A STATISTICAL FRAMEWORK FOR DYNAMIC COGNITIVE DIAGNOSIS IN DIGITAL LEARNING ENVIRONMENTS

A PREPRINT

Yawen Ma[*]
Centre for Health Informatics, Computing, and Statistics
Lancaster Medical School, Lancaster University
Lancaster, LA1 4YW, United Kingdom
y.ma24@lancaster.ac.uk

Anastasia Ushakova
Centre for Health Informatics, Computing, and Statistics
Lancaster Medical School, Lancaster University
Lancaster, LA1 4YW, United Kingdom
a.ushakova@lancaster.ac.uk

Kate Cain
Department of Psychology
Lancaster University
Lancaster, LA1 4YF, United Kingdom
k.cain@lancaster.ac.uk

Gabriel Wallin
School of Mathematical Sciences
Lancaster University
Lancaster, LA1 4YF, United Kingdom
g.wallin@lancaster.ac.uk

June 18, 2025

ABSTRACT

Reading is foundational for educational, employment, and economic outcomes, but a persistent proportion of students globally struggle to develop adequate reading skills. Some countries promote digital tools to support reading development, alongside regular classroom instruction. Such tools generate rich log data capturing students' behaviour and performance. This study proposes a dynamic cognitive diagnostic modeling (CDM) framework based on restricted latent class models to trace students' time-varying skills mastery using log files from digital tools. Unlike traditional CDMs that require expert-defined skill-item mappings (Q-matrix), our approach jointly estimates the Q-matrix and latent skill profiles, integrates log-derived covariates (e.g., reattempts, response times, counts of mastered items) and individual characteristics, and models transitions in mastery using a Bayesian estimation approach. Applied to real-world data, the model demonstrates practical value in educational settings by effectively uncovering individual skill profiles and the skill-item mappings. Simulation studies confirm robust recovery of Q-matrix structures and latent profiles with high accuracy under varied sample sizes, item counts and different sparsity of Q-matrices. The framework offers a data-driven, time-dependent restricted latent class modeling approach to understanding early reading development.

Keywords Cognitive Diagnostic Models · Deterministic Input, Noisy "And" Gate (DINA) Models · Educational Game Application · Log Files · Q-matrix Estimation

1 Introduction

Early literacy is widely recognized as essential for educational success and lifelong development [1]. Yet, despite substantial investments in literacy education, recent data indicate persistent global challenges. For example, according to the Progress in International Reading Literacy Study, 86% of fourth-grade pupils in England reached the Intermediate International Benchmark, compared to 81% in the United States, 37% in Brazil, and only 9% in South Africa, highlighting substantial global disparities in basic literacy achievement [2, 3]. To address gaps in literacy skills, some countries advocate the integration of evidence-based digital reading support in classrooms, to support students with specific education needs and English language learners [4]. Building on the demonstrated benefits of individualised and

[*] Corresponding author.

real-time feedback in digital learning environment [5], this study addresses the critical need to empirically evaluate how digital reading tools can support early literacy development. Although these educational technologies are now widely adopted, relatively few studies have utilised the rich log files they generate to capture fine-grained learning processes. Such log files enable detailed tracking of student interactions, such as response times and learning trajectories, which in turn facilitates statistical modeling of students' learning development.

Many existing approaches to modeling students' learning development employ latent variable frameworks, in which unobservable proficiencies are inferred from observable item responses. A widely used class of such models in educational and psychological assessment is the restricted latent class models, also known as cognitive diagnostic models (CDMs), originally proposed by [6] and further termed by [7]. [6] initially referred to these models as binary skills models, which classify students into one of 2^K latent profiles based on their mastery or non-mastery of skills (also called "attributes"), where K denotes the number of skills of interest. These models typically account for guessing (responding correctly by chance despite lacking mastery) and slipping (responding incorrectly despite possessing the skill) behaviours [8, 7]. CDMs provide diagnostic feedback for teachers to help them make informed decisions regarding targeted instruction or interventions [9].

While CDMs and their variants have demonstrated effectiveness across diverse applications, traditional implementations face several limitations that researchers have sought to address. First, these models are typically restricted to static, single time point assessments. To overcome this constraint, researchers have developed various extensions to capture learning over time. Different approaches have integrated CDMs with transition models to track temporal skill development. For example, latent transition analysis (LTA; [10]) combined with CDMs can assess shifts in mastery across repeated assessments [11]. Building on this framework, [12] introduced a bias-corrected three-step method for latent transition CDMs to evaluate covariate effects. Several studies have incorporated hidden Markov models with CDMs to track skill acquisition in computer-based spatial learning interventions [13, 14, 15, 16]. More recently, [17] proposed a Bayesian longitudinal extension of restricted latent class models - structured as a directed graphical model [18] - which accommodates polytomous attributes and allows covariates to influence transitions between latent states.

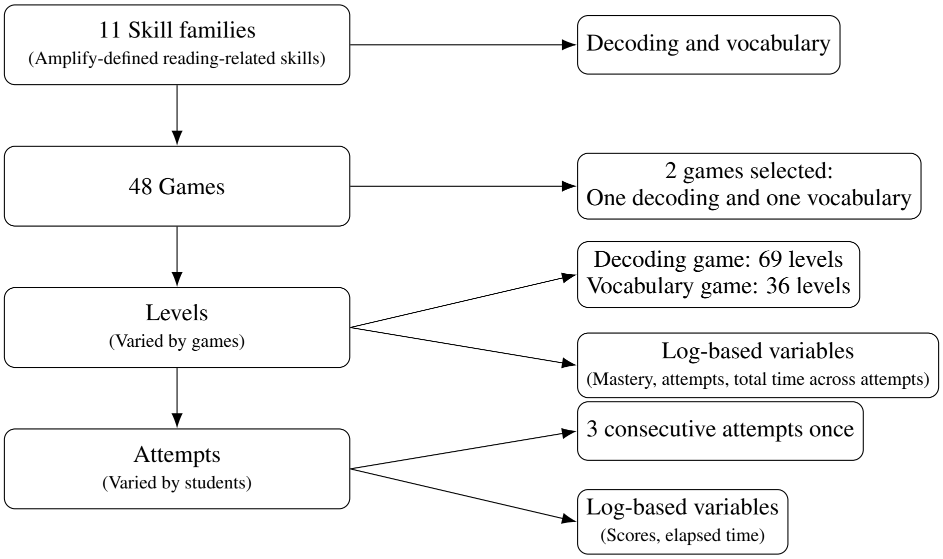
Second, CDMs incorporate a design matrix, known as the Q-matrix, which specifies the mapping between the test items and the attributes of the underlying skills. Each row of the Q-matrix corresponds to an item and each column to a skill, with entries of 1 indicating that the item requires the skill and 0 otherwise. Although the Q-matrix itself only defines the item-skill relationships, the complete CDM framework uses this structure to generate diagnostic inferences about individual skill mastery. However, a misspecified Q-matrix can significantly bias parameter estimates and diagnostic classifications [19]. Many applications of CDMs rely on a predefined Q-matrix provided by domain experts [13, 9], or focus on validating and modifying pre-specified expert Q-matrices ([20]; [21]; [22]). To overcome the limitations of such confirmatory approaches, researchers have developed data-driven, exploratory estimation methods. For instance, [23] proposed a regularised estimator for the Q-matrix in a cross-sectional setting, whereas [24] developed a Bayesian approach. [17] compared their longitudinal model's Q-matrix estimates to a prior confirmatory analysis by [25]. Prior work has furthermore established identifiability conditions for the Q-matrix [23, 26, 24].

Third, most CDMs assume that students' attributes are binary, which may be overly simplistic for domains such as reading development, where a broader range of proficiency levels are necessary to sensitively capture ability. Polytomous skill models have been developed to address this limitation, incorporating partial response accuracy [27, 28, 29] or jointly modeling response accuracy and response times to account for accuracy-speed trade-offs [30, 31, 32]. These polytomous approaches have been subsequently adapted to dynamic frameworks to track learning progression over time.

Finally, while there is growing interest in integrating log files into CDMs, most existing applications either target different populations or use alternative modeling approaches. For example, [33] employed machine learning techniques on gameplay data to detect reading difficulties. Other studies have explored process data in university-level reading subskills using CDMs [34], or adult problem-solving skill within technology rich environments [35]. Although these studies illustrated the value of log files, the application of CDMs to log files in early literacy development remains limited. Yet, log files provide opportunities to track not only response correctness but also dynamic insights into time-related learning processes, such as response times, item durations, and session timestamps. This represents a significant missed opportunity to understand learning processes in a finer temporal grain, particular in real-world digital environments designed for young learners.

Our approach in this study was motivated by our use of digital learning log files, which included multiple time-points, an unknown Q-matrix, log-derived behaviour indicators and individual characteristics for each student. This resulted in a novel framework for time dependent restricted latent class modeling that integrated the various extensions discussed above into a single unified approach. While previous research has addressed individual aspects, such as temporal dynamics, data-driven Q-matrix estimation, or integration of log-derived covariates and individual characteristics, our model synthesised these advances within one comprehensive framework. We jointly estimated all components,

Figure 1: The hierarchical structure of the log files. The left column shows the full structure of Boost Reading (skill families, games, levels, and attempts). The right column highlights the subset selected for analysis, including two skill families, one game from each, and relevant levels and attempts.



including the item-skill relationship (Q-matrix), time-varying latent skill profiles, and transition parameters, within a single integrated system. The Q-matrix was thus inferred directly from the data, based on the dependence structure among item responses rather than assuming that it was known. Simultaneously, the framework modeled the dynamic evolution of skills over time while incorporating rich log-derived behavioural indicators and individual characteristics, including reattempts, response times, counts of mastered items, demographics, and game and learning characteristics, leveraging the full information available in digital learning environments.

The remainder of this paper is structured as follows. Section 2 provides the background and description of the digital learning data used in our analysis. Section 3 details our methodological framework, including the full specification of the time-dependent restricted latent class model. In Section 4, we present an empirical study applying our model to real-world educational data, and Section 5 evaluates the model's performance through simulation studies. Section 6 discusses the implications of our findings, limitations, and directions for future research. Finally, Section 7 provides a summary of our contributions and their significance for both statistical methodology and educational practice.

2 Data Background

The data for this study originate from the Boost Reading digital program, developed by Amplify, a U.S.-based education technology company (`https://amplify.com`). Founded in 2000, Amplify now reaches more than 5,000 school districts, serving over 15 million students in 2024. Amplify offers a wide range of curriculum programmes in literacy (Boost Reading is one of them), science, and mathematics, providing schools and educators with digital tools to support effective teaching and learning. Boost Reading (previously Amplify Reading) is a research-informed classroom-based digital reading supplement consisting of multiple literacy-focused games targeting core reading skills such as phonological awareness, decoding, vocabulary, and sentence comprehension. These games are categorised into distinct research-informed defined skill families, each aligned with a foundational reading skill, specifically designed for students in kindergarten through 5th grade (K-5). The diagram in Figure 1 summarises the hierarchical structure of the log files from Boost. The left column displays the complete structure (11 skill families, 48 games, levels, attempts), while the right column specifies the subset analysed in this paper, focusing on decoding and vocabulary skill families and one game from each. The impact of Boost games and student interaction in K–2 is reported in [36].

Within each skill family, multiple games include different levels. Students achieve either "mastery" or "no mastery" at each level, where mastery is defined as achieving approximately 80% accuracy on the items they engage with that they can either get right or wrong. Students may attempt each level an unlimited number of times. However, after three consecutive non-mastery attempts for a level students are directed to other games to support related skills before they reattempt the challenging levels of that game. Students do not choose which game or level to engage with; instead, access to games and levels is determined by their initial ability and on-going in-game performance. Within each game, levels are unlocked sequentially, progressing from foundational to advanced levels. As a result, students follow different learning trajectories and encounter varying sets of levels. In addition to detailed log files for specific games, we also obtained students' initial reading performance measured by the Dynamic Indicators of Basic Early Literacy Skills (DIBELS; [37]), administered prior to the Boost program and used for placement.

3 Methodological Framework

Using the data structure outlined above, we developed a comprehensive framework for analysing student learning trajectories in digital environments. Our approach addressed the limitations of traditional CDMs by jointly estimating three key components: (1) the underlying item–skill mappings (Q-matrix), (2) item-specific parameters and individual skill profiles across time points, and (3) the effects of log-derived covariates and individual characteristics on both initial skills mastery and their transitions.

3.1 Model Overview

The proposed model integrated cognitive diagnostic measurement with temporal dynamics in analysing how students' proficiency in specific skills evolved over time. At each time point t, student i's observed responses to learning items were denoted by $\mathbf{Y}_{i,t} = (Y_{i,1,t}, \ldots, Y_{i,J,t})$, where $Y_{i,j,t} \in \{0, 1\}$ indicated an incorrect or correct response to item j. These observed responses were governed by latent skill mastery patterns and item-specific parameters, as detailed in the following subsections.

3.2 Cognitive Diagnostic Model

CDMs are restricted latent class models designed to infer students' mastery of specific skills from their observed responses. Under these models, each student i at time t is characterised by a binary attribute vector $\boldsymbol{\alpha}_{i,t} = (\alpha_{i,1,t}, \ldots, \alpha_{i,K,t})$, where $\alpha_{i,k,t} \in \{0, 1\}$ indicates mastery (1) or non-mastery (0) of attribute k.

Among various CDMs, we employed the Deterministic Inputs, Noisy "And" Gate model (DINA; [6, 38]). The DINA model assumes a non-compensatory structure, meaning that students must master all required attributes to successfully complete an item. Formally, the ideal response indicator $\eta_{i,j,t}$ for student i on item j at time t is given by:

$$\eta_{i,j,t} = \prod_{k=1}^{K} \alpha_{i,k,t}^{Q_{j,k}}, \tag{1}$$

where $\eta_{i,j,t} = 1$ indicates full mastery of all required attributes, and $\eta_{i,j,t} = 0$ otherwise.

To account for response uncertainty, the DINA model incorporates two parameters: slipping ($s_{j,t}$) and guessing ($g_{j,t}$). The probability of a correct response is modeled as:

$$P\big(Y_{i,j,t} = 1 \mid \boldsymbol{\alpha}_{i,t}, \mathbf{Q}\big) = (1 - s_{j,t})^{\eta_{i,j,t}} g_{j,t}^{1-\eta_{i,j,t}}, \tag{2}$$

$$s_{j,t} = P(Y_{i,j,t} = 0 \mid \eta_{i,j,t} = 1), \quad g_{j,t} = P(Y_{i,j,t} = 1 \mid \eta_{i,j,t} = 0). \tag{3}$$

Thus, a student who has mastered all required attributes may still respond incorrectly due to slipping ($s_{j,t}$), while a student lacking mastery may respond correctly by guessing ($g_{j,t}$). We selected the DINA model for its straightforward interpretations, smaller sample size requirements for accurate parameter estimation [39], and its flexibility for extension to more general CDMs.

3.3 The Q-Matrix

A critical component of our framework is the Q-matrix, a binary matrix that specifies which skills are required for each item:

$$Q_{jk} = \begin{cases} 1 & \text{if item } j \text{ requires skill } k, \\ 0 & \text{otherwise.} \end{cases}$$

Unlike many applications where this structure is pre-specified by domain experts, in Boost Reading, the mapping between items and skills [2] is not explicitly defined. We therefore treated each element Q_{jk} as a parameter to be

[2] Here, "skill" is used differently from the definition used in Boost Reading.

estimated from the data. Table 3 presents an example of an estimated Q-matrix from our application. The structure of the Q-matrix plays a critical role in model identifiability [40]. To ensure proper estimation, we implemented several established identifiability constraints. Under the DINA model, a Q-matrix is considered complete if it contains a $K \times K$ identity submatrix I_K up to column permutation [20]. Following [23] and [26], we fixed two such I_K submatrices (i.e., $2K$ single attribute items) in each Q_t when J exceeded 15 items. For the remaining elements, we imposed a structured sparsity pattern. Each item was constrained to measure one or two attributes, and each attribute had to be measured by at least three items [41]. The global sparsity of the estimated Q-matrix is quantified by:

$$\theta = \frac{\text{number of non-zero entries in } Q_t - 2K}{J \times K}.$$

3.4 Transition Model with Covariates

An important aspect of the modeling framework is how students transition between mastery states over time, and how these transitions are influenced by learning behaviours captured in log files and individual characteristics. We employed logistic regression models to link covariates $\boldsymbol{Z}$ to both initial skill mastery and transitions between states.

The initial attribute mastery probabilities at the first time point are modeled as:

$$\text{logit}\left(P(\alpha_{i,k,t=1} = 1)\right) = \beta_{0,k} + \sum_{c=1}^{C} \beta_{k,c} Z_{i,c}, \tag{4}$$

where $\beta_{0,k}$ represents the baseline log-odds of mastering attribute k initially, $\beta_{k,c}$ quantifies the effect of covariate c on this probability, and $Z_{i,c}$ denotes the value of covariate c for student i. For subsequent time points, we modeled transitions between latent attribute states from time t to $t+1$ as:

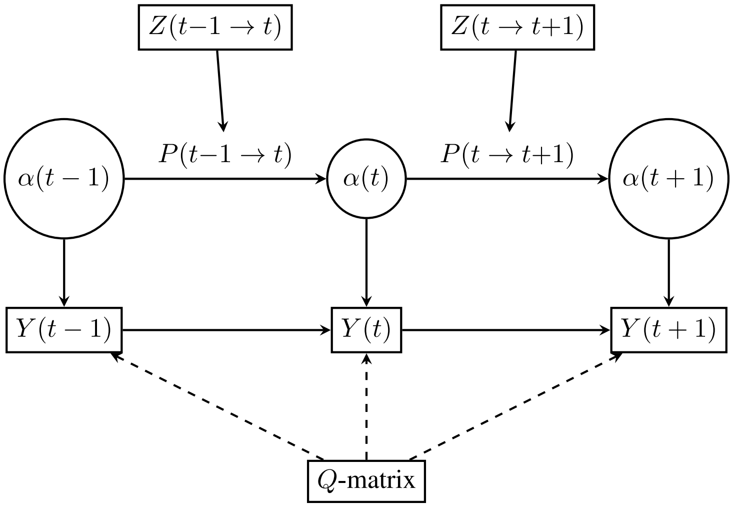
$$\text{logit}\left(P(\alpha_{i,k,t+1} = 1 \mid \alpha_{i,k,t} = 0)\right) = \gamma_{01,k,0} + \sum_{c=1}^{C} \gamma_{01,k,c} Z_{i,c}, \tag{5}$$

$$\text{logit}\left(P(\alpha_{i,k,t+1} = 0 \mid \alpha_{i,k,t} = 1)\right) = \gamma_{10,k,0} + \sum_{c=1}^{C} \gamma_{10,k,c} Z_{i,c}, \tag{6}$$

Here, $\gamma_{01,k,0}$ and $\gamma_{01,k,c}$ quantify the baseline and covariate effects on transitioning from non-mastery to mastery of attribute k, while $\gamma_{10,k,0}$ and $\gamma_{10,k,c}$ quantify the effects on transitions from mastery back to non-mastery. By incorporating log-derived covariates and individual characteristics, our model was able to identify which specific covariates were most predictive of skill acquisition and retention.

The relationships between covariates Z, latent attributes α, and observed responses Y across time points are summarised in Figure 2.

Figure 2: The relationships between covariates Z, latent variables α, and responses Y across three time points.



3.5 Inference Procedure

To estimate all model parameters simultaneously, we employed a Bayesian approach using Markov Chain Monte Carlo (MCMC). This enabled us to quantify uncertainty in all parameters while incorporating prior knowledge where available.

3.5.1 Prior Specifications

To encourage sparsity in the Q-matrix, we adopted a hierarchical Bernoulli–Beta prior:

$$Q_{jk} \sim \text{Bernoulli}(\theta), \tag{7}$$

$$\theta \sim \text{Beta}(\alpha, \beta), \tag{8}$$

where the prior mean $\frac{\alpha}{\alpha+\beta}$ equals the non–zero proportion of the true Q-matrix and the concentration defined as $\alpha + \beta$. This structure permits the data to inform the overall sparsity level. For the empirical analysis, the $\text{Beta}(24, 6)$ prior had a mean of 0.8 and a variance of approximately 0.0052, placing roughly two-thirds of its mass between 0.65 and 0.92. Additionally, we conducted a sensitivity analysis with alternative priors reported in supplementary material A (posterior means of g and s differed by less than 0.04), finding that $\text{Beta}(24, 6)$ provided the best balance between model fit and interpretability of the resulting Q-matrix structure. This prior is informative enough to guide estimation toward meaningful structures while remaining flexible enough to adapt to the data.

To further ensure identifiability, we fixed two $K \times K$ identity submatrices in each Q_t when $J > 15$, following established procedures [23, 26]. The remaining elements were estimated from the data, with the constraint that each attribute was measured by at least three items to satisfy sufficient conditions for identifiability [41].

For the guessing and slipping parameters, we used weakly informative priors following [42]:

$$g_{j,t} \sim \text{Beta}(1, 1) \tag{9}$$

$$s_{j,t} \sim \text{Beta}(1, 1) \tag{10}$$

We initialised these parameters with draws from $\text{Uniform}(0, 0.3)$ to reflect the empirical observation that item-level guessing and slipping parameters rarely exceed 0.30 in applied settings [43, 44]. For the regression coefficients in the attribute and transition models (i.e., $\boldsymbol{\beta}_Z$, γ_{01}, and γ_{10}), we specified weakly informative priors assuming a standard normal distribution, $N(0, 1)$. In order to test the sensitivity of the choice of priors, we varied them and re-ran the analysis. In doing so, we found little difference in both classification performance and posterior densities of hyperparameters.

3.5.2 Joint Posterior Distribution

The joint posterior distribution of all parameters given the observed data is:

$$\begin{aligned} P(\mathbf{Q}, \mathbf{g}, \mathbf{s}, \beta, \gamma_{01}, \gamma_{10}, \alpha_1, \ldots, \alpha_T \mid \mathbf{Y}_1, \ldots, \mathbf{Y}_T, \mathbf{Z}) \propto & \Big(\prod_{t=1}^{T} P(\mathbf{Y}_t \mid \mathbf{Q}, \mathbf{g}_t, \mathbf{s}_t, \alpha_t)\Big) \cdot \\ & \Big(\prod_{t=2}^{T} P(\alpha_t \mid \alpha_{t-1}, \gamma_{01}, \gamma_{10}, \mathbf{Z}_{t-1})\Big) \cdot \\ & P(\alpha_1 \mid \beta, \mathbf{Z}_0) \cdot P(\mathbf{Q}) \cdot \\ & P(\mathbf{g}, \mathbf{s}) \cdot P(\beta, \gamma_{01}, \gamma_{10}) \end{aligned} \tag{11}$$

The first component, $P(\mathbf{Y}_t \mid \mathbf{Q}, \mathbf{g}_t, \mathbf{s}_t, \alpha_t)$, represents the likelihood of observed responses at time t based on the DINA model defined in Equation 2. The second component, $P(\alpha_t \mid \alpha_{t-1}, \gamma_{01}, \gamma_{10}, \mathbf{Z}_{t-1})$, captures the transition probabilities between attribute states from time $t-1$ to t as defined in Equations 5 and 6. The third component, $P(\alpha_1 \mid \beta, \mathbf{Z}_0)$, models the initial attribute mastery probabilities at time $t = 1$ using the logistic regression in Equation 4. The remaining terms, $P(\mathbf{Q})$, $P(\mathbf{g}, \mathbf{s})$, and $P(\beta, \gamma_{01}, \gamma_{10})$, represent the prior distributions for the Q-matrix, the guessing and slipping parameters, and the regression coefficients, respectively, as specified in the previous section.

3.5.3 Computation

The model was implemented using JAGS (Just Another Gibbs Sampler) through the R package RJags [45], guided by the approach in [46]. For each model, we ran three parallel chains with different starting values. After a burn-in period of 5,000 iterations, we collected 10,000 samples from each chain and assessed convergence using the potential scale reduction factor ($\hat{R}$) and trace plot inspection.

The complete code for reproducing all results in this study is available on GitHub[3]. Additionally, our research plan has been preregistered via the Open Science Framework at `https://osf.io/5mv3r`.

4 Empirical Study

4.1 Data

We selected two games that support the development of two core components of reading comprehension: decoding skills, which support word recognition, and vocabulary, which supports language comprehension [47]. Theoretically, word recognition and language comprehension are considered as separate constructs, and earlier analysis of a related dataset confirms that student engagement and performance on each are separable [48]. In our empirical study, we treated decoding and vocabulary as two target attributes. To ensure sufficient engagement and a balanced sample, we included 263 students who had interacted with both games corresponding to these skills. In our analysis, we examined the beginning of Year 1 and Year 2 of the study dataset ($T = 2$ in Equation 11). We extracted six responses per time point from the student log files from the first level that they interacted with for the decoding game (items 1-3) and for the vocabulary game (items 4-6). The items and levels differed by individual student due to different initial start points and in-game trajectories. Detailed descriptions of the game-level content are provided in the Appendix, and exploratory data analysis on level engagement variability (presented in supplementary materials B) confirmed that the majority of students engaged with a consistent set of comparable game levels, thereby supporting the simplification adopted in our study.

To control for individual differences, we considered 12 covariates informed by prior research [48]. These comprised six log-based variable—average number of attempts, number of levels mastered, and average response times—calculated separately for each of the two games, as presented in Table 1. Additionally, we included six individual characteristics—four demographic variables (gender, race, English language status, and special educational needs) and two variables related to game and learning characteristics (initial literacy ability and engagement group)—as summarised in Table 2. The engagement group variable reflected students' overall participation across multiple literacy games in Boost, based on a set of theory- and data-driven engagement indicators [48]. That study identified nine distinct engagement profiles, and students in our sample were assigned to three of these profiles. Specifically, students in the high vocabulary group engaged more with games targeting vocabulary skills, those in the high decoding group engaged more with decoding-focused games, and the balanced group demonstrated comparable engagement across both skill families.

Table 1 presents the summary statistics for the log-based variables. In each academic year, students completed six items in our sample, comprising three levels per game, with a maximum of three levels mastered in each game. Compared to the vocabulary game, students in the decoding game exhibited slightly higher average levels mastered, longer average response times, and lower variability across measures. In contrast, students showed higher average reattempts in the vocabulary game than in the decoding game, suggesting greater effort was needed in the vocabulary game.

Table 1: Summary statistics for log-based continuous variables.
Note: NLM = average number of levels mastered; REAT = average number of reattempts; RT = average response time. For REAT and RT, individual means were first computed across game levels for each student and then averaged across the entire sample. For NLM, the number of levels mastered was first counted at the individual level before calculating the group average.

Variable	Mean (SD)	Quantiles: (1st, 3rd)
Log-based Variables		
NLM decoding game	1.87 (0.90)	(1, 3)
NLM vocabulary game	1.76 (0.93)	(1, 2)
REAT decoding game	2.14 (1.41)	(1, 3)
REAT vocabulary game	2.41 (1.55)	(1, 3)
RT decoding game	2.14 (0.58)	(1.78, 2.47)
RT vocabulary game	1.80 (1.25)	(1.13, 2.06)

As shown in Table 2, approximately half of the students belonged to the high vocabulary group. Most students in this group were non-English language learners (non-ELL) and did not have special education needs (non-SEN). The gender distribution was relatively balanced, with a slightly higher proportion of male students. Regarding initial literacy ability, students were distributed across different benchmark levels on DIBELS, with a notable proportion classified as Well Below Benchmark or At Benchmark defined by Amplify criteria. The sample included students from diverse racial and

[3] `https://github.com/Yawen-Ma/Q-matrix.git`

ethnic backgrounds, with Hispanic or Latino, White, and Black or African American students comprising the largest groups.

Table 2: Summary statistics for categorical demographic and game and learning characteristics variables. Values represent the number of students in each category, with the corresponding proportion.
Note: ELL = English language learner; SEN = special educational needs; B = Black or African American; H = Hispanic or Latino; M = Multiracial/Other; W = White; NS = Not Specified; AI = American Indian; AN = Alaskan Native; AS = Asian.

Variable	**Summary**
Demographic Variables	
ELL	non-ELL: 196 (74.5%); ELL: 67 (25.5%)
SEN	non-SEN: 233 (88.6%); SEN: 30 (11.4%)
Gender	Female: 117 (44.5%); Male: 146 (55.5%)

Race	AS: 15 (5.7%); B: 34 (12.9%); H: 63 (24%); M: 27 (10.3%);
	W: 58 (22.1%); NS: 47 (17.9%); Other (AI & AN): 3 (1.2%);
	Missing: 16 (6.1%)

Game and Learning Characteristics	

Engagement Group	High Vocabulary Group: 142 (54%);
	High Decoding Group: 26 (9.9%);
	Balanced Group: 95 (36.1%)

Initial Literacy Ability	Above Benchmark: 53 (20.2%); At Benchmark: 57 (21.7%);
	Below Benchmark: 38 (14.4%);
	Well Below Benchmark: 75 (28.5%); NA: 40 (15.2%)

4.2 Model Diagnostics

We applied the proposed model to the dataset and assessed convergence following [49]. From 60,000 total iterations (3 chains with 20,000 for each), the first half were discarded as warm-up. Diagnostics indicated the maximum potential scale reduction factor ($\hat{R}$) with 1.018. The minimum effective sample size (ESS) were 5,661— above the recommended threshold of 400—indicating efficient sampling. Running time for the empirical analysis (MCMC chain length = 30,000) was approximately 25 minutes, conducted on a MacBook Pro (13-inch, M1, 2020) equipped with an Apple M1 chip (8-core: 4 performance and 4 efficiency cores) and 16 GB of unified memory.

4.3 Estimation and Evaluation of the Q-Matrix and Item Parameters

The estimated Q-matrices at Time 1 and Time 2 are shown in Table 3. At Time 1, the Q-matrix aligned with the test design: Items 1–3 targeted decoding, and Items 4–6 targeted vocabulary, reflecting the game structure. Note that the game structure was treated as unknown when we estimated the model, adding evidence that the model was able to correctly identify the item-skill relationship. At Time 2, Items 2, 3, 4, and 6 loaded on both decoding and vocabulary, indicating increased skill integration.

To evaluate the estimated Q-matrices, we adopted the proportion of variance accounted for (PVAF), a criterion based on the G-DINA discrimination index (GDI; [50]). For item j, let $K_j^* = \sum_{k=1}^{K} q_{jk}$ denote the number of required attributes, and let $\boldsymbol{\alpha}_{lj}^*$ denote the lth reduced attribute pattern among the $L = 2^{K_j^*}$ possible combinations. The GDI is then defined as the variance of success probabilities across these patterns given a possible q-vector $\mathbf{q}$:

$$\zeta_j^2(\mathbf{q}) = \sum_{l=1}^{2^{K_j^*}} p(\boldsymbol{\alpha}_{lj}^* \mid \mathbf{q}) \left[P(Y=1 \mid \boldsymbol{\alpha}_{lj}^*, \mathbf{q}) - \bar{P}(Y=1 \mid \mathbf{q})\right]^2,$$

where the average success probability is:

$$\bar{P}(Y=1 \mid \mathbf{q}) = \sum_{l=1}^{2^{K_j^*}} p(\boldsymbol{\alpha}_{lj}^*) P(Y=1 \mid \boldsymbol{\alpha}_{lj}^*, \mathbf{q}).$$

The PVAF was computed as the ratio of the GDI for a particular q-vector to the maximum GDI. As noted by [50, p. 261], the maximum GDI is achieved when all the attributes are specified. Thus, the PVAF expresses how well a candidate q-vector accounts for the variability in success probabilities relative to this maximum. In our analysis, two Time 1 items met revision criteria (PVAF < 0.8). After modification, both exceeded the threshold: Item 3 was updated to include vocabulary, and Item 5 to include decoding. No changes were indicated at Time 2. As validation methods for dynamic models with covariates are limited, additional simulation studies were conducted to support these findings.

Table 3 summarises the guessing (g), and slipping (s) parameters for each item across Time 1 and Time 2. Compared to real-data analyses reported by [32], where guessing rates often exceeded 0.5 while slipping rates were often below 0.4, the guessing and slipping parameters estimated in the current study were moderate. Specifically, guessing rates were slightly higher at Time 2 than at Time 1, particularly for items involving both decoding and vocabulary skills (items 2, 3, and 6; see Table 3). Slipping rates showed small decreases over time for items requiring only one skill (items 1, 4, and 5).

Table 3: Estimated Q-matrix, guessing (g), and slipping (s) parameters for each item at Time 1 and Time 2. Q-matrix values indicate whether each item requires decoding (A_1) and/or (A_2) vocabulary skills. Bolded values in the table highlight the maximum estimates of g and s across items.

Item	**Time 1**				**Time 2**			
	A_1	A_2	g	s	A_1	A_2	g	s
1	1	0	0.328	0.248	1	0	0.354	0.232
2	1	0	0.244	0.143	1	1	0.323	0.149
3	1	0	0.280	**0.351**	1	1	0.287	0.311
4	0	1	0.161	0.336	0	1	0.334	0.302
5	0	1	**0.376**	0.344	0	1	0.313	**0.317**
6	0	1	0.118	0.256	1	1	**0.365**	0.269

4.4 Identified Attribute Profiles, Initial Attribute Mastery and Attribute Transition with Covariates

Table 4 presents the distribution and transitions of attribute profiles across two time points. At Time 1, approximately one quarter of the students had not mastered either skill. By Time 2, only about 5% of the students remained in the non-mastery group, with most transitioning to mastery of vocabulary alone or to mastery of both skills. The number of students who mastered both skills at Time 2 was nearly twice that at Time 1. Students who had mastered at least one skill at Time 1 were more likely to achieve mastery of the other skill or both skills by Time 2.

Table 4: Transition matrix of attribute profiles from Time 1 (rows) to Time 2 (columns) for 263 students. Each element shows the number of students (proportion). Row sums represent Time 1 distributions; column sums represent Time 2 distributions. Profile labels: 00 = no mastery, 10 = decoding skill only, 01 = vocabulary skill only, 11 = mastery of both skills.

		Time 2				**Totals**
		00	10	01	11	
Time 1	00	12(4.563%)	3(1.141%)	38(14.449%)	21(7.985%)	74(28.136%)
	10	1(0.380%)	1(0.380%)	17(6.464%)	25(9.506%)	44(16.730%)
	01	1(0.380%)	1(0.380%)	46(17.490%)	37(14.068%)	85(32.319%)
	11	1(0.380%)	2(0.760%)	16(6.084%)	41(15.589%)	60(22.814%)
Totals		15(5.703%)	7(2.662%)	117(44.487%)	124(47.148%)	263(100%)

Table 5 presents the posterior means of odds ratios for initial mastery (β_z) and transition probabilities (γ_{01} and γ_{10}), estimated separately by attribute K. Covariates included six log-based variables, four demographics, and two game and learning characteristics listed in Tables 1 and 2. Covariates without statistically significant effects are not shown the table. All odds ratios and confidence intervals for all covariates are provided in supplementary material C. A sparsity criterion was applied following [17], whereby odds ratios with 95% credible intervals including 1 were considered inactive.

For initial mastery (β_z), more reattempts on decoding game were negatively associated with mastery in the decoding (K=1), while a larger number of levels mastered showed a strong positive association. For vocabulary (K=2), greater level mastery and longer response times were positively associated with mastery, and boys were less likely than girls to achieve mastery. For transition probabilities, significant effects were also observed. For the transition from non-mastery to mastery (γ_{01}), more reattempts were negatively associated with achieving mastery in decoding, while greater level mastery increased the likelihood of transitioning. Boys, compared to girls, were less likely to transition to mastery of vocabulary skill. For the transition from mastery to non-mastery (γ_{10}), students performing well below benchmark in initial literacy ability were more likely to lose mastery compared to those at benchmark. In the absence of covariate effects, the probability of losing mastery was low.

Table 5: Significant covariates for β_z (initial mastery), γ_{01}, and γ_{10} by attribute (K). Only covariates with statistically significant odds ratios (OR) are shown. *Note*: K = attribute; (Intercept) = model intercept; NRA = number of reattempts; NLM = number of levels mastered; RT = response time; Gender = female (0), male (1); ILA = initial literacy ability (at benchmark: reference level); WB = well below benchmark; Decoding = decoding game; Vocabulary = vocabulary game.

Initial mastery			Transition probabilities					
	β_z			γ_{01}			γ_{10}	
K	Covariates	OR	K	Covariates	OR	K	Covariates	OR
1	NRA Decoding	0.085	1	NRA Decoding	0.363	1	ILA-WB	6.003
1	NLM Decoding	15.281	1	NLM Decoding	4.603	2	(Intercept)	0.158
2	NLM Vocabulary	19.328	2	Gender	0.582			
2	RT Vocabulary	3.808						
2	Gender	0.373						

5 Simulation Study

The main objectives of this simulation study were to assess the model performance under various conditions, and to evaluate the model's robustness under settings that can be present within the area of application. We designed 18 simulation settings across three experimental factors: sample size (N), the number of items (J), and Q-matrix sparsity level (θ).

5.1 Simulation Design and Validations

Inspired by our empirical study, we considered two tests administered at two time points ($T = 2$), each measuring two binary attributes ($K = 2$). Students responded to the same number of items at each time point, and the total number of items across both tests was denoted by $J = \sum_{t=1}^{2} J_t$. The six covariates were generated from a multivariate standard normal distribution, $\boldsymbol{Z} \sim \mathcal{N}(\mathbf{0}, I_6)$, where I_6 is the 6×6 identity matrix. These covariates were designed to influence both initial mastery and attribute transitions across time (Equation 4–6). Following [32], the guessing and slipping parameters were generated from $\texttt{Uniform}(0.05,\ 0.20)$, while all regression coefficients in the attribute and transition models (i.e., $\boldsymbol{\beta}_Z$, γ_{01}, and γ_{10}) were assigned using estimates obtained from our empirical analysis to simulate the data.

The true Q-matrices used under each simulation condition are provided in supplementary material D. We considered both a sparse scenario and a dense scenario. To ensure that the identifiability conditions were not violated, we avoided using extensively dense Q-matrices [20, 41]. The density level adopted in the dense condition was close to the upper limit reported in prior studies [19, 23, 26]. The sparse Q-matrix contains approximately 58–61% nonzero entries, and the dense Q-matrix includes about 67% nonzero entries. Overall recovery accuracy was defined as the proportion of correctly classified Q-matrix entries. For comparison, simulation results assuming a known Q-matrix for a small sample size (N=200) are provided in supplementary material E.

We assessed model performance using both parameter estimation accuracy and classification accuracy. For item parameters and regression coefficients in the attribute and transition models, we computed the mean bias (MBias) and root mean square error (RMSE) across replications, given by

$$\text{MBias} = \frac{1}{R}\sum_{r=1}^{R}(\hat{p}^{(r)} - p), \quad \text{RMSE} = \sqrt{\frac{1}{R}\sum_{r=1}^{R}(\hat{p}^{(r)} - p)^2},$$

where p is the true parameter and R is the number of replications.

The classification accuracy of students' attribute profiles, in terms of each individual attribute at each time point, was calculated as follows:

$$\text{AAR}_k = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}(\hat{\boldsymbol{\alpha}}_{ik} = \boldsymbol{\alpha}_{ik}),$$

where N is the number of students, K is the number of attributes, $\boldsymbol{\alpha}_i$ represents the true attribute profile of student i, $\hat{\boldsymbol{\alpha}}_i$ denotes its estimated counterpart, and $\mathbb{I}(\cdot)$ is the indicator function, taking the value 1 if the condition holds and 0 otherwise. We further evaluated the recovery of the Q-matrix by comparing the estimated and true Q-matrix entries across simulation replications. Specifically, we computed the false positive rate, false negative rate, and classification accuracy based on true positive and true negative for each element of the Q-matrix.

5.2 Results

We conducted 25 replications per simulation condition, using three independent Markov chains of 20,000 iterations each, with different initial values. Following [51], convergence was assessed using the potential scale reduction factor ($\hat{R}$). After discarding the first 10,000 burn-in iterations, all maximum values of $\hat{R}$ remained below 1.1, indicating acceptable convergence.

For the sparsity prior on the Q-matrix, the prior mean $\frac{\alpha}{\alpha+\beta}$ matched the true sparsity level (proportion of non-zero elements), and the concentration $\alpha + \beta$ was scaled with the number of items to stabilise estimation (20, 70, and 120 for $J = 6$, 18, and 30, respectively). Sensitivity analyses further varied the prior specifications, and recovery of the Q-matrix and attribute classifications showed little difference (see supplementary material F), confirming the robustness of the model to reasonable prior misspecification.

Table 6 presents attribute agreement rates (AARs) across conditions. Across all combinations of N and J_t, accuracy at Time 2 was consistently higher than at Time 1, except for small sample (N=200), indicating improved attribute recovery over time. For a fixed N, AARs tended to increase slightly as J_t increased. Nevertheless, accuracy remained high in all conditions (AARs ≥ 0.913). As sample size increased, dense Q matrices showed clearer gains, with AARs at Time 2 reaching up to 1.000 for $N = 600$, $J_t = 30$.

Table 6: Recovery of attribute profiles measured by attribute agreement rates (AAR_1 and AAR_2) across time points (T), under varying sparsity levels (θ), sample sizes (N), and number of items (J_t). Bold values indicate the smallest AARs for each attribute across all conditions.

			Sparse Q		Dense Q	
N	J_t	T	AAR_1	AAR_2	AAR_1	AAR_2
200	6	1	0.956	**0.928**	0.949	0.924
		2	**0.927**	0.953	0.930	0.913
	18	1	0.994	0.994	**0.990**	**0.993**
		2	0.996	0.994	0.995	0.992
	30	1	0.999	1.000	0.998	0.999
		2	0.999	1.000	0.999	0.999
400	6	1	0.970	0.956	0.953	0.958
		2	0.941	0.961	0.941	0.923
	18	1	0.996	0.996	0.993	0.993
		2	0.996	0.996	0.995	0.992
	30	1	1.000	1.000	0.999	1.000
		2	0.999	1.000	0.999	0.999
600	6	1	0.972	0.947	0.955	0.955
		2	0.930	0.956	0.941	0.935
	18	1	0.997	0.997	0.994	0.995
		2	0.996	0.996	0.994	0.993
	30	1	1.000	0.999	0.999	1.000
		2	1.000	0.999	0.999	0.999

As shown in Table 7, Q-matrix recovery accuracy was consistently higher at Time 2 than Time 1 across all conditions. When controlling for sample size, increasing the number of items slightly reduced accuracy, although classification remained generally high. Under denser Q-matrices, larger sample sizes yielded higher accuracy.

Table 7: Recovery of Q-matrix across time points (T) under different values of sparsity (θ), sample size (N), and number of items (J_t). Bold values indicate the highest scores per column. FPR = FP / [FP + TN], FNR = FN / [TP + FN], Acc. = (TP + TN) / (TP + FP + FN + TN). Bold values indicate the highest scores per column.

			Sparse Q			Dense Q		
N	J_t	T	FPR	FNR	Acc.	FPR	FNR	Acc.
200	6	1	0.000	0.120	0.925	0.000	0.042	0.969
		2	0.000	0.040	0.975	0.000	0.083	0.938
	18	1	0.000	0.067	0.957	0.000	0.020	0.986
		2	0.000	0.050	0.968	0.000	0.035	0.975
	30	1	0.000	0.117	0.927	0.000	0.070	0.951
		2	0.000	0.108	0.932	0.000	0.018	0.987
400	6	1	0.000	0.080	0.950	0.000	0.083	0.938
		2	0.000	0.040	0.975	0.000	0.000	**1.000**
	18	1	0.000	0.111	0.929	0.000	0.020	0.986
		2	0.000	0.039	0.975	0.000	0.010	0.993
	30	1	0.000	0.104	0.935	0.000	0.044	0.969
		2	0.000	0.092	0.943	0.000	0.018	0.987
600	6	1	0.000	0.000	**1.000**	0.000	0.000	**1.000**
		2	0.000	0.000	**1.000**	0.000	0.042	0.969
	18	1	0.000	0.106	0.932	0.000	0.015	0.989
		2	0.000	0.039	0.975	0.000	0.030	0.979
	30	1	0.000	0.121	0.924	0.000	0.026	0.982
		2	0.000	0.108	0.932	0.000	0.018	0.987

Table 8 summarises the estimation accuracy of guessing and slipping parameters (g_{jt} and s_{jt}) under the sparse Q. Overall, bias was minimal, with RMSEs for g_{jt} generally below 0.035 and for s_{jt} below 0.075, except in the small-sample condition ($N = 200$, $J_t = 6$). Accuracy improved at Time 2 in most cases, and larger sample sizes yielded more stable estimates, especially under sparse Q. Results for the dense Q-matrix are presented in supplementary material G, where bias remained minimal across all conditions ($g_{jt} \leq 0.035$, $s_{jt} \leq 0.073$). Item-level RMSE comparisons for g_{jt} and s_{jt} under different simulation conditions are reported in supplementary material H.

Table 8: Estimation accuracy of item parameters (g_{jt} and s_{jt}), evaluated by mean bias (MBias), and root mean square error (RMSE), under sparse Q-matrix, sample sizes (N), and number of items (J_t). Bold values indicate the highest estimates across conditions.

Sparsity of Q	N	J_t	T	g_{MBias}	g_{RMSE}	s_{MBias}	s_{RMSE}
Sparse Q	200	6	1	**0.037**	**0.072**	**0.065**	**0.115**
			2	0.006	0.027	-0.002	0.041
		18	1	0.004	0.027	0.012	0.057
			2	0.003	0.033	0.001	0.036
		30	1	-0.000	0.026	0.004	0.050
			2	-0.000	0.031	0.004	0.037
	400	6	1	0.017	0.048	0.031	0.071
			2	-0.005	0.023	-0.005	0.027
		18	1	0.001	0.022	0.004	0.042
			2	-0.003	0.023	-0.002	0.026
		30	1	-0.005	0.021	-0.003	0.034
			2	-0.001	0.023	-0.002	0.022
	600	6	1	-0.000	0.019	-0.003	0.027
			2	-0.013	0.031	-0.002	0.025
		18	1	-0.002	0.017	0.002	0.039
			2	-0.001	0.018	-0.003	0.019
		30	1	-0.004	0.016	-0.002	0.033
			2	-0.003	0.017	-0.004	0.021

Table 9 presents the estimation accuracy for initial mastery probabilities ($\beta_{0,k}$) and covariate effects ($\beta_{Z,k}$). For both attributes, the estimation bias tended to decrease slightly as sample size (N) or the number of items (J_t) increased. Similarly patterns are evident in Table 10, the bias in estimating covariate effects on attribute transitions was generally small across conditions.

Table 9: Estimation accuracy of initial covariate effects on attribute (k) mastery ($\boldsymbol{\beta}$), evaluated by mean bias (MBias), and root mean square error (RMSE), under varying sparsity levels (θ), sample sizes (N), and number of items (J_t). Each $\beta_{k,z}$ quantifies the impact of covariate z on the initial mastery probability of attribute k at time $T = 1$. Bold values indicate the largest absolute MBias and the largest RMSE across conditions.

			Sparse Q				Dense Q			
N	J_t	Metric	$\beta_{0,1}$	$\beta_{0,2}$	$\beta_{Z,1}$	$\beta_{Z,2}$	$\beta_{0,1}$	$\beta_{0,2}$	$\beta_{Z,1}$	$\beta_{Z,2}$
200	6	Bias	**0.077**	**0.374**	-0.002	**-0.059**	0.038	-0.027	0.013	-0.027
		RMSE	**0.192**	**0.404**	**0.247**	**0.317**	**0.154**	0.088	**0.218**	**0.283**
	18	Bias	-0.076	-0.042	-0.017	-0.002	-0.063	-0.040	-0.014	-0.003
		RMSE	0.110	0.131	0.167	0.185	0.107	0.132	0.167	0.190
	30	Bias	0.021	-0.047	0.008	0.028	0.018	-0.053	0.007	0.028
		RMSE	0.135	0.180	0.169	0.181	0.136	**0.178**	0.192	0.198
400	6	Bias	0.015	0.204	**-0.030**	0.019	-0.017	**0.104**	**-0.053**	0.009
		RMSE	0.132	0.226	0.126	0.240	0.074	0.138	0.112	0.187
	18	Bias	0.043	0.020	0.009	0.029	0.055	0.022	0.007	**0.031**
		RMSE	0.117	0.071	0.122	0.165	0.124	0.052	0.100	0.150
	30	Bias	-0.073	-0.016	-0.009	0.003	-0.070	-0.015	-0.009	0.002
		RMSE	0.135	0.096	0.165	0.140	0.132	0.098	0.166	0.139
600	6	Bias	0.062	0.053	0.020	-0.005	**0.082**	0.027	0.017	-0.007
		RMSE	0.077	0.092	0.105	0.115	0.095	0.087	0.121	0.109
	18	Bias	0.016	-0.003	-0.004	-0.013	0.029	-0.011	0.001	-0.016
		RMSE	0.088	0.072	0.130	0.125	0.102	0.088	0.116	0.106
	30	Bias	-0.038	-0.070	0.008	0.032	-0.039	-0.067	0.008	**0.031**
		RMSE	0.101	0.099	0.122	0.141	0.101	0.097	0.122	0.140

6 Discussions and Future Directions

This study introduces a temporal cognitive diagnostic modeling framework for analysing student learning from digital educational tools. Application of the framework to log data from a research-informed digital reading program demonstrates its flexibility, interpretability, and alignment with theoretical models of early reading development. Simulation studies further confirm the accuracy and robustness of the proposed model across various conditions.

In this study, we examined decoding and vocabulary as two subcomponents of the critical constructs proposed in the Simple View of Reading [47]: word recognition and language comprehension, respectively. Decoding supports the development of word recognition, while vocabulary supports language comprehension. These are theoretically considered distinct constructs. Our findings reinforce this distinction: transitions from non-mastery to mastery within each skill (decoding or vocabulary) followed expected developmental trajectories, and some students mastered one skill without the other. The theoretical independence, together with our empirical observation that a substantial proportion of students mastered only one of the two skills, justified the use of a non-compensatory modeling framework.

However, the estimation of the Q-matrix for this specific dataset also revealed evidence of interrelations between word recognition and vocabulary skills. Specifically, students appeared to require vocabulary knowledge to support their performance in word recognition skill games, consistent with findings reported by [52]. Students who demonstrated mastery in both skills may have acquired them in parallel or may have benefited from prior familiarity with the digital learning environment. Log-based covariates and demographics contributed meaningfully to profile classification. Higher achievement (i.e., more levels mastered), fewer reattempts, and faster response times were positively associated with mastery, in line with prior evidence on the roles of practice and fluency in early literacy development [53, 54]. Students with weaker initial literacy skills were less likely to master decoding, possibly due to persistent challenges in acquiring letter–sound correspondence. Furthermore, boys exhibited a lower likelihood of mastering vocabulary compared to girls, consistent with previous findings on gender differences in vocabulary development [55]. By integrating multiple log-derived behavioural indicators and individual characteristics, rather than relying solely on a single measure, our modeling approach more effectively captures the dynamic, multidimensional nature of early reading development, thus

Table 10: Estimation accuracy of covariate effects on attribute (k) transitions ($\boldsymbol{\gamma}$), including both mastery acquisition (γ_{01}) and loss of mastery (γ_{10}), evaluated by mean bias (MBias), and root mean square error (RMSE). The results are reported under varying θ, N, and J_t. Each $\gamma_{01,k,c}$ and $\gamma_{10,k,c}$ quantifies the impact of covariate c on the probability of attribute k transitioning between states from time $T = 1$ to $T = 2$. Bold values indicate the largest absolute MBias and the largest RMSE across conditions.

			Sparse Q				Dense Q			
N	J_t	Metric	$\gamma_{01,1}$	$\gamma_{01,2}$	$\gamma_{10,1}$	$\gamma_{10,2}$	$\gamma_{01,1}$	$\gamma_{01,2}$	$\gamma_{10,1}$	$\gamma_{10,2}$
200	6	Bias	**-0.031**	**0.083**	-0.029	-0.049	0.002	0.027	-0.010	**-0.130**
		RMSE	**0.264**	**0.272**	0.320	**0.416**	**0.223**	**0.283**	0.307	**0.532**
	18	Bias	-0.004	-0.034	-0.034	-0.070	-0.007	-0.032	-0.020	-0.073
		RMSE	0.204	0.197	0.367	0.293	0.201	0.198	0.363	0.290
	30	Bias	-0.007	-0.037	0.029	-0.063	-0.003	-0.029	0.031	-0.048
		RMSE	0.194	0.202	**0.442**	0.321	0.186	0.200	**0.420**	0.321
400	6	MBias	-0.019	0.055	**-0.082**	-0.041	**-0.030**	**0.081**	**-0.082**	-0.059
		RMSE	0.178	0.171	0.316	0.262	0.187	0.209	0.302	0.329
	18	MBias	-0.009	0.016	-0.024	-0.051	-0.008	0.020	-0.037	-0.046
		RMSE	0.146	0.126	0.232	0.218	0.147	0.135	0.237	0.218
	30	MBias	-0.006	0.006	-0.024	-0.004	-0.006	0.005	-0.023	-0.005
		RMSE	0.133	0.138	0.259	0.223	0.134	0.138	0.258	0.225
600	6	MBias	-0.009	-0.023	0.032	-0.048	0.013	-0.030	0.029	-0.069
		RMSE	0.146	0.121	0.220	0.208	0.128	0.136	0.203	0.189
	18	MBias	-0.002	-0.001	-0.014	-0.016	-0.008	-0.001	-0.019	-0.025
		RMSE	0.096	0.095	0.197	0.199	0.096	0.096	0.192	0.201
	30	MBias	0.010	-0.003	0.030	**-0.072**	0.009	-0.003	0.028	-0.071
		RMSE	0.093	0.106	0.173	0.254	0.093	0.109	0.173	0.252

better reflecting the complexity of real-world cognitive processes. Additionally, an important contribution of our study relies in its ability to empirically validate theoretically specified item-skill mappings (Q-matrix). Thus, our approach aligns with expert design while improving the robustness and validity of digital reading tools in practice.

In simulations, the proposed MCMC algorithm reliably recovered latent profiles, item-skill mapping (Q-matrix), and parameters across a range of conditions, supporting the model's practical utility in real-world applications where the Q-matrix structure is unknown. Notably, Q-matrix recovery remained consistently high even under relatively small designs (e.g., 200 students and 6 items), diverging from earlier recommendations (e.g., [21]; [32]) and highlighting the model's robustness in recovering item–skill mappings from limited data. However, static validation methods such as PVAF [50] appear less suited for dynamic models that incorporate covariates and attribute transitions. In contrast, increases in sample size or item numbers—regardless of the sparsity of the Q matrix—consistently improved parameter recovery, in line with prior findings by [32] and [56]. Furthermore, our classification accuracy for attribute profiles closely matched the accuracy of recovery the unknown parameters reported in [32], despite not assuming the Q-matrix to be known. These results indicate that high accuracy can still be achieved even when the Q-matrix is estimated rather than fixed.

Several promising directions emerge for extending the current modeling framework. First, our application focused on time-invariant covariates to explain individual differences. Time-varying covariates could easily be incorporated to capture dynamic influences on learning trajectories. Second, the model recovered meaningful item–attribute relationships under a known number of skills based on game structure. Future work could estimate the dimensionality of skills using tools such as parallel analysis, supporting more flexible model specifications. Third, the individual attribute profile was assumed to be fixed over a short observation window to demonstrate empirical applicability. Extending the framework to longer-term or more frequent assessments could reveal detailed developmental patterns and detect change points in learning. Fourth, the model assumed similarity in items taken across game levels. Introducing greater item variation and modeling item-level heterogeneity could improve generalizability. Fifth, the proposed framework highlights timing as an important factor influencing reading ability—one that has been linked to fluency in the literature. Building on this, future work could extend the framework to accommodate polytomous skill states, such as partial mastery or graded proficiency levels (e.g., fluency). Sixth, this study reflects the diversity of covariates commonly available in digital settings, including both log-based (e.g., number of attempts, response time) and socio-demographic (e.g., gender, ELL) variables. This suggests that future research could consider incorporating covariate selection methods. Lastly, although decoding and vocabulary were modeled using a non-compensatory framework, many other skills learning contexts may involve compensatory or partially compensatory relationships. Extensions of this framework to models such as deterministic input, noisy or gate (DINO) model [57], or the generalized DINA model [50] could reveal diverse patterns of skill integration and individual differences.

7 Conclusion

This study presents a temporal cognitive diagnostic modeling framework designed for analysing student learning through digital educational tools. By applying this framework to log files from a research-informed digital reading program, we demonstrated its flexibility, interpretability, and strong alignment with established theoretical models of early reading development. Simulation studies further confirmed the model's accuracy and robustness in jointly estimating the unknown item-skill mappings, latent attribute profiles, item parameters, and covariate effects across diverse conditions. From a policy perspective, our approach supports educational initiatives such as ESSA, emphasising personalized instruction and evidence-based digital tools. Practically, the ability to dynamically track student transitions between non-mastery and mastery using response accuracy, log-derived behavioural indicators, and individual characteristics reflects the multidimensional complexity of reading development in real-world settings. The capacity to identify item-skill relationships solely from response data is an important contribution that can help educators to validate targeted educational tools, monitor their effectiveness longitudinally, and support students' reading development.

8 Acknowledgments

The authors declare no competing interests. This research was supported by the Economic and Physical Research Council, which funded the first author through a PhD studentship. The authors acknowledge the support of Boost Reading at Amplify for providing the dataset used in this analysis.

9 Data availability

Due to the commercial sensitivity of these data, our data sharing agreement with the company who provided the dataset requires that the raw data remain confidential and cannot be shared.

References

[1] Anthony Cree, Andrew Kay, and June Steward. The economic & social cost of illiteracy: A snapshot of illiteracy in a global context. World Literacy Foundation, September 2023. Final report.

[2] Ariel Lindorff, Jamie Stiff, and Heather Kayton. Pirls 2021: National report for england. research report. Research report, UK Department for Education, April 2024. Published by the University of Oxford for the UK Department for Education.

[3] Ina V. S. Mullis, Matthias von Davier, Pierre Foy, Bethany Fishbein, Katherine A. Reynolds, and Erin Wry. *PIRLS 2021 international results in reading*. TIMSS & PIRLS International Study Center, Lynch School of Education and Human Development, Boston College, and IEA, Chestnut Hill, MA, 2023.

[4] U.S. Department of Education. Every student succeeds act (essa). Public Law No. 114-95, 2015.

[5] Uwe Maier and Christian Klotz. Personalized feedback in digital learning environments: Classification framework and literature review. *Computers and Education: Artificial Intelligence*, 3:100080, 2022.

[6] Edward Haertel. An application of latent class models to assessment data. *Applied Psychological Measurement*, 8(3):333–346, 1984.

[7] Jonathan Templin, Robert A Henson, et al. *Diagnostic measurement: Theory, methods, and applications*. Guilford press, 2010.

[8] André A Rupp and Jonathan L Templin. Unique characteristics of diagnostic classification models: A comprehensive review of the current state-of-the-art. *Measurement*, 6(4):219–262, 2008.

[9] Peida Zhan, Hong Jiao, and Dandan Liao. Cognitive diagnosis modelling incorporating item response times. *British Journal of Mathematical and Statistical Psychology*, 71(2):262–286, 2018.

[10] Bethany C. Bray, Stephanie T. Lanza, and Linda M. Collins. Modeling relations among discrete developmental processes: A general approach to associative latent transition analysis. *Structural Equation Modeling*, 17(4):541–569, 2010.

[11] Yasemin Kaya and Walter L. Leite. Assessing change in latent skills across time with longitudinal cognitive diagnosis modeling: An evaluation of model performance. *Educational and Psychological Measurement*, 77(3):369–388, 2017.

[12] Qianru Liang, Jimmy de la Torre, and Nancy Law. Latent transition cognitive diagnosis model with covariates: A three-step approach. *Journal of Educational and Behavioral Statistics*, 48(6):690–718, 2023.

[13] Shiyu Wang, Yan Yang, Steven Andrew Culpepper, and Jeffrey A Douglas. Tracking skill acquisition with cognitive diagnosis models: A higher-order, hidden markov model with covariates. *Journal of Educational and Behavioral Statistics*, 43(1):57–87, 2018.

[14] Shu Wang, Si Zhang, and Yanjun Shen. A joint modeling framework of responses and response times to assess learning outcomes. *Multivariate Behavioral Research*, 55:49–68, 2019.

[15] Yinghan Chen, Steven Andrew Culpepper, Shiyu Wang, and Jeffrey Douglas. A hidden markov model for learning trajectories in cognitive diagnosis with application to spatial rotation skills. *Applied Psychological Measurement*, 42(1):5–23, 2018.

[16] Shiyu Wang, Yiling Hu, Qi Wang, Bian Wu, Yawei Shen, and Martha Carr. The development of a multidimensional diagnostic assessment with learning tools to improve 3-d mental rotation skills. *Frontiers in Psychology*, 11:305, 2020.

[17] Eric Alan Wayman, Steven Andrew Culpepper, Jeff Douglas, and Jesse Bowers. A restricted latent class hidden markov model for polytomous responses, polytomous attributes, and covariates: Identifiability and application, 2025. arXiv preprint.

[18] Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.

[19] Andre A Rupp and Jonathan Templin. The effects of Q-matrix misspecification on parameter estimates and classification accuracy in the dina model. *Educational and Psychological Measurement*, 68(1):78–96, 2008.

[20] Chia-Yi Chiu, Jeffrey A Douglas, and Xiaodong Li. Cluster analysis for cognitive diagnosis: Theory and applications. *Psychometrika*, 74:633–665, 2009.

[21] Jimmy de la Torre. The generalized DINA model framework. *Psychometrika*, 76(2):179–199, 2011.

[22] Wenchao Ma and Jimmy de la Torre. An empirical q-matrix validation method for the sequential generalized dina model. *British Journal of Mathematical and Statistical Psychology*, 73(1):142–163, 2020.

[23] Yunxiao Chen, Jingchen Liu, Gongjun Xu, and Zhiliang Ying. Statistical analysis of Q-matrix based diagnostic classification models. *Journal of the American Statistical Association*, 110(510):850–866, 2015.

[24] Yinghan Chen, Steven Andrew Culpepper, Yuguo Chen, and Jeffrey Douglas. Bayesian estimation of the DINA Q matrix. *Psychometrika*, 83:89–108, 2018.

[25] Fang Tang and Peida Zhan. Does diagnostic feedback promote learning? Evidence from a longitudinal cognitive diagnostic assessment. *AERA Open*, 7:23328584211060804, 2021.

[26] Gongjun Xu and Zhuoran Shang. Identifying latent structures in restricted latent class models. *Journal of the American Statistical Association*, 113(523):1284–1295, 2018.

[27] Jingchen Chen and Jimmy de la Torre. A general cognitive diagnosis model for expert-defined polytomous attributes. *Applied Psychological Measurement*, 37(6):419–437, 2013.

[28] Xinxin Xu, Siyu Ren, Xiaoyu Shan, and Dan Zhang. A polytomous extension of the higher-order, hidden markov model with covariates and hierarchical learning trajectories. *Journal of Educational and Behavioral Statistics*, 0(0), 2025.

[29] Peng Zhan, Wei-Chung Wang, and Xuelong Li. A partial mastery, higher-order latent structural model for polytomous attributes in cognitive diagnostic assessments. *Journal of Classification*, 37:328–351, 2020.

[30] Shu Wang and Yue Chen. Using response times and response accuracy to measure fluency within cognitive diagnosis models. *Psychometrika*, 85(3):600–629, 2020.

[31] Zichu Liu, Shiyu Wang, Shumei Zhang, and Tao Qiu. A mixture fluency model using responses and response times with cognitive diagnosis model framework. *Behavior Research Methods*, 56(4):3396–3451, 2024.

[32] Zichu Liu, Shiyu Wang, Houping Xiao, Shumei Zhang, and Tao Qiu. A general dynamic learning model framework for cognitive diagnosis. *British Journal of Mathematical and Statistical Psychology*, 2025.

[33] Njål Foldnes, Per Henning Uppstad, Steffen Grønneberg, and Jenny M Thomson. School entry detection of struggling readers using gameplay data and machine learning. In *Frontiers in Education*, volume 9, page 1487694. Frontiers Media SA, 2024.

[34] Huilin Chen, Yuyang Cai, and Jimmy de la Torre. Investigating second language (l2) reading subskill associations: A cognitive diagnosis approach. *Language Assessment Quarterly*, 20(2):166–189, 2023.

[35] Mehdi Rajeb, Wenchao Ma, Qiwei He, and Qingzhou Shi. Incorporating process information into cognitive diagnostic models: A four-component joint modeling approach. *Journal of Educational and Behavioral Statistics*, page 10769986251334788, 2023.

[36] Stephen Newton, Harrison Gamble, Yu Su, Jennifer Zoski, and Danielle Damico. *Examining the impact of Amplify Reading on student literacy in Grades K–2: 2019 report*. Technical Report ED604917, ERIC, 2019. Available from ERIC (Education Resources Information Center).

[37] University of Oregon. 8th edition of dynamic indicators of basic early literacy skills (dibels). `https://dibels.uoregon.edu`, 2018. Accessed: 2025-05-29.

[38] Brian W Junker and Klaas Sijtsma. Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement*, 25(3):258–272, 2001.

[39] G Rojas, J de la Torre, and J Olea. Choosing between general and specific cognitive diagnosis models when the sample size is small. In *annual meeting of the National Council of Measurement in Education, Vancouver, British Columbia, Canada*, 2012.

[40] Guanhua Fang, Jingchen Liu, and Zhiliang Ying. On the identifiability of diagnostic classification models. *Psychometrika*, 84(1):19–40, 2019.

[41] Yuqi Gu and Gongjun Xu. Sufficient and necessary conditions for the identifiability of the q-matrix. *Statistica Sinica*, 31(1):449–472, 2021.

[42] Shiyu Wang, Susu Zhang, Jeff Douglas, and Steven Culpepper. Using response times to assess learning progress: A joint model for responses and response times. *Measurement: Interdisciplinary Research and Perspectives*, 16(1):45–58, 2018.

[43] Susu Zhang and Shiyu Wang. Modeling learner heterogeneity: A mixture learning model with responses and response times. *Frontiers in psychology*, 9:2339, 2018.

[44] Steven Andrew Culpepper. Revisiting the 4-parameter item response model: Bayesian estimation and application. *Psychometrika*, 81(4):1142–1163, 2016.

[45] Martyn Plummer, Alexey Stukalov, and Matt Denwood. *rjags: Bayesian graphical models using MCMC*, 2025. R package version 4-17.

[46] Peida Zhan, Hong Jiao, Kaiwen Man, and Lijun Wang. Using jags for bayesian cognitive diagnosis modeling: A tutorial. *Journal of Educational and Behavioral Statistics*, 44(4):473–503, 2019.

[47] Philip B. Gough and William E. Tunmer. Decoding, reading, and reading disability. *Remedial and Special Education*, 7(1):6–10, 1986.

[48] Yawen Ma, Kate Cain, and Anastasia Ushakova. Application of cluster analysis to identify different reader groups through their engagement with a digital reading supplement. *Computers & Education*, 214:105025, 2024.

[49] Aki Vehtari, Andrew Gelman, Daniel Simpson, Bob Carpenter, and Paul-Christian Bürkner. Rank-normalization, folding, and localization: An improved $\hat{R}$ for assessing convergence of mcmc (with discussion). *Bayesian Analysis*, 16(2):667–718, 2021.

[50] Jimmy de la Torre and Chia-Yi Chiu. A general method of empirical q-matrix validation. *Psychometrika*, 81(2):253–273, 2016.

[51] Andrew Gelman and Donald B Rubin. Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4):457–472, 1992.

[52] Language and Reading Research Consortium. Learning to read: Should we keep things simple? *Reading Research Quarterly*, 50(2):151–169, 2015.

[53] David LaBerge and S Jay Samuels. Toward a theory of automatic information processing in reading. *Cognitive psychology*, 6(2):293–323, 1974.

[54] William Dee Nichols, William H Rupley, and Timothy Rasinski. Fluency in learning to read for meaning: Going beyond repeated readings. *Literacy Research and Instruction*, 48(1):1–13, 2008.

[55] Janellen Huttenlocher, Wendy Haight, Anthony Bryk, Michael Seltzer, and Thomas Lyons. Early vocabulary growth: relation to language input and gender. *Developmental psychology*, 27(2):236, 1991.

[56] Xiaolin Yu, Peng Zhan, and Qian Chen. Don't worry about the anchor-item setting in longitudinal learning diagnostic assessments. *Frontiers in Psychology*, 14:1112463, 2023.

[57] Jonathan L Templin and Robert A Henson. Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods*, 11(3):287, 2006.

10 Supplementary Material

Supplementary material is available online at *Journal of the Royal Statistical Society: Series A*.

11 Supplementary Material A

For sparsity in the Q-matrix,

$$Q_{jk} \sim \text{Bernoulli}(\theta),$$
$$\theta \sim \text{Beta}(\alpha, \beta),$$

where the prior mean $\frac{\alpha}{\alpha+\beta}$ equals the non–zero proportion of the true Q-matrix and the concentration defined as $\alpha + \beta$.

To evaluate the sensitivity of model estimates to the prior specification of θ, we compared item parameters results from the empirical analysis under varied concentration levels ($\alpha + \beta$) in Table 11 and varied priors mean ($\frac{\alpha}{\alpha+\beta}$) in Table 12, relative to the reference prior $\text{Beta}(24, 6)$ used in the main analysis. Tables 13 and 14 summarises the changes in the posterior means of the guessing (g) and slipping (s) parameters under alternative priors.

Results showed that item parameters remained qualitatively unchanged across different prior means, supporting the robustness of the proposed approach. Across all alternative settings, changes in the posterior means of g and s parameters were generally less than 0.04, with most differences below 0.02. The sensitivity analysis indicates that the choice of prior concentration and mean for θ had minimal impact on the estimation of item parameters. The $\text{Beta}(24, 6)$ prior thus provides a reasonable balance between interpretability and flexibility for empirical applications.

Table 11: Alternative prior specifications for θ with fixed means $\frac{\alpha}{\alpha+\beta} = 0.8$ and varying concentration ($\alpha + \beta$).

Prior	Prior Mean	Concentration	Variance	95% Interval
$\text{Beta}(24, 6)$	0.8	30	0.0052	(0.65, 0.92)
$\text{Beta}(8, 2)$	0.8	10	0.0145	(0.55, 0.93)
$\text{Beta}(40, 10)$	0.8	50	0.0031	(0.70, 0.90)

Table 12: Alternative prior specifications for θ with fixed concentration ($\alpha + \beta = 30$) and varying means $\frac{\alpha}{\alpha+\beta}$.

Prior	Prior Mean	Concentration	Variance	95% Interval
$\text{Beta}(24, 6)$	0.8	30	0.0052	(0.65, 0.92)
$\text{Beta}(21, 9)$	0.7	30	0.0067	(0.55, 0.85)
$\text{Beta}(27, 3)$	0.9	30	0.0031	(0.78, 0.96)

Table 13: Posterior means of guessing (g) and slipping (s) parameters for each item at Time 1 and Time 2 under three alternative Beta priors (concentration 10, 30, and 50) whereas fixed mean 0.8. Bold values indicate the highest estimates across conditions.

Item	Prior	Time 1		Time 2	
		g	s	g	s
1	$\text{Beta}(24, 6)$	0.328	0.248	0.354	0.232
	$\text{Beta}(8, 2)$	0.332	0.254	0.356	0.229
	$\text{Beta}(40, 10)$	0.329	0.252	0.353	0.231
2	$\text{Beta}(24, 6)$	0.244	0.143	0.323	0.149
	$\text{Beta}(8, 2)$	0.229	0.153	0.305	0.135
	$\text{Beta}(40, 10)$	0.237	0.148	0.310	0.146
3	$\text{Beta}(24, 6)$	0.280	0.352	0.287	0.311
	$\text{Beta}(8, 2)$	0.242	**0.359**	0.273	0.300
	$\text{Beta}(40, 10)$	0.264	0.356	0.285	0.307
4	$\text{Beta}(24, 6)$	0.161	0.336	0.334	0.302
	$\text{Beta}(8, 2)$	0.148	0.344	0.350	0.295
	$\text{Beta}(40, 10)$	0.156	0.339	0.336	0.301
5	$\text{Beta}(24, 6)$	0.376	0.344	0.313	0.317
	$\text{Beta}(8, 2)$	**0.379**	0.346	0.281	**0.337**
	$\text{Beta}(40, 10)$	0.377	0.345	0.307	0.320
6	$\text{Beta}(24, 6)$	0.118	0.256	**0.365**	0.269
	$\text{Beta}(8, 2)$	0.113	0.259	0.330	0.282
	$\text{Beta}(40, 10)$	0.115	0.257	0.363	0.270

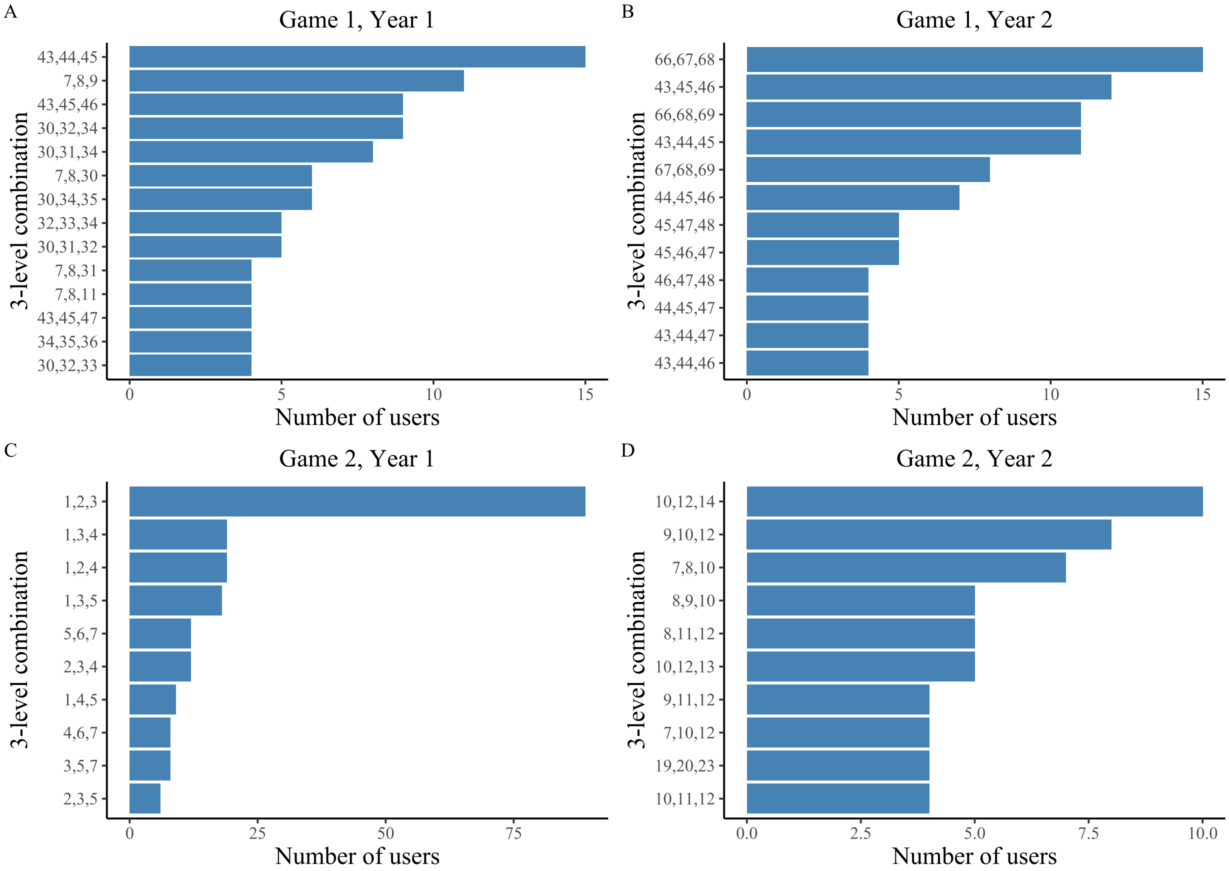
Table 14: Posterior means of guessing (g) and slipping (s) parameters for each item at Time 1 and Time 2 under three alternative Beta priors (mean 0.7, 0.8, and 0.9; concentration fixed at 30). Bold values indicate the highest estimates across conditions.

Item	Prior	Time 1		Time 2	
		g	s	g	s
1	$\text{Beta}(24, 6)$	0.328	0.248	**0.354**	0.232
	$\text{Beta}(21, 9)$	0.329	0.246	0.353	0.238
	$\text{Beta}(27, 3)$	**0.330**	0.251	0.348	0.242
2	$\text{Beta}(24, 6)$	0.244	0.143	0.323	0.149
	$\text{Beta}(21, 9)$	0.232	0.143	0.315	0.135
	$\text{Beta}(27, 3)$	0.267	0.150	0.316	0.152
3	$\text{Beta}(24, 6)$	0.280	0.352	0.287	0.311
	$\text{Beta}(21, 9)$	0.269	**0.359**	0.281	0.300
	$\text{Beta}(27, 3)$	0.254	0.351	0.284	0.311
4	$\text{Beta}(24, 6)$	0.161	0.336	0.334	0.302
	$\text{Beta}(21, 9)$	0.153	0.340	0.351	0.287
	$\text{Beta}(27, 3)$	0.155	0.341	0.348	0.291
5	$\text{Beta}(24, 6)$	0.376	0.344	0.313	0.317
	$\text{Beta}(21, 9)$	0.378	0.341	0.316	**0.322**
	$\text{Beta}(27, 3)$	0.374	0.344	0.342	0.309
6	$\text{Beta}(24, 6)$	0.118	0.256	0.365	0.269
	$\text{Beta}(21, 9)$	0.123	0.266	0.341	0.279
	$\text{Beta}(27, 3)$	0.119	0.274	0.360	0.269

12 Supplementary Material B

The exploratory data analysis in this section identifies and visualizes the ten most frequent three levels that students completed in each game and year. These panels highlight the variety of levels engaged, driven by differences in students' initial starting points and in-game progression. Panels A and B display the top ten combinations from decoding game (Years 1 and 2), while Panels C and D show the corresponding results for vocabulary game (Years 1 and 2).

Figure 3: Top ten three-level combinations engaged by students in each game and year. For example, panel A shows the most frequent level sequences for decoding game in Year 1. Each bar represents the number of students who completed the corresponding three-level combination.



13 Supplementary Material C

The posterior means and 95% confidence intervals (CIs) of the odds ratios (ORs) for β_z (initial mastery) by attribute (K), as well as the transition probabilities γ_{01} and γ_{10}, are reported in Tables 15–17.

Table 15: Posterior means of odds ratios (OR) for β_z (initial mastery) by attribute (K), with 95% confidence intervals (CI). Statistically significant results (CI excluding 1) are shown in **bold**. Part 1 of 2.
Note: K = attribute; (Intercept) = model intercept; nra = number of reattempts; nlm = number of levels mastered; rt = response time; gender = female (0), male (1); SEN = special education needs (1); ELL = English language learner (1); ILA = initial literacy ability (at benchmark: reference level); WB = well below benchmark; BB = below benchmark; AB = above benchmark; group = engagement group (group 5: reference); Others = American Indian, Alaskan Native, Black or African American, Hispanic or Latino, Multiracial/Other, and not Specified (reference group: White); DG = decoding game; VG = vocabulary game; HVG = high vocabulary group; HDG = high decoding group.

K	Measure	(Intercept)	nra DG	nra VG	nlm DG	nlm VG	rt DG	rt VG
1	OR	0.685	**0.085**	0.782	**15.281**	1.110	1.060	1.454
	CI	(0.149, 3.250)	(0.024, 0.260)	(0.355, 1.802)	(6.567, 42.138)	(0.508, 2.478)	(0.416, 2.453)	(0.694, 3.533)
2	OR	1.281	1.549	0.542	0.988	**19.328**	0.752	**3.808**
	CI	(0.326, 5.809)	(0.756, 3.523)	(0.248, 1.126)	(0.416, 2.326)	(7.382, 56.053)	(0.323, 1.837)	(1.379, 10.493)

Table 15 (continued): Part 2 of 2.

K	Measure	gender	SEN	ELL	ILA-WB	ILA-BB	ILA-AB	HVG	HDG	Asian	Others
1	OR	0.937	0.553	0.746	0.890	0.782	1.130	0.888	0.937	0.667	0.544
	CI	(0.278, 3.036)	(0.152, 1.999)	(0.174, 3.316)	(0.248, 3.146)	(0.170, 3.501)	(0.243, 4.624)	(0.220, 3.321)	(0.221, 4.138)	(0.112, 4.160)	(0.137, 2.086)
2	OR	**0.373**	0.468	0.627	1.255	0.896	1.543	0.931	1.228	0.872	0.642
	CI	(0.108, 0.706)	(0.128, 1.664)	(0.156, 2.485)	(0.357, 4.222)	(0.163, 5.149)	(0.361, 6.549)	(0.214, 3.908)	(0.316, 4.809)	(0.157, 4.628)	(0.170, 2.179)

Table 16: Posterior means of odds ratios (OR) for γ_{01} by attribute (K), with 95% confidence intervals (CI). Statistically significant results (CI excluding 1) are shown in **bold**. Part 1 of 2.
Note: K = attribute; (Intercept) = model intercept; nra = number of reattempts; nlm = number of levels mastered; rt = response time; gender = female (0), male (1); SEN = special education needs (1); ELL = English language learner (1); ILA = initial literacy ability (at benchmark: reference level); WB = well below benchmark; BB = below benchmark; AB = above benchmark; group = engagement group (group 5: reference); Others = American Indian, Alaskan Native, Black or African American, Hispanic or Latino, Multiracial/Other, and not Specified (reference group: White); DG = decoding game; VG = vocabulary game; HVG = high vocabulary group; HDG = high decoding group.

K	Measure	(Intercept)	nra DG	nra VG	nlm DG	nlm VG	rt DG	rt VG
1	OR	3.451	**0.363**	1.353	**4.603**	1.556	1.129	1.171
	CI	(0.103,0.978)	(0.342,4.397)	(1.306,17.705)	(0.489,4.934)	(0.339,4.079)	(0.257,5.271)	(0.249,4.140)
2	OR	1.034	1.203	2.336	1.563	1.218	0.666	0.865
	CI	(0.207,5.319)	(0.330,5.358)	(0.490,14.480)	(0.340,6.170)	(0.316,4.974)	(0.130,3.676)	(0.157,5.223)

Table 16 (continued): Part 2 of 2.

K	Measure	gender	SEN	ELL	ILA-WB	ILA-BB	ILA-AB	HVG	HDG	Asian	Others
1	OR	0.997	1.081	0.694	0.511	0.780	1.275	0.370	2.129	1.077	0.829
	CI	(0.663, 17.677)	(0.259, 4.505)	(0.117, 3.890)	(0.111, 2.401)	(0.168, 3.728)	(0.240, 6.760)	(0.084, 1.736)	(0.441, 11.264)	(0.166, 7.341)	(0.193, 3.604)
2	OR	**0.582**	1.159	0.948	0.855	1.281	0.486	1.266	1.220	1.167	1.856
	CI	(0.110, 0.671)	(0.224, 5.919)	(0.159, 5.816)	(0.157, 5.866)	(0.204, 8.096)	(0.085, 3.495)	(0.215, 7.532)	(0.243, 6.961)	(0.180, 7.819)	(0.341, 10.185)

Table 17: Posterior means of odds ratios (OR) for γ_{10} by attribute (K), with 95% confidence intervals (CI). Statistically significant results (CI excluding 1) are shown in **bold**. Part 1 of 2.
Note: K = attribute; (Intercept) = model intercept; nra = number of reattempts; nlm = number of levels mastered; rt = response time; gender = female (0), male (1); SEN = special education needs (1); ELL = English language learner (1); ILA = initial literacy ability (at benchmark: reference level); WB = well below benchmark; BB = below benchmark; AB = above benchmark; group = engagement group (group 5: reference); Others = American Indian, Alaskan Native, Black or African American, Hispanic or Latino, Multiracial/Other, and not Specified (reference group: White); DG = decoding game; VG = vocabulary game; HVG = high vocabulary group; HDG = high decoding group.

K	Measure	(Intercept)	nra DG	nra VG	nlm DG	nlm VG	rt DG	rt VG
1	OR	0.230	2.677	0.412	1.371	0.895	2.019	0.715
	CI	(0.042,1.063)	(0.565,12.403)	(0.102,1.244)	(0.334,5.161)	(0.289,2.761)	(0.695,7.510)	(0.174,2.534)
2	OR	**0.158**	1.056	1.807	0.968	0.308	1.216	0.641
	CI	(0.024,0.877)	(0.232,7.325)	(0.350,6.838)	(0.214,4.878)	(0.056,1.848)	(0.303,5.583)	(0.145,1.970)

Table 17 (continued): Part 2 of 2.

K	Measure	gender	SEN	ELL	ILA-WB	ILA-BB	ILA-AB	HVG	HDG	Asian	Others
1	OR	0.868	1.297	0.824	**6.003**	0.960	1.000	1.404	0.553	1.034	0.625
	CI	(0.191, 3.565)	(0.254, 6.121)	(0.149, 4.333)	(1.073, 31.329)	(0.159, 5.475)	(0.178, 5.534)	(0.289, 7.260)	(0.116, 2.673)	(0.169, 6.330)	(0.138, 2.910)
2	OR	0.637	0.771	0.921	1.458	0.705	0.784	1.339	0.541	0.783	0.572
	CI	(0.111, 3.524)	(0.120, 4.447)	(0.131, 6.252)	(0.208, 8.778)	(0.113, 4.322)	(0.126, 5.026)	(0.195, 7.906)	(0.090, 3.111)	(0.113, 5.220)	(0.106, 3.158)

14 Supplementary Material D

The true Q-matrices are presented under different levels of sparsity and varying numbers of items in Tables 18–20. Each Q-matrix assumes $K = 2$ latent attributes and was held fixed across all simulation replicates.

Table 18: True Q-matrices for $J_t = 6$ under sparse (left) and dense (right) attribute patterns.

Item	**Sparse Matrix**				**Dense Matrix**			
	Time 1		Time 2		Time 1		Time 2	
	A_1	A_2	A_1	A_2	A_1	A_2	A_1	A_2
1	1	0	0	1	1	0	0	1
2	0	1	1	0	0	1	1	0
3	1	0	0	1	1	0	0	1
4	0	1	1	0	0	1	1	0
5	1	0	0	1	1	1	1	1
6	1	1	1	1	1	1	1	1

Table 19: True Q-matrices for $J_t = 18$ under sparse (left) and dense (right) attribute patterns.

Item	**Sparse Matrix**				**Dense Matrix**			
	Time 1		Time 2		Time 1		Time 2	
	A_1	A_2	A_1	A_2	A_1	A_2	A_1	A_2
1	1	0	0	1	1	0	0	1
2	0	1	1	0	0	1	1	0
3	1	0	0	1	1	0	0	1
4	0	1	0	1	0	1	0	1
5	1	0	0	1	1	0	0	1
6	0	1	0	1	0	1	0	1
7	1	0	0	1	1	0	0	1
8	0	1	0	1	0	1	1	0
9	1	0	1	0	1	0	1	0
10	0	1	1	0	0	1	1	0
11	1	0	1	0	1	0	1	0
12	0	1	1	0	0	1	1	0
13	1	0	1	0	1	1	1	1
14	0	1	1	0	1	1	1	1
15	1	1	1	1	1	1	1	1
16	1	1	1	1	1	1	1	1
17	1	1	1	1	1	1	1	1
18	1	1	1	1	1	1	1	1

Table 20: True Q-matrices for $J_t = 30$ under sparse (left) and dense (right) attribute patterns.

Item	**Sparse Matrix**				**Dense Matrix**			
	Time 1		Time 2		Time 1		Time 2	
	A_1	A_2	A_1	A_2	A_1	A_2	A_1	A_2
1	1	0	0	1	1	0	0	1
2	0	1	1	0	0	1	1	0
3	1	0	0	1	1	0	0	1
4	0	1	0	1	0	1	0	1
5	1	0	0	1	1	0	0	1
6	0	1	0	1	0	1	0	1
7	1	0	0	1	1	0	0	1
8	0	1	0	1	0	1	0	1
9	1	0	0	1	1	0	0	1
10	0	1	0	1	0	1	0	1
11	1	0	0	1	1	0	0	1
12	0	1	0	1	0	1	1	0
13	1	0	0	1	1	0	1	0
14	0	1	1	0	0	1	1	0
15	1	0	1	0	1	0	1	0
16	0	1	1	0	0	1	1	0
17	1	0	1	0	1	0	1	0
18	0	1	1	0	0	1	1	0
19	1	0	1	0	1	0	1	0
20	0	1	1	0	0	1	1	0
21	1	0	1	0	1	1	1	1
22	0	1	1	0	1	1	1	1
23	1	0	1	0	1	1	1	1
24	0	1	1	0	1	1	1	1
25	1	1	1	1	1	1	1	1
26	1	1	1	1	1	1	1	1
27	1	1	1	1	1	1	1	1
28	1	1	1	1	1	1	1	1
29	1	1	1	1	1	1	1	1
30	1	1	1	1	1	1	1	1

15 Supplementary Material E

This section gives the simulation results when Q-matrix is known to show our recovery of parameters did not lost accuracy brings with Q-matrix is unknown. The attribute profile classification accuracy is provided in Table 21. Estimation accuracy of item parameters (g_{jt} and s_{jt}), initial covariate effects on attribute mastery (β) and attribute acquisition (γ_{01}) and attribute lossing (γ_{10}) are given in Tables 22–24.

Table 21: Average Attribute Recovery (AAR) across different number of items.

N	J	AAR T1 K1	AAR T1 K2	AAR T2 K1	AAR T2 K2
200	6	0.998	0.993	0.993	0.999
200	18	1.000	1.000	1.000	1.000
200	30	1.000	1.000	1.000	1.000

Table 22: RMSE and bias for guessing (g) and slipping (s) parameters.

N	J	Measure	g (T1)	g (T2)	s (T1)	s (T2)
200	6	Bias	0.029	0.031	0.034	0.032
		RMSE	0.031	0.032	0.035	0.034
200	18	Bias	0.030	0.032	0.034	0.032
		RMSE	0.032	0.034	0.036	0.034
200	30	Bias	0.029	0.030	0.039	0.035
		RMSE	0.031	0.032	0.041	0.038

Table 23: Average bias of β parameters.

N	J	β_0 Bias (K1)	β_0 Bias (K2)	β_Z Bias (K1)	β_Z Bias (K2)
200	6	-0.027	-0.052	0.019	0.029
200	18	-0.077	-0.026	-0.067	0.039
200	30	-0.011	0.065	-0.065	0.011

Table 24: Average bias of transition parameters γ_{01} and γ_{10}.

N	J	γ_{01} Bias (K1)	γ_{01} Bias (K2)	γ_{10} Bias (K1)	γ_{10} Bias (K2)
200	6	0.031	0.138	0.353	-0.165
200	18	-0.052	0.086	0.414	-0.189
200	30	0.012	0.100	0.303	-0.219

16 Supplementary Material F

For sparsity in the Q-matrix,

$$\begin{aligned} Q_{jk} &\sim \text{Bernoulli}(\theta), \\ \theta &\sim \text{Beta}(\alpha, \beta), \end{aligned}$$

where the prior mean $\frac{\alpha}{\alpha+\beta}$ equals the non–zero proportion of the true Q-matrix and the concentration defined as $\alpha + \beta$. We examined the robustness of the model performance to alternative prior specifications for θ, the sparsity parameter of the Q-matrix, under the setting with 6 items with 200 users. Specifically, we varied the prior means ($\frac{\alpha}{\alpha+\beta}$) at 0.5, 0.7, and 0.9 and separately varied the concentration levels ($\alpha + \beta$) at 20, 40, and 60.

Across 25 replications under each prior condition, we assessed the recovery of the Q-matrix, the bias in the posterior means of item parameters (g and s), and the attribute classification accuracy (Tables 25 and 26). Results showed that the changes in posterior estimates remained small across different prior settings, with absolute biases for g and s parameters generally below 0.04. The Q-matrix recovery rates ranged from 75% to 100%, and the attribute-level accuracy rates exceeded 84% in most cases. These findings demonstrate that the model's estimation of Q-matrix, item parameters, and attribute profiles was robust to reasonable prior misspecification.

Table 25: Q-matrix recovery and parameter bias by varying priors.
Note: PriMean = Prior mean ($\frac{\alpha}{\alpha+\beta}$); Conc. = Concentration ($\alpha + \beta$); g = guessing parameter; s = slipping parameter; Q acc. = Q-matrix accuracy. Bold values indicate the highest accuracy and minimum bias across conditions.

Prior			**Time 1**			**Time 2**		
PriMean	Conc.	Beta(α, β)	Q acc.(%)	g_{Bias}	s_{Bias}	Q acc.(%)	g_{Bias}	s_{Bias}
0.5	20	Beta(10, 10)	86.100	0.033	0.037	82.200	0.011	0.024
0.5	40	Beta(20, 20)	91.700	**-0.000**	0.030	86.100	-0.002	0.021
0.5	60	Beta(30, 30)	97.200	**-0.000**	0.046	**100.000**	-0.020	0.017
0.7	20	Beta(12, 8)	94.400	0.001	0.014	91.700	0.008	-0.002
0.7	40	Beta(24, 16)	**100.000**	0.007	**0.007**	**100.000**	**0.005**	**0.007**
0.7	60	Beta(36, 24)	**100.000**	0.006	0.016	97.200	-0.005	0.019
0.9	20	Beta(18, 2)	94.400	0.008	0.065	97.200	0.031	0.070
0.9	40	Beta(36, 4)	**100.000**	0.017	0.107	**100.000**	0.043	-0.009
0.9	60	Beta(54, 6)	97.200	0.001	0.037	97.200	0.029	0.040

Table 26: Attribute-level accuracy rate (AAR) by by varying priors.
Note: PriMean = Prior mean ($\frac{\alpha}{\alpha+\beta}$); Conc. = Concentration ($\alpha + \beta$); T1 = Time 1; T2 = Time 2; A1 = Attribute 1; A2 = Attribute 2. Bold values indicate the highest estimates across conditions.

Prior			**AAR (%)**			
PriMean	Conc.	Beta(α, β)	T1–A1	T1–A2	T2–A1	T2–A2
0.5	20	Beta(10, 10)	96.500	95.300	92.500	94.800
0.5	40	Beta(20, 20)	97.200	95.200	**94.500**	92.700
0.5	60	Beta(30, 30)	96.800	93.500	92.500	95.500
0.7	20	Beta(12, 8)	**97.300**	92.000	93.000	95.500
0.7	40	Beta(24, 16)	96.800	**96.800**	90.800	94.300
0.7	60	Beta(36, 24)	95.300	93.800	93.200	**96.300**
0.9	20	Beta(18, 2)	94.700	92.200	92.500	94.500
0.9	40	Beta(36, 4)	96.500	91.300	90.000	93.000
0.9	60	Beta(54, 6)	95.700	94.000	92.000	89.800

17 Supplementary Material G

Table 27 includes the estimation bias of item parameters for dense Q-matrix.

Table 27: Estimation accuracy of item parameters (g_{jt} and s_{jt}), evaluated by mean bias (MBias), and root mean square error (RMSE), under dense Q-matrix, sample sizes (N), and number of items (J_t). Bold values indicate the highest absolute estimates across conditions.

Sparsity of Q	N	J_t	T	g_{MBias}	g_{RMSE}	s_{MBias}	s_{RMSE}
Dense Q	200	6	1	0.004	0.025	**0.035**	**0.073**
			2	**0.007**	**0.036**	-0.008	0.044
		18	1	0.005	0.027	0.016	0.061
			2	0.003	0.033	0.003	0.037
		30	1	0.003	0.028	0.011	0.061
			2	0.001	0.031	0.006	0.037
	400	6	1	-0.001	0.017	0.015	0.053
			2	**-0.007**	0.024	0.006	0.036
		18	1	0.001	0.019	0.004	0.041
			2	-0.003	0.022	0.000	0.027
		30	1	-0.003	0.020	0.001	0.039
			2	0.000	0.023	0.000	0.024
	600	6	1	0.003	0.016	-0.003	0.029
			2	-0.008	0.028	0.002	0.025
		18	1	-0.002	0.014	0.003	0.037
			2	-0.000	0.017	-0.002	0.019
		30	1	-0.002	0.016	0.000	0.034
			2	-0.001	0.017	-0.002	0.022

18 Supplementary Material H

The detailed item parameters (g and s) shown in Figure 4 showed the RMSD root mean of squres and Mbias (mean bias) for item parameters for each item.

Panels A and B display the item-level RMSE of guessing parameters from two test conditions: $(N = 200, J = 6)$ and $(N = 600, J = 30)$, respectively. Panels C and D show the corresponding RMSEs for slipping parameters under the same conditions. Triangular and square markers represent time points $T = 1$ and $T = 2$, respectively.

Figure 4: Root Mean Squared Errors (RMSE) of item-level guessing and slipping parameter estimates. Panels A and B correspond to the guessing parameters under conditions with $(N = 200, J = 6)$ and $(N = 600, J = 30)$, respectively. Panels C and D show the slipping parameters under the same conditions.

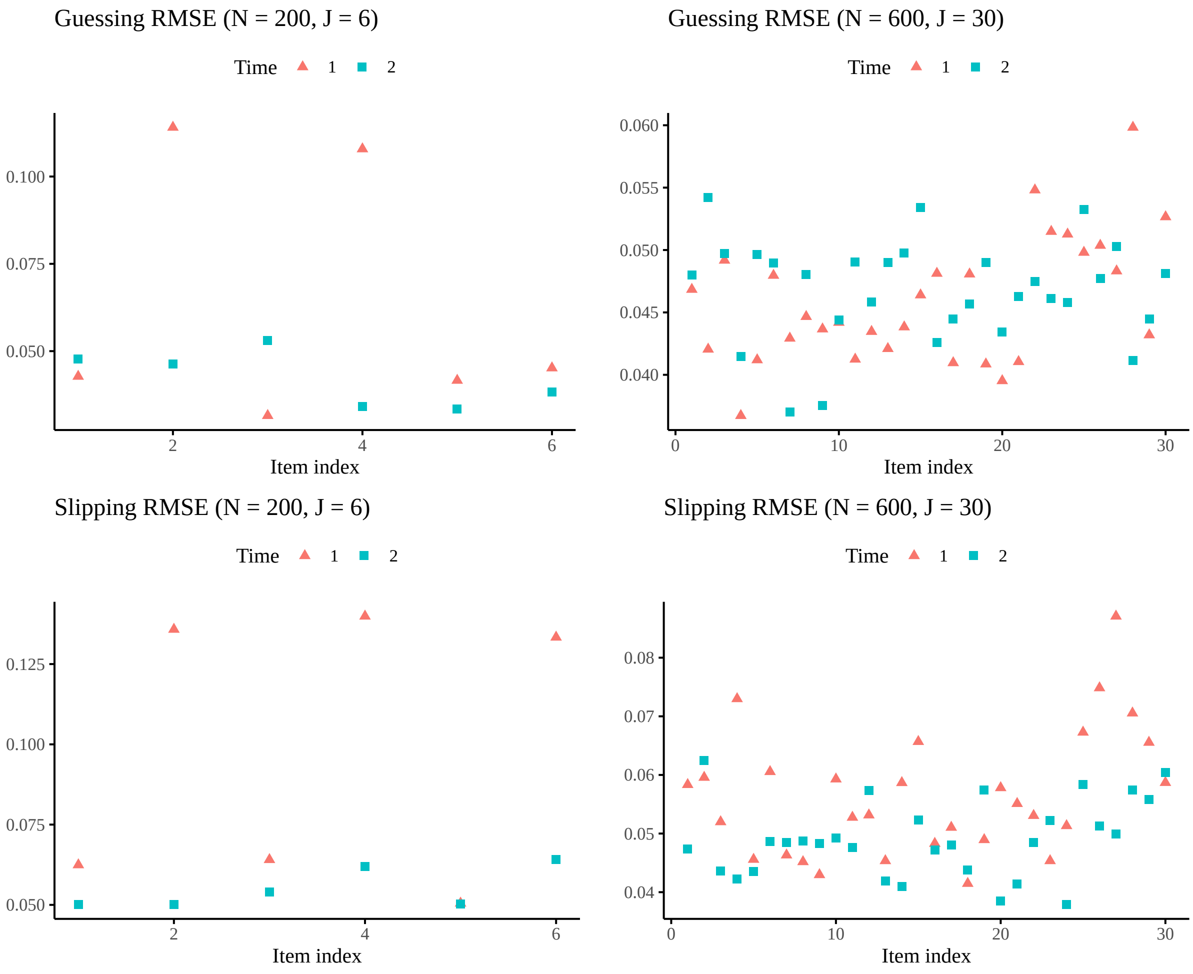
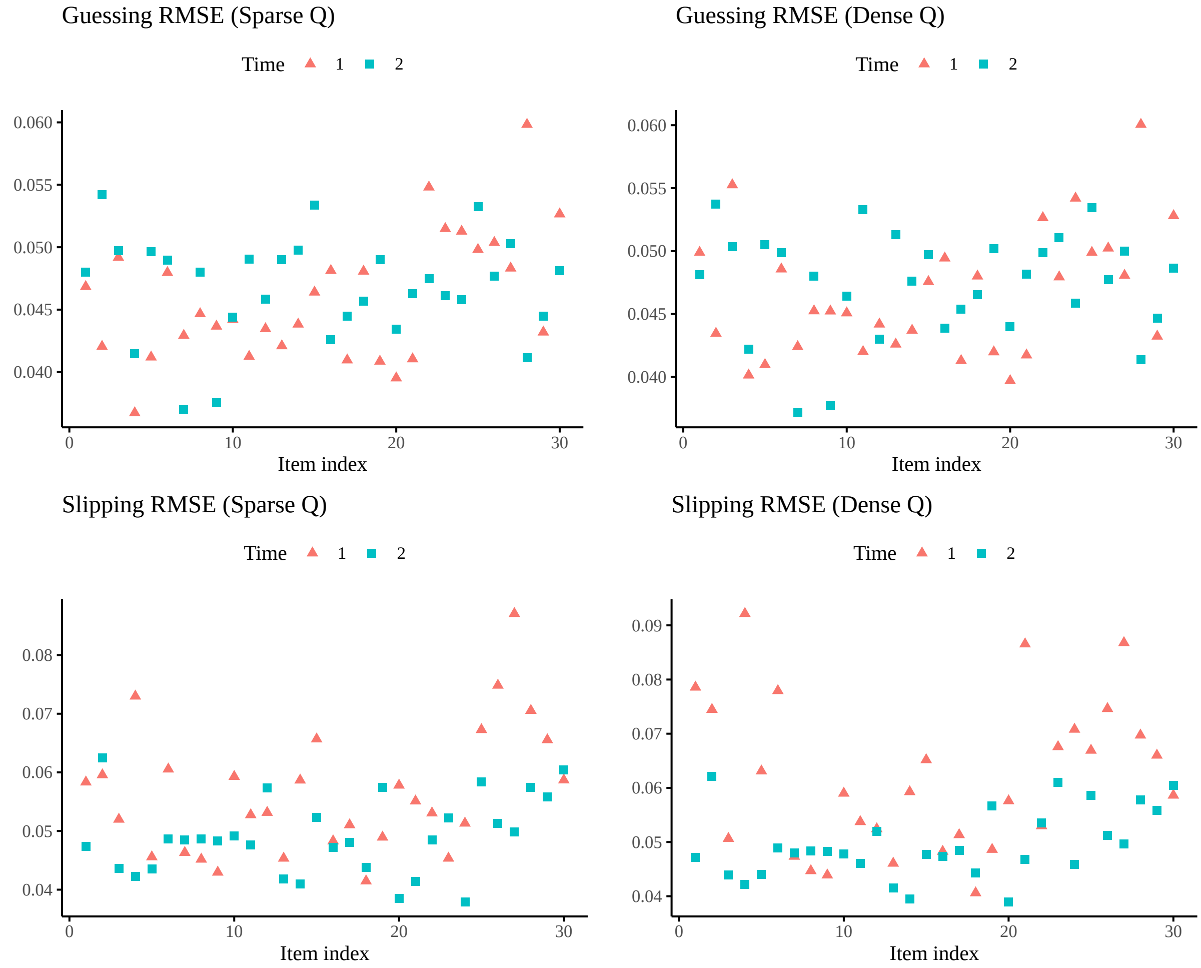


Figure 5 compares the item-level root mean squared errors (RMSE) of the estimated guessing and slipping parameters under two different prior settings for θ (0.5 and 0.7), with fixed test conditions $(N = 600, J = 30)$. Panels A and B present the RMSEs for guessing parameters, while Panels C and D present the results for slipping parameters. The triangular and square markers indicate estimates at time points $T = 1$ and $T = 2$, respectively. The comparison shows how prior informativeness impacts parameter recovery, particularly in terms of consistency across items and time points.

19 Appendix

See the file *Game content* for the full table of participant content codes across 12 items.

Figure 5: RMSE of item-level guessing and slipping parameter estimates under different θ values. Panels A and B show guessing RMSEs for $\theta = 0.5$ and $\theta = 0.7$, respectively. Panels C and D show slipping RMSEs under the same conditions.


Guessing RMSE (Sparse Q)
Time ▲ 1 ■ 2
Item index
Guessing RMSE (Dense Q)
Time ▲ 1 ■ 2
Item index
Slipping RMSE (Sparse Q)
Time ▲ 1 ■ 2
Item index
Slipping RMSE (Dense Q)
Time ▲ 1 ■ 2
Item index