Examinee Variance in Strategy Shifts During Testing: An Explanatory IRT Approach

A Dissertation
Presented to
The Academic Faculty

by

Clifford Erhardt Hauenstein IV

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Psychology

Georgia Institute of Technology
December 2022

Examinee Variance in Strategy Shifts During Testing: An Explanatory IRT Approach

Approved by:

Dr. Susan Embretson, Advisor
School of Psychology
Georgia Institute of Technology

Dr. Michael Hunter
School of Human Development and Family Studies
Pennsylvania State University

Dr. James Roberts
School of Psychology
Georgia Institute of Technology

Dr. Matthew Johnson
Research and Development
Educational Testing Service

Dr. Rick Thomas
School of Psychology
Georgia Institute of Technology

Date Approved: October 28, 2022

This project is dedicated to the legacy and memory of Dr. Benjamin Drake Wright, whose passion and conviction continue to inspire my commitment to improving psychological assessment.

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my advisor, Dr. Susan Embretson, as well as the rest of my committee for their invaluable guidance and feedback during the course of this work. I would also like to thank the Partnership for Advanced Computing Environment (P.A.C.E.) at Georgia Tech for supplying the high performance computing resources necessary to complete the Monte Carlo simulations. A great deal of gratitude is especially warranted for the assistance of the PACE Service Desk, who provided immense help in troubleshooting and rectifying technical issues. Lastly, this work could not have been completed without the assistance from the passionate academics and scholars who have generously invested countless hours in developing freely accessible modules and responding to questions submitted across public forums.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS AND ABBREVIATIONS

2PL	Two Parameter Logistic Model
2PLC	Two Parameter Logistic Constrained Model
3PL	Three Parameter Logistic Model
AIC	Akaike Information Criterion
BIC	Bayesian Information Criterion
DIC	Deviance Information Criterion
DP(α, *H*)	Dirichlet process prior with associated concentration parameter (α) and base distribution (*H*)
HM	Hidden Markov
HM-2PLC	Hidden Markov Two Parameter Logistic Constrained Model
MHM	Mixture Hidden Markov
MHM-2PLC	Mixture Hidden Markov Two Parameter Logistic Constrained Model
ICL	Integrated Complete Likelihood
IRT	Item Response Theory
LLTM	Linear Logistic Test Model
Mix-2PLC	Mixture Two Parameter Logistic Constrained Model
q_i, m_i	Observed item complexity features for item *i*
RMSE	Root Mean Square Error
SB(α)	Stick-breaking prior, with associated clustering parameter (α)
A	*K* x *K* matrix containing the full set of transition probabilities for *K* total states
γ_{f_k}	Item feature weight on discrimination for feature *f* and strategy *k*
δ_j	Transition matrix assignment for subject *j*

ζ_t Marginal distribution of strategies at time point *t*.

η_{f_k} Item feature weight on difficulty for feature *f* and strategy *k*

θ_{jk} Latent ability for subject *j* on strategy *k*.

λ_{j_t} Row vector of length *K*, such that each *kth* element is a dichotomous indicator variable designating whether or not subject *j* visited strategy *k* at time *t*.

ν_g Probability of belonging to transition matrix *g*.

ξ_j Vector populated by the entire chain of strategies invoked for subject *j*.

π_{I_1} Probability of invoking strategy 1 at time 1 (i.e. initial state probability)

$\pi_{kk'}$ Probability of transitioning from strategy *k* to strategy *k'*

ϕ_{j_t} Strategy invoked by subject *j* at time point *t*.

SUMMARY

Considerable work (both within the psychometric and cognitive literature) has explored the tendency for response strategies to shift within the course of a single assessment. The current project proposes a novel, flexible item response model that intends to capture the pattern of these strategy shifts, as well as any latent classes defined by differences in these shifting patterns. The novel model represents an integration of hidden Markov techniques within the framework of Explanatory Item Response Modeling; a family of psychometric models that constrain parameters as functions of observed variables. Such an integration affords the opportunity to evaluate various hypotheses of response strategies and shifting patterns, while still retaining the desirable measurement properties of item response models. The feasibility of such a modeling approach is evaluated via parameter recovery with a set of Monte Carlo simulations, and its practical utility demonstrated with an empirical example involving response data from a spatial reasoning task.

Results from the simulation study show that the structural parameters (transition probabilities, initial strategy probabilities, and IRT parameters) are generally well estimated, with the exception of the case when an absorbing state is present. In these cases, IRT parameters for the transient state are estimated with a considerable degree of bias and imprecision, although all transition and initial strategy probabilities are still recovered well. In all cases, however, person level parameters were estimated with substantial error and bias. A correction to the deviance statistic is proposed in order to account for imprecision in the person level parameter estimates, and it is shown that theis corrected deviance statistic performs reasonably well for model selection procedures.

When applied to the empirical response data from a spatial reasoning test, the model seems to account for guessing processes, making salient its degree of flexibility. If guessing is reframed as a subject centered response style, then it can be captured within a broader class of response process models; in this case, a class of temporally variant mixture models. In fact, many traditional standalone IRT models and traditions (3PL model, mixture IRT models) may be subsumed by this more generalizable, coherent framework.

CHAPTER 1: INTRODUCTION

The standards and expectations for measurement practice have increased markedly over the last 20 years. Because of the widespread use and increased stakes of standardized testing in the educational community, measurement instruments are being evaluated with much more rigor and are increasingly expected to provide more discrete and nuanced information regarding examinee ability (Huff & Goodman, 2007; Mislevy, Almond, & Lukas, 2004). Evidence-centered test design is representative of the current zeitgeist in the educational community, in which items are expected to be developed and selected according to rigorous empirical and theoretical notions regarding the trait of interest. The objective is to construct tests that allow for a more refined set of hypotheses and conclusions to be drawn regarding the limits of skill acquisition and mastery.

This sets the stage for an approach to testing that transcends a pure measurement and scaling focus to an appraisal of *how* specific examinee properties interact with item features to produce observed test behavior. Calls to establish response process validity and identify examinees with fundamental differences in problem solving approaches were made explicit in the measurement community 20 years ago (Standards for educational and psychological testing, 1999) and continue to be reverberated (Embretson & Gorin, 2001; Noble, Rosebery, Suarez, Warren & O’Connor, 2014; Therese N. Hopfenbeck & Andrew Maul, 2011;).

In maintaining a contemporaneous position with the prevailing conceptual frameworks, measurement and psychometric research has begun to align its focus towards the particular strategies and cognitive processes that actually underly test performance and capturing the individual differences therein. One particular modeling approach that affords an integration of measurement theory and cognitive theory falls within the psychometric tradition of explanatory IRT modeling (De Boeck & Wilson, 2004). In essence, the explanatory IRT framework represents a family of models that reparametrize traditional IRT parameters as functions of observed person, item, or environmental features. This provides a relatively intuitive and straightforward means of reframing psychometric model parameters in a manner that is directly interpretable with respect to both basic and applied research questions. For example, although the structure of the Rasch model is rigorously justified in terms of measurement theory, it represents a saturated model with respect to item difficulty parameters, and therefore is atheoretical with respect to notions of response process and invoked strategies. In contrast, the explanatory IRT analogue of the Rasch model reparametrizes item difficulty as a weighted sum of relevant item features, allowing for hypothesis testing with respect to response processes. If a cognitive theory can explicate the impact of specific item features on response processes, then such a psychometric model has the flexibility to evaluate the validity of a proposed solution process while still conforming to basic Rasch measurement principles of additive conjoint measurement and specific objectivity (see Fischer, 1987 for a formal justification).

This reparameterization of the Rasch model is formally referred to as the linear logistic test model (LLTM; Fischer, 1973) and can be considered the seminal explanatory IRT model. It provided a powerful generalization of the Rasch model to accommodate validity questions of response processes (the Rasch is simply an application of the LLTM with a saturated design matrix), and established the foundation for reframing IRT model parameters in the context of relevant covariates. However, the LLTM is still highly restrictive in the sense that it assumes a singular, temporally invariant response process; and furthermore, equal loadings of items on the latent trait. In turn, the basic LLTM tenet of regressing IRT parameters on observed contextual features has continued to be integrated into progressively more general and complex IRT models to alleviate these restrictions, thus establishing the tradition of explanatory IRT modeling.

In the following sections, I briefly review a few noteworthy models within the explanatory IRT framework, with an emphasis on a proposed model from Rijmen et al. (2005) that allows for temporal variance in strategy use. Subsequently, I propose a generalized form of this model that is able to capture different patterns of temporal invariance; additionally, I present and explore several approaches to estimating model parameters (both parametric and non-parametric) that represent alternatives to the EM algorithm originally adopted by Rijmen and colleagues. A simulation study is then proposed in order to evaluate the conditions under which model parameters (for both Rijmen's model and the generalized form introduced here) may be reasonably recovered. Lastly, an application of the generalized model to an empirical data set is proposed in order to exemplify its potential utility in applied settings.

1.1 The Explanatory IRT Modeling Tradition

1.1.1 Rasch and LLTM Foundations

As mentioned, the seminal explanatory IRT model, the LLTM, has its roots in the Rasch modeling tradition. Formally, the Rasch model establishes that, given some monotonic transformation, g, the expected value of an observed response for person j on item *i* can be represented as an additive function of latent person (θ_j) and item difficulty parameters (β_i):

$$f(E(r_{ij}|\theta_j, \beta_i)) = \theta_j - \beta_i$$

For dichotomously scored items with a Bernoulli error distribution and expected value p, the logistic transformation is often invoked for *f*:

$$f\left(E(r_{ij}|\theta_j, \beta_i)\right) = \ln(\frac{p(x_{ij} = 1|\theta_j, \beta_i)}{1 - p(x_{ij} = 1|\theta_j \beta_i)}) = \theta_j - \beta_i$$

In contrast, the LLTM applies a linear constraint to item difficulty, with observed item features serving as predictors:

$$\ln(\frac{p(x_{ij} = 1|\boldsymbol{q_i \eta}, \theta_j)}{1 - p(x_{ij} = 1|\boldsymbol{q_i \eta}, \theta_j)}) = \theta_j - \left(\sum_{f=1}^{F} q_{f_i} \eta_f + \eta_0\right) \qquad \textit{(1)}$$

where q_{f_i} is the observed value for feature *f* on item *i*, η_f is the associated weighting parameter for feature *f*, and η_0 is an intercept. This reframes the Rasch approach within both a measurement model *and* response process model framework, setting the foundation for a family of models that can make predictions regarding cognitive processes and strategies.

1.1.2 When Examinees Invoke Different Strategies – the Mixture LLTM

In 1990, Mislevy and Verhelst noted that if the LLTM is capable of validating theories of response processes, then it could serve as the basis for deriving latent classes defined by fundamentally different cognitive strategies. Consider, for example, the oft-replicated finding of a strong linear relationship between degree of required rotation for spatial reasoning tasks and both response time and accuracy (Berg, Hertzog, & Hunt, 1982; Farrell & Robertson, 1998; Shepard & Metzler, 1975; Hertzog & Rypma, 1991; Shiima, Saito, & Suzuki, 1997). The finding implies that Rasch item difficulties are dimensionally over-specified for such items, and that item difficulties may instead be adequately represented along a single dimension related to rotational requirement. In this case, the LLTM offers utility as a method of dimensionality reduction *and* validating the relationship between required rotation and task difficulty.

However, a number of studies have demonstrated some preference for an alternative, verbal-analytical solution strategy that operates independently of spatial complexity (Bethell-Fox & Shepard, 1988; Gluck & Fitting, 2003; Schultz, 1991). Thus, a two-dimensional solution for item difficulty may be required in order to account for both spatial rotation and analytical complexity of items:

$$\ln(\frac{p_{ij}(x_{ij} = 1|\boldsymbol{q_i\eta}, \theta_j)}{1 - p_{ij}(x_{ij} = 1|\boldsymbol{q_i\eta}, \theta_j)}) = \theta_j - (q_{spatial_i}\eta_{spatial} + q_{analytical_i}\eta_{analytical} + \eta_0)$$

where $q_{spatial_i}$and $q_{analytical_i}$ represent scored spatial and analytical complexity measures for item *i*. However, if individual differences in strategy preference are reliable and remain fixed across time, then model parameters are obfuscated by a mixture of response processes. To rectify the issue, responses from each strategy group would need to be modeled with a separate set of LLTM parameters:

$$\ln(\frac{p_{ij}(x_{ij} = 1|\boldsymbol{q_i}, \boldsymbol{\eta_k}, \theta_{j_k}, k)}{1 - p_{ij}(x_{ij} = 1|\boldsymbol{q_i}, \boldsymbol{\eta_k}, \theta_{j_k}, k)}) = \theta_{j_k} - (q_{spatial_i}\eta_{spatial_k} + q_{analytical_i}\eta_{analytical_k} + \eta_{0_k})$$

where the subscript *k* corresponds to a class specific indicator. The probability of the full response vector for person *j*, conditional on strategy class and latent ability, is defined as follows:

$$p(\boldsymbol{x_j}|\boldsymbol{q_i},\boldsymbol{\eta_k},\theta_{j_k},k) = \prod_{i=1}^{I} \frac{\exp(x_{ij}[\theta_{j_k} - (q_{spatial_i}\eta_{spatial_k} + q_{analytical_i}\eta_{analytical_k} + \eta_{0_k})])}{1 + \exp(\theta_{j_k} - (q_{spatial_i}\eta_{spatial_k} + q_{analytical_i}\eta_{analytical_k} + \eta_{0_k}))} \tag{2}$$

and if a prior density function for the distribution of latent abilities is established, then the full log-likelihood is:

$$\mathcal{L}(\Omega|\boldsymbol{x}) = \sum_{j=1}^{J} \log[p(\phi_j = k|\boldsymbol{\pi})] + \sum_{j=1}^{J} \log[p(\boldsymbol{x_j}|\boldsymbol{q_i},\boldsymbol{\eta_k},\theta_{j_k},k_j)] + \sum_{j=1}^{J} \log[F(\theta_{j_k}|\psi_k)] \tag{3}$$

Where ϕ_j is a parameter that classifies examinees into one of *K* possible strategy classes, $\boldsymbol{\pi}$ is a vector of length *K*, $\{\pi_k \in \mathbb{R}: 0 \leq \pi \leq 1\}$ and $\sum_{k=1}^{K} \pi_k = 1$, representing strategy class proportions; and ψ_k is a vector of class specific hyperpriors for the presumed ability distributions.

Of course, the complication for traditional maximum likelihood parameter estimation is that the ϕ_j values are not known a priori. Mislevy and Verhelst recommend an EM algorithm, treating class designation and ability as missing values. Expected values of class designation and ability are derived, conditional on a set of provisional structural parameters ($\eta_{spatial_k}, \eta_{analytical_k}, \pi_k , \psi_k$) and the observed responses. Once these expected values are obtained, the structural parameters can be estimated via traditional maximum likelihood approaches. The new structural estimates are then treated as fixed, and a new set of expected values for group designation and ability are derived. The process iterates until some convergence criterion is achieved. Mislevey and Verhelst then re-estimate the group designation and ability values via an empirical Bayes solution, treating the EM estimated structural parameters as fixed and known.

More recently, fully Bayesian methods of parameter estimation for the mixture LLTM have been utilized (see Choi and Wilson, 2015 for an application with MCMC methods and Gibbs sampling) and shown to be superior for smaller sample sizes and smaller item pools when applied to Mixture Rasch models (Finch & French, 2012). Given the likelihood function defined in (3) and a set of prior distributions for $\eta_{f_k}, \theta_{j_k}, \psi_k, k_j, \boldsymbol{\pi}$, the joint posterior of model parameters can be derived. Choi and Wilson adopted Gaussian priors for the item feature weights and person abilities, with all hyperparameters fixed for model identifiability except ability variance (which was drawn from a gamma prior). The strategy class designations, k_j, were drawn from a categorical distribution:

$$\phi_j \sim cat(\pi_1, \pi_2 \dots \pi_K)$$

with the vector of class probabilities drawn from a Dirichlet distribution with fixed hyperparameters:

$$\boldsymbol{\pi} \sim Dirichlet(\frac{1}{K}, \frac{1}{K}, \dots \frac{1}{K})$$

Note that the Dirichlet distribution represents the conjugate prior to the categorical distribution, simplifying the estimation of the posterior.

1.1.3 Strategy Specific Item Loading Patterns – 2PL Constrained and Mixture 2PL Constrained

In 1999, Embretson proposed a 2PL analogue of the LLTM, whereby both discrimination and difficulty parameters are regressed on a set of observed contextual features:

$$\ln(\frac{p_{ij}(x_{ij}=1|\boldsymbol{q_i},\boldsymbol{\eta},\boldsymbol{m_i},\boldsymbol{\gamma},\theta_j)}{1-p_{ij}(x_{ij}=1|\boldsymbol{q_i},\boldsymbol{\eta},\boldsymbol{m_i},\boldsymbol{\gamma},\theta_j)}) = \left(\sum_{l=1}^{L} m_{l_i}\gamma_l + \gamma_0\right)(\theta_j - \left(\sum_{f=1}^{F} q_{f_i}\eta_f + \eta_0\right)) \qquad (4)$$

with m_{li} representing the scored value for an item feature, *l*, that is assumed to be related to item loading; and γ_l its associated weight. The approach was empirically justified with a set of matrix reasoning items, whereby both the number and types of rules dictating the matrix patterns were shown to be related to the slope and location of item characteristic curves. Embretson proposed a Joint Maximum Likelihood approach to estimating model parameters, though marginal maximum likelihood approaches may also be suitable if an appropriate prior for ability may be assumed. The model has particular utility when there is interest in automatic item generation for a test battery.

In 2021, Hauenstein and Embretson adopted the 2PLC as a basis for identifying latent clusters, introducing a mixture 2PLC (Mix-2PLC) with fully Bayesian estimation to evaluate performance on spatial reasoning items across several learning interventions.

Consistent with previously noted motivations for IRT mixture modeling, difficulty was regressed on both analytical and spatial indicators of complexity; however, discrimination was regressed on number of interventions received. Results showed a reliable clustering of examinees by differential mean changes in item discrimination across learning interventions, implying that interventions were effective in reducing construct irrelevant variance for some examinees, but not others.

1.1.4 Temporal Variance in Strategy Use – the Hidden Markov LLTM

While the Mix-LLTM and Mix-2PLC removed the constraint of strategy invariance across groups, temporal invariance in strategy use was still assumed. Although the assumption of temporal invariance may be tenable in some measurement applications, it is certainly unjustified in certain contexts. Using a previous example, Michaelides (2003) has found that not only will both spatial and analytic solution strategies be invoked for spatial reasoning items, but that examinees may shift between the two approaches within a single testing window. Additionally, dynamic testing paradigms are predicated on incurring both ability gains and strategy shifts over the course of a test through the use of scaffolding and feedback (Embretson, 1987; Resing & Elliot, 2011; Resing, Tunteler, & Elliot, 2015). And in the context of formative assessment, where ability tests are administered periodically over a longer time course, it is certainly expected that maturation and development would effect strategy shifts. It is this latter notion in particular that motivated Rijmen et. al to propose a hidden Markov formulation of the LLTM which directly accounts for these strategy shifts. Formally, the hidden Markov LLTM (HM-LLTM) defines the probability of an observed response vector for person *j*, *x*, as follows:

$$p(\boldsymbol{x_j}|\boldsymbol{q_i},\boldsymbol{\eta_k},\boldsymbol{\xi_j},\theta_{j_k}) = \prod_{k=1}^{K}\prod_{i:x_{ij}\in I_{jk}}^{I}\frac{\exp(x_{ij}[\theta_{j_k}-(\sum_{f=1}^{F}q_{f_i}\eta_{f_k}+\eta_{0_k})])}{1+\exp\left(\theta_{j_k}-(\sum_{f=1}^{F}q_{f_i}\eta_{f_k}+\eta_{0_k})\right)} \qquad (5)$$

where $\boldsymbol{\xi_j}=(\phi_{j,t=1},\phi_{j,t=2}\ldots\phi_{j,t=T})$ and represents the chain of strategy classes visited by subject *j* over the entire item set, and I_{j_k} reflects the subset of items responded to by subject *j* while visiting state *k*. Thus, the inner product function calculates the conditional probability of a set of item responses for which a particular state was visited; and the outer product function computes the joint probability across all state specific response vectors. Note that the parameterization is identical to that of the mixture LLTM in (1), with the only distinction being the adjustment to account for multiple states being visited by a single individual. This implies a time variant extension of the mixture LLTM, such that individuals are allowed to 'switch' from one strategy class to another over time. In particular, the model assumes a Markov process for these strategy shifts:

$$P(\phi_{j_t}|\phi_{j_{t-1}},\phi_{j_{t-2}},\ldots\phi_{j_{t-\infty}})=P(\phi_{j_t}|\phi_{j_{t-1}}) \qquad (6)$$

Practically, this constraint dictates that the strategy invoked at any given time point is dependent *only* on the strategy invoked in the immediately previous time point (what exactly constitutes a time point will be discussed shortly). Although the degree to which this assumption is tenable may be debated, it drastically reduces the burden for parameter estimation and represents a parsimonious means of tracking longitudinal shifts in response processes over time. For the simple case of two possible strategies, the longitudinal shifting pattern across *any* number of time points is governed simply by a set of three parameters:

$$\pi_{I_1}, \pi_{12}, \pi_{21}$$

where π_{I_1} is the initial probability of invoking strategy 1 at time point *t=1*, π_{12} is the probability of transitioning to strategy 2 at time *t+1*, given a classification of strategy 1 at time *t*, and π_{21} is the probability of transitioning to strategy 1 at time *t+1*, given a classification of strategy 2 at time *t*. The complementary events (e.g. $\pi_{I_2}, \pi_{11}, \pi_{22}$) do not represent free parameters since the probabilities must sum to 1. If the transition probabilities are collected into a transition matrix ($\boldsymbol{A}$):

$$\boldsymbol{A} = \begin{pmatrix} 1 - \pi_{12} & \pi_{12} \\ \pi_{21} & 1 - \pi_{21} \end{pmatrix} \tag{7}$$

such that the *kth* row represents the full set of transition probabilities for moving from state *k* to another state (or remaining in state *k*), then the probability of the full chain of classes visited by examinee *j*, given the transition matrix and initial class probabilities, can be calculated as follows:

$$p(\boldsymbol{\xi_j}|A, \boldsymbol{\lambda}, \pi_I) = \pi_I \left(\lambda_{j_1}\right)^{TP} \prod_{t=1}^{T} \lambda_{j_t} A(\lambda_{j_{t+1}})^{TP} \tag{8}$$

where λ_{j_t} is a row vector of length *K*, such that each *kth* element is a dichotomous indicator variable designating whether or not subject *j* visited strategy *k* at time *t*. Note that the *TP* superscript here refers to the transpose function.

Given matrix *A*, the marginal distribution of states at any given time point (ζ_t) is also easily obtained:

$$\zeta_t = \zeta_{t-1}\boldsymbol{A} \tag{9}$$

For a motivated example, the authors reference the developmental stages proposed by Piaget, and describe how such a model can capture discontinuous transitions from one stage to another across testing windows. In such an application, a particular Piagetian stage would be reflected in the pattern of relations between item features and difficulty; and furthermore, that sudden shifts in these relations are indicative of stage transitions.

Taking into account both the LLTM measurement model and the Markov state transition process, the complete likelihood function for the HM-LLTM can be defined:

$$\mathcal{L}(\Omega|\boldsymbol{x}) = \sum_{j=1}^{J} \log[\,p(\boldsymbol{\xi_j}|\boldsymbol{\pi_I},\boldsymbol{A})] \tag{10}$$
$$+\sum_{j=1}^{J} \log[p(\boldsymbol{x_j}|\boldsymbol{q_i},\boldsymbol{\eta_k},\theta_j,\boldsymbol{\xi_j}\,)] + \sum_{j=1}^{J}\sum_{k=1}^{K} \log[F(\theta_{j_k}|\psi_k)]$$

where $\boldsymbol{\pi_I}$ is a vector of length *K* representing strategy class proportions at *t*=1, $\boldsymbol{A}$ is the transition matrix populated by the full set of state transition probabilities, and ψ_k is again a vector of class specific hyperpriors for the presumed ability distributions. For parameter estimation, the Rijmen et al. adopt the same EM approach as Mislevy and Verhelst in the mixture LLTM, with each examinee's vector of strategy specific latent abilities $\left(\theta_{j_{k=1}},\theta_{j_{k=2}},\dots\theta_{j_{k=K}}\right)$ and chain of strategy classes ($\boldsymbol{\xi_j}$, or equivalently $\boldsymbol{\lambda_j}$) treated as missing data in the EM algorithm.

The approach of Rijmen et al. provides one of the most flexible and generalized approaches to explanatory IRT modeling, accomplished by removing the constraints of inter-cluster invariance and temporal invariance in model parameters. Nonetheless, their presentation leaves several critical issues unaddressed. First, the motivated example (which also constitutes the empirical application provided in their report) represents a very restrictive testing scenario. In such a scenario, time points are represented by whole tests; multiple tests need to be administered over an extended period of time (in their case, 11 moderate length tests over the course of a year) with little attrition. Although such a paradigm is relevant for formative assessment protocols and developmental applications, it is at odds with situations where the response process is expected to be variant *within* a test (whether due to spontaneous or experimenter controlled learning). Neither trial and error learning nor insight (a cognitive restructuring of the problem by uncovering novel relations among stimuli and integrating these relations to construct a new solution hypothesis; see Greeno, 1978) are expected to be restricted to between testing windows. The authors do mention briefly that the model has the potential to capture strategy shifts within a single test, whereby each item represents a single time point, but no further discussion regarding the testing conditions that provide reasonable parameter recovery is provided. Furthermore, while the authors cite the potential use of penalized fit statistics to determine the optimal number of strategies to model, no further guidance is offered in terms of the conditions under which such metrics are accurate and reliable. Additionally, the use of the LLTM as a basis for identifying states ignores the potential for items to exhibit differential loading patterns across strategies (see Embretson, 1999 for more discussion on this point). Lastly, their approach assumes that all subjects follow a uniform pattern of state transition probabilities; this may be an unreasonable constraint in certain testing applications. For example, some examinees may uniformly transition from one state to another with no reversals (such is the case for an ‘absorbing’ state), whereas others may be more fluid in their response process and periodically transition between states during the course of testing. Adopting the same motivated example as the authors to address this point, substantial variability in the timing of Piaget stage transitions has been observed, and in applied clinical work there is an understandable interest in identifying clusters of developmentally anomalous children with delayed transitions. Modeling cluster specific transitions is relatively straightforward to accommodate within the modeling framework of Rijmen et al. *if the clusters governed by different transition matrices are known a priori*. However, identification of latent clusters with different transition patterns requires a novel modeling approach.

The goal of the current project is to address these issues, and in doing so, propose an extended form of Rijmen’s model that can account for latent clusters with distinct state transition patterns. This novel model conforms the class of mixture hidden Markov models, which have been implemented in other domains, but are new to the IRT tradition. Additionally, through a simulation study, the current project evaluates the conditions under which model parameters are estimated with reasonable precision, both for the generalized form (distinct transition matrices for latent classes) and the restricted form proposed by Rijmen (singular transition matrix). Also through simulation, the utility of several model selection procedures is evaluated; specifically the use of more traditional metrics (AIC, BIC, DIC), as well as the use of a ‘stick-breaking’ process in the prior in order to infer the number of mixture components from the posterior. This is a more contemporary, non-parametric approach that avoids the necessity of fixing the number of mixture components a priori and obviates the need for using penalized fit statistics to identify the optimal number of mixture components. The adoption of a stick breaking process into a hidden Markov IRT model is novel, and provides an opportunity for improved time efficiency in exploratory applications. I also consider a parametric alternative to the stick-breaking process that carries the same benefits with respect to time economy. Lastly, the proposed, extended form of Rijmen's model is applied to an empirical dataset to demonstrate its usefulness in applied measurement settings. In the following section, a formal presentation of the model and proposed estimation procedures is provided, along with an account of several model selection procedures to be considered. Next, a description of the simulation study and manipulated conditions is given; and finally a description of the empirical dataset to which the proposed model will be applied.

CHAPTER 2: A NOVEL MODEL FOR CLUSTER VARIANCE IN STRATEGY TRANSITIONS

As mentioned previously, the proposed extension of the Rijmen et al. model can be effectively characterized as an application of a mixture hidden Markov model; and specifically, a mixture hidden Markov IRT model where the states are defined by different relations of item features to item parameters via the 2PLC model. For the remainder of this proposal, this model will be referred to as the Mixture Hidden Markov 2PLC, or MHM-2PLC (and the single transition cluster variant as the hidden Markov 2PLC, or HM-2PLC).

If the chain of strategy classes is known for subject j, then the probability a response vector in the MHM-2PLC is identical to that in (5), except that discrimination values are no longer constrained to unity and are allowed to vary according to a linear model:

$$p(\boldsymbol{x_j}|\boldsymbol{q_i},\boldsymbol{\eta_k},) = \prod_{k=1}^{K}\prod_{i:x_{ij}\in I_{jk}}^{I} \frac{\exp\left(x_{ij}\left(\left(\sum_{l=1}^{L} m_{l_i}\gamma_{l_k}+\gamma_{0_k}\right)\left[\theta_{jk}-\left(\sum_{f=1}^{F} q_{f_i}\eta_{f_k}+\eta_{0_k}\right)\right]\right)\right)}{1+\exp(\left(\sum_{l=1}^{L} m_{l_i}\gamma_{l_k}+\gamma_{0_k}\right)\left(\theta_{jk}-\left(\sum_{f=1}^{F} q_{f_i}\eta_{f_k}+\eta_{0_k}\right)\right))} \qquad \textit{(11)}$$

The primary distinction in the MHM-2PLC is the presence of *group specific* initial and transition probabilities, so that the probability of a subject’s sequence of state visitations is conditional on latent group assignment, g:

$$p(\boldsymbol{\xi_j}|A_g, \pi_I, g) = \pi_{I_g}\left(\lambda_{j_1}\right)^{TP} \prod_{t=1}^{T} \lambda_{j_t} A_g(\lambda_{j_{t+1}})^{TP} \tag{12}$$

which produces cluster specific marginal state distributions for each time point:

$$\zeta_{t_g} = \zeta_{t-1,g}\boldsymbol{A_g} \tag{13}$$

So that every group is defined by unique set of transition probabilities (note the *g* subscript on the vector of initial state probabilities and the transition matrix). The full log-likelihood function for the MHM-2PL is then be defined:

$$\begin{aligned}\mathcal{L}(\Omega|\boldsymbol{x}) = {} & \sum_{j=1}^{J} \log[p_j(\delta_j|\nu)] \\ & + \sum_{j=1}^{J} \log[\, p_j\left(\xi_j\middle|\pi_{1_g}, A_g\right)] \\ & + \sum_{j=1}^{J} \log[p_{.j}(x_{.j}|q_{.i}, \eta_k, \theta_{j_k}, k_j)] + \sum_{j=1}^{J}\sum_{c=1}^{C} \log[F(\theta_{j_k}|\psi_k)]\end{aligned} \tag{14}$$

where $\boldsymbol{\nu}$ is a vector of probability values that defines the distribution of examinees across the *g* transition clusters. Each subject's transition cluster membership (e.g. which transition matrix governs their pattern of strategy shifts) may be indexed by parameter δ_j; if the δ_j's are known, then the EM algorithm proposed by Rijmen et al. can be directly applied to estimate model parameters. If the δ_j's are not known, then a novel estimation approach is required. One possibility is a further extension of Rijmen et al.'s approach that considers δ_j, ξ_j, and θ_{j_k} all as missing data. Alternatively, a fully Bayesian method can be adopted to estimate structural and person parameters simultaneously, similar to the approach implemented by Choi and Wilson (2016) and described previously. For this project, fully Bayesian methods are implemented due to improved parameter recovery for small sample sizes (Finch & French, 2012) and because they afford the opportunity to develop non-parametric and pseudo non-parametric hidden Markov models that learn the number of strategy classes and transition clusters from the data. In the next section, the general structure of prior distributions for the MHM-LLTM is described; and following, different strategies for identifying the appropriate number of strategy classes and transition clusters are presented.

2.1 Prior Structure for Bayesian Estimation of MHM-2PLC

Given the likelihood function in (14) and a full set of prior distributions for the structural and person parameters, EAP estimates may be obtained for all parameters. As with the Mix-LLTM and Mix-2PLC described previously, all item feature weights are assumed to be Gaussian distributed with fixed hyperparameters:

$$\eta_{f_k} \sim N(0,1)$$

$$\gamma_{f_k} \sim N(0,1)$$

And on the person side, the class specific ability estimates also follow a Gaussian prior with fixed hyperparameters to anchor the scale:

$$\theta_{f_k} \sim N(0,1)$$

The strategy class designations at time *t* follow a categorical distribution, conditional on the class designation at *t-1* to conform to the Markov process constraint. For a particular strategy class *k* at *t-1*

$$\phi_{j_t}|(\phi_{j_{t-1}} = k) \sim cat(\pi_{k1_g}, \pi_{k2_g} \dots \pi_{kK_g})$$

Where the vector $(\pi_{k1_g}, \pi_{k2_g} \dots \pi_{kK_g})$ represents the full set of *cluster specific* transition probabilities for moving from state *k* at *t-1* to any designated state at *t*. These transition probabilities are collected into the *kth* row of the transition matrix A_g, and are themselves distributed according some designated prior distribution, *C* with parameter vector $\boldsymbol{\alpha}$:

$$(\pi_{k1_g}, \pi_{k2_g} \dots \pi_{kK_g}) \sim C(\boldsymbol{\alpha_0})$$

And the actual transition cluster assignments for examinees, δ_j follow a categorical prior:

$$\delta_j \sim cat(\nu_1, \nu_2, \dots \nu_G)$$

With the vector of $\boldsymbol{\nu}$ values, representing the proportional distribution of examinees across the *G* transition matrices, also being distributed according to some chosen prior distribution *C*:

$$(\nu_1, \nu_2, \dots \nu_G) \sim D(\boldsymbol{\alpha_1})$$

Note that if *G* is fixed to 1, then the MHM-2PLC simply reduces to the HM-2PLC, providing an alternative to the EM approach of Rijmen et al. for estimating parameters in hidden Markov explanatory IRT models. Note furthermore that this presentation has thus far neglected to establish any specific form to the prior distribution on the class and cluster probabilities. The reason is that several choices are available for the forms of $C(\boldsymbol{\alpha_0})$ and $D(\boldsymbol{\alpha_1})$**,** each based on different theoretical representations of the clustering process and representing different biases with respect to the degree of clustering. In the following sections, I review three common choices for $C(\boldsymbol{\alpha_0})$ and $D(\boldsymbol{\alpha_1})$: 1) the **finite**

Dirichlet, 2) the **stick-breaking process**, and 3) the **sparse finite Dirichlet,** and discuss their implications for identifying the number of strategy classes and transition clusters.

2.1.1 The Finite Dirichlet

One exploratory method of identifying the number of strategy classes and transition clusters in the MHM-2PLC is to simply run a series of models that each specify a different set of fixed values for K and G; and then subsequently evaluate penalized fit statistics (e.g. AIC, BIC, DIC) to select an optimal solution. For this method, both $C(\boldsymbol{\alpha_0})$ and $D(\boldsymbol{\alpha_1})$ are set to Dirichlet priors, with concentration parameters fixed to some reasonably large value. The reason for this decision on the concentration parameters is to force the model to populate the full set of *K* strategies and *G* clusters in each run (e.g. avoid vacant classes/clusters). In this case, each penalized fit statistic has a direct interpretation with respect to a specific clustering solution. Frühwirth-Schnatter (2006) recommends fixing full set of concentration parameters to 4 to achieve this, and this guidance is followed in the current analysis.

2.1.2 The Dirichlet Process/Stick-Breaking Process

A disadvantage of the approach above pertains to time economy; Bayesian estimation procedures for hidden Markov models, relying on Monte Carlo sampling from the posterior, can be quite time consuming. It becomes prohibitively time consuming when fit statistics must be derived for an entire set of models that exhaust a set of *K* x *G* combinations.

An alternative approach is to estimate the number of strategy and transition classes from the observed data, given a reasonably specified prior distribution. In particular, the *Dirichlet process* (Ferguson, 1973; and specifically, its *stick-breaking* representation), can be exploited to simultaneously consider a range of *N* possible strategy classes and *N* possible transition clusters, where *N* is the total number of subjects. This bypasses the need to run multiple analyses with differing numbers of class/cluster solutions.

The Dirichlet process distribution is an infinite extension of the standard Dirichlet distribution and can be thought of as a 'distribution of distributions'; that is, its realizations are themselves entire distributions. Specifically, the Dirichlet process defines a distribution of partitions over a measurable set, given two parameters: a concentration parameter (α) and a base distribution (H):

$$V \sim DP(\alpha, H)$$

Formally, the Dirichlet process is represented as draws from a Dirichlet distribution, with the Dirichlet parameters defined as a function of the base distribution and alpha parameter. For a measurable set, *S*, and *K* finite partitions (or clusters) of *S*, $A_1, A_2, \dots A_K$:

$$\left(V(A_1), V(A_2), \dots V(A_K)\right) \sim Dirichlet(\alpha H(A_1), \alpha H(A_2), \dots \alpha H(A_K)) \qquad (15)$$

where $H(A_i)$ is the probability mass associated with cluster *I* according to base distribution, *H*, and the vector $\left(V(A_1), V(A_2), \dots V(A_K)\right)$ is a single realization of the Dirichlet process; also note that for *I* in 1:*K*, $V(A_k) \geq 0$ and $\sum_{i=1}^{k} V(A_k) = 1$. Effectively, the Dirichlet process transforms a continuous distribution into a discrete distribution with a countably infinite number of point masses. This property is what establishes the Dirichlet process as a useful prior for cluster and mixture modeling applications; in fact, data clustering algorithms represent one of the most widely used applications of the Dirichlet process.

A simple, intuitive way to think of the Dirichlet process is through a classification procedure, whereby each new value is either classified into an existing cluster (with probability proportional to the cluster size, n_k) or into its own new cluster (with probability proportional to the concentration parameter, α). Thus, the concentration parameter controls the degree of clustering; when α is large, new clusters are favored and $V(A) \rightarrow H(A)$. Conversely, if α is sufficiently small, V reflects a degenerate distribution. This has obvious implications for the application of the Dirichlet process prior in mixture modeling, as α can be tuned according to one's prior beliefs regarding the degree of clustering. Alternatively, α can itself be derived empirically from the data evidence.

This intuitive perspective of the Dirichlet process has often been illustrated through the metaphor of the 'Chinese Restaurant Process'. The basic idea is that a restaurant may have as many tables (clusters) as customers (observations). As customer x_i enters, that customer may be seated at a particular table with probability proportional to $\frac{n_k}{N+\alpha}$, where n_k is the number of customers already seated at table A_k, N is the total number of customers seated in the restaurant, and α is the concentration parameter described previously. However, customer x_j may also be seated at a new, unoccupied table with probability proportional to $\frac{\alpha}{N+\alpha}$, and the process repeats itself as a new customer enters. Essentially, tables (or clusters) may be continually added to accommodate new customers according to parameter α and existing cluster membership:

$$p(\phi_i = k|\boldsymbol{\phi}_{-i}, \alpha) = \begin{cases} \frac{n_k}{N+\alpha} \; for\ existing\ cluster \\ \frac{\alpha}{N+\alpha} for\ new\ cluster \end{cases} \tag{16}$$

The fact that tables can continually be added as needed underscores the utility of the Dirichlet prior in mixture modeling applications; it represents a non-parametric approach with a theoretically infinite number of components. The number of components is learned from the data rather than being set a priori.

The Chinese Restaurant process can be extended to the situation where cluster assignments are associated with some relevant, observed value. If each table corresponds to a particular value, θ_k, then each customer seated at the table would be assigned that value. Conversely, if a customer is seated at a new table, they adopt a value drawn from the base distribution, H. This is consistent with the representation from Blackwell and MacQueen (1973), and reflects the posterior probability of the assigned θ value for the *nth* customer, given all previous assignments for *n-1* customers.

Another means of representing the Dirichlet process is through a ‘stick-breaking process’, and it is this conceptualization which will be directly reflected in the choice for $C(\boldsymbol{\alpha})$ here (see the WinBugs code for the MHM-2PLC in Appendix A). The ‘stick-breaking’ moniker refers to the use of a beta distribution to recursively ‘break’ segments of unit length; with the relative lengths of these components used as a prior for cluster probabilities. First, the parameter defining each class/cluster is distributed according to a base distribution as before:

$$\theta_k \sim H$$

Then the stick-breaking process to obtain cluster probabilities is defined:

$$\pi_k = \beta_k \prod_{l=1}^{k-1} (1 - \beta_l)$$

$$\beta_k \sim Beta(1, \alpha) \tag{17}$$

where the parameter, α, controls the degree of clustering as discussed previously. For the sake of simplifying presentation, the stick breaking process in (17) will be given the following representation:

$$C(\boldsymbol{\alpha}) = SB(\alpha)$$

If some a priori notion regarding the degree of clustering can be established, then α may be fixed accordingly (smaller values favor fewer clusters). In a fully exploratory approach, however, it makes sense to allow α to be freely estimated from the data. As mentioned, a common choice is the gamma distribution:

$$\alpha \sim Gamma(\varepsilon, \tau)$$

And with cluster probabilities now derived from the stick-breaking process, V is defined:

$$V = \sum_{k=1}^{K} \pi_k \delta_{\theta_k}$$

where δ_{θ_k} is the Dirac delta function. This construction is equivalent to the $V \sim DP(\alpha, H)$ described earlier.

For applications of the MHM-2PLC, a hierarchical version of the stick breaking process is applied, conforming to the hierarchical Dirichlet process presented by Teh, Jordan, Beal, and Blei (2006):

$$V_g \sim DP(\alpha_g, L)$$

$$L \sim DP(\alpha_0, H)$$

so that the result of a lower order Dirichlet process forms the base distribution for a higher order Dirichlet process. This hierarchical structure to the Dirichlet process essentially constrains the cluster specific parameter values to be constant across higher order latent groups, but allows the probability weights associated with these parameter values to differ across the groups. In the case of the MHM-2PLC specifically, this amounts to constraining the item feature weights for all strategy classes to be the same for all transition clusters (eg $\eta_{f_{k_g}} = \eta_{f_{k_{-g}}}$ and $\gamma_{f_{k_g}} = \gamma_{f_{k_{-g}}}$ for all *f*, *k* and *g*), but allowing the transition probabilities to vary by cluster. Put another way, the higher order transition clusters only differ in the pattern of strategy shifts, not the strategies themselves. Given the structure of prior distributions presented in section 2.1, the hierarchical stick-breaking process for the MHM-2PLC is easily constructed:

$$C(\boldsymbol{\alpha_0}) = SB(\alpha_0)$$

$$D(\boldsymbol{\alpha_1}) = SB(\alpha_1)$$

In the case where no assumptions are made regarding the number of strategy classes and transition clusters, and the posterior distributions of α_0 and α_1 are derived from the data evidence with a gamma distributed prior, the choice for the shape and rate parameters on the prior become crucial to ensure the number of classes/clusters is not under or over specified. This issue is discussed further in section 3 (simulation).

2.1.3 The Sparse Finite Dirichlet

The sparse finite Dirichlet approach to clustering (Frühwirth-Schnatter, & Malsiner-Walli, 2019) represents a compromise between the finite Dirichlet in 2.1.1 and the infinite, non-parametric Dirichlet in 2.2.2. The approach is not built on the theoretical concept of potentially infinite clusters (as the Dirichlet process is); rather it retrofits a finite Dirichlet prior to function similar to a Dirichlet process prior. This is accomplished by fixing K and G to large (overspecified) values and adopting conversative distributions for α_0 and α_1 that are biased towards smaller values. The distinction between the Dirichlet process prior and sparse Dirichlet prior is salient when considering the probability of forming new clusters. As mentioned, the Dirichlet process generates a new cluster with probability proportional to α:

$$p\left(\phi_i = k_{new} \middle| \boldsymbol{\phi}_{-\boldsymbol{i}}, \alpha\right) = \frac{\alpha}{N+\alpha}$$

so that new clusters are generated independently of the number of non-vacant clusters, K^+. Given that the expected number of non-vacant clusters in the Dirichlet process increases logarithmically with sample size, $K^+ = \alpha \log(N)$, the Dirichlet process favors a solution with a small number of highly populated clusters and a large number of sparsely populated clusters. Conversely, adopting a finite Dirichlet prior will produce a new cluster with probability that increases with α and K, and decreases with K^+:

$$p\left(\phi_i = k_{new} \middle| \boldsymbol{\phi}_{-\boldsymbol{i}}, \alpha, K\right) = \frac{\alpha(K-K^+)}{N+\alpha K}$$

such that the probability of generating a new cluster approaches 0 as K^+ increases. Consequently, the finite Dirichlet prior favors a solution with a moderate number of clusters that is more robust to increases in sample size.

Thus, a finite Dirichlet prior can also have utility alongside the Dirichlet process prior for exploratory clustering applications, with the proviso that it remains more conservative in estimating K^+. Of course, care needs to be taken in the choice for α, as large values will still favor a solution where $K^+ = K$. Once again, for exploratory applications it makes sense to freely estimate α from the data with a more conservative prior distribution. Choices for the parametrization of the prior distribution will be further discussed in section 3.

2.2 Constraints and Assumptions

For the HM-2PLC and MHM-2PLC, a location and scale indeterminacy exists within strategy classes similar to the 2PL model:

$$p(x_{ij} = 1 | \theta_j, k) = \frac{\exp(\frac{\sum q_{f_i} \gamma_{f_k} + \gamma_{0_k}}{s} (s\theta_{j_k} + c - s \sum q_{fi} \eta_{f_k} + s\eta_{0_k} + c))}{1 + \exp(\frac{\sum q_{f_i} \gamma_{f_k} + \gamma_{0_k}}{s} (s\theta_{j_k} + c - s \sum q_{fi} \eta_{f_k} + s\eta_{0_k} + c))} \tag{18}$$

such that equivalent fit will be exhibited for any value of s and c. A simple means of resolving this indeterminacy is to set the difficulty model intercept, η_{0_k}, and the discrimination model intercept, γ_{0c}, to some fixed value (for example, 0 and 1, respectively) for all strategy classes. Alternatively, equivalent scale anchoring can be achieved by fixing the parameters of the ability distributions across classes: $\theta_{j_k} \sim N(0,1)$, and freely estimating η_{0_k} and γ_{0_k} for all *k*. For this application, this latter constraint is chosen (and reflected in the fixed hyperparameters in 2.1). The following assumptions are also reflected in the previous presentation of the model, but are explicated here:

1) The transition matrix in (7) has no time subscript and is therefore time invariant. The $\pi_{kk'}$ transition probabilities remain constant across time points for all *k*.
2) The class specific emission probabilities are time invariant, conditional on θ_{j_k}; that is, the class specific item feature weights (η_{f_k}, γ_{f_k}) remain constant across all time points. Put another way, strategy transitions occur, but not gradual shifts in strategy. If a particular strategy evolves sufficiently over time, this would be modeled as a novel, distinct strategy with a new set of item feature weights.

The assumptions above are similar to what Rijmen and colleagues established for their hidden Markov IRT model. The MHM-2PLC carries the additional assumption that each examinee is governed by a single transition matrix during the course of testing. Just as the transition matrix in (7) is populated by values that govern the pattern of strategy transitions *within a single cluster*, a higher order transition matrix can be defined that governs *the shifting pattern from one transition cluster to another*:

$$\begin{pmatrix} \tau_{11} & \tau_{12} & \cdots & \tau_{1G} \\ \tau_{21} & \tau_{22} & \cdots & \tau_{2G} \\ \vdots & \vdots & \ddots & \vdots \\ \tau_{G1} & \tau_{G2} & \cdots & \tau_{GG} \end{pmatrix} \tag{19}$$

so that the element in row 1 and column 2, for example, represents the probability of an examinee migrating from transition matrix 1 to transition matrix 2, $A_1 \rightarrow A_2$. If this matrix is freely estimated, then an examinee's pattern of strategy transitions is allowed to shift over time to another pattern of transitions. For this application, however, I restrict this higher order transition matrix to an identity matrix:

$$\begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}$$

Future work may explore the data conditions under which this restriction may be relaxed and the matrix in (19) freely estimated.

CHAPTER 3: METHODS

3.1 Simulation Study

To evaluate the conditions under which precise and accurate parameter estimates are obtained for both the HM-2PLC and HMH-2PLC, a simulation study will be carried out across two manipulated factors: the dissimilarity of the strategy classes and the type of transition matrix. Dissimilarity of the invoked strategies was indexed by Euclidean distance between the sets of item parameters defining each strategy class. The dissimilar strategy condition was defined by a Euclidean distance of 1.414; this reflects a situation where the discrimination and difficulty parameters between strategy classes differ by an average margin of ~1, given *z* scored item features. The similar strategy class condition cut this margin by half. For the transition matrix type, 4 basic transition processes were considered: Slow regression, rapid regression, slow absorption, and rapid absorption:

$$\begin{array}{cccc} \begin{pmatrix} .8 & .2 \\ .2 & .8 \end{pmatrix} & \begin{pmatrix} .2 & .8 \\ .8 & .2 \end{pmatrix} & \begin{pmatrix} .8 & .2 \\ 0 & 1 \end{pmatrix} & \begin{pmatrix} .2 & .8 \\ 0 & 1 \end{pmatrix} \\ \textit{Slow Regression} & \textit{Fast Regression} & \textit{Slow Absorption} & \textit{Fast Absorption} \end{array}$$

For the HM-2PLC, each of the transition matrices above was considered a separate level of the manipulated transition matrix factor and fully crossed with similarity of strategy, producing eight total conditions. For the MHM-2PLC, which considers more than one transition matrix operating in the data, various pairs of the above matrices were considered as a manipulated factor. Specifically, the following three pairs were considered: Slow Regression/Fast Regression, Slow Absorption/Slow Regression, and

Fast Absorption/Slow Absorption. Similar to the HM-2PLC analysis, these manipulated conditions were fully crossed with similarity of strategy, producing six total conditions. Note additionally that across these 14 conditions (eight for HM-2PLC and six for MHM-2PLC), three types of modeling approaches were considered: parametric modeling, non-parametric modeling with a stick breaking prior, and non-parametric with a sparse finite Dirichlet prior. Thus, across all manipulations, a total of 42 conditions were considered. Sample size and number of items were fixed at 770 and 28, respectively, to reflect the properties of the empirical dataset. Tables 1 and 2 provide a summary of both the manipulated and fixed factors for the HM-2PLC and MHM-2PLC, respectively.

Table 1

Manipulated and Fixed Conditions for HM-2PLC Simulation

	Factor	Conditions
Manipulated Factors	Dissimilarity of Strategy Classes	*d*=1.414 *d*=.707
	Transition Matrix Type	Slow Regression Fast Regression Slow Absorption Fast Absorption
	Modeling Approach	Parametric (finite Dirichlet prior) Non-parametric (stick breaking prior) Non-Parametric (sparse finite Dirichlet)
Fixed Factors	Number of Subjects	770
	Number of Items	28
	Number of Strategy Classes	2

Table 2

Manipulated and Fixed Conditions for MHM-2PLC Simulation

	Factor	Conditions
Manipulated Factors	Dissimilarity of Strategy Classes	*d*=1.414 *d*=.707
	Transition Matrix Type	Slow Reg./Fast Reg. Slow Abs./Slow Reg. Fast Abs./Slow Abs.
	Modeling Approach	Parametric (finite Dirichlet prior) Non-parametric (stick breaking prior) Non-Parametric (sparse finite Dirichlet)
Fixed Factors	Number of Subjects	770
	Number of Items	28
	Number of Strategy Classes	2

The pattern of transition probabilities is an important consideration, since determines how much support is provided for estimation of IRT parameters. In the fast absorption case, for example, subjects quickly transition from strategy 1 to strategy 2 without ever regression back to strategy 1. In this case, there will be very few instances where strategy 1 is invoked, resulting in parameter estimates with diminished reliability for that strategy. The dissimilarity of strategy classes was also considered an important condition to evaluate, since it allows for an evaluation of the model's degree of sensitivity in discriminating strategy classes when the differences may be more subtle.

For both the HM-2PLC and MHM-2PLC, the number of invoked strategies will be fixed to two, reflecting common two-strategy hypothesis testing in discriminating spatial vs analytical reasoning (Milevy & Verhelst, 1990), engaged vs guessing behavior (Yamamato, 1989), and fast vs slow processing (Molenaar, Bolsinova, Rozsa & De Boeck, 2016).

As mentioned previously, an issue with exploratory applications of the presented models is selecting the appropriate number of strategy and transition clusters. Three approaches were described with respect to the prior placed on the class/cluster probabilities – use of a finite Dirichlet prior, a stick-breaking process prior, and a spare finite Dirichlet prior. Use of the finite Dirichlet prior usually carries with it a dependence on penalized fit statistics for multiple model specifications. While this approach is widely implemented and has demonstrated reasonable success in identifying the appropriate number of classes/clusters (Finch & French, 2012; Frühwirth-Schnatter, 2006), it can be prohibitively time consuming for the MHM-2PLC. The steak-breaking prior and the sparse finite Dirichlet offer a more time efficient alternative, and as thus will be preferred approaches unless shown to be inferior in parameter recovery. Theoretically, the stick-breaking process prior and sparse Dirichlet prior should perform somewhat similarly, though with the stick breaking process prior favoring more clusters with smaller representation. Whether there is a distinction between the two approaches in practice with the MHM-2PLC (and hierarchical hidden Markov models more generally) remains an open question, however. To resolve these questions, I evaluate parameter recovery and class/cluster recovery across all manipulated conditions with each of the three classification approaches. Table 3 provides detailed information regarding the choices for fixed parameter values; these choices were guided by the recommendations of Frühwirth-Schnatter (2019) for proper cluster identification in mixture models.

Table 3

Hyperpriors and Hyperparameters for Implemented Clustering Approaches at K =5

Clustering Approach	Prior Structure
Finite Dirichlet	$\boldsymbol{\pi} \sim Dirichlet(\alpha_1, \alpha_2, \ldots \alpha_K)$ $\epsilon = \alpha_1 = \alpha_2 \ldots = \alpha_K$ $\epsilon_0 = 4$
Stick-Breaking	$\pi_k = \beta_k \prod_{l=1}^{k-1}(1 - \beta_l)$ $\beta_k \sim Beta(1, \alpha)$ $\alpha \sim Gamma(1{,}10)$
Sparse Finite Dirichlet	$\boldsymbol{\pi} \sim Dirichlet(\alpha_1, \alpha_2, \ldots \alpha_K)$ $\epsilon = \alpha_1 = \alpha_2 \ldots = \alpha_K$ $\epsilon \sim Gamma(1{,}10)$

In all cases, posterior distributions were obtained through the Gibbs sampling procedure as implemented in either the OpenBugs or JAGS software packages.

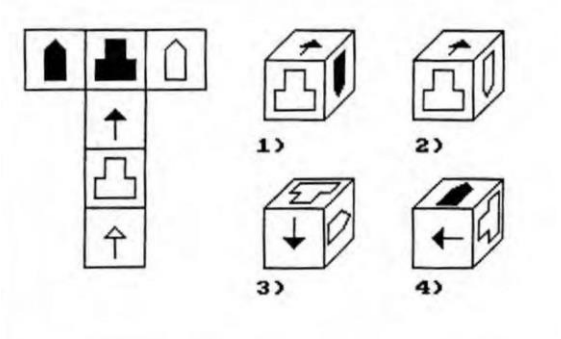
3.2 Empirical Study

To demonstrate the practical utility of the MHM-2PLC with data from cognitive and educational batteries, I apply the MHM-2PLC to a set of item responses from the Spatial Learning Ability Test (SLAT; Embretson 1994). Specifically, responses to 28 items from

769 Air Force recruits were considered for the analysis. Items required examinees to correctly match an unfolded cube to its folded counterpart (see Figure 1).

Figure 1

Example SLAT item



The expected strategy is a physical-analogue process, whereby the mental solution process reflects the physical process of converting the unfolded stem to the folded, cube form. This analogue process involves a two step procedure: in step one, two adjacent sides from the unfolded stem are mentally overlayed onto a response option. In step two, the remaining surfaces are folded to determine whether the remaining visible side of the response option matches the folded stem. Step one difficulty is affected by the degree of rotation required to overlay the stem onto each response option (0° 90° or 180°), and step two difficulty is related to the number of stem surfaces that must be folded in order to match the third side of the response option (one, two, or three folds). This dataset has been evaluated in a number of previous studies (Embretson, 1997; Embretson, 1997; Embretson, 2007) with various item response models. However, apart from one study which considered growth in the latent trait across time (Embretson, 1991), none of these previous analyses have considered longitudinal shifts in the response process. Given the items were generated according to a specific strategy approach which most examinees may not initially be familiar with, it seems reasonable to consider whether or not longitudinal changes are observed in applying this strategy. For example, many examinees may initially exhibit guessing behavior, then transition to a response process representative of the physical-analogue approach with increasing familiarity with the items. Additionally, there may be cluster differences in how quickly this transition is made, or there may be a small of cluster of examinees who invoke the physical-analogue strategy immediately on the first item. A previous analysis using the mixture Rasch model (Embretson, 2007) found evidence for four distinct classes of examinees, based on differing patterns of item difficulty. The physical-analogue process described previously was well represented in one class of examinees, while the remaining classes were comprised of an analytic-verbal group, a guessing group, and a partial analogue processing group ('partial analogue' since this latter group only seemed to engage in the analogue process for a subset of items).

I carry out a reanalysis of the data with a MHM IRT model in order to determine whether previous analyses were overly restrictive in assuming temporally invariant use of strategies. In particular, the following questions are examined:

- Do strategy class transitions occur, such that examinees shift from the guessing class to a more process oriented strategy class, with increasing familiarity with items? Do similar transitions occur within process oriented classes (e.g. transitioning from the partial analogue processing class to the full analogue processing class)?

- Do clusters of examinees exist that show differential patterns of transitions? For example, can a cluster of examinees who never transition out of the guessing class be distinguished from a cluster of examinees who do?

Consistent with a previous analysis (Embretson, 2007) the item difficulty predictors include degrees of rotation, number of surfaces carried, and the interaction thereof. Novel to the 2007 analysis, I also consider whether these item properties are related to item loading patterns within a particular strategy class. Note additionally that no a priori notion regarding the number of relevant strategies or transition matrices is established, I apply a MHM IRT model in a completely exploratory fashion with respect to the number of strategy classes and transition patterns.

CHAPTER 4: RESULTS

Generally speaking, the structural parameters of both the HM-2PLC and MHM-2PLC were recovered well from the posterior distributions for simpler models. Note, however, that when states were visited only fleetingly (as is the case when subjects rapidly transition from one state to another with no return path), IRT parameter estimates were imprecise and biased, though this is expected given the lack of data support. However, even in the case of states with only fleeting visitations, transition probabilities were estimated well.

Person level parameter estimates (strategy specific abilities, item by item strategy classifications, and transition cluster membership) exhibited a high degree of entropy in the Bayesian sampling chain, with inaccurate modal and expected values. These inaccuracies in turn produced erroneous deviance values, which in turn invalidated the use of any deviance based fit statistic for model selection in exploratory applications. As a correction, I propose an adjusted deviance statistic that simply models the overall transition cluster proportions, the change in strategy proportions at each time point, and integrates out the strategy specific abilities. This effectively produces a likelihood statistic that evaluates fit without reference to any person level parameter. This adjusted fit statistic can then be penalized by model complexity by a weighting convention of choice for model selection purposes.

Alternatively, one may wish to adopt one of the proposed non-parametric approaches for exploratory applications. However, the non-parametric, stick-breaking formulation of the MHM-2PLC demonstrated a tendency to overselect the number of strategies, as well as elevated error in the IRT parameters. Furthermore, these non-parametric approaches exhibited unreliable implementation across multiple Bayesian software packages and operating systems (Table 4 summarizes the capabilities of both OpenBUGS and JAGS for implementing these models across various operating systems).

Table 4

Feasibility of Implementing parametric and non-parametric HM-2PLC and MHM-2PLC

	OpenBUGS (Windows operating system)	OpenBUGS (Linux operating system)	JAGS (Windows operating system)	JAGS (Linux operating system)
Parametric (finite Dirichlet)	Compiles within a few minutes arrives at an optimal solution within 15,000 iterations	Compiles within a few minutes and arrives at an optimal solution within 15,000 iterations	Compiles after 12 hours and arrives at an optimal solution within 15,000 iterations	Compiles after 12 hours and arrives at an optimal solution within 15,000 iterations
Non-paramatric (stick breaking process)	Compiles within 12 hours and arrives at an optimal solution within 15,000 iterations	Does not compile within 5 days	Compiles after 1 day but does not arrive at an optimal solution within 50,000 iterations	Compiles after 1-2 days but does not arrive at an optimal solution within 50,000 iterations
Non-parametric (sparse finite Dirichlet)	Does not compile	Does not compile	Compiles but sampling stalls	Compiles but sampling stalls

And although the non-parametric form with the stick-breaking prior can be implemented OpenBUGS with a Windows operating system, its complexity results in a marked increase in runtime. Even if the ceiling for both the number of strategy classes and transition clusters is restricted to just five, the expected runtime is roughly 24 to 48 hours to obtain a stable posterior distribution. In contrast, the parametric form of the model for two strategy classes and two transition clusters requires roughly two to three hours to achieve a stable posterior distribution. Thus, it is recommended that interested users adopt the parametric forms of the HM-2PLC and MHM-2PLC for exploratory applications, and utilize the proposed adjusted deviance statistic for model selection. This recommendation is particularly strong if one expects a lower number of strategies and transition clusters in the data and has access to multiple computing cores for parallel processing. In the following sections, I review the simulation results for both HM-2PLC and MHM-2PLC in more detail and provide a more detailed account of the proposed adjusted deviance statistic.

4.1 Parameter Recovery for Parametric (finite Dirichlet) HM-2PLC

Tables 5 through 8 display bias and RMSE values for structural parameters when applying the parametric HM-2PLC to simulated data according to four different transition matrices and two levels of strategy similarity (see methods for full description of conditions). Generally, item difficulty estimates exhibited less bias and higher precision than the item discrimination estimates. Additionally, accuracy and precision were attenuated when the strategies were more similar; though still within a reasonable range for the conditions with slow regression and fast regression transition matrices. However, both the conditions with slow absorbing and fast absorbing transition matrices exhibited at least a moderate (and sometimes severe) degree of bias and imprecision. As noted previously, this is due to the lack of data to support accurate IRT parameter estimates when a state acts solely as a transition point to another state with no chance for revisitation (e.g. a transitory state along the path to an absorbing state).

Tables 9 shows parameter recovery for the transition probabilities and starting state probabilities. Note that across all conditions, the initial state and transition probability estimates exhibited reasonable precision; even when the transition matrices included an absorbing state. This is important not just for matters of substantive interpretation; it allows users to quickly determine how much trust may be placed in the difficulty and discrimination estimates. For transition matrices with either vary slow transitions to an absorbing state or regressions to previous states, the difficulty and discrimination parameters may be interpreted with a high degree of confidence. However, if a transition matrix with an absorbing state is produced (particularly with a fast transition to that absorbing state), the user should exercise extreme caution when using the difficulty and discrimination parameters to characterize the nature of the identified strategies. It is also worth noting that although the presence of an absorbing state compromises recovery of IRT parameters for the transitory state, it does lead to fairly accurate strategy classifications for persons at the item level (see Table 10) since within a few short items almost all subjects are invoking the same strategy.

As can be expected, strategy specific ability estimates were more accurate and precise for absorbing strategy states (see Table 11); this is simply due to the increased number of response observations within that strategy to support estimation. Although there was no overall bias in the ability estimates (averaged within and over replications) there was enough imprecision in the estimates for transitory states to question the model's utility for deriving individual level parameters. The inconsistencies in classifying individual strategies also reinforce this point. The model's seeming limitation at providing any reliable estimates outside of the structural parameters is a point that will be returned to in section 5.3.1 when model selection procedures are discussed for exploratory applications.

Table 5

IRT Parameter Recovery for HM-2PLC; Slow Regression Transition

	Dissimilar Strategy			Similar Strategy		
IRT Parm	True Value	Bias	RMSE	True Value	Bias	RMSE
γ_{0_1}	2	-0.033	0.132	2	0.039	0.183
γ_{1_1}	1	-0.021	0.079	.75	0.003	0.137
γ_{2_1}	0	0.005	0.052	.25	-0.005	0.128
γ_{0_2}	2	0.027	0.158	2	-0.042	0.133
γ_{1_2}	0	-0.006	0.075	.25	0.006	0.097
γ_{2_2}	1	0.006	0.093	.75	-0.044	0.151
η_{0_1}	0	0.018	0.052	0	-0.004	0.084
η_{1_1}	1	0.031	0.054	.75	-0.022	0.114
η_{2_1}	0	0.002	0.029	.25	0.029	0.122
η_{0_2}	0	0.028	0.051	0	-0.043	0.122
η_{1_2}	0	-0.001	0.032	.25	0.0423	0.137
η_{2_2}	1	-0.018	0.054	.75	-0.033	0.162

Table 6

IRT Parameter Recovery for HM-2PLC; Slow Absorption Transition

	Dissimilar Strategy			Similar Strategy		
IRT Parm	True Value	Bias	RMSE	True Value	Bias	RMSE
γ_{0_1}	2	-0.120	0.245	2	0.005	0.299
γ_{1_1}	1	-0.010	0.180	.75	-0.208	0.398
γ_{2_1}	0	-0.010	0.140	.25	0.035	0.262
γ_{0_2}	2	0.014	0.123	2	0.022	0.061
γ_{1_2}	0	0.005	0.125	.25	0.004	0.024
γ_{2_2}	1	0.003	0.070	.75	0.006	0.037
η_{0_1}	0	0.060	0.067	0	0.004	0.092
η_{1_1}	1	0.050	0.0270	.75	0.014	0.110
η_{2_1}	0	-0.015	0.038	.25	-0.016	0.078
η_{0_2}	0	-0.004	0.037	0	-0.003	0.035
η_{1_2}	0	-0.010	0.016	.25	-0.003	0.013
η_{2_2}	1	0.005	0.034	.75	-0.001	0.024

Table 7

IRT Parameter Recovery for HM-2PLC; Fast Regression Transition

	Dissimilar Strategy			Similar Strategy		
IRT Parm	True Value	Bias	RMSE	True Value	Bias	RMSE
γ_{0_1}	2	0.037	0.188	2	-0.003	0.132
γ_{1_1}	1	0.016	0.114	.75	-0.042	0.103
γ_{2_1}	0	-0.017	0.076	.25	0.017	0.157
γ_{0_2}	2	-0.001	0.096	2	-0.011	0.140
γ_{1_2}	0	0.002	0.052	.25	0.034	0.122
γ_{2_2}	1	0.003	0.070	.75	-0.040	0.157
η_{0_1}	0	-0.010	0.049	0	-0.019	0.078
η_{1_1}	1	0.003	0.034	.75	-0.018	0.125
η_{2_1}	0	0.007	0.030	.25	0.039	0.107
η_{0_2}	0	0.008	0.045	0	0.026	0.070
η_{1_2}	0	-0.003	0.027	.25	0.027	0.118
η_{2_2}	1	0.008	0.051	.75	-0.032	0.109

Table 8

IRT Parameter Recovery for HM-2PLC; Fast Absorption Transition

	Dissimilar Strategy			Similar Strategy		
IRT Parm	True Value	Bias	RMSE	True Value	Bias	RMSE
γ_{0_1}	2	-0.750	0.819	2	-0.801	0.856
γ_{1_1}	1	-1.056	1.216	.75	-0.793	1.037
γ_{2_1}	0	0.529	0.998	.25	-0.173	0.680
γ_{0_2}	2	0.064	0.096	2	0.079	0.107
γ_{1_2}	0	0.005	0.026	.25	0.008	0.031
γ_{2_2}	1	0.034	0.046	.75	0.040	0.052
η_{0_1}	0	0.304	0.806	0	0.072	0.635
η_{1_1}	1	-0.739	0.927	.75	-0.370	0.723
η_{2_1}	0	1.050	1.393	.25	0.311	0.621
η_{0_2}	0	-0.013	0.045	0	-0.001	0.044
η_{1_2}	0	<.001	0.014	.25	-0.005	0.013
η_{2_2}	1	-0.009	0.037	.75	-0.008	0.029

Table 9

Parameter Recovery for HM-2PLC Initial and Transition Probabilities

			Dissimilar Strategy		Similar Strategy	
Matrix	Parameter	True Value	Bias	RMSE	Bias	RMSE
Slow Regression	π_{I_1}	.5	-.003	.046	.019	.066
	π_{11}	.8	.001	.013	-.010	.024
	π_{22}	.8	-.003	.014	.006	.022
Slow Absorption	π_{I_1}	.5	.016	-.035	-.005	.034
	π_{11}	.8	-.012	.002	-.023	.033
	π_{22}	1	-.002	.02	-.004	.004
Fast Regression	π_{I_1}	.5	.008	.045	-.007	.057
	π_{11}	.2	-.001	.022	.001	.024
	π_{22}	.2	.007	.031	-.002	.024
Fast Absorption	π_{I_1}	.5	.009	.067	.025	.056
	π_{11}	.2	.051	.066	.050	.069
	π_{22}	1	-.014	.014	-.013	.013

Table 10

Strategy Classification Accuracy for HM-2PLC

	Dissimilar Strategy	Similar Strategy
Matrix	Strategy Classification Accuracy	Strategy Classification Accuracy
Slow Reg	.716	.621
Slow Abs	.955	.934
Fast Reg	.718	.716
Fast Abs	.944	.982

Table 11

Parameter Recovery for HM-2PLC Person Abilities

	Dissimilar Strategy						Similar Strategy					
	Bias		RMSE		$\rho_{\hat{\theta}\theta}$		Bias		RMSE		$\rho_{\hat{\theta}\theta}$	
Matrix	θ_1	θ_2	θ_1	θ_2	θ_1	θ_2	θ_1	θ_2	θ_1	θ_2	θ_1	θ_2
Slow Reg	.008	.014	.663	.669	.747	.746	-.010	-.016	.760	.747	.649	.663
Slow Abs	-.005	-.005	.915	.343	.403	.956	.003	.003	.925	.335	.379	.942
Fast Reg	-.002	-.002	.611	.615	.791	.787	.008	.008	.663	.669	.747	.746
Fast Abs	-.004	-.004	1.00	.337	.157	.917	-.004	-.004	.972	.305	.242	.954

4.2 Parameter Recovery for Parametric (finite Dirichlet) MHM-2PLC

For the novel and more general MHM-2PLC, two transition clusters were considered, with different pairings of transition matrices across the two clusters serving as the manipulated condition. Three pairings were considered: slow regression against fast regression, slow absorption against slow regression, and fast absorption against slow absorption. Additionally, these three conditions were crossed with strategy similarity. Across all conditions, Bayesian estimation of the MHM-2PLC behaved at least as well as the HM-2PLC (see Tables 12 through 14). In fact, for the MHM-2PLC, the presence of a fast absorbing state was less consequential for IRT parameter recovery, since a second cluster of examinees with different transition patterns compensated for the limited information produced by an absorbing state. For example, recall that for the HM-2PLC, IRT parameter estimates for a transitory strategy/state was severely biased in the presence of an absorbing state. However, in the case where one cluster is dominated by a fast absorbing process, but another cluster is dominated by a slower absorbing process, IRT parameter estimates are not so severely biased. This is because responses from the slower absorbing subjects provide the data support missing from the subjects who visit the transitory state only fleetingly; see Table 14 where difficulty and discrimination parameters are not severely biased when response processes are defined by both fast absorbing and slow absorbing transition matrices.

Table 12

IRT Parameter Recovery for MHM-2PLC, Slow Regression/Fast Regression Clusters

	Dissimilar Strategy			Similar Strategy		
IRT Parm	True Value	Bias	RMSE	True Value	Bias	RMSE
γ_{0_1}	2	0.065	0.176	2	0.058	0.196
γ_{1_1}	1	0.018	0.104	.75	0.005	0.115
γ_{2_1}	0	0.004	0.073	.25	0.023	0.083
γ_{0_2}	2	-0.025	0.146	2	0.011	0.162
γ_{1_2}	0	0.005	0.069	.25	0.004	0.105
γ_{2_2}	1	-0.032	0.093	.75	-0.006	0.094
η_{0_1}	0	0.028	0.068	0	-0.001	0.077
η_{1_1}	1	0.021	0.056	.75	0.029	0.066
η_{2_1}	0	-0.002	0.042	.25	0.009	0.038
η_{0_2}	0	-0.009	0.056	0	-0.001	0.072
η_{1_2}	0	-0.004	0.040	.25	-0.013	0.046
η_{2_2}	1	0.007	0.045	.75	-0.010	0.057

Table 13

IRT Parameter Recovery for MHM-2PLC, Slow Absorption/Fast Regression Clusters

	Dissimilar Strategy			Similar Strategy		
IRT Parm	True Value	Bias	RMSE	True Value	Bias	RMSE
γ_{0_1}	2	0.089	0.227	2	0.157	0.214
γ_{1_1}	1	0.039	0.130	.75	0.001	0.097
γ_{2_1}	0	0.003	0.079	.25	0.028	0.084
γ_{0_2}	2	0.013	0.109	2	0.071	0.117
γ_{1_2}	0	0.001	0.039	.25	0.011	0.059
γ_{2_2}	1	0.006	0.062	.75	0.034	0.048
η_{0_1}	0	-0.013	0.066	0	-0.029	0.075
η_{1_1}	1	-0.004	0.0688	.75	-0.017	0.073
η_{2_1}	0	-0.010	0.044	.25	0.004	0.041
η_{0_2}	0	0.005	0.058	0	0.001	0.056
η_{1_2}	0	<.001	0.020	.25	-0.029	0.035
η_{2_2}	1	0.013	0.044	.75	0.003	0.038

Table 14

IRT Parameter Recovery for MHM-2PLC, Fast Absorption/Slow Absorption Clusters

	Dissimilar Strategy			Similar Strategy		
IRT Parm	True Value	Bias	RMSE	True Value	Bias	RMSE
γ_{0_1}	2	-0.128	0.266	2	-0.121	0.250
γ_{1_1}	1	-0.121	0.212	.75	0.001	0.149
γ_{2_1}	0	0.078	0.207	.25	-0.032	0.114
γ_{0_2}	2	0.017	0.087	2	0.104	0.149
γ_{1_2}	0	-0.004	0.014	.25	0.002	0.015
γ_{2_2}	1	0.021	0.044	.75	0.059	0.069
η_{0_1}	0	0.077	0.134	0	0.029	0.097
η_{1_1}	1	-0.109	0.178	.75	-0.007	0.086
η_{2_1}	0	0.036	0.062	.25	0.011	0.087
η_{0_2}	0	-0.006	0.034	0	0.013	0.037
η_{1_2}	0	0.014	0.024	.25	-0.007	0.017
η_{2_2}	1	-0.025	0.038	.75	0.005	0.044

Similar to the HM-2PLC, transition probabilities were estimated precisely and accurately, regardless of the type of transition matrix or the similarity of strategy used to generate data (see Table 15). There is no relation between the accuracy and precision of transition probability estimates and IRT parameter estimates; reliable estimates for transition probabilities may be derived even in the face of biased IRT parameter estimates. Transition cluster proportions were also estimated quite well (see Table 16).

Table 15

Parameter Recovery for MHM-2PLC Initial and Transition Probabilities

			Dissimilar Strategy		Similar Strategy	
Matrix	Parameter	True Value	Bias	RMSE	Bias	RMSE
Slow Reg./ Fast Reg.	$\pi_{I_{1_1}}$	1	-0.141	0.154	-0.200	0.209
	π_{11_1}	.2	0.007	0.038	-0.021	0.042
	π_{22_1}	.2	0.023	0.041	0.020	0.037
	$\pi_{I_{1_2}}$	.5	0.007	0.084	0.020	0.086
	π_{11_2}	.8	-0.010	0.033	0.010	0.037
	π_{22_2}	.8	-0.005	0.027	-0.028	0.053
Slow Abs./ Slow Reg.	$\pi_{I_{1_1}}$	1	-0.143	0.182	-0.299	0.332
	π_{11_1}	.8	0.048	0.057	0.057	0.102
	π_{22_1}	.8	0.037	0.055	0.011	0.040
	$\pi_{I_{1_2}}$	.5	0.089	0.133	0.092	0.140
	π_{11_2}	.8	0.011	0.012	0.034	0.058
	π_{22_2}	1	0.061	0.076	0.068	0.091

Table 15 - continued

			Dissimilar Strategy		Similar Strategy	
Matrix	Parameter	True Value	Bias	RMSE	Bias	RMSE
Fast Abs./ Slow Abs.	$\pi_{I_{1_1}}$	1	-	-	-	-
	π_{11_1}	.2	-0.026	0.045	0.051	0.080
	π_{22_1}	1	0.040	0.042	-0.055	0.070
	$\pi_{I_{1_2}}$	1	-	-	-	-
	π_{11_2}	.8	0.053	0.057	-0.041	0.052
	π_{22_2}	1	0.004	0.004	-0.007	0.008

Table 16

Parameter Recovery for MHM-2PLC Transition Cluster Probabilities

	Dissimilar Strategy		Similar Strategy	
Matrix	Bias	RMSE	Bias	RMSE
Slow Reg./Fast Reg.	0.021	0.058	0.004	0.040
Slow Abs./Slow Reg.	-0.014	0.027	-0.034	0.065
Fast Abs./ Slow Abs.	-0.012	0.057	-0.116	0.159

However, initial state probabilities were less well estimated (see Table 15 again), exhibiting moderate bias when data was generated with some initial states being vacant. Recall that for the parametric form of the HM-2PLC and MHM-2PLC, a symmetric set of concentration parameters with large values was fixed for the Dirichlet prior (based on recommendations from Frühwirth-Schnatter, 2006). Fixing concentration parameters to be symmetric and large establishes a rather strong assumption that all states and clusters are populated at all times; this makes sense for the confirmatory approach that the parametric form is meant to serve. If, however, some classes or clusters remain unpopulated, then the model is mis-specified and parameters will be biased towards the prior. Note, that while the initial state probabilities are susceptible to this bias, the transition probabilities seem more robust to these types of overspecifications (showing about half as much bias or less, see Table 15 for bias and RMSE values for strategy states with a retention rate of 1).

As far as parameter estimates at the person level, recovery showed similar patterns as with the HM-2PLC. Strategy specific abilities were inconsistently estimated, with absorbing strategies showing satisfactory recovery but transitory strategies exhibiting increased RMSE and low correlations with true values (see Table 17). Strategy and transition classifications were also inconsistent (Tables 18 and 19), sometimes falling at just above chance values. As with the HM-2PLC, the model seems to be limited at deriving reliable structural parameters; with enough error in individual level parameter estimates to severely limit their utility or render them unusable for practical interpretation.

Table 17

Parameter Recovery for MHM-2PLC Person Abilities

	Dissimilar Strategy						Similar Strategy					
	Bias		RMSE		$\rho_{\hat{\theta}\theta}$		Bias		RMSE		$\rho_{\hat{\theta}\theta}$	
Matrix	θ_1	θ_2	θ_1	θ_2	θ_1	θ_2	θ_1	θ_2	θ_1	θ_2	θ_1	θ_2
Slow Reg/ Fast Reg	.014	.003	.663	.662	.751	.753	-.003	-.001	.715	.711	.698	.707
Slow Abs/ Slow Reg	-.010	.009	.810	.552	.595	.834	-.015	<.001	.847	.606	.545	.799
Fast Abs/ Slow Abs	-.008	-.004	.820	.340	.553	.940	.016	.014	.799	.334	.611	.943

Table 18

Strategy Classification Accuracy for MHM-2PLC

	Strategy Classification Accuracy, Dissimilar Strategy	Strategy Classification Accuracy, Similar Strategy
Slow Reg./ Fast Reg.	.710	.650
Slow Abs./Slow Reg.	.857	.800
Fast Abs./Slow Abs.	.930	.925

Table 19

Transition Cluster Classification Accuracy for MHM-2PLC

	Transition Cluster Classification Accuracy, Dissimilar Strategy	Transition Cluster Classification Accuracy, Similar Strategy
Slow Reg/ Fast Reg	.673	.618
Slow Abs/Slow Reg	.819	.730
Fast Abs/Slow Abs	.660	.592

4.3 Model Selection in Exploratory Applications

The previous section focused solely on parameter recovery for the parametric form of the model, which fixes the concentration parameters of the Dirichlet prior to be symmetrical and fairly large. This specification assumes more clustering in the solution, with all specified strategy classes and transition clusters being populated. Thus, this formulation is conducive to situations where a strong a priori notion is held regarding the nature of strategy use and transition patterns.

For exploratory applications, where no strong a priori assumptions are made regarding the response process, one generally has two options. The first is to apply a series of confirmatory models, each with a different specification, and then examine all fits to determine an optimal representation of the data. The other option is learn the concentration/clustering parameter from the data itself.

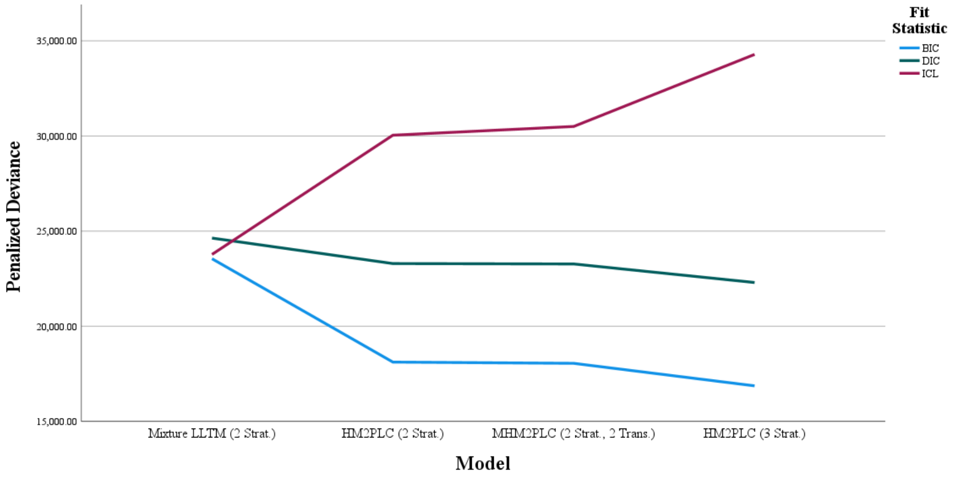
The former approach is predicated on deriving an accurate likelihood estimate to serve as the basis for a fit statistic of choice. However, in the previous section, concern was noted regarding the model’s ability to accurately and reliably derive person level parameters, even if structural parameters are estimated well. Since the likelihood estimate is based on both structural and person parameters (see equation 14), evaluating relative fit of several models no longer becomes a straightforward endeavor. In the following section, I elaborate on this challenge more fully and propose a potential solution.

4.3.1 Developing a Fit Statistic for Model Comparisons

Consider Figures 2 and 3, which show a series of fit statistics for a single replication of the HM-2PLC and MHM-2PLC, respectively. For Figure 2, data was generated with a two strategy HM-2PLC, and evaluated with the properly specified model, as well as an underspecified model where transitions were constrained to 0 (essentially a Mixture 2PLC) and two over-specified models (HM-2PLC with three strategies, and MHM-2PLC with two strategies and two transition clusters). For Figure 3, data was generated with a two strategy, two transition cluster MHM-2PLC and evaluated with the properly specified model, an underspecified model (two strategy HM-2PLC), and two over-specified models (two strategy, three transition cluster MHM-2PLC and a three strategy two transition cluster MHM-2PLC). Critically, note that in both cases, the BIC and DIC favor over-specified models, while the integrated complete likelihood statistic (ICL) favors underspecified models.

Figure 2

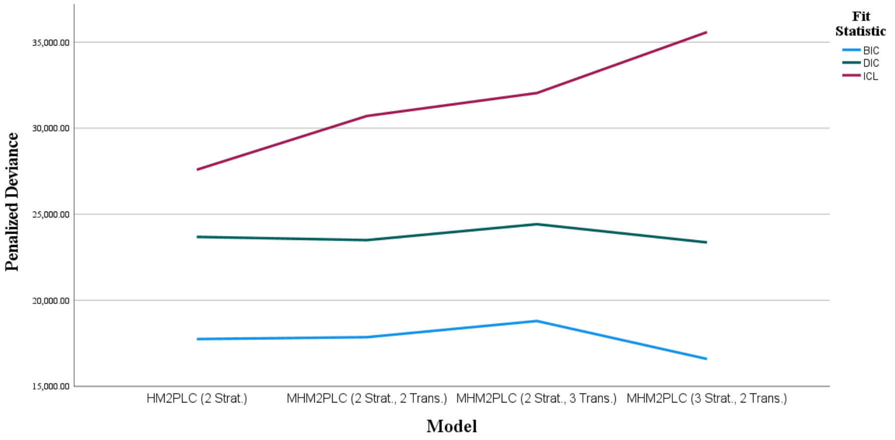
Penalized Deviances for HM-2PLC Simulated Data with Several Hidden Markov IRT Models



Note. Models are ordered with increasing complexity. The HM-2PLC (2 Strat.) represents the correctly specified model.

Figure 3

Penalized Deviances for MHM-2PLC Simulated Data with Several Hidden Markov IRT Models



Note. Models are ordered with increasing complexity. The MHM-2PLC (2 Strat., 2 Trans.) represents the correctly specified model.

The tendency to select over-specified models is not just a quirk specific to that set of generated data. Across a set of 25 replications with the HM-2PLC, the average drop in deviance between a correctly specified model with two strategies (15 structural parameters) and an over-specified model with three strategies (26 structural parameters) was 17305. For the MHM-2PLC, the average drop between a correctly specified model with two transition matrices and two strategies (19 structural parameters) and an over-specified model with two transition matrices and three strategies (35 structural parameters) was 1399. This is a substantial margin that will be difficult to compensate for with any penalty term for complexity. In fact, neither the penalties applied with BIC, DIC or WAIC were able to overcome this deficit across 25 replications, with over-specified models still being favored.

A possible explanation for the substantial drops in deviance for over-specified models is the inability to precisely and accurately estimate person level parameters. If these estimates are fraught with error, then any resulting fit statistics are based on mis-specified parameters and reflect mostly fit to error variance. As components are added to the model, relative fit seems to improve since there is no tradeoff between capturing systemic and error variance. This also aligns with the tendency for the ICL to prefer simpler and under-specified models. The ICL is essentially a traditional BIC statistic penalized by average entropy in the cluster and class assignments. Because the model produces imprecise estimates for person level parameters with little information, more complex models receive large penalties for their uncertainty in these classifications.

Thus, a simple correction would be to define a likelihood statistic that is only concerned with the fit of structural parameters, and ignores person level parameters. One possible approach to obtain this correction is to define a likelihood statistic that is marginalized over ability, strategy classifications, and transition cluster assignments. This can be accomplished by averaging response probabilities over all possible chains of strategy visitations and all possible transition clusters for each person; and furthermore by treating ability as a random effect and integrating it out within each strategy chain.

The only complication with this approach is computational feasibility. If a test with 28 items and 770 subjects is considered with two strategies, this amounts to averaging over 2^{28} possible chains of strategy visitations for each subject's response vector. Furthermore, ability must be integrated out from each of the 2^{28} chains for each person. Even with a small number of quadrature points (say 11 per ability dimension), this amounts to over 1.4 quadrillion calculations(!). The number of calculations increases exponentially if three strategies are considered. For most personal computing machines, this calculation becomes untenable. An alternative approach is to reconsider the model as a mixture model, but with a time varying mixture proportion;

$$p(\boldsymbol{x_j}) = \prod_{i=1}^{I}\sum_{k=1}^{K}\zeta_{k_i} * p(x_{ij}|q_i,\eta_k,\gamma_k,\theta_{jk}) \tag{20}$$

Note that this conforms to the structure of a traditional mixture IRT model, with the exception of an item subscript on the mixing proportion, ζ_{i_k}. For ease of estimation, one can apply the constraint that the mixing proportion evolves with time according to a Markov process:

$$\zeta_{k_i} = \zeta_{k_{i-1}}\boldsymbol{\pi_{k'k}}$$

where $\pi_{k'k}$ represents the transition probability of moving from strategy *k*' to *k* and is collected into the transition matrix, *A*. Note that this is equivalent to the HM-2PLC described previously, since the strategy transition probabilities at the individual level will be reflected in the temporal changes in value of the overall mixing proportion. The exception here is that no effort is made to derive individual strategy classifications. For the MHM-2PLC, the mixing proportion shows a different pattern of evolution across latent clusters, and is identified with a cluster specific subscript, g:

$$p(\boldsymbol{x_j}) = \prod_{i=1}^{I} \sum_{k=1}^{K} \zeta_{k_{i_g}} * p(x_{ij}|q_i, \eta_k, \gamma_k, \Theta_{jk}) \tag{21}$$

and cluster specific transition probabilities are defined:

$$\zeta_{k_{i_g}} = \zeta_{k_{i-1_g}} \boldsymbol{\pi_{k'k_g}}$$

This approach eliminates the need to derive individual strategy classifications, but still depends on other person level parameters (ability and transition cluster assignment). These can simply be marginalized out:

$$p(\boldsymbol{x_j}) = \sum_{g=1}^{G} v_g \iint \prod_{i=1}^{I} \sum_{k=1}^{K} \zeta_{k_{i_g}} * p(x_{ij}|q_i, \eta_k, \gamma_k, \Theta_{jk}) g(\Theta_1, \Theta_2|\mu, \Sigma) d\Theta_1 d\Theta_2 \tag{22}$$

and the product taken over all subjects in order to derive a likelihood statistic that only depends on structural parameters. This reduces the number of calculations to around 5 million for 28 items, 770 subjects, two strategies, and 11 quadrature points. This is much more manageable set of calculations, taking roughly five minutes on a personal computer.

To evaluate the performance of these adjusted likelihood values for model comparison and selection, I simulated data from the HM-2PLC with two strategies (25 replications), and from the MHM-2PLC with two strategies and two transition matrices (25 replications). For all replications of the HM-2PLC two strategy data, I applied the correctly specified model, as well as an underspecified model (two strategies with no transitions, i.e. a Mixture 2PLC) and two over-specified models (HM-2PLC with three strategies, and a MHM-2PLC with two strategies and two transition matrices). I then evaluated model selection with both unadjusted and adjusted BIC values (that is, BIC terms based on either the unadjusted or adjusted likelihood values previously discussed). For the MHM-2PLC generated data, a similar analysis was carried out, with several models applied to the simulated data: a correctly specified model, one underspecified model (HM-2PLC with two strategies) and two over-specified models (HM-2PLC with three strategies and MHM-2PLC with two strategies and three transition matrices).

Table 20 shows the results for the comparison of unadjusted and adjusted BIC values across all applied models with the two strategy HM-2PLC data. Note that the unadjusted BIC values select the three strategy HM-2PLC 100% of the time, while the adjusted BIC value selects the correctly specified model 88% of the time. Table 21 shows the results for the same analysis with the two strategy, two transition matrix MHM-2PLC. Again, the unadjusted BIC values select an over-specified model 100% of the time, with a vast improvement in performance observed for the adjusted BIC values. However, in the case of the MHM-2PLC, the adjusted BIC statistic is somewhat too conservative, selecting the correctly specified model 68% of the time and the underspecified model 32% of the time.

Table 20

Model Selection Performance of Unadjusted and Adjusted BIC for HM-2PLC

	Mixture 2PLC	2 Strategy HM-2PLC	2 Transition/2 Strategy MHM-2PLC	3 Strategy HM-2PLC
Parameters	13	15	19	26
Selection Rate, Unadjusted BIC	0%	0%	0%	100%
Selection Rate, Adjusted BIC	0%	88%	22%	0%

Note. Green heading indicates correctly specified model.

Table 21

Model Selection Performance of Unadjusted and Adjusted BIC for MHM-2PLC

	2 Strategy HM-2PLC	2 Transition/2 Strategy MHM-2PLC	3 Transition/2 Strategy MHM-2PLC	2 Transition/3 Strategy MHM-2PLC
Parameters	15	19	23	35
Selection Rate, Unadjusted BIC	0%	0%	0%	100%
Selection Rate, Adjusted BIC	32%	68%	0%	0%

Note. Green heading indicates correctly specified model.

4.3.2 Model Selection and Parameter Recovery for the Non-parametric MHM-2PLC

For exploratory applications, one can alternatively, attempt to learn the clustering/concentration parameters from the data. This amounts to sampling the clustering/concentration parameter itself from the posterior, rather than fixing it to a prespecified value. However, removing this constraint drastically increases runtime, both in terms of model compilation and posterior sampling. As mentioned, there are two ways to go about freely estimating the clustering/concentration parameter – with a stick breaking process prior or a sparse finite Dirichlet prior. The latter approach proved untenable with both OpenBUGS and JAGS across Linux and Windows operating systems. The stick-breaking process approach was able to be implemented in OpenBUGS on a Windows machine; however, this eliminated the possibility of running a full set of simulations across all conditions within a reasonable timeframe. Because a single run with a stick-breaking non-parametric formulation requires between three and four days, the total runtime across all simulations is roughly 3.8 years (!). Carrying out these simulations only becomes feasible with access to high performance computing resources where analyses can be distributed across 350 cores or more. While high performance computing resources were able to be utilized for the simulations with parametric forms of the model, the inability to implement the stick breaking non-parametric formulation on a Linux system excluded the possibility of submitting these analyses to a supercomputer. However, since the stick breaking non-parametric approach was at least able to be implemented on a Windows operating system, a limited simulation was carried out on a local machine. For this analysis, the simulated data was generated from a MHM-2PLC model with two strategies and two transition matrices. One transition matrix was represented by a slow regression process, and the other transition matrix was represented by a fast regression process. Additionally, the data was generated according to a pair of dissimilar strategies. As with the other simulations, sample size was set to 770 and number of items to 28 in order to reflect the properties of the empirical dataset.

First, model selection with the non-parametric MHM-2PLC entails analyzing initial and transition strategy probabilities to determine a reasonable solution. The model produces sparser populations in the various strategy classes to the extent that they are less relevant to the observed data. In an ideal situation, only relevant strategies would be populated, and any additional strategies specified would be left vacant. In practice, discretion and judgment must be exercised to determine how minimally a strategy needs to be populated before it is considered irrelevant. And as mentioned previously in section 2.1.3, the stick breaking approach to non-parametric hidden Markov modeling is expected to over-select the number of clusters. This expectation was borne out in the current analysis, with a required transition probability threshold of $\pi_{kk'} < .11$ set in order to select the appropriate number of strategy transitions in the data. In other words, any estimated transitions falling below .11 had to be ignored in order to arrive at the correct solution. However, a more conservative threshold was able to be set for selecting the number of starting states and transition clusters, with $\pi_{I_1} < .03$ and ν <.003, respectively.

As far as bias and RMSE values for IRT parameter estimates, the stick-breaking form of the MHM-2PLC exhibits moderately higher bias and RMSE in the discrimination weight parameter estimates, relative to the same analysis with a parametric MHM-2PLC (see table 22 for results from the non-parametric analysis and Table 12 for the comparison, parametric analysis). However, difficulty weights were estimated reasonably well for the stick-breaking MHM-2PLC and comparable to the parametric approach.

Table 22

IRT Parameter Recovery for Stick-breaking MHM-2PLC, Slow Regression/Fast Regression Clusters

	Dissimilar Strategy		
IRT Parm	True Value	Bias	RMSE
γ_{0_1}	2	0.321	0.463
γ_{1_1}	1	0.158	0.260
γ_{2_1}	0	0.030	0.060
γ_{0_2}	2	0.045	0.346
γ_{1_2}	0	-0.031	0.053
γ_{2_2}	1	0.023	0.155
η_{0_1}	0	0.001	0.043
η_{1_1}	1	-0.059	0.088
η_{2_1}	0	-0.016	0.028
η_{0_2}	0	0.018	0.065
η_{1_2}	0	0.015	0.040
η_{2_2}	1	-0.004	0.034

Table 23 displays bias and RMSE results for recovery of initial state and transition probabilities, as well as the transition cluster probabilities ($\boldsymbol{\pi_{I,}}\boldsymbol{\pi_{k'k}}, \boldsymbol{\nu}$, respectively).

Again, bias and RMSE values are somewhat elevated relative to the parametric solution (see Table 15). The start values in particular show very elevated RMSE values.

Table 23

Initial, Transition, and Cluster Probability Parameter Recovery for Stick-breaking MHM-2PLC, Slow Regression/Fast Regression Clusters

			Dissimilar Strategy	
Matrix	Parameter	True Value	Bias	RMSE
Slow Reg./ Fast Reg.	$\pi_{I_{1_1}}$	1	-0.081	0.121
	$\pi_{I_{2_1}}$	0	0.066	0.107
	π_{11_1}	.2	0.017	0.078
	π_{12_1}	.8	-0.048	0.114
	π_{21_1}	.8	-0.063	0.078
	ν_1	.5	.069	0.183
	π_{22_1}	.2	0.057	0.070
	$\pi_{I_{1_2}}$	.5	0.012	0.458
	$\pi_{I_{2_2}}$	.5	-0.022	0.461
	π_{11_2}	.8	-0.039	0.051
	π_{12_2}	.2	-0.028	0.077

Table 23 - continued

			Dissimilar Strategy	
Matrix	Parameter	True Value	Bias	RMSE
Slow Reg./ Fast Reg.	π_{21_2}	.2	-0.033	0.053
	π_{22_2}	.8	-0.021	0.027
	ν_2	.5	-.072	0.184

4.4 Findings from the Empirical Analysis

In order to demonstrate the practical utility of these modeling approaches in applied settings, response data from a spatial reasoning test was reanalyzed with a mixture hidden Markov IRT model (see Embretson, 2007 for a full description of the spatial reasoning test, as well as a summary in section 4 of this manuscript). Recall that items were generated according to a physical-analogue response process model, which defines a mental solution strategy that mirrors a set of physical manipulations to identify the correct answer. In the first stage, the unfolded cube in the stem is mentally rotated to align with the various response options; in the second stage, the sides are folded around the response option in order to determine if a correct match is made. Previous analyses did not consider temporal variance in the response process, and the goal here was to implement the mixture hidden Markov IRT modeling framework to identify longitudinal changes in the use of the physical-analogue strategy or alternative strategies; and furthermore, if different clusters of examinees demonstrated different patterns of longitudinal change.

No a priori notion regarding the number strategies or transition matrices operating in the data was established; this work remains completely exploratory. Thus, the non-parametric form of the MHM-2PLC with the stick breaking prior was applied in order to learn the number of strategies and transition patterns producing the response data. However, given the tendency for the stick-breaking approach to over-select the number of strategies, results were confirmed with a set of confirmatory model comparisons (utilizing adjusted BIC, see section 5.3.1). Additionally, a preliminary analysis indicated little variance in item discrimination values, so discrimination values were fixed to unity within each strategy class.

Table 24 displays results for the estimated number of transition matrices from the non-parametric approach. The model strongly identifies only a single transition cluster, with the first cluster accounting for strategy transitions in 99.87% of the examinees, and the remaining clusters together accounting for less than half a percent.

Table 24

Transition Cluster Proportions from the Non-parametric MHM-2PLC

Transition Cluster	Prob. Estimate	Posterior St. Dev.
1	.9987	.0027
2	.001196	.0027
3	<.001	<.001
4	<.001	<.001
5	<.001	<.001

Tables 25 shows the estimated initial probabilities (for up to five strategies) and Table 26 shows the transition probabilities for the first transition cluster. Note that while the model was set to explore up to five strategies, transitions for only the three most relevant strategies are displayed here for the sake of clarity of presentation. From Table 25, only two primary initial strategies are identified, with others demonstrating only trivial representation.

Table 25

Starting State Probabilities from the Non-parametric MHM-2PLC

Strategy	Prob. Estimate	Posterior St. Dev.
1	.636	.003
2	.353	.003
3	.008	<.001
4	.0013	<.001
5	.0013	<.001

Table 26

Transition Probabilities from the Non-parametric MHM-2PLC

	Transition	Prob. Estimate	Posterior St. Dev.
Strategy 1 Transitions	1 → 1	.768	.015
	1 → 2	.232	.015
	1 → 3	<.001	<.001
	1 → 4	<.001	<.001
	1 → 5	<.001	<.001
Strategy 2 Transitions	2 → 1	.713	.032
	2 → 2	.265	.026
	2 → 3	.015	.019
	2 → 4	.006	.011
	2 → 5	<.001	<.001
Strategy 3 Transitions	3 → 1	.900	.205
	3 → 2	.093	.191
	3 → 3	.009	.034
	3 → 4	<.001	.001
	3 → 5	<.001	<.001

After completing the first item, the subjects who had invoked strategy one continued to maintain that strategy with a probability of .768, and those who invoked strategy two maintained it into the next item with a probability of .713. Note that for those using strategy two, there was a very small but non-ignorable probability of transitioning into strategy three (.015 probability); however, strategy three is a fluid state; subjects immediately transition back to strategy one or two. There remains almost zero probability of subjects maintaining strategy three beyond a single item. Thus, responses seem primarily driven by two distinct strategies, with only a single transition matrix governing the patterns of strategy shifts. Note that this solution from the non-parametric Mixture Hidden Markov IRT model was also confirmed with a series of confirmatory models and adjusted BIC values (Table 27).

Table 27

Adjusted BIC Values for Hidden Markov IRT Models with Increasing Complexity

Model	Adjusted BIC
Mixture 2PLC	26606.917
2 Strategy HM-2PLC	26097.906
2 Strategy/2 Transition MHM-2PLC	26298.070
3 Strategy HM-2PLC	26158.690

Note. The model printed in green text corresponds to the solution identified with the non-parametric form. Both the non-parametric approaches and parametric approaches select the same model solution.

With a two strategy/single transition matrix solution identified, the remaining task is to characterize the nature of these strategies in terms of underlying response processes. The differences in the discrimination and difficulty weights across strategies can provide some insight into these underlying response processes. Note that no absorbing states were observed in the estimated transition matrix; thus IRT parameter estimates across all strategies can be assumed to be reasonably reliable.

. Table 28 shows the estimated values for these weights across the two strategy classes. For strategy one, the assumed physical-analogue response process seems to be represented, with both required rotation and surfaces carried being positively related to item difficulty. Note, however, that surfaces carried shows a stronger effect on difficulty; almost double the weight of spatial rotation. Thus, for the purposes of discriminating subjects across a wide range of ability levels, manipulating the number of required surfaces to be carried in an item is more effective.

Table 28

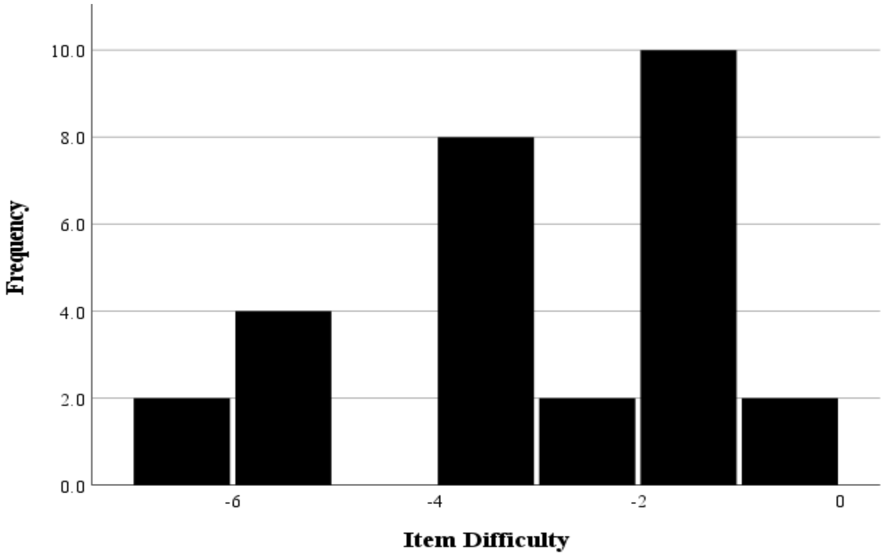
IRT Parameter Estimates for the First Two Strategies

	IRT Parm.	Estimate	Posterior St. Dev.
Strategy 1	γ_{0_1}	2.326	.5712
	η_{0_1}	.2479	.1065
	$\eta_{Rotation_1}$	.4216	.0613
	$\eta_{Surfaces_1}$	.8091	.0581
	$\eta_{Rotation*Surfaces_1}$	.03568	.03932
Strategy 2	γ_{0_2}	1.269	.9175
	η_{0_2}	-2.957	.500
	$\eta_{Rotation_2}$	-1.736	.4939
	$\eta_{Surfaces}$	-.3382	.2352
	$\eta_{Rotation*Surfaces}$	-.4893	.2465

Conversely, the physical-analogue strategy does not seem to be represented in strategy two. In fact, both the required degrees of rotation and the number of surfaces carried exhibit a *negative* relationship with difficulty. However, note the large, negative value for the intercept. The item feature scores are centered around zero, so the large negative intercept indicates all items are centered around a low difficulty value (see the histogram of produced item difficulties in Figure 4).

Figure 4

Distribution of Item Difficulties for Strategy 2



The range of item difficulties is so low that even subjects falling at the lower end of the latent trait scale are expected to have at least a moderate degree of success on any given item. Note also that items show much less discriminating power in strategy two, relative to strategy one. Thus, strategy two seems to reflect a state where items are correctly answered regardless of ability. This can be interpreted in one of two ways. One interpretation is that strategy two represents a sort of ‘hack’ to solving the items, similar to the strategy of identifying parallel end pieces to solve spatial rotation items (see Vandenberg and Kruse, 1978). However, one would expect such insight to result in an absorbing state; it would not make sense for examinees to return to an earlier strategy that requires much more cognitive effort. Recall that examinees here were found to transition back to strategy 1 from strategy 2 with an estimated probability of .713. Thus, a more logical conclusion is that the model is essentially capturing correct guessing behavior as a traditional 3PL model would. This interpretation seems even more plausible when one considers that each item is presented with four response options, and that the probability of transitioning into strategy two (.232) closely reflects the probability of selecting a correct answer at random (.25). Furthermore, the probability of remaining in strategy 2 (.265) also closely reflects the probability of randomly choosing a correct answer.

CHAPTER 5: DISCUSSION AND FUTURE DIRECTIONS

This work was concerned a flexible approach to explanatory IRT modeling that allows for temporal variance in response strategies. In particular, I consider a hidden Markov IRT approach, where strategies are defined by differential relationships of observed item features to IRT item parameters. Anchored by the work of Rijmen and colleagues (2005), who integrate the traditional LLTM within a hidden Markov framework, I address a few important questions in developing and applying temporally variant explanatory IRT models. These questions may be summarized as follows:

1. How do explanatory IRT hidden Markov models behave, in terms of parameter recovery, when time points are represented by individual items? In other words, can these models be applied when it is assumed that examinees may transition strategies after any given item? Note that Rijmen and colleagues only considered a more restrictive case where time points were represented by blocks of items.
2. Embretson (1999) has demonstrated that it may be important to define strategies in terms of the relationships between item features and *both* item difficulty and discrimination. This effectively establishes a 2PL explanatory IRT model. Can the hidden Markov LLTM from Rijmen and colleagues be extended to its 2PL counterpart with reasonable parameter recovery?
3. Acknowledging that in many applications it would be reasonable to assume cluster-level differences in the pattern of strategy transitions, is a *mixture* hidden Markov IRT model tenable? Such a novel model would essentially define cluster specific transition probabilities.

4. For exploratory applications, where the number of strategies is not known a priori, what is the optimal and most practical approach to identifying the number of strategies? Can non-parametric approaches, which learn the number of strategies form the data, provide an expedient and reliable solution?

In essence, the works represents somewhat of a ‘testing the limits’ evaluation of hidden Markov explanatory IRT models. Can the extant models be extended to 2PL and cluster-variant representations, while limiting the information available for each time point to a single item?

In general, results demonstrate that while the model is not completely untenable under these circumstances, many conditions must be considered and limitations acknowledged. In addressing questions one through three, the general finding is that while the structural parameters of the HM-2PLC and MHM-2PLC (item parameters, transition and initial strategy state probabilities, and transition cluster proportions) generally show reasonable parameter recovery, individual level parameters are often quite inaccurate and imprecise. Thus, when it is assumed that transitions may occur after any item, interpretation should be limited to characterizing processes at the population or cluster-level. The model should not be used when there is interest in characterizing the strategies and abilities of specific individuals. Even for interpreting structural parameters at the cluster-level, caution needs to be exercised. For the HM-2PLC, Fast movement to an absorbing state limits the volume of response data available for the transient state, and greatly compromises the reliability of the IRT parameters (item difficulties and discriminations) for the transient state. This issue is less prevalent for the MHM-2PLC though, as long as the same fast absorbing state is not present across all transition clusters. The transition probabilities are the only parameters that demonstrated reasonable parameter recovery across all conditions. As pointed out previously, this provides a critical diagnostic tool for assessing the reliability of item parameters. If an absorbing state is observed with a fast pathway, then much caution should be exercised in evaluating response process hypotheses with the item parameters.

For the fourth question, while non-parametric approaches seem to generally identify the correct number of strategies and transition clusters, bias and RMSE values are somewhat elevated relative to the parametric approach, especially for item discrimination parameters and starting state probabilities. However, the most salient challenge with the non-parametric approaches pertains to their difficult implementation. In this study I found that the sparse finite Dirichlet form of the non-parametric hidden Markov IRT model was essentially impossible to employ across the most widely available Bayesian software packages and operating systems. Implementation of the non-parametric form with the stick breaking prior was not much easier, though was at least viable using OpenBUGS on a Windows machine. Given these challenges, the protracted runtimes of non-parametric approaches (roughly three and a half days, even with a modest ceiling of five strategies and five transition clusters), and the elevated bias and RMSE parameter estimates, parametric approaches with a finite Dirichlet prior are recommended. At most, the stick-breaking non-parametric approach may provide some utility as a preliminary analysis to narrow the scope of models that should be considered in a set of confirmatory runs. If high performance computing resources are available, with multiple cores accessible and the opportunity to run models in parallel, then the preference for the parametric approaches becomes unequivocal.

However, a parametric approach is not without its challenges either. Choosing an appropriate fit statistic, penalized by model complexity, that reliably selects an appropriate number of strategies and transition clusters is not straightforward. Because of the inaccurate and imprecise person level parameter estimates, the likelihood values from the posterior are unreliable and tend to favor over-specified models. A potential solution to this issue, with some encouraging results, is to redefine the likelihood in a manner that only considers fit at the level of the structural parameters (see equation 22). Using simulated data, these adjusted likelihoods produced BIC statistics with vastly improved accuracy in model selection, though still were not perfect (in the case of the HM-2PLC, the rate of selecting over-specified models was reduced from 100% to 22%; and in the case of the MHM-2PLC, the adjusted BIC was too conservative at a rate of 32%, while the unadjusted BIC selected an over-specified 100% of the time). Note, however, that only 11 quadrature points were used per ability dimension when integrating out ability from the likelihood function. A reanalysis with improved precision in ability weights may show improved performance with the adjusted BIC. In any case, given the challenges of implementing non-parametric approaches, it will be imperative for to develop reliable fit statistics and procedures for model selection in future work.

It is worth noting here that the improved performance of the adjusted likelihood statistic in model selection points more generally to an alternative estimation procedure that should be considered. The Bayesian approach with Dirichlet priors pursued in this work is of limited utility for two reasons: 1) when each time point is represented by a single item, person level parameters are extremely error prone and unreliable, so estimates of structural parameters should not be conditioned on their values and 2) the principle advantage of a Bayesian approach is the availability of non-parametric clustering algorithms. Since the non-parametric approaches demonstrated exceedingly difficult implementation with questionable expediency advantages over parametric approaches, there remains little basis to justify their use. Alternatively, the adjusted likelihood statistic proposed earlier suggests the feasibility of a marginal maximum likelihood approach to estimation. This eliminates all person level parameters from the estimation procedure, and therefore avoids issues with conditioning IRT parameters, transition probabilities, and cluster proportions on erroneous person level parameters. Furthermore, estimation times may be drastically improved, since a cumbersome sampling procedure from a posterior across structural and person parameters is no longer necessary.

Of course, these points all relate to more technical aspects of implementation and estimation feasibility. A pressing question pertains to the model’s practical utility. How do the proposed approaches improve psychometric modeling in an applied sense? An example application with empirical data was carried in order to demonstrate the flexibility of mixture hidden Markov IRT models to address a wide range of response process questions. The results of the empirical analysis may initially seem uninspiring, given that only a single transition pattern was identified, with a solution that conformed to essentially a 3PL model. However, an alternative perspective is that these results speak to the flexibility of a mixture hidden Markov approach, and the utility of reconceptualizing previously defined IRT models within a broader hierarchy of temporally variant mixture models. In fact, this is a point addressed by Matthias von Davier in 2009. He notes that the guessing component of the 3PL model can easily be reparametrized, without any loss of meaning, and integrated into broader classes of mixture IRT models, discrete mixture distribution models, and multinomial processing tree models. From this perspective, the 3PL model becomes somewhat antiquated and no longer needed as a standalone model; rather, guessing processes can be identified simply as specific types of processes within a broader set of process oriented models. This effectively reduces the required ‘toolkit’ for a psychometrician, since broader modeling frameworks can be implemented to identify a wide range of processes and components in place of a series of standalone models. In this analysis, it was shown how guessing processes can be implemented in an even broader framework that extend mixture IRT models to their temporally variant counterparts. For example, the Hybrid model from Yamamato (1989), and its extension to a change point model (Yamamoto, 1995) have often been utilized to capture guessing behavior within a broader hierarchy of mixture IRT modeling. Both of these models may be subsumed by the HM-2PLC and MHM-2PLC approaches presented here. The HM-2PLC removes the constraint of one directional movement in the change point hybrid model, and the MHM-2PLC further allows for different clusters to move in and out of guessing states with different patterns.

One last point is worth mentioning for future developments within the mixture hidden Markov IRT framework. The biggest shortcoming of the approach is the inability to extract accurate person level parameters when individual items represent time points. With only dichotomous responses available for each time point, it becomes difficult to extract information regarding strategy use, ability, and transition cluster membership for each person. However, by implementing response data into the model that exists on a continuous scale, it may be easier to extract person level parameters because of the increased information available for clustering. For example, response time data may be easily incorporated into the model to improve strategy identification at the individual level.

APPENDIX A. WINBUGS/JAGS CODE FOR HM-2PLC/MHM-2PLC

A.1 HM-2PLC (finite Dirichlet, K=2, G=2)

```
model{
for (j in 1:J){
for (i in 1:I){
Response[j,i]~dbern(prob[j,i])

logit(prob[j,i]) <- (a[class[j,i],1]+a[class[j,i],2]*q[i,1]+a[class[j,i],3]*q[i,2])*(theta[class[j,i], j]-
(b[class[j,i],1]+b[class[j,i],2]*q[i,1]+b[class[j,i],3]*q[i,2]))
}
}

for (j in 1:J){
theta[1,j]~dnorm(0,1)
theta[2,j]~dnorm(0,1)
class[j,1]~dcat(start[1:2])
}

for (j in 1:J){
for (i in 2:I){
class[j,i]~dcat(phi[class[j,i-1],1:2])
}
}


phi[1,1:2]~ddirch(alpha[])
phi[2,1:2]~ddirch(alpha[])
start[1:2]~ddirch(alpha[])

alpha[1]<-4
alpha[2]<-4

for (g in 1:2){
a[g,1]~dlnorm(0,.5)
a[g,2]~dnorm(0,1)
a[g,3]~dnorm(0,1)
b[g,1]~dnorm(0,1)
b[g,2]~dnorm(0,1)
b[g,3]~dnorm(0,1)
}
}
```

A.2 MHM-2PLC (finite Dirichlet, K=2, G=2)

```
model{
for (j in 1:J){
for (i in 1:I){
Response[j,i]~dbern(prob[j,i])
logit(prob[j,i]) <- (a[class[j,i],1]+a[class[j,i],2]*q[i,1]+a[class[j,i],3]*q[i,2])*(theta[class[j,i], j]-
(b[class[j,i],1]+b[class[j,i],2]*q[i,1]+b[class[j,i],3]*q[i,2]))
}
}

for (j in 1:J){
theta[1,j]~dnorm(0,1)
theta[2,j]~dnorm(0,1)
class[j,1]~dcat(start[cluster[j],1:2])
}

for (j in 1:J){
for (i in 2:I){
class[j,i]~dcat(phi[cluster[j],class[j,i-1],1:2])
}
}

for (j in 1:J){
cluster[j]~dcat(xi[1:2])
}

xi[1:2]~ddirch(alpha[])
lambda[1]<-4
lambda[2]<-4

for (h in 1:2){
phi[h,1,1:2]~ddirch(alpha[])
phi[h,2,1:2]~ddirch(alpha[])
start[h,1:2]~ddirch(alpha[])
}

alpha[1]<-4
alpha[2]<-4

for (g in 1:2){
a[g,1]~dlnorm(0,.5)
a[g,2]~dnorm(0,1)
a[g,3]~dnorm(0,1)
b[g,1]~dnorm(0,1)
b[g,2]~dnorm(0,1)
b[g,3]~dnorm(0,1)
}
}
```

A.3 HM-2PLC (Stick-Breaking Process for K=10, G=10)

```
model{
for (j in 1:J){
for (i in 1:I){
Response[j,i]~dbern(prob[j,i])
logit(prob[j,i]) <- (a[class[j,i],1]+a[class[j,i],2]*q[i,1]+a[class[j,i],3]*q[i,2])*(theta[class[j,i], j]-
(b[class[j,i],1]+b[class[j,i],2]*q[i,1]+b[class[j,i],3]*q[i,2]))
}
}

for (j in 1:J){
class[j,1]~dcat(start[1:10])
}

for (j in 1:J){
for (i in 2:I){
class[j,i]~dcat(phi[class[j,i-1],1:10])
}
}

for (g in 1:10){
for (j in 1:J){
theta[g,j]~dnorm(0,1)
}
{

for (h in 1:10){
phi[h,1] <- r[h,1]
for (j in 2:10) {
phi[h,j] <- r[h,j] * (1- r[h,j-1]) * (phi[h,j -1] / r[h,j - 1])
}
phi.sum[h] <- sum(phi[h,])
for (j in 1:10){
r[h,j] ~ dbeta(1,10)
# scaling to ensure sum to 1
pi[h,j] <- phi[h,j] / phi.sum[h]
}
}

start[1] <- r0[1]
for (j in 2:10) {
start[j] <- r0[j] * (1- r0[j-1]) * (start[j -1] / r0[j - 1])
}
start.sum <- sum(start[])
for (j in 1:10){
r0[j] ~ dbeta(1,10)
# scaling to ensure sum to 1
pi0[j] <- start[j] / start.sum
}

for (g in 1:10){
a[g,1]~dlnorm(0,.5)
a[g,2]~dnorm(0,1)
a[g,3]~dnorm(0,1)
b[g,1]~dnorm(0,1)
b[g,2]~dnorm(0,1)
b[g,3]~dnorm(0,1)
}
}
```

A.4 MHM-2PLC (Stick-Breaking Process for K=10, G=10)

```
model{
for (j in 1:J){
for (i in 1:I){
Response[j,i]~dbern(prob[j,i])
logit(prob[j,i]) <- (a[class[j,i],1]+a[class[j,i],2]*q[i,1]+a[class[j,i],3]*q[i,2])*(theta[class[j,i], j]-
(b[class[j,i],1]+b[class[j,i],2]*q[i,1]+b[class[j,i],3]*q[i,2]))
}
}

for (j in 1:J){
class[j,1]~dcat(start[cluster[j],1:10])
cluster[j]~dcat(xi[1:10])
}

for (j in 1:J){
for (i in 2:I){
class[j,i]~dcat(phi[cluster[j],class[j,i-1],1:10])
}
}

for (g in 1:10){
for (j in 1:J){
theta[g,j]~dnorm(0,1)
}
{

for (g in 1:10){
for (h in 1:10){
phi[g,h,1] <- r[g,h,1]
for (j in 2:10) {
phi[g,h,j] <- r[g,h,j] * (1- r[g,h,j-1]) * (phi[g,h,j -1] / r[g,h,j - 1])
}
phi.sum[g,h] <- sum(phi[g,h,])
for (j in 1:10){
r[g,h,j] ~ dbeta(1,10)
# scaling to ensure sum to 1
pi[g,h,j] <- phi[g,h,j] / phi.sum[g,h]
}
}
}

for (g in 1:10){
start[g,1] <- r0[g,1]
for (j in 2:10) {
start[g,j] <- r0[g,j] * (1- r0[g,j-1]) * (start[g,j -1] / r0[g,j - 1])
}
start.sum[g] <- sum(start[g,])
for (j in 1:10){
r0[g,j] ~ dbeta(1,10)
# scaling to ensure sum to 1
pi0[g,j] <- start[g,j] / start.sum[g]
}
}
xi[1] <- x0[1]
for (j in 2:10) {
```

```
xi[j] <- x0[j] * (1- x0[j-1]) * (xi[j -1] / x0[j - 1])
}
xi.sum <- sum(xi[])
for (j in 1:10){
x0[j] ~ dbeta(1,10)
# scaling to ensure sum to 1
xpi0[j] <- xi[j] / xi.sum
}

for (g in 1:10){
a[g,1]~dlnorm(0,.5)
a[g,2]~dnorm(0,1)
a[g,3]~dnorm(0,1)
b[g,1]~dnorm(0,1)
b[g,2]~dnorm(0,1)
b[g,3]~dnorm(0,1)
}
}
```

A.5 HM-2PLC (sparse finite Dirichlet for K = 10, G=10)

```
model{
for (j in 1:J){
for (i in 1:I){
Response[j,i]~dbern(prob[j,i])
logit(prob[j,i]) <- (a[class[j,i],1]+a[class[j,i],2]*q[i,1]+a[class[j,i],3]*q[i,2])*(theta[class[j,i], j]-
(b[class[j,i],1]+b[class[j,i],2]*q[i,1]+b[class[j,i],3]*q[i,2]))
}
}

for (j in 1:J){
class[j,1]~dcat(start[1:10])
}

for (j in 1:J){
for (i in 2:I){
class[j,i]~dcat(phi[class[j,i-1],1:10])
}
}

for (g in 1:10){
for (j in 1:J){
theta[g,j]~dnorm(0,1)
}
{

for (h in 1:10){
phi[h,1:10]~ddirch(alpha[h,1:10])
}

for (h in 1:10){
for (n in 1:10){
alpha[n,h]<-E[n]
}
}

for (n in 1:10){
E[n]~dgamma(1,200)
}

start[1:10]~ddirch(beta[])
for (h in 1:10){
beta[h]<-S
}

S~dgamma(1,10)

for (g in 1:10){
a[g,1]~dlnorm(0,.5)
a[g,2]~dnorm(0,1)
a[g,3]~dnorm(0,1)
b[g,1]~dnorm(0,1)
b[g,2]~dnorm(0,1)
b[g,3]~dnorm(0,1)
}
}
```

A.6 MHM-2PLC (sparse finite Dirichlet for K = 10, G=10)

```
model{
for (j in 1:J){
for (i in 1:I){
Response[j,i]~dbern(prob[j,i])
logit(prob[j,i]) <- (a[class[j,i],1]+a[class[j,i],2]*q[i,1]+a[class[j,i],3]*q[i,2])*(theta[class[j,i], j]-
(b[class[j,i],1]+b[class[j,i],2]*q[i,1]+b[class[j,i],3]*q[i,2]))
}
}

for (j in 1:J){
class[j,1]~dcat(start[cluster[j],1:10])
cluster[j]~dcat(xi[1:10])
}

for (j in 1:J){
for (i in 2:I){
class[j,i]~dcat(phi[cluster[j], class[j,i-1],1:10])
}
}

for (g in 1:10){
for (j in 1:J){
theta[g,j]~dnorm(0,1)
}
{

for (g in 1:10){
for (h in 1:10){
phi[g,h,1:10]~ddirch(alpha[g,h,1:10])
}
}

for (g in 1:10){
for (h in 1:10){
for (n in 1:10){
alpha[g,h,n]<-E[g,h]
}
}
}

for (g in 1:10){
for (h in 1:10){
E[g,h]~dgamma(1,10)
}
}

for (g in 1:10){
start[g,1:10]~ddirch(beta[g,1:10])
}

for (g in 1:10){
for (h in 1:10){
beta[g,h]<-S[g]
}
}

for (g in 1:10){
S[g]~dgamma(1,10)
}

xi[1:10]~ddirch(delta[])
for (g in 1:10){
delta[g]<-C
}

C~dgamma(1,10)

for (g in 1:10){
a[g,1]~dlnorm(0,.5)
a[g,2]~dnorm(0,1)
a[g,3]~dnorm(0,1)
b[g,1]~dnorm(0,1)
b[g,2]~dnorm(0,1)
b[g,3]~dnorm(0,1)
}
}
```

REFERENCES

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. American Educational Research Association.

Berg, C., Hertzog, C., & Hunt, E. (1982). Age differences in the speed of mental rotation. *Developmental Psychology*, *18*(1), 95.

Bethell-Fox, C. E., & Shepard, R. N. (1988). Mental rotation: Effects of stimulus complexity and familiarity. *Journal of Experimental Psychology: Human Perception and Performance*, *14*(1), 12.

Blackwell, D., & MacQueen, J. B. (1973). Ferguson distributions via Pólya urn schemes. *The annals of statistics*, *1*(2), 353-355.

Choi, I. H., & Wilson, M. (2015). Multidimensional classification of examinees using the mixture random weights linear logistic test model. *Educational and Psychological Measurement*, *75*(1), 78-101.

Embertson, S. E. (1987). Improving the measurement of spatial aptitude by dynamic testing. *Intelligence*, *11*(4), 333-358.

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

Embretson, S. (1994). Applications of cognitive design systems to test development. In *Cognitive assessment* (pp. 107-135). Springer, Boston, MA.

Embretson, S. E. (1997). The factorial validity of scores from a cognitively designed test: The Spatial Learning Ability Test. *Educational and Psychological Measurement*, 57(1), 99-107.

Embretson, S. E. (1999). Generating items during testing: Psychometric issues and models. *Psychometrika*, *64*(4), 407-433.

Embretson, S. E. (2007). Mixed Rasch models for measurement in cognitive psychology. In *Multivariate and mixture distribution Rasch models* (pp. 235-253). Springer, New York, NY.

Embretson, S., & Gorin, J. (2001). Improving construct validity with cognitive psychology principles. *Journal of Educational Measurement*, *38*(4), 343-368.

Farrell, M. J., & Robertson, I. H. (1998). Mental rotation and automatic updating of body-centered spatial relationships. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*(1), 227.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *The annals of statistics*, 209-230.

Finch, W. H., & French, B. F. (2012). Parameter estimation with mixture item response theory models: A Monte Carlo comparison of maximum likelihood and Bayesian methods. *Journal of Modern Applied Statistical Methods*, *11*(1), 14.

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

Fischer, G. H. (1987). Applying the principles of specific objectivity and of generalizability to the measurement of change. *Psychometrika*, *52*(4), 565-587.

Frühwirth-Schnatter, S., & Frèuhwirth-Schnatter, S. (2006). *Finite mixture and Markov switching models* (Vol. 425). New York: Springer.

Frühwirth-Schnatter, S., & Malsiner-Walli, G. (2019). From here to infinity: sparse finite versus Dirichlet process mixtures in model-based clustering. *Advances in data analysis and classification*, *13*(1), 33-64.

Gluck, J., & Fitting, S. (2003). Spatial strategy selection: Interesting incremental information. *International Journal of Testing*, *3*(3), 293-308.

Greeno, J. G. (1978). Natures of problem-solving abilities. *Handbook of learning and cognitive processes*, *5*, 239-270.

Hauenstein, C. E., & Embretson, S. E. (2021). A Mixture Explanatory IRT Model for Dynamic Processes. *Multivariate Behavioral Research*, 1-1.

Hertzog, C., & Rypma, B. (1991). Age differences in components of mental-rotation task performance. *Bulletin of the Psychonomic Society*, *29*(3), 209-212.

Hopfenbeck, T. N., & Maul, A. (2011). Examining evidence for the validity of PISA learning strategy scales based on student response processes. *International Journal of Testing*, *11*(2), 95-121.

Huff, K., & Goodman, D. P. (2007). The demand for cognitive diagnostic assessment. *Cognitive diagnostic assessment for education: Theory and applications*, 19-60.

Michaelides, M. P. (2003). Age and gender differences in performance on a spatial rotation test. In proceedings of Annual Meeting of the American Educational Research Association, Chicago, IL.

Mislevy, R. J., & Verhelst, N. (1990). Modeling item responses when different subjects employ different solution strategies. *Psychometrika*, *55*(2), 195-215.

Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2004). A Brief Introduction to Evidence-Centered Design. CSE Report 632. *US Department of Education*.

Molenaar, D., Oberski, D., Vermunt, J., & De Boeck, P. (2016). Hidden Markov item response theory models for responses and response times. *Multivariate behavioral research*, *51*(5), 606-626.

Noble, T., Rosebery, A., Suarez, C., Warren, B., & O'Connor, M. C. (2014). Science assessments and English language learners: Validity evidence based on response processes. *Applied Measurement in Education*, *27*(4), 248-260.

Resing, W. C., & Elliott, J. G. (2011). Dynamic testing with tangible electronics: Measuring children's change in strategy use with a series completion task. *British Journal of Educational Psychology*, *81*(4), 579-605.

Resing, W. C., Tunteler, E., & Elliott, J. G. (2015). The effect of dynamic testing with electronic prompts and scaffolds on children’s inductive reasoning: A microgenetic study. *Journal of Cognitive Education and Psychology*, *14*(2), 231-251.

Rijmen, F., Boeck, P. D., & van der Maas, H. L. (2005). An IRT model with a parameter-driven process for change. *Psychometrika*, *70*(4), 651-669.

Schultz, K. (1991). The contribution of solution strategy to spatial performance. *Canadian Journal of Psychology/Revue canadienne de psychologie*, *45*(4), 474.

Shiina, K., Saito, T., & Suzuki, K. (1997). Analysis of problem solving process of a mental rotations test-performance in shepard-metzler tasks. *Journal for Geometry and Graphics*, *1*(2), 185-193.

Teh, Y. W., Jordan, M. I., Beal, M. J., & Blei, D. M. (2006). Hierarchical dirichlet processes. *Journal of the american statistical association*, *101*(476), 1566-1581.

von Davier, M. (2009). Is there need for the 3PL model? Guess what? *Measurement*, 7, 110–114.

Wilson, M., & Boeck, P. D. (2004). Descriptive and explanatory item response models. In *Explanatory item response models* (pp. 43-74). Springer, New York, NY.

Yamamoto, K.Y. (1989). HYBRID model of IRT and latent class models. *ETS Research Report RR89-41*. Educational Testing Service: Princeton, NJ.

Yamamoto, K. (1995). Estimating the effects of test length and test time on parameter estimation using the HYBRID model. *ETS Research Report Series*, 1995(1), i-39.