between the states, we compared the difference statistics with the between-knowledge differences, as parameterized by the standard deviations (SDs) of the knowledge-level random effects, δ. The random-effects SDs were estimated as $\hat{\sigma}_{\delta_1} = .497$ (95% credible interval [.366, .676]) for the time spent, $\hat{\sigma}_{\delta_2} = 1.387$ (95% CI [1.019, 1.858]) for errors, and $\hat{\sigma}_{\delta_3} = 1.529$ (95% CI [1.069, 2.128]) for hints. Observe that the estimated SDs were much smaller than the corresponding differences between the states (i.e., -1.493, -4.193, -12.275).

[4]Precursory analysis suggested strong nonzero random variance in the measurement effects and the measurement model was uniformly fit by including the random-effect terms

[5]The ML estimation took about 15 minutes to converge and Bayesian estimation base on MCMC took over five days when run in parallel on an AMD Ryzen Threadripper 3.5 GHz 16-Core Processor.

Table 3 Parameters of the measurement model of the final LCM

Indicator	Parameter	State 1	State 2	Difference
Time	μ_{1m}	−.500 (.117)	.992 (.117)	−1.493 (.008)
	σ_m	.757 (.003)	.838 (.005)	−.080 (.005)
Error	μ_{2m}	−2.092 (3.551)	2.101 (3.551)	−4.193 (.038)
	c_{21}	2.126 (3.557)		
	c_{22}	3.807 (3.557)		
Hint	μ_{3m}	−6.131 (3.807)	6.145 (3.804)	−12.275 (3.046)
	c_{31}	8.312 (3.881)		
	c_{32}	9.888 (3.882)		
	$P(S_0)$	.810 (.003)	.190 (.003)	

Note. The measurement model parameters were defined in Eqs. 4 and 5. μ_{1m}: Mean of log interaction time at State m. σ_{1m}: SD of log interaction time at State m. $\mu_{\cdot m}$: Mean of logit error/hint at State m. $c_{\cdot h}$: Step decrease in logit error/hint in category h. Within the parentheses are standard errors. All estimated differences in the final column were significant at $p = .001$

This suggests that the difference between the states was much more important than the differences between the knowledge components.

Covariate effects Table 4 reports effects of the student-level covariates. The estimation algorithm treated State 1 (i.e., the flow state) as a baseline and the reported coefficients represent loadings on State 2. A large value means that a student with the corresponding characteristic was more likely to work in the out-of-flow state. Some distinct patterns in Table 4 are worth noting. Students with higher pretest scores and gain scores (i.e., whose standardized test scores increased the most from the beginning to the end of the study) were more likely to work in State 1; students with lower pretest and gain scores were more inclined to work in State 2. Students with different racial backgrounds also showed disparate patterns. Black, multiethnic (RaceBN), Hispanic and Native (RaceHN) students, and to a lesser extent males, were more likely to work in State 2 than White or Asian students and females. Lastly, students who were eligible for free or reduced-price lunches were more apt to work in State 1 though the trend was somewhat weak.

As we examine the SD of the student random effects, $\hat{\sigma}_\eta = .590$ (95% CI [.560, .622]), we found that the estimate was larger in magnitude than any of the covariate coefficients in Table 4, and also larger than the standard deviation of the student-level log-odds of working a problem in State 1 predicted by covariates, $\mathrm{SD}(\hat{\boldsymbol{\beta}}^\top \boldsymbol{Z}) \approx .268$. This suggests that there may be important unmeasured student characteristics that predict the latent states, which we speculate to be the transitioning of latent states.

Table 4 Loading of covariates on the out-of-flow state

Covariate	Est	SE	p
Pretest score	−.258	.020	.000
Gain score	−.132	.020	.000
RaceBM	.274	.051	.000
RaceHN	−.195	.040	.000
Sex	.081	.032	.016
Free lunch	−.020	.036	.579

Note. RaceBM = Black & Multiethnic. RaceHN = Hispanic & Native America. Sex: Male = 1, Female = 0. Free lunch: Yes = 1, No = 0

Flow across tutoring stages

As the latent class analysis revealed heterogeneity in the students' workings, subsequent analysis was performed to examine the evolution of learning behaviors over tutoring sessions. For the analysis, we partitioned the observed data into different tutoring stages and applied LTMs to track development of latent states over time.

Random-intercept latent transition model

The LTM was formulated similarly to the LCM but additionally included a transition model. The structural model took a similar form with Eq. 1 but now models the initial state probability only, $P(S_0 = m)$ $(m = 1, \ldots, M)$, or $P(S_{i0} = m \mid \boldsymbol{Z}_i = \boldsymbol{z}_i)$ $(m = 1, \ldots, M)$ if covariates are applicable. The transition model then describes the probability of ensuing states as a function of the preceding state(s). Assuming a first-order Markov process, the probability of a latent state at time t is modeled as $P(S_t \mid S_0, S_1, \ldots, S_{t-1}) = P(S_t \mid S_{t-1})$. The functional form of $P(S_t \mid S_{t-1})$ commonly takes

multinomial logistic regression. If covariates are available, the transition probability can be modeled as

$$P(S_{it} = m' \mid S_{i(t-1)} = m,\ \boldsymbol{Z}_i = \boldsymbol{z}_i) = \frac{\exp\left(\gamma_{0m'} + \gamma_{mm'} + \boldsymbol{\beta}_m^\top \boldsymbol{z}_i\right)}{\sum_{l=1}^{M} \exp\left(\gamma_{0l} + \sum_{l'=1}^{M-1} \gamma_{ll'} d_{l'} + \boldsymbol{\beta}_m^\top \boldsymbol{z}_i\right)}, \tag{6}$$

where m and m' denote distinct latent states in $\mathcal{S}$; $\gamma_{0m'}$ ($m' = 1, \ldots, M-1$) gives an intercept of the logit of the transition probability; $\gamma_{mm'}$ ($m, m' = 1, \ldots, M-1$) models logit change between the two states; and $\boldsymbol{\beta}_m$ models the effect of covariate on the logit. The $d_{l'}$ in the denominator is a dummy variable that indicates the first $(M-1)$ states (i.e., $l' = 1, \ldots, M-1$; the last state is a reference category).

The measurement model describes the conditional probability of manifest indicators as a function of a latent state. One of the important assumptions in formulating the measurement model for transition analysis is longitudinal measurement invariance. Since the state variable is evaluated across the multiple time points, the problems must have constant effects on the measurement outcomes so that the variance in the observed performance can be attributed to the underlying latent state. As alluded to above, this assumption is generally not tenable in CTA because problems were administered adaptively according to the students' skill levels. To address the variance in the problem assignments and the measurement properties thereof, an additional means needs to be arranged.

Our strategy in this study for addressing the measurement noninvariance was to adopt student-level random intercept factors. The random-intercept LTM (RI-LTM; Muthén and Asparouhov 2020) introduces person-level random intercepts to measurement variables so that they can explain away extra variance present in the measurement outcomes. For example, in the context of CTA modeling, a measurement model for indicator j can be formulated as

$$P_j(Y_{ijt} \mid S_{it} = m, \eta_i) = f_j\left(\boldsymbol{\psi}_{jm} + \lambda_j \eta_i\right), \tag{7}$$

where f_j gives the functional form of the probability measure for the indicator, $\boldsymbol{\psi}_{jm}$ gives the kernel function indicating the effect of state m on j, η_i denotes the student-level random intercept, and λ_j models the extent to which students induce extra variance to indicator j. The random-intercept factor η_i is set to follow $\mathcal{N}(0, 1)$ so that λ_j models the size of extra variance in each measurement. The parameterization in Eq. 7 can be seen as decomposing the variance into within- and between-subject variance. The slope coefficient λ_j models the average magnitude of between-subject variance across the problems. Note that, in introducing the random-effect terms, the LCM and LTM assume different parameterizations (see Fig. 1 for a graphical comparison). In LCM, η_i is assumed to follow $\mathcal{N}(0, \sigma_\eta^2)$ and have unit slope whereas in RI-LTM, η_i follows $\mathcal{N}(0, 1)$ and has distinct slopes. The latter parameterization is to allow flexibility in modeling the measurement noninvariance. For example, the slope coefficient, λ_j in Eq. 7 can be reparameterized to allow time-specific loading (i.e., λ_{jt}) when a variable induces time-varying between-subject variance.

Integrating the sub-models, the random-intercept LTM is formulated as

$$P(\boldsymbol{Y} \mid \boldsymbol{Z}) = \prod_{i=1}^{N} \sum_{S_t \in \mathcal{S}} P(S_0 \mid \boldsymbol{Z}_i) \left(\prod_{t=1}^{T} P(S_{it} \mid S_{i(t-1)},\ \boldsymbol{Z}_i)\right) \times \left(\prod_{t=1}^{T} \prod_{j=1}^{J} P(Y_{ijt} \mid S_{it},\ \eta_i)\right). \tag{8}$$

The marginal model Eq. 8 can be estimated using a regular marginal ML estimator. In this study, we apply Mplus (Muthén & Muthén, 2017) to perform the estimation. We note that Mplus does not support intensive time-series data, and data cleaning is necessary to accommodate the constraints of the software. Next section presents strategies applied in the CTA analysis.

Flow transition in CTA

As with the preceding analysis, we used the data from one unit, *es*1, that are observed on the same days. The observed data consisted of intensive time series with a maximum of 151 problems. Currently available estimation programs for LTMs (e.g., Mplus, PROC LTA) assume longitudinal data with a small number of time points (e.g., 10 at most) and have limited capacity in modeling intensive time-series data. For evaluating flow in CTA, it was necessary to reshape the raw data and reduce the number of measurement points.

Our strategy in this study was to segment tutoring into several sessions and aggregate problems within each segment to create small sets of problem series. Since the problems in the evaluation data uniformly measured the same equation-solving skills, we assumed that the problems would exhibit weak measurement invariance. For the event that the problems induce significant measurement noninvariance, we also experimented with more general RI-LTMs that allow time-variant slopes when modeling the random intercepts.

The data partition was achieved as follows. We first determined the number of evaluation points as three, four, and five, balancing the granularity of description and convergence of model estimation. We then split the data from each student into three-, four-, and five-point time series by allocating approximately equal numbers of

observations to each series. For example, when a student received 56 problems, we sequentially aggregated (18, 18, 20) problems to create three-point time series. Within each partition, observations were averaged and placed on the calibration scale. The timing variable was placed on the log scale after being averaged. The count variables were placed on the ordinal scale after the average (i.e., indicating no, one, and more than one error/hint). As we conduct transition analysis in all three data sets, we found that the results generally suggest similar patterns on the flow development. The results differed only in the model convergence and the specific parameter estimates. Given these findings, we present results from the four-point time-series data as a representative example.

Results

Model comparison Tables 5, 6, 7 and 8 report fit statistics of the models that are evaluated in the four-time-point interaction data. Table 5 compares one- and two-state models. The subsequent tables refine the formulation of the resultant model. In Table 5, the comparison reveals that the two-state model achieved substantially better fit. All criterion measures consistently preferred the two-state model over the one-state counterpart, suggesting heterogeneity in the students' behaviors. The adjusted likelihood ratio test (Lo et al., 2001) similarly indicated significantly better fit of the two-state model ($\chi^2 = 23320.293, df = 6, p < .001$).

As we explore more complex LTMs that assume more latent states, we found that the models with the more states generally lead to better fit. The fitted outcomes however had generally similar bearing on the out-of-flow state. For example, when the LTM was fit with three states, the state identified as out-of-flow in the two-state model remained the same in the three-state model, and the flow state identified in the two-state model was separated into two distinct states in the three-state model. It appeared that increasing the number of latent states in LTMs tends to disintegrate the flow state and differentiate behavioral strategies of the regular problem-solving mode. The inference on the deviant state remained the same across the models that assumed different numbers of latent states.

As we decide on the two-state model, subsequent analysis was performed to examine the variants of the two-state model and determine the final model for CTA. We in particular investigated the models that differ in the three key assumptions: (i) measurement invariance, (ii) covariate effects, and (iii) transition probability modeling. Below details the analyses performed.

Measurement invariance Table 6 compares three models that assume different degrees of measurement (non)invariance: (i) the model that assumes longitudinal measurement invariance (labeled as MI), (ii) the model that assumes time-invariant between-subject residual variance (i.e., the model that includes the random intercept; labeled as RI), and (ii) the model that assumes time-variant between-subject residual variance (i.e., the model that allows time-varying effects of the random intercept; labeled as VRI). Recall that the random intercept in LTM was devised to describe between-subject residual variance present at each time point. The loading of an indicator on the random intercept models the magnitude of variance caused by the measurement stimuli at the time of evaluation. The nonzero loading coefficient will indicate that the problem sets induced unignorable variance in the students' behaviors. If the estimated loading coefficients are approximately equal across time, it will indicate that the problems entailed similar amounts of variance across time and the constant variance can be modeled by time-invariant loading coefficients. If the estimated loading coefficients differ substantially over time, it will signify that the problem sets entailed different amounts of variance across time and they must be modeled by time-contingent loading coefficients.

In Table 6, comparison of LTMs with and without the random intercept reveals that the random-intercept model achieved significantly better fit ($\chi^2_{(\mathrm{MI,RI})} = 589.69, df = 3, p < .001$). This suggests that the problem sets administered across the tutoring indeed induced nonzero variance in the outcome variables. When the indicators were allowed to have different loadings over time (i.e., RI vs. VRI), the model achieved even greater fit, suggesting that the measurements induced different amounts of residual variance over time. Among the three indicators, the interaction time showed the largest variability across time

Table 5 Fit statistics of LTMs: one- vs. two-state models

M	df	LL	AIC	CAIC	BIC	ABIC
1	15	−20727.74	41485.48	41485.70	41571.05	41523.4
2	21	−8995.53	18033.06	18033.48	18152.86	18086.14

Note. M = Number of states at each time point. df = Number of free parameters. LL = −2 log likelihood. AIC = Akaike information criterion. CAIC = Corrected AIC. BIC = Bayesian information criterion. ABIC = Adjusted BIC

Table 6 Fit statistics of LTMs: measurement invariance

Condition	df	LL	AIC	CAIC	BIC	ABIC
MI	21	−8995.53	18033.06	18033.48	18152.86	18086.14
RI	24	−8700.68	17449.37	17449.91	17586.28	17510.03
VRI	33	−8674.55	17415.11	17416.14	17603.37	17498.52

Note. df = Number of free parameters. LL = -2 log likelihood. AIC = Akaike information criterion. CAIC = Corrected AIC. BIC = Bayesian information criterion. ABIC = Adjusted BIC. MI = LTM with measurement invariance. RI = LTM with the random intercept. VRI = LTM with the random intercept with time-varying effects

(i.e., measurement noninvariance), followed by the number of errors, and the number of hints. All in all, the outcomes of the three models suggested that the indicators induced different amounts of measurement variance over time and it is sensible to include the random intercepts and allow time- and indicator-specific effects.

Covariate effects In Table 7 we compare models with and without the covariates to evaluate covariate effects. The results show that including covariates consistently improved the model fit. All pairwise comparisons preferred the models that allowed covariate effects (e.g., $\chi^2_{(\text{MI,MI-C})} = 133.07$, $df = 6$, $p < .001$). The information criteria similarly preferred the covariate-integrated models, producing smaller fit statistics. Based on the observations made here, we retain the covariates in the final model and subsequently examine the transition model allowing covariate effects.

Transition probability The last assumption, the stationarity of transition probabilities, was evaluated by comparing models with different temporal effects. Table 8 reports fit statistics of the evaluated models—the models that assume stationary and temporal state transitions. The stationary transition model (labeled as ST) assumes that all predictors had constant effects on the transition probability over time. The time-variant transition models assume that the latent states and covariates had differential effects across tutoring. Three possible scenarios were examined in the time-variant transition model: (i) when both the preceding latent state and covariates have time-varying effects on the transition probability (labeled as VTP), (ii) when only the latent state has time-variant effect and the covariates have fixed effects (labeled as VTS), and (iii) when only the covariates have time-varying effects and the latent state has a constant effect (labeled as VTC).

Comparison of the models in Table 8 suggests that the time-variant transition model is generally preferred over the stationary model. All criteria achieved the best results when the models allowed time-varying effects. Among the three models that allowed the time-variant effects, the model that assumed time-varying state effects and constant covariate effects demonstrated the best fitness. All in all, the results seemed to suggest that the preceding latent states had substantially different influences on the transition likelihood as time progressed while the effects of the covariates may be modeled by constants if the state effects are modeled with time-varying coefficients.

Comparison of various LTMs suggested that the model that allows time-variant effects for the random intercept and latent states (labeled as VRI-C-VTS) demonstrates the best fit for the CTA data. Based on this finding, we subsequently examined the outcomes of the final model to infer students' latent flow states. We note that the model identified here also yielded the best results in the other data sets—i.e., when the data were partitioned into three and five segments.

Table 7 Fit statistics of LTMs: covariate effects

Condition	df	LL	AIC	CAIC	BIC	ABIC
(MI, -)	21	−8995.53	18033.06	18033.48	18152.86	18086.14
(MI, C)	27	−8928.99	17911.99	17912.68	18066.02	17980.23
(RI, -,)	24	−8700.68	17449.37	17449.91	17586.28	17510.03
(RI, C)	30	−8634.15	17328.30	17329.15	17499.44	17404.13
(VRI, -)	33	−8674.55	17415.11	17416.14	17603.37	17498.52
(VRI, C)	39	−8608.02	17294.04	17295.47	17516.53	17392.62

Note. df = Number of free parameters. LL = -2 log likelihood. AIC = Akaike information criterion. CAIC = Corrected AIC. BIC = Bayesian information criterion. ABIC = Adjusted BIC. MI = LTM with measurement invariance. C = Covariates. RI = LTM with the random intercept. VRI = LTM with the random intercept with time-varying effects

Table 8 Fit statistics of LTMs: stationarity of transition probabilities

Condition	df	LL	AIC	CAIC	BIC	ABIC
(VRI, C, ST)	39	−8608.02	17294.04	17295.47	17516.53	17392.62
(VRI, C, VT)	61	−8581.60	17285.19	17288.70	17633.19	17439.38
(VRI, C, VTS)	43	−8591.04	17268.08	17269.82	17513.38	17376.77
(VRI, C, VTC)	59	−8588.22	17294.44	17297.72	17631.03	17443.58

Note. df = Number of free parameters. LL = -2 log likelihood. AIC = Akaike information criterion. CAIC = Corrected AIC. BIC = Bayesian information criterion. ABIC = Adjusted BIC. VRI = LTM with the random intercept with time-varying effects. C = Covariates. ST = LTM with stationary transition probability. VT = LTM with time-varying transition probability. VTS = Time-varying transition due to latent states (covariate effects fixed). VTC = Time-varying transition due to covariates (state effects fixed)

Measurement model Table 9 reports loading coefficients of the latent variables. The estimates reveal some distinct patterns in the indicator variables between the two states. The first state entailed shorter task interaction time, no help requests, and less erroneous attempts. The second state led to distinctly longer interaction time, nonzero help requests, and more errors. As we examine the prevalence of the states, we found that the first state appeared more frequently than the second (68.58% vs. 31.42%). The frequency of the second state tended to decrease in the first three quarters of the tutoring (36.73%, 29.02%, 27.13%) and increased in the final phase (32.81%). The observations made here overall seem to suggest that the first state reflects the normal problem-solving state and the second state the state of out-of-flow. Since CTA was assigned only infrequently and students tended to receive a large number of problems once started (50.236 problems on average and 151 maximum), we surmised that the distinct patterns of the second state is possibly due to warm-up or fatigue.

In Table 9, the coefficients related to the random intercept reveal that the indicator variables entailed different amounts of measurement noninvariance. The coefficients for the task interaction time showed the largest variance across time, suggesting that the variable induced the largest amount of longitudinal measurement noninvariance. The discretized error and hint variables showed relatively smaller and constant loadings, suggesting that they induced little and approximately equal amounts of measurement noninvariance. We note that constraining the loading coefficients of the error and hint variables to constants did not improve the model fitness. We therefore retain the final model allowing time-variant loadings on the random intercepts.

Transition model In Table 10 we report transition probabilities estimated from the final model. The first entry of each transition matrix consistently showed high probability (.733 on average), suggesting that students tended to maintain the same flow state once in place. Students exhibiting the deviant state (i.e., State 2) tended to switch to the flow state in the early phase of tutoring but gradually showed a stronger tendency to stay in the same out-of-flow state as the tutoring progressed. The present pattern supports our hypothesis on the deviant state. Earlier, we stated that the deviant state may reflect a warm-up or fatigue effect. The results from the transition probabilities suggest that the aberrant state identified in the early phase of tutoring is likely to reflect a warm-up or a learning period; the state

Table 9 Measurement model parameters of the final LTM (VRI, C, VTS): loading of latent variables on indicators

	Latent state		Random intercept			
Par	$\alpha(S=1)$	$\alpha(S=2)$	$\lambda(t=1)$	$\lambda(t=2)$	$\lambda(t=3)$	$\lambda(t=4)$
Time	3.788	4.141	−.387	−.261	−.246	−.224
(SE)	(.010)	(.013)	(.037)	(.031)	(.025)	(.026)
Error	.686	1.171	−.038	−.005	.006	−.035
(SE)	(.007)	(.008)	(.020)	(.020)	(.019)	(.021)
Hint	.000	1.021	−.010	−.003	−.006	−.014
(SE)	(.000)	(.004)	(.005)	(.003)	(.003)	(.005)

Note. α = Loading on the latent state. λ = Loading on the random intercept. The subscripts within the brackets indicate the latent state or the time point at which the variable was conditioned

Table 10 Transition probabilities in the final LTM

	$P(S_2 \mid S_1,\ \mathbf{Z})$		$P(S_3 \mid S_2,\ \mathbf{Z})$		$P(S_4 \mid S_3,\ \mathbf{Z})$		Overall	
From \To	$S_2 = 1$	$S_2 = 2$	$S_3 = 1$	$S_3 = 2$	$S_4 = 1$	$S_4 = 2$	$S_t = 1$	$S_t = 2$
$S_{t-1} = 1$	.730	.270	.752	.248	.718	.282	.733	.267
$S_{t-1} = 2$*	.676	.324	.674	.326	.550	.450	.633	.367

Note. S_t = Latent state at time t. $\mathbf{Z}$ = Covariates. * Suspected as the state of disengagement. $P(S_0 = 1) = .633$, $P(S_0 = 2) = .367$

appearing in the later stage is likely related to loss of motivation (e.g., fatigue, exhaustion). The high transition rate from the deviant to flow state in the beginning in particular suggests that students who showed slow progress at the outset began to engage in learning as they become familiar with the contents and problems. The decreased transition probability in the later phase suggests that weary students were unlikely to shift back to the flow state in the subsequent phases. The individual students' transition patterns further corroborated our hypothesis. About 63.68% of the students who showed early aberrance received *es*1 as the first unit on the day. About 86.13% of the students who showed aberrancy in the final stage received *es*1 as the last unit of the day.

Table 11 reports loading of latent states and covariates on the transition probabilities. During the estimation, the second state was treated as a baseline, and the coefficients were obtained for the first state (i.e., normal state). Observe that the first state had constantly positive loadings on the subsequent first state. This means that students who showed the normal state in the earlier problem set tended to maintain the same state in the following problem set. The increasing loading coefficients suggest that the tendency to stay in the same flow state intensified as time progressed.

The effects of covariates on the transition likelihood varied in both the direction and size. The negative loadings of the two ethnicity groups suggest that Black and Multiethnic, and Hispanic and Native American students were less likely to transition to the flow state than were the White and Asian students. The strong negative loading of the Hispanic-native American students implies that students in this ethnic group showed a greater tendency to exhibit a deviant state. The positive loadings in the other covariates suggest that students with higher scores had a stronger tendency to move to the first flow state. For example, male students showed greater likelihood to move to the flow state than female students. Students with higher pre- and gain-scores showed a greater tendency to transition to the flow state as they receive next problem set(s). Among others, the results pointed to a distinct tendency in the Hispanic and Native American female students. The effect coefficients suggested that these students were at greater risk of showing out-of-flow states, signaling the need for remedial interventions.

State transition We conclude the latent transition analysis with a summary of the most pronounced state patterns. Among the 16 transitional patterns, the largest group of students (27.27%) showed continued attention to the tasks with the estimated state pattern $\boldsymbol{S}$ = (1, 1, 1, 1). The second group of students (12.80%; $\boldsymbol{S}$ = (2, 1, 1, 1)) showed a transition from the deviant to the flow state in the early stage. The third group (9.19%; $\boldsymbol{S}$ = (1, 1, 1, 2)) showed a shift from flow to out-of-flow in the final stage. The next frequent groups changed midway, showing state patterns (1, 2, 1, 1) (8.70%) and (1, 1, 2, 1) (6.17%). These results suggest that, despite the self-regulated tutoring, many students performed conscientiously in most problems and exhibited deviant behaviors only occasionally. We remark that the present analysis was performed on the aggregated problem sets. The results do not inform specific problems at which a student showed deviant behaviors nor whether the aberrant behaviors occurred continuously across the

Table 11 Effects of latent states and covariates on transition probabilities

	$S_t[1]$ on $S_{t-1}[1]$			$S_t[1]$ on Covariates (t = 1, 2, 3, 4)					
	$t = 2$	$t = 3$	$t = 4$	Pre	Gain	RaceBM	RaceHN	Sex	FRL
Est	.262	.385	.745	.375	.198	−.081	−.126	.276	.021
(SE)	(.097)	(.103)	(.100)	(.042)	(.037)	(.059)	(.075)	(.049)	(.055)

Note. The base category of Race was White and Asian. RaceBM = Black and mutiethnicity. RaceHN = Hispanic and native American. The baseline category of Sex was Female. FRL = Free lunch. Pre = Prescore. Gain = Gain score. All latent states and covariates except for RaceBM ($p = .207$) and free lunch ($p = .901$) were significant predictors of the transition probabilities ($p < .01$)

partition or on a subset of problems. In the next section, we examine individual students' transitional patterns across the problems to draw finer-grained information.

Flow across problems

The last analysis applied hidden Markov models (Burke & Rosenblatt, 1958) to examine flow transition across the individual problems.

Hidden Markov model

As with the LTM, the HMM uses a sub-model to describe the transition between the latent states. Consider a latent stochastic process, $\boldsymbol{S}$, that evolves over discrete time points. The person-level subscript i is omitted as the analysis is performed at each student level (see Fig. 1(c) for graphical illustration). A series of manifest indicators is then modeled jointly with the corresponding latent states, $P(\boldsymbol{Y}_i,\ \boldsymbol{S})$. If covariates are available, $P(\boldsymbol{Y}_i,\ \boldsymbol{S})$ can be modeled as

$$P(\boldsymbol{Y}_i,\ \boldsymbol{S} \mid \boldsymbol{Z}_i) = P(S_0 \mid \boldsymbol{Z}_i) \prod_{t=1}^{T} P(S_t \mid S_{t-1};\ \boldsymbol{Z}_i) \prod_{t=1}^{T} P(\boldsymbol{Y}_{it} \mid S_t). \quad (9)$$

The formulation of Eq. 9 implies three sub-models: the (i) structural model, (ii) transition model, and (ii) measurement model. The structural model defines the probability of an initial state, $\boldsymbol{\pi} = (\pi_m = P(S_0 = m) : m = 1, \ldots, M)$. The transition model describes the transition probability between the latent states, $\mathbb{P} = (p_{mm'} = P(S_t = m \mid S_{t-1} = m') : m, m' = 1, \ldots, M)$. Observe that the transition model assumes a homogeneous first-order Markov process. The LTM similarly assumed the first-order Markov process but it allowed the transition probabilities to vary across times. The last component, the measurement model describes the emission probability of an indicator given the latent state, $P(Y_{jt} \mid S_t)$. Some commonly used models in HMM applications include: Gaussian, Poisson, binomial, Gamma, and multinomial models. It is important to note that, in modeling the emission probabilities, HMM does not typically allow measurement-level parameters; instead, it treats all observations from the same latent state as homogeneous observations from the same emission model and estimates the state-specific emission probabilities (or corresponding distributional parameters) that apply to all within-state observations.

Let $\boldsymbol{\tau}$ contain the parameters of the emission probabilities for all indicator variables and states. HMM is then defined by a triple, $\boldsymbol{\theta} = (\boldsymbol{\pi},\ \mathbb{P},\ \boldsymbol{\tau})$. For estimating $\boldsymbol{\tau}$, we use an `R` package, depmixS4 (Visser & Speekenbrink, 2010). The model is commonly fit to the person-level data under the weak measurement invariance assumption. Although the program allows sample-level fitting, our empirical analysis based on the simulated and real data suggests that the estimation tends to experience recurrent convergence problems. To make the inference adequately reliable across the analyses, we apply HMMs to the individual student-level data. This approach also follows the existing practice in the HMM application. Based on our empirical analysis, we discuss some possible consequences of fitting at the person level when the problems have distinct measurement effects.

Flow progression in CTA

The analysis based on the HMM was performed as follows. The input data included observations from $N = 2219$ students identified at the planning stage. The analysis used the same indicator variables and covariates as before but applied minimal transformation given the flexibility in the estimation software. Specifically, we applied the log transformation to the task interaction time and used the other count variables as collected.

Several distinctions made in the HMM analysis are worth mentioning. In the preceding analyses, we categorized the count variables into three levels to accommodate the constraints in the estimation programs. The HMM analysis however used the raw data and is expected to show greater sensitivity to the manifest variables. In some cases, the large variation in the indicator values can lead to overextraction of latent states because of the increased heterogeneity.

Second, unlike the other models, HMMs were fit to the individual student-level data and do not account for systematic effects of the problems. In CTA, students received different problem sets in various orders, and it made it difficult to unravel the variance in the measurement outcomes with the current HMM estimation software. Our preliminary study based on the simulated data suggests that ignoring measurement variance can result in overextraction of latent states. The current study attempted to alleviate this tendency by (i) focusing on a single unit of homogeneous problems (i.e., *es*1), (ii) conducting analysis in an exploratory manner, (iii) using a conservative information criterion in the final model selection, and (iv) giving greater priority to the models with clear binary interpretation. Specifically, when fitting the HMMs, we assumed one to ten latent states at each time point and determined a final model applying the most restrictive criterion, CAIC (Burnham & Anderson, 2002; Sugiura, 1978). The range of the number of latent states was determined in the precursory analysis by weighing the interpretability and generalizability of the models. While more complex models are conceivable, we did not explore these possibilities because our primary goal in this study was to make a binary decision on the latent states that can

be characterized as in- and out-of-flow. When the initial model fitting suggested multiple states, we refit the two-state HMM to see if the estimated states can be simplified to two states.

Third, recall that all covariates used in this study were time-invariant. When HMMs are fit to the person-level data with the time-invariant covariates, the covariates become constant and play no part in estimating the transition probabilities. Our supplementary analysis suggests that, although the constant covariates function as incidental variables, they can help choose a parsimonious solution. As another way of mitigating the measurement noninvariance problem, we therefore retain the covariates when fitting the HMMs though no particular inference could be attached to the covariate effects.

The outcomes of the HMM analysis were evaluated in three aspects: (i) intricacy of the latent process, (ii) progression of latent states across tutoring (e.g., when does a student most likely lose attention?), and (iii) consistency with the preceding analyses. Below, we present results of the HMM analysis and give three representative example students that showed typical transition patterns.

Results

As with the other analyses, we applied the information criteria to determine a final model. Among the criteria evaluated, ABIC and AIC suggested the most complex models with a minimum of one and a maximum of ten states. CAIC and BIC suggested simpler models with a maximum of four states. In the following, we present results from CAIC that led to the most parsimonious and clear model solutions. The outcomes of CAIC suggest that about 39.97% students (N = 887) displayed a homogeneous latent state as they advance problems, 51.83% (N = 1150) displayed two states, 8.07% (N = 179) three states, and .14% students (N = 3) four latent states. There was weak correlation between the estimated number of states and the number of observations (r = .241), suggesting that students with more problem attempts tended to display more diverse states.

Figures 2, 3 and 4 present three example students that were suspected of showing out-of-flow during tutoring. Each student was identified as displaying two, three, and four states across the evaluated measurement times. The first three plots at the top of each figure present observation series from the indicators. The last plot shows the sequence of latent states estimated across time. The dotted vertical lines indicate the points at which the problems were demarcated in the LTM analysis. Along with the figures, we also present estimates of the emission and transition model parameters in Table 12.

The student in Fig. 2 worked on a total of 57 problems over one hour and 43 minutes. As can be noted, the student maintained the same state in most cases but displayed somewhat deviant behavior toward the end of tutoring. In the early stage, the student solved most problems on his own and rarely made errors (.087 on average), resulting in average interaction time 54.49 seconds (per problem). As tutoring progressed, the student made more errors (6.909 on average) and requested more help (2.727), spending average time of 2.794 minutes. We restate that the problems assigned in CTA were adapted to the student's algebraic ability, and it is not generally expected to have a radical change in the outcome values. The continued digression from the normal state suggests that the student was either paying little attention or struggling with the problems. The transition probability estimates in Table 12 suggest that once the student adopted a particular mode, he tended to maintain the same mode in the subsequent problems. The probabilities of staying in the same state were .934 and .798 for each state.

Figure 3 presents a transition pattern of a student who displayed three states. The student worked on 79 problems in 63 minutes. As evident from the figure, the student displayed distinct behaviors in each state. When in State 1, he seldom made mistakes (.143 on average), asked for no hints, and spent little time on the problems (21.224 seconds on average). As he approaches the midpoint of tutoring, he tended to make more errors and asked for nonzero help. Specifically, when the student was in State 2 (28th–51st problems), he made .417 errors and asked for 1.792 hints on average. When he was in State 3, he showed a stronger tendency to make errors (4.167) and ask for help (3.833), and spent substantially longer time on the problems (3.584 minutes). Across the evaluated time periods, he showed State 1 most frequently (62.03%, T = 49 items), and States 2 (30.38%, T = 24) and 3 (7.59%, T = 6) in sequence. Combining the results, we conclude that States 2 and 3 reflect the state of out-of-flow. In both states, the student spent more time than usual and made distinctly more trials. These two states differed only in the magnitude of the indicators with those from State 3 reflecting greater detachment. As we fit the HMM assuming two latent states, States 2 and 3 indeed merged together and formed one deviant state, corroborating our hypothesis. The transition model in Table 12 suggests that the student showed a strong propensity to maintain the same S_1 and S_2 attitudes once adopted. The staying probabilities in these states equaled .959 and .957, respectively, indicating high

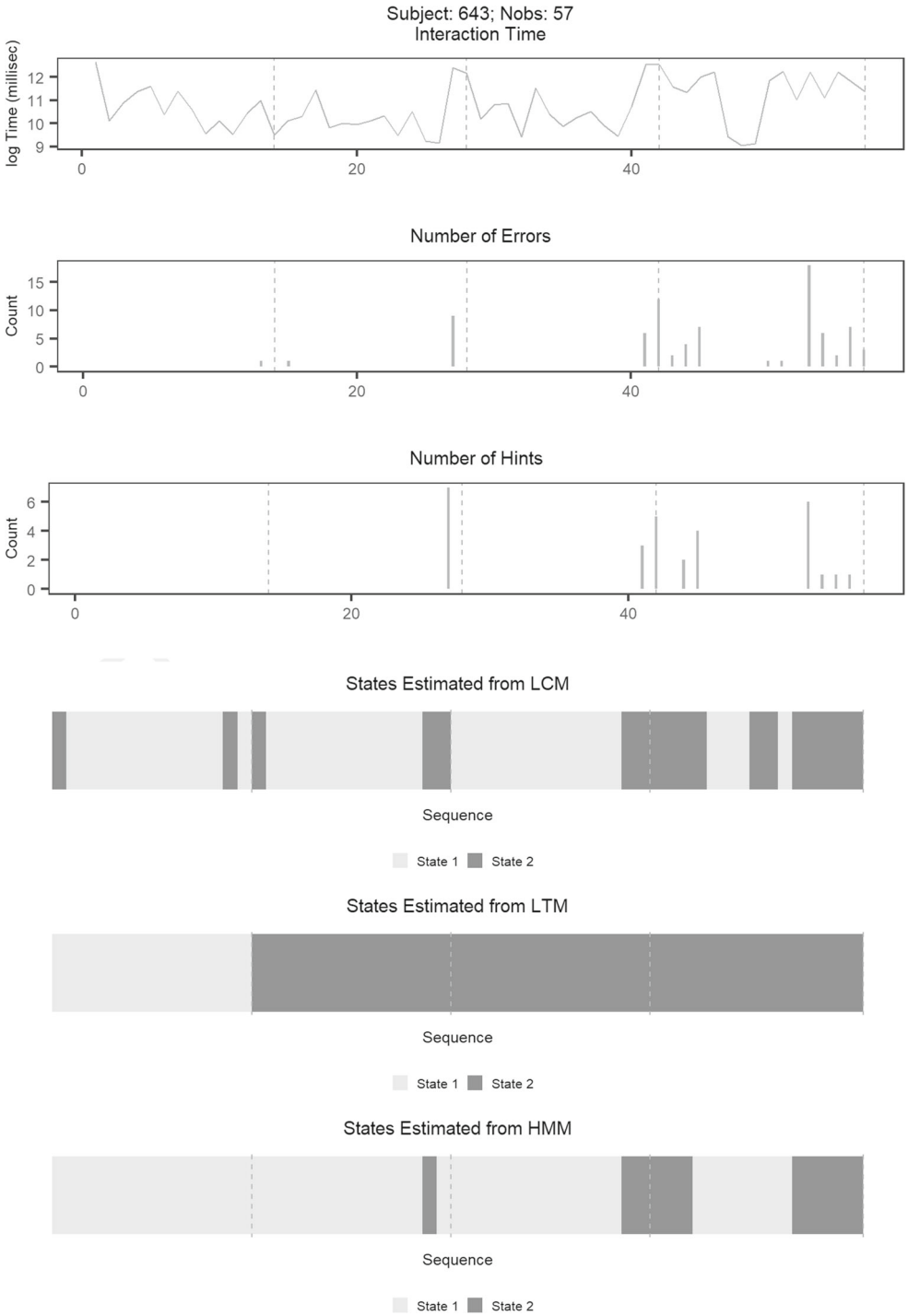


Fig. 2 An example student displaying two states. *Note*. State 1 was considered a flow state. The dotted vertical lines indicate the points at which the problems were demarcated in the LTM analysis

stationarity. When the student was in State 3, he maintained the same state with .598 probability and moved to State 2 or 1 with probabilities of .204 and .198, respectively.

Figure 4 presents a state transition pattern of a student displaying four attitudes. The student attempted 50 problems in 51 minutes. Figure 4 shows that the student rarely made errors and asked for only a few hints when working on the first three-quarters of the unit. As he worked on the last quarter, he began to show deviating behaviors, making many mistakes and asking for multiple hints. The indicator patterns in each state suggest that when the student was in State 1 or 2, he made no or very few mistakes (.250

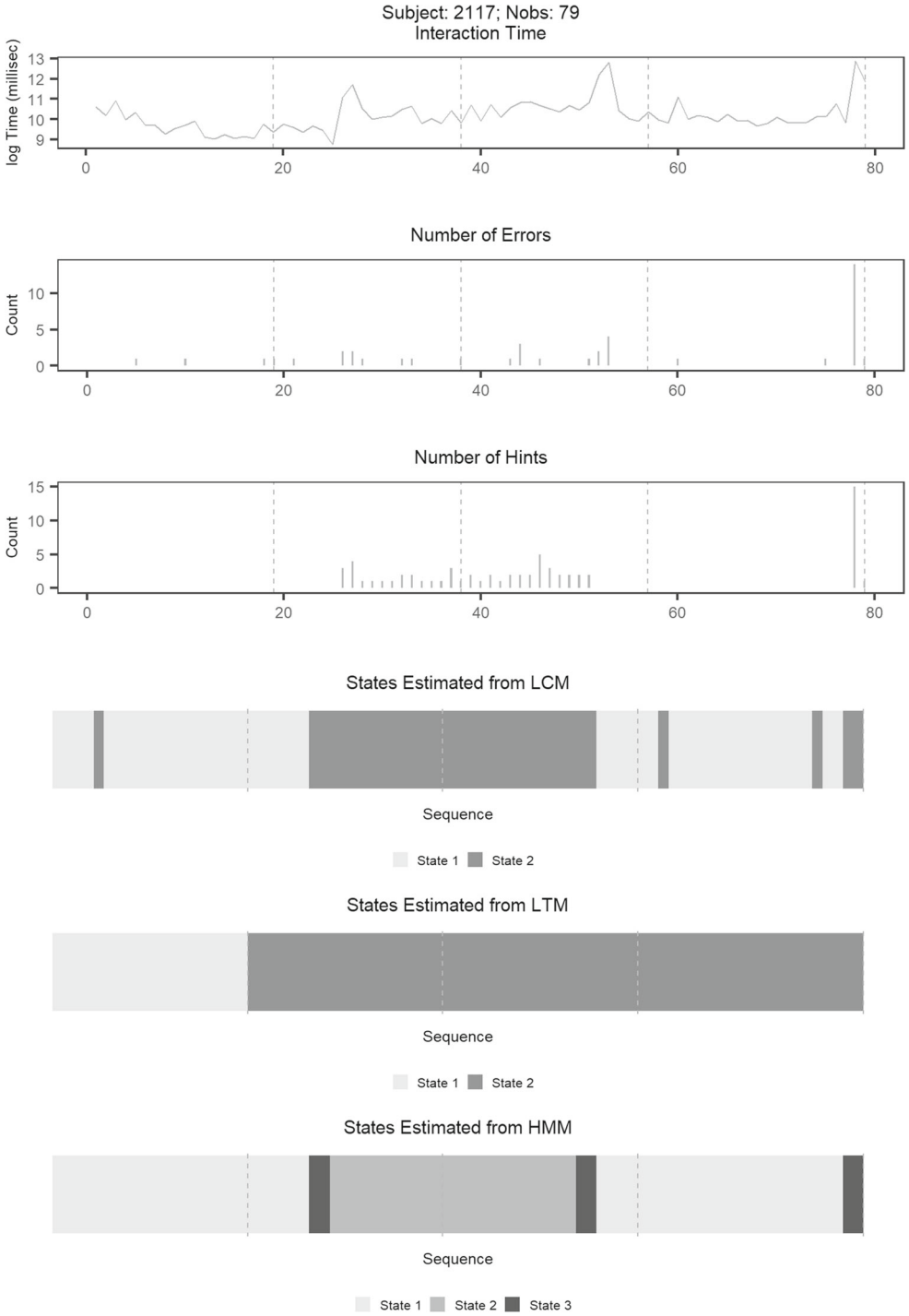


Fig. 3 An example student displaying three states. *Note*. State 1 was considered a flow state. The dotted vertical lines indicate the points at which the problems were demarcated in the LTM analysis

on average) and asked for zero help. When in State 3, the student made 5.333 errors and asked for more help (.833 hints on average). In State 4, the student showed a sharp increase in both indicators, making 24.333 mistakes and asking five hints on average. The total task interaction time similarly indicated different degrees of lassitude. In each of the four states, the student spent on average 19.873 seconds, 49.216 seconds, 1.470 minutes, and 5.829 minutes on the problems, respectively, indicating that the student spent substantially longer time when in States 3 and 4. Across the 50 problems attempted, the student displayed States 1 and 2 most frequently (42%, $T = 21$; 40%, $T = 20$ each) and

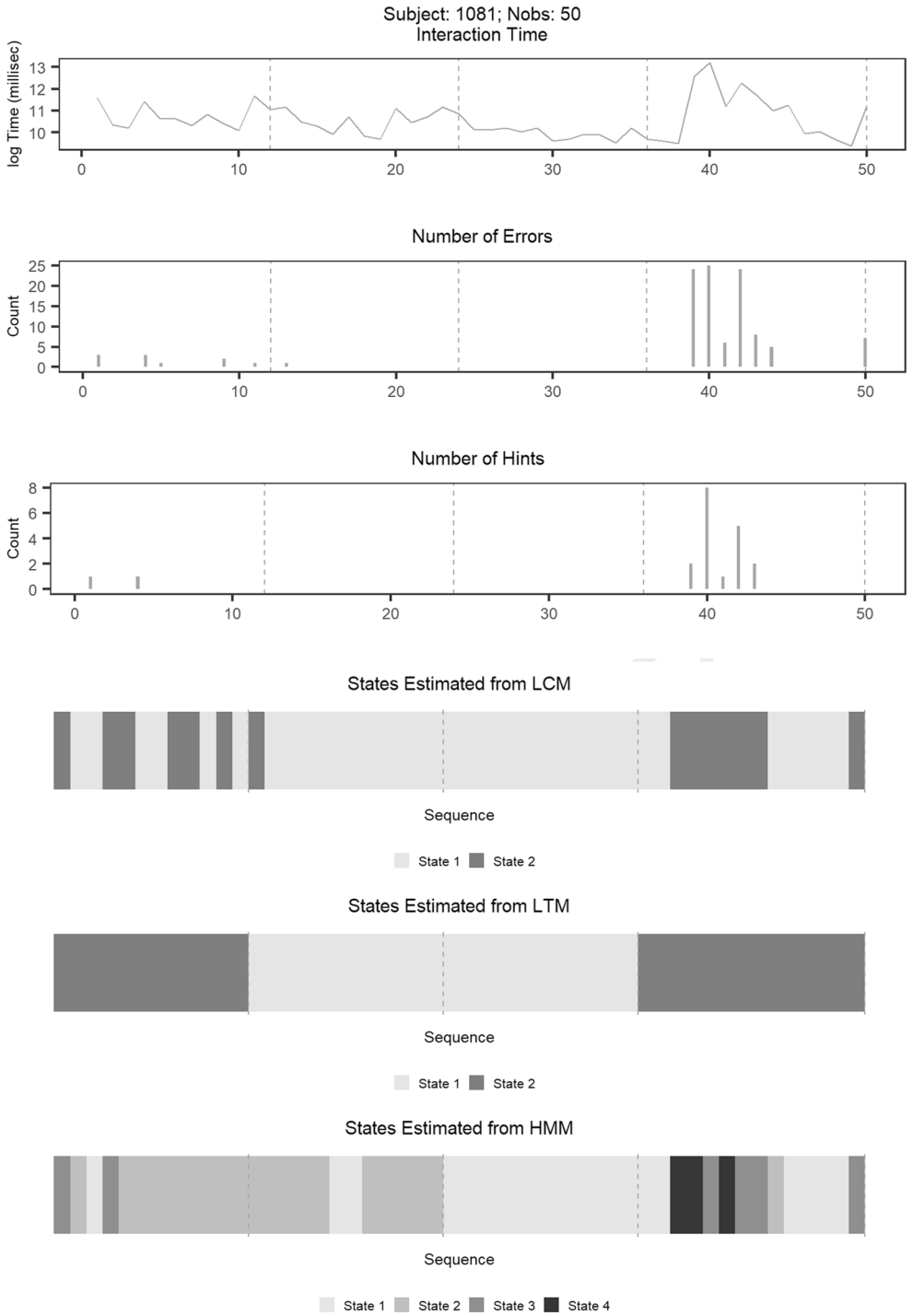


Fig. 4 An example student displaying four states. *Note*. State 1 was considered a flow state. The dotted vertical lines indicate the points at which the problems were demarcated in the LTM analysis

States 3 and 4 less frequently (12%, $T = 6$; 6%, $T = 3$ each). When the two-state HMM was fit to the same data, States 1 and 2 were grouped together to form one state, describing 82% of the observations ($T = 41$ problems). States 3 and 4 combined together to form the other state, explaining 18% of the observations ($T = 9$ problems). The present results suggest that States 3 and 4 are likely to indicate the deviant state. In both states, the student tended to make distinctly large numbers of errors and lingered on the problems longer. The transition probability estimates in Table 12 suggest that the student showed high stationarity when in State 1 or 2 (.780 probability on average). When he was in State 3 or 4,

Table 12 Emission and transition model parameters of HMMs

	Emission			Transition				
	Task Time	Num Err	Num Hint	From \To	$S_t = 1$	$S_t = 2$	$S_t = 3$	$S_t = 4$
Student 1 (M = 2; Grade 9, White-Asian, Male, Free lunch, No ESL)								
State 1	$\mathcal{N}(10.448, .907^2)$	Pois(.000)	Pois(.088)	$S_{t-1} = 1$	.934	.066		
State 2	$\mathcal{N}(11.917, .491^2)$	Pois(2.731)	Pois(6.916)	$S_{t-1} = 2$	.202	.798		
Student 2 (M = 3; Higher grade, Black-Multi, Male, Free lunch, ESL)								
State 1	$\mathcal{N}(9.840\ .487^2)$	Pois(.000)	Pois(.143)	$S_{t-1} = 1$	.959	.000	.041	
State 2	$\mathcal{N}(10.375, .351^2)$	Pois(1.791)	Pois(.413)	$S_{t-1} = 2$	.000	.957	.043	
State 3	$\mathcal{N}(12.082, .640^2)$	Pois(3.820)	Pois(4.152)	$S_{t-1} = 3$	.198	.204	.598	
Student 3 (M = 4; Higher grade, White-Asian, Male, Free lunch, No ESL)								
State 1	$\mathcal{N}(9.875, .267^2)$	Pois(.000)	Pois(.000)	$S_{t-1} = 1$	.788	.068	.097	.047
State 2	$\mathcal{N}(10.706, .428^2)$	Pois(.000)	Pois(.251)	$S_{t-1} = 2$	.228	.772	.000	.000
State 3	$\mathcal{N}(11.360, .248^2)$	Pois(.824)	Pois(5.286)	$S_{t-1} = 3$	.000	.605	.197	.197
State 4	$\mathcal{N}(12.687, .387^2)$	Pois(5.000)	Pois(24.333)	$S_{t-1} = 4$	.000	.000	.667	.333

Note. M = Number of latent states

he showed great mobility shifting to each State 2 (.605) or 3 (.667).

Conclusion

The purpose of this study was to explore a latent variable modeling approach for tracking learning flow in the ITS data. The study considered three models that give discrete profiles of latent states and applied to log data from CTA to demonstrate the application. In drawing learning flow, the study especially focused on three aspects: (i) the progression of flow states across tutoring, (ii) interaction modalities under in- and out-of-flow states, and (iii) relation between the flow and student's demographic profiles. The experimental application suggests that the models can reveal substantive information about students' flow process. Despite the difference in the assumptions and data constraints, the models suggested consistent findings on the flow pattern and students' learning behaviors under different flow states.

The models were applied in three gradual phases. The first stage applied the latent class models to identify latent profiles underlying the knowledge-level data. The study in particular applied the random-effect LCMs to account for extra variance in the students and measurement stimuli. For fitting the models, the evaluation data were rearranged at the knowledge level to regulate the dimensionality of measurement times. The results from the latent class analysis suggested that students generally showed uniform behaviors when working on the CTA. A small group of students (18.97%) displayed deviant behaviors, spending distinctly longer time, asking for more help, and making more errors. The latent profile estimates suggested that students with lower pre- or gain-scores, students in minority groups, and male students were more prone to exhibit deviating behaviors from the flow state.

As the latent class analysis revealed that students displayed distinct behaviors under different learning states, subsequent analysis examined the evolution of latent states across tutoring windows. The state profiles identified in the latent class analysis gave only temporal portrayal of the states that underlie at each measurement time and do not provide a full picture of the flow development across tutoring. The second analysis was conducted to evaluate such progression of flow states across the tutoring sessions. Specifically, the study applied the latent transition models to explicitly model the evolution of flow states across the problem series. To model different amounts of measurement noninvariance, the models were modified to include person-level random intercepts. The results from the transition analysis suggested that the majority of students showed adequate persistence in general (about 68.58% of the time), spending a reasonable amount of time and making conscientious efforts. At other times, students showed deviating behaviors (31.42%), dawdling on the problems and making distinctly more attempts. The deviant behaviors appeared more prevalently in the beginning and final stages of tutoring, possibly suggesting warm-up or fatigue. The transition probability estimates suggested that students with

higher pre- and gain-scores and male students had a stronger tendency to stay in flow. Female students and students in minority ethnic groups were more likely to transition to the deviant state.[6]

The last analysis examined flow development across individual problems using the hidden Markov models. The latent transition analysis revealed heterogeneity in the students' behaviors and their flow transition over time but the analysis was performed on the aggregated data to accommodate the constraints in the estimation software. The HMM allows a finer-grained description of flow transition at the individual measurement level and is more flexible in modeling different kinds of indicator variables. In this study, HMMs were applied to examine flow states underlying the individual problem solvings. The models were fit to the student-level data to address the possible measurement noninvariance. The outcomes of the model generally indicated consistent findings with those from LCM and LTM while giving more detailed descriptions of the flow progression and revealing specific problems at which points students deviated from the flow.

As described, the findings from the three analyses were generally consistent and indicated similar conclusions on the flow states and related patterns. The information from each model was however unique enough to warrant separate attention and appeared to complement each other such that they together give a comprehensive overview of the students' behaviors and underlying states.

While the present study demonstrated the potential utility of the latent variable models for describing the ITS data, it also revealed room for further improvement. As illustrated throughout the analyses, the models required additional modifications to accommodate the assumptions and estimation constraints. The latent class model needed random-effect terms to account for extra variance inherent in the students and measurement stimuli. Introducing a large number of random-effect terms however created a challenge in fitting the models and necessitated corrective steps. In addition, the model did not explicitly describe the progression of latent states, limiting the inference to each measurement time. The latent transition model improved some of the limitations of the latent class model, for example, by explicitly modeling the state transition and applying a more affordable estimation routine. The model however placed strict constraints on the number of measurements and the measurement invariance. The currently available estimation programs are not generally suited for intensive time-series data and require transformation or recoding of indicator variables to meet the distributional assumptions of the estimation program. Among the three latent variable models, the hidden Markov model was most flexible in modeling different kinds of indicator variables and placed the least constraint. This flexibility was however achieved at the cost of ignoring systematic effects of measurement stimuli, and consequently, it was more susceptible to random fluctuations in the indicators and was likely to overextract latent states.

Our empirical analysis of CTA data suggests that, while the current latent variable models have some merits, the development of a more comprehensive and systematic modeling framework is generally more desirable. For example, the current modeling frameworks can be extended to accommodate various indicator variables and parametric distributions that are commonly observed in the interactive ITS. The framework can also be further advanced using a more efficient estimation program that supports the analysis of intensive time-series data and allows integration of covariate information as well as prior information on the latent states. Together with the efficient estimation routine, the model can also be used online to signal a change in behaviors in real time.

Open Practices Statement:

The program code and data that support the findings of this study are available from the authors' GitHub pages: https://github.com/HyeonahKang/LatentVariableModel-IntelligentTutor; https://github.com/adamSales/SELS.

Appendix A: Measurement properties of the problems within units

Given the lack of supporting tools, summary statistics of the problem indicators were used to examine the measurement invariance. For count data (i.e., number of hints/errors), we examined the mean and standard deviation (SD) of the observed values, $E(X_{it} : i = 1, \ldots, N)$ and $SD(X_{it} : i = 1, \ldots, N)$, where i and t each indexes students and problems (administered across time). For timing data (i.e., interaction time), we examined the mean and SD of the log-transformed values, $E(\log X_{it} : i = 1, \ldots, N)$ and $SD(\log X_{it} : i = 1, \ldots, N)$. The (dis)similarity of the statistics across the problems within a unit was then evaluated by the variance across the problems. Table below reports the variance of the summary statistics observed from each unit. As can be seen, $es1$—the unit examined in this study—, showed the smallest variance on both the evaluation criteria. The unit was administered to a fairly large sample of students and yet showed small variance in the measurement summary statistics.

[6]The flow pattern in gender differed between the latent class and transition analyses because in LCMs the covariates were used to predict state probability at a given time point whereas in LTMs the covariates were used to model transition probability between the states.

Table 13 Variance of summary statistics of the problems within each unit

		Mean				SD			
Unit	*N*	nerr	nhint	ltime	Avg	nerr	nhint	ltime	Avg
cta1 01	2841	10.258	1.285	.222	3.921	5.658	.908	.005	2.190
cta1 02	2469	55.882	4.542	.073	20.166	62.809	4.092	.002	22.301
cta1 04	2067	7.970	.800	.290	3.020	5.820	.869	.002	2.230
cta1 06	1531	53.186	3.725	.319	19.076	66.601	4.380	.000	23.661
cta1 08	1045	6.631	.705	.084	2.473	6.239	.769	.001	2.336
cta1 10	1129	3.076	.420	.071	1.189	2.126	.448	.001	.858
cta1 13	1677	3.349	.370	.112	1.277	1.901	.423	.001	.775
cta1 14	974	4.517	.749	.119	1.795	6.102	1.522	.004	2.543
cta1 12	702	12.883	1.280	.225	4.796	3.459	.944	.011	1.471
es 01	2270	.354	.021	.690	.355	1.391	.351	.028	.590
es 02	1646	1.000	.177	.442	.540	4.193	2.336	.037	2.189
es 03	1315	2.365	.144	.528	1.012	7.169	3.761	.021	3.650
es 04	1232	.491	.058	.624	.391	2.814	.799	.017	1.210
es 05	701	1.134	.070	.647	.617	3.554	.518	.016	1.363
es 07	784	14.207	4.673	.760	6.547	11.050	2.566	.015	4.544
General linear form	743	10.329	2.088	.065	4.160	14.270	.942	.051	5.088
Linear inequal graphing	516	2.671	.158	.079	.969	17.273	2.077	.004	6.451
Quad add area alg1	422	22.198	2.445	.053	8.232	29.401	4.255	.033	11.230
Quad vertical motion	326	38.662	3.878	.076	14.205	103.537	4.516	.007	36.020
expt product simp a-es	401	.413	.034	.463	.303	4.730	.549	.028	1.769
expt quotient simp a-es	397	7.444	.708	.510	2.888	9.441	1.545	.004	3.663
Polynomial arith es	279	.614	.012	.058	.228	1.371	.049	.009	.476
Exponential functions	61	31.586	3.208	.095	11.630	47.612	6.037	.006	17.885
Unit conversions	1								
Inequality systems	1								
glf modeling	1								
Ratiol irratiol numbers	1								

Note. nerr = Number of errors. nhint = Number of hints. ltime = Logarithm of interaction time. Avg = Average

Appendix B: Fitting the random-effect latent class model in Stan

The latent class model described in "Flow across tutoring stages" was fit in Stan, called from R via the `rstan` package. In this appendix, we will first give the full likelihood equation and prior model, and then the Stan code defining the model. Complete replication code may be found the in this manuscript's companion GitHub repository.

The full likelihood is given by

$$
\begin{aligned}
&P(\boldsymbol{Y} \mid \boldsymbol{\eta},\ \boldsymbol{Z},\ \boldsymbol{\delta}, \boldsymbol{\beta},\ \boldsymbol{\mu},\ \boldsymbol{\sigma},\ \Sigma_\delta,\ \boldsymbol{\sigma}_\eta) \\
&\quad = \prod_{i=1}^{N}\prod_{t=1}^{T_i}\left\{\sum_{m=1}^{M} P(S_{it}=m \mid \eta_i,\ \boldsymbol{Z}_i=\boldsymbol{z}_i,\ \boldsymbol{\beta}) \prod_{j=1}^{3} P_j(Y_{ijt} \right. \\
&\qquad \left. = y_{ijt} \mid S_{it}=m,\ \delta_{jk[t]})\right\} \times \cdots \\
&\qquad \times \phi\left(\frac{\beta_0}{5}\right) \prod_{j=1}^{3}\left\{\phi(\sigma_{\delta_j}) \prod_{m=1}^{2} \phi\left(\frac{\mu_{jm}}{5}\right)\right\} \\
&\qquad \times \prod_{i=1}^{N} \phi\left(\frac{\eta_i}{\sigma_\eta}\right) \prod_{k=1}^{18} \boldsymbol{\phi}_3\left(\Sigma_\delta^{1/2}\boldsymbol{\delta}_k\right) \prod_{l=1}^{6} \phi(\beta_l)
\end{aligned}
$$

where $\phi(\cdot)$ is the standard normal density, $\phi_3(\cdot)$ is the standard multivariate normal density with three components, and $\boldsymbol{\sigma}_\delta = (\sigma_{\delta_j} : j = 1, \ldots, 3)$ represents the three diagonal elements of the covariance matrix Σ_δ.

The Stan code is written as

```
data {
  int<lower=1> nworked; //number of worked items--rows of data
  int<lower=1> nprob;      // number of items
  int<lower=1> nstud;      // number of respondents
  int<lower=1> ncov;        // number of person-level covariates
  int<lower=0> hint[nworked];
  int<lower=0> err[nworked];
  real ltime[nworked];
  int<lower=1,upper=nprob> prob[nworked];
  int<lower=1,upper=nstud> stud[nworked];
  matrix[nstud,ncov] X;
  vector[3] zeros;
  real<lower=0> sigStud;
}
parameters {
  real meanTime[2];
  real<lower=0> sigTime[2];
  real effHint[2];
  real effErr[2];
  vector[3] probEff[nprob]; // hint, err, time
  corr_matrix[3] OmegaProb;
  vector<lower=0>[3] sigProb;
  real alpha;
  vector[nstud] studEff;
  vector[ncov] beta;
  ordered[2] cHint;
  ordered[2] cErr;
}
transformed parameters {
  vector[nstud] nu=inv_logit(alpha+X*beta+studEff);
  cov_matrix[3] SigmaProb=quad_form_diag(OmegaProb, sigProb);
}
model{
// priors
 meanTime~normal(0,5);
 sigTime~normal(0,5);
 effHint~normal(0,5);
 effErr~normal(0,5);
 sigProb~normal(0,1);
 to_vector(beta)~normal(0,1);
 studEff~normal(0,sigStud);
 probEff~multi_normal(zeros,SigmaProb);
 //measurement model
 for(w in 1:nworked)
   target += log_sum_exp(
    log(nu[stud[w]])+
    ordered_logistic_lpmf(hint[w]|probEff[prob[w]][1]+effHint[1],cHint)+
    ordered_logistic_lpmf(err[w]|probEff[prob[w]][2]+effErr[1],cErr)+
    normal_lpdf(ltime[w]| probEff[prob[w]][3]+meanTime[1],sigTime[1]),
    log(1-nu[stud[w]])+
    ordered_logistic_lpmf(hint[w]|probEff[prob[w]][1]+effHint[2],cHint)+
    ordered_logistic_lpmf(err[w]|probEff[prob[w]][2]+effErr[2],cErr)+
    normal_lpdf(ltime[w]| probEff[prob[w]][3]+meanTime[2],sigTime[2])
   );
}
```

Table 14 Estimated posterior means

	μ_1		σ		μ_2		μ_3	
Chain	Class 1	Class 2	Class 1	Class 2	Class 1	Class 2	Class 1	Class 2
1	−0.50	0.99	0.76	0.84	−1.76	2.43	−6.29	6.00
2	0.91	−0.58	0.84	0.76	1.93	−2.26	6.18	−6.17
3	1.02	−0.48	0.84	0.76	1.85	−2.34	6.14	−6.11
4	−0.46	1.03	0.76	0.84	−2.07	2.12	−5.95	6.32
5	−0.50	0.99	0.76	0.84	−2.60	1.60	−6.14	6.14
6	1.00	−0.49	0.84	0.76	2.53	−1.66	5.70	−6.53
7	−0.51	0.98	0.76	0.84	−1.85	2.34	−5.78	6.57
8	−0.48	1.01	0.76	0.84	−2.19	2.00	−5.95	6.24

This model syntax was saved in a file called `lca2class.stan`, and the data was encoded in `R` in an object named `sdat`. Then we fit the model with the code:

```
mod <- stan('lca2class.stan',data=sdat,
iter=2000,chains=8)
```

That is, we fit eight separate chains, beginning from randomly-chosen initial values, running for 2,000 iterations each. By default, the first 1,000 of these iterations were denoted as "warm-up," and discarded. The hope is that the Markov chains would have each converged on the posterior distribution during the warm-up iterations. We checked the convergence of the algorithm by examining traceplots and the Gelman-Rubin $\hat{R}$ statistic (Gelman et al., 2013, § 11.4). Inference was based on the remaining 1,000 iterations in each chain, which we consider as samples from the posterior distribution.

Relabeling classes in LCM Two different runs of a Markov Chain Monte Carlo estimation technique can return essentially equivalent posterior estimates, but with the labels of the latent classes switched. Here we illustrate our *ad hoc* solution to label switching.

First, some parameters in the model are invariant to label switching, such as the optimized log posterior, denoted `lp__` in `rstan`. We began assessing model convergence by inspecting `lp__`. If this parameter has converged, we re-label the classes and assess convergence on the remaining parameters.

To re-label, we first examined the posterior means for each of the four measurement parameters, μ_1, μ_2, and μ_3, location parameters for the distributions of time spent, errors made, and hints requested, respectively, and σ, parameterizing the scale of the distribution of time spent, for each of the two latent classes, as estimated by the 8 Markov chains. These are reproduced in Table 14.

Note that for parameter μ_1, chains 1, 4, 5, 7, and 8 have means between -.51 and -.46 for class 1 and between 0.98 and 1.03 for class 2, but that these values are reversed for chains 2, 3, and 6. Similar patterns hold for measurement parameters σ, μ_2, and μ_3. This discrepancy between the chains leads to the unacceptably large $\hat{R}$ values in the first row of Table 15. However, if we switch the labels of classes 1 and 2 for chains 2, 3, and 6, all chains roughly agree on all four parameters, leading to acceptable $\hat{R}$ values in the 2nd row of Table 15. Importantly, we switch the labels on the same chains—2, 3, and 6—for all parameters.

Once we have determined the chains whose labels need to be switched, we relabel those chains' estimates for $Pr(S_{it} = m|\eta_i, Z_i = z_i, \boldsymbol{\beta})$ by subtracting the initial draws from 1, and we correct posterior draws for structural model coefficients $\boldsymbol{\beta}$ by multiplying the original draws by -1. Once the draws have been suitably transformed, we pooled the draws across all 8 chains to calculate the estimates in Tables 3 and 11.

Code for this full process can be found on the Github site.

Table 15 Gelman-Rubin $\hat{R}$

	μ_{11}	μ_{12}	σ_1	σ_2	μ_{21}	μ_{22}	μ_{31}	μ_{32}
Original	6.88	7.10	11.32	10.22	1.18	1.18	1.91	1.96
Relabeled	1.05	1.05	1.00	1.00	1.00	1.00	1.00	1.00

Funding The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305D210036. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

References

Agresti, A. (2012). *Categorical data analysis*, (3rd ed.). New York, NY: Wiley.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In Petrov, B. N., & Csáki, F. (Eds.) *Proceedings of the second international symposium on information theory*, (pp. 267–281). Budapest: AkadémiaiKiadó.

Burke, C. J., & Rosenblatt, M. (1958). A Markovian function of a Markov chain. *The Annals of Mathematical Statistics*, *29*(4), 1112–1122.

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach*, (2nd ed.). New York: Springer.

Csíkszentmihályi, M. (1990). *Flow: The psychology of optimal experience*. New York, NY: Harper & Row.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). Bayesian data analysis (3rd ed.). CRC press.

Hoffman, M. D., & Gelman, A. (2014). The no-u-turn sampler: Adaptively setting path lengths in hamiltonian monte carlo. *Journal of Machine Learning Research*, *15*(1), 1593–1623.

Lo, Y., Mendell, N. R., & Rubin, D. B. (2001). Testing the number of components in a normal mixture. *Biometrika*, *88*(3), 767–778.

Muthén, B., & Asparouhov, T. (2020). Latent transition analysis with random intercepts (RI-LTA). Psychological Methods., Advance online publication, pp. 1–18.

Muthén, L. K., & Muthén, B. O. (2017). Mplus: Statistical analysis with latent variables: User's guide (Version 8). Los Angeles, CA.

Nocedal, J., & Wright, S. J. (2006). *Numerical Optimization*, (2nd ed.). Berlin: Springer.

Pane, J. F., Griffin, B. A., McCaffrey, D. F., & Karam, R. (2014). Effectiveness of Cognitive Tutor Algebra I at scale. *Educational Evaluation and Policy Analysis*, *36*(2), 127–144.

Qu, Y., Tan, M., & Kutner, M. H. (1996). Random effects models in latent class analysis for evaluating accuracy of diagnostic tests. *Biometrics*, *52*(3), 797–810.

R Core Team (2020). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from https://www.R-project.org/

Ritter, S., Anderson, J. R., Koedinger, K. R., & Corbett, A. (2007). Cognitive tutor: Applied research in mathematics education. *Psychonomic Bulletin & Review*, *14*(2), 249–255.

Schwartz, G. E. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*(2), 461–464.

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

Sugiura, N. (1978). Further analysis of the data by Akaike?s information criterion and the finite corrections. *Communications in Statistics Theory Methods*, *7*(1), 13–26.

Visser, I., & Speekenbrink, M. (2010). depmixS4: An R package for hidden markov models. *Journal of Statistical Software*, *36*(7), 1–21. Retrieved from http://www.jstatsoft.org/v36/i07/