Group-Aware Cognitive Diagnosis via Variational Item Response Theory

Dongxuan Han[1,2], Shuanghong Shen[3,*], Siqi Lei[1,4], Bin Feng[1], Xinjie Sun[1], Xin Zhang[1], Shiwei Tong[1], Wei Huang[5]

[1]State Key Laboratory of Cognitive Intelligence, University of Science and Technology of China, Hefei, China
[2]Southwest University of Science and Technology, Mianyang, China
[3]Institute of Artificial Intelligence, Hefei Comprehensive National Science Center, Hefei, China
[4]City University of Hong Kong, Hongkong, China
[5]Item Bank Department, National Education Examinations Authority, Beijing, China
[*]Correspondence to *shshen@iai.ustc.edu.cn*

Abstract

Group-level cognitive diagnosis is fundamental in educational assessment, where examinees are organized into groups like schools or classes. However, most methods rely on mean-field assumptions that treat students independently, ignoring dependencies induced by shared resources. This oversight yields unstable estimates, particularly under sparse observations. We propose **Group-aware Latent Ability Diagnosis (GLAD)**, a hierarchical variational framework that models within-group dependence via a latent group context while retaining scalable inference. We decompose student ability into a global mean, a group effect, and an individual residual, yielding three variants: a group-conditioned prior, a deterministic effect, and a stochastic effect. For efficient inference, we develop a permutation-invariant encoder to aggregate student representations. Experiments on real-world Eedi data demonstrate consistent improvements over baselines across multiple CDMs.

Keywords: Cognitive Diagnosis; Item Response Theory; Hierarchical Bayesian Modeling; Educational Data Mining

Introduction

Cognitive diagnosis aims to infer learners' latent proficiency from their responses to assessment items, and is fundamental to educational measurement, psychological surveys, and questionnaires (Jiang et al., 2024; Q. Liu et al., 2018; Wang, Huang, et al., 2023; G. Yu et al., 2025). A canonical approach is Item Response Theory (IRT) (Lord, 1952), which models the probability of a correct response as a function of latent student ability and item parameters (e.g., difficulty and discrimination), typically under two assumptions: (i) students are conditionally independent given their abilities, and (ii) item properties are invariant across examinees. These latent variables are estimated from historical response logs and serve as the basis for prediction and diagnosis.

In practice, educational data are inherently hierarchical: students are nested within classes, schools, or regions, and share common influences such as instructional quality, curricular coverage, peer effects, and sociocultural context. Such group structure induces within-group dependence and cross-group heterogeneity, directly challenging the independence assumptions of standard IRT. Classical multilevel IRT addresses this issue via group-level random effects, providing a principled Bayesian treatment but often relying on inference procedures that are difficult to scale to modern, sparse, and large datasets. Recent learning-based cognitive diagnosis models further increase modeling flexibility, which need a structured, scalable inference that can exploit group information without sacrificing probabilistic interpretability.

Variational inference (VI) has emerged as a practical solution for large-scale Bayesian IRT (Wu et al., 2020), yet existing variational IRT methods adopt mean-field posteriors that factorize across students. As a result, group effects are often incorporated only indirectly. To address this, we propose *Group-aware Latent Ability Diagnosis* (GLAD), a hierarchical variational framework for group-aware cognitive diagnosis with multidimensional abilities.

The key idea is to introduce a group-level latent context that captures shared influences and to decompose each student's ability into a global mean, a group component, and an individual residual. This yields a unified probabilistic view that supports three variants: (1) a group-specific prior over student abilities (GLAD-GP), (2) a deterministic group effect (GLAD-FE), and (3) a stochastic group effect via a shared latent context (GLAD-SE). Importantly, GLAD-SE provides a principled mechanism to induce within-group dependence in the marginal ability distribution through a shared latent cause, while remaining compatible with standard cognitive diagnosis models(CDMs).

For inference, we develop permutation-invariant amortized encoders: a student encoder for individual residuals and a group encoder for group contexts, trained end-to-end by maximizing an ELBO.

Our contributions are summarized as follows:

- We propose a unified hierarchical generative framework for group-aware ability estimation that explicitly separates global, group, and individual components, and clarifies the roles of group-conditioned priors, deterministic group effects, and stochastic group effects.
- We develop permutation-invariant amortized inference with a student encoder and a group encoder with pooling aggregation, enabling scalable variational learning on large grouped response logs.
- We conduct extensive experiments on a real-world grouped response dataset, showing consistent gains in predictive performance, likelihood, and calibration over independent variational baseline; additionally, we analyze robustness under shrinking group sizes and provide empirical evidence on group cohesiveness and predictive uncertainty.

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

Related Works

Cognitive Diagnosis. IRT (Lord, 1952) models responses via student ability and item parameters. Classic CDMs like DINA (De La Torre, 2009) extend this to multidimensional concepts. Recently, neural approaches, e.g., NCD, KaNCD (Wang, Liu, et al., 2023) have improved expressiveness using deep networks. Current trends also harness Large Language Models (LLMs) for interpretability (Wang et al., 2025) and introduce causal reasoning to mitigate bias (Zhang et al., 2024). Yet, most methods treat students as independent, ignoring inherent educational dependencies.

Group Effects. Recognizing nested student structures, recent works explore group-aware diagnosis. Early attempts utilized multi-task learning or subgroup discovery (Huang et al., 2021; S. Liu et al., 2023). Advanced methods like RDGT (X. Yu et al., 2024) employ relation-guided graph transformers to explicitly capture student-group interactions. While effective for sparse data, these models typically rely on deterministic embeddings, neglecting group-level uncertainty.

Variational Methods. Bayesian methods quantify uncertainty but struggle with scalability (Bürkner, 2021). Variational Inference (VI) offers an efficient alternative with flexible response functions (Chen et al., 2019; Wu et al., 2020). However, standard variational IRT assumes a mean-field posterior that factorizes across students. This work extends variational IRT to hierarchical group structures, bridging scalable inference with group-level dependency modeling.

Preliminaries

Problem Definition

Let $\mathcal{G} = \{g_1, \ldots, g_L\}$ be a set of disjoint groups. Each student i belongs to a group $g(i) \in \mathcal{G}$. For a group g, denote its students as $\mathcal{S}_g = \{i : g(i) = g\}$ with size N_g. Let $\mathcal{E} = \{1, \ldots, M\}$ be items and $\mathcal{C} = \{1, \ldots, K\}$ be knowledge concepts. We observe binary interactions $\mathcal{I} = \{(i, j, r_{ij})\}$, where $r_{ij} \in \{0, 1\}$ indicates whether student i answers item j correctly.

For each student, let $\mathbf{r}_i = \{(j, r_{ij}) : (i, j, r_{ij}) \in \mathcal{I}\}$ be the set of observed responses, and let $\mathcal{J}(i) = \{j : (i, j, r_{ij}) \in \mathcal{I}\}$ be the set of attempted items. For each group, let $R_g = \{(i, j, r_{ij}) \in \mathcal{I} : g(i) = g\}$ be the interactions in group g.

The Q-matrix $\boldsymbol{Q} \in \{0, 1\}^{M \times K}$ encodes which concepts each item assesses; denote $\boldsymbol{q}_j \in \{0, 1\}^K$ as the j-th row.

Definition 1 (Student Performance Prediction) *Given a student i (with group membership $g(i)$) and an item j, the goal is to predict the probability of a correct response, $p(r_{ij} = 1 \mid \mathcal{I})$, using inferred student ability and item parameters.*

We consider a general form of decoder in CDMs:

$$p(r_{ij} = 1 \mid \boldsymbol{\theta}_i, j) = \sigma\big(f_\beta(\boldsymbol{\theta}_i, j)\big), \tag{1}$$

where $\boldsymbol{\theta}_i \in \mathbb{R}^K$ is the K-dimensional ability vector aligned with the concept space, $\sigma(\cdot)$ is the sigmoid function, and

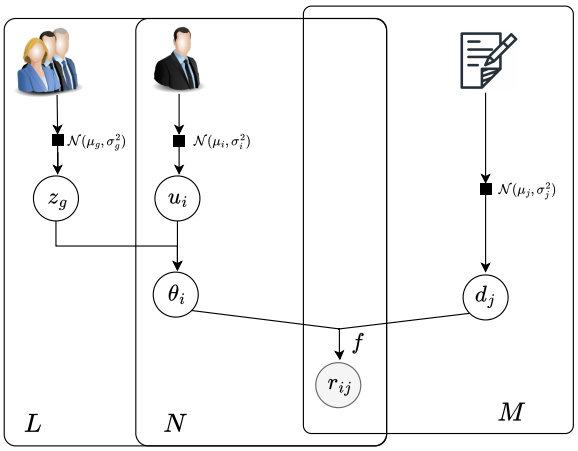

Figure 1: Unified hierarchical probability model of group-aware cognitive diagnosis.

$f_\beta(\cdot)$ is a logit function parameterized by decoder parameters β (e.g., item parameters).

For multidimensional IRT (MIRT) logit, f_β is obtained with a Q-matrix masked :

$$f_\beta(\boldsymbol{\theta}_i, j) = (\boldsymbol{a}_j \odot \boldsymbol{q}_j)^\top \boldsymbol{\theta}_i - b_j, \tag{2}$$

where $\boldsymbol{a}_j \in \mathbb{R}^K$ is an item discrimination vector, $b_j \in \mathbb{R}$ is item difficulty, and $\odot$ masks irrelevant dimensions.

Variational Inference

Estimating latent abilities requires marginalizing $\boldsymbol{\theta}_i$ under the nonlinear likelihood in Eq. (1), which is generally intractable. Following variational approaches such as VIBO (Wu et al., 2020), we introduce an amortized variational posterior $q_\varphi(\boldsymbol{\theta}_i \mid \mathbf{r}_i)$ parameterized by an inference network (encoder) with parameters φ. Maximizing the marginal log-likelihood is then approximated by maximizing the Evidence Lower Bound (ELBO), a metric that balances data reconstruction accuracy with prior regularization:

$$\log p(\mathcal{I}) \geq \mathcal{L}_{\text{ELBO}}(\varphi) = \mathbb{E}_{q_\varphi}\Big[\sum_{(i,j,r_{ij}) \in \mathcal{I}} \log p(r_{ij} \mid \boldsymbol{\theta}_i, j)\Big] - \sum_i \mathrm{KL}\big(q_\varphi(\boldsymbol{\theta}_i \mid \mathbf{r}_i) \,\|\, p(\boldsymbol{\theta}_i)\big), \tag{3}$$

where the first term is the reconstruction term and the second term regularizes the posterior toward a prior, typically $p(\boldsymbol{\theta}_i) = \mathcal{N}(\mathbf{0}, \mathbf{I}_K)$ under the independence assumption.

Methodology

We propose GLAD, a hierarchical Bayesian framework that explicitly models within-group dependence via scalable permutation-invariant amortized inference.

Unified Hierarchical Probability Model

Recall that each student i belongs to a group $g(i)$ and interactions are observed as $(i, j, r_{ij}) \in \mathcal{I}$. We model student ability

as a sum of a global mean, a group effect, and an individual residual:

$$\boldsymbol{\theta}_i = \boldsymbol{\mu}_\theta + \Delta_{g(i)} + \boldsymbol{u}_i, \qquad \boldsymbol{u}_i \sim \mathcal{N}(\mathbf{0}, \mathrm{diag}(\boldsymbol{\sigma}_u^2)). \quad (4)$$

Given $\boldsymbol{\theta}_i$, responses are generated by the decoder in Eq. (1) (e.g., the masked MIRT instantiation in Eq. (2)).

Different variants correspond to different parameterizations of the group effect Δ_g.

Variant 1: Group-Conditioned Prior

This variant GLAD-GP injects group information through a group-conditioned prior on abilities, which:

$$p(\boldsymbol{\theta}_i \mid g(i) = g) = \mathcal{N}(\boldsymbol{\mu}_g, \mathrm{diag}(\boldsymbol{\sigma}_g^2)), \quad (5)$$

where $(\boldsymbol{\mu}_g, \boldsymbol{\sigma}_g)$ are learned per group (e.g., via group embeddings). We use a mean-field variational posterior $q_\varphi(\boldsymbol{\theta}_i \mid \mathbf{r}_i)$ and optimize

$$\begin{aligned}\mathcal{L}_{\mathrm{GP}} = \mathbb{E}_{q_\varphi}\Big[\sum_{(i,j,r_{ij})\in\mathcal{I}} \log p(r_{ij} \mid \boldsymbol{\theta}_i, j)\Big] \\ - \sum_i \mathrm{KL}\big(q_\varphi(\boldsymbol{\theta}_i \mid \mathbf{r}_i) \,\|\, p(\boldsymbol{\theta}_i \mid g(i))\big)\,.\end{aligned} \quad (6)$$

Variant 2: Fixed Effect Model

This variant GLAD-FE models the group effect as a deterministic embedding: $\Delta_g = B\boldsymbol{e}_g$, where $\boldsymbol{e}_g \in \mathbb{R}^p$ is a learned group embedding and $B \in \mathbb{R}^{K\times p}$ is shared across groups. We perform VI for $\boldsymbol{u}_i$ with a prior $p(\boldsymbol{u}_i) = \mathcal{N}(\mathbf{0}, \mathrm{diag}(\boldsymbol{\sigma}_u^2))$:

$$\begin{aligned}\mathcal{L}_{\mathrm{FE}} = \mathbb{E}_{q_\varphi}\Big[\sum_{(i,j,r_{ij})\in\mathcal{I}} \log p(r_{ij} \mid \boldsymbol{\theta}_i, j)\Big] \\ - \sum_i \mathrm{KL}\big(q_\varphi(\boldsymbol{u}_i \mid \mathbf{r}_i) \,\|\, p(\boldsymbol{u}_i)\big)\,,\end{aligned} \quad (7)$$

where $\boldsymbol{\theta}_i = \boldsymbol{\mu}_\theta + B\boldsymbol{e}_{g(i)} + \boldsymbol{u}_i$.

Variant 3: Stochastic Effect Model

GLAD-SE introduces a shared *stochastic* group effect via a latent group context:

$$\boldsymbol{z}_g \sim \mathcal{N}(\mathbf{0}, I_p), \qquad \Delta_g = B\boldsymbol{z}_g. \quad (8)$$

Thus $\boldsymbol{\theta}_i = \boldsymbol{\mu}_\theta + B\boldsymbol{z}_{g(i)} + \boldsymbol{u}_i$. We use the structured variational family

$$q_\varphi(\{\boldsymbol{z}_g\}, \{\boldsymbol{u}_i\} \mid \mathcal{I}) = \prod_g q_\varphi(\boldsymbol{z}_g \mid R_g) \prod_i q_\varphi(\boldsymbol{u}_i \mid \mathbf{r}_i), \quad (9)$$

with diagonal Gaussian factors. The ELBO is

$$\begin{aligned}\mathcal{L}_{\mathrm{SE}} = \mathbb{E}_{q_\varphi}\Big[\sum_{(i,j,r_{ij})\in\mathcal{I}} \log p(r_{ij} \mid \boldsymbol{\theta}_i, j)\Big] \\ - \sum_g \mathrm{KL}\big(q_\varphi(\boldsymbol{z}_g \mid R_g) \,\|\, p(\boldsymbol{z}_g)\big) - \sum_i \mathrm{KL}\big(q_\varphi(\boldsymbol{u}_i \mid \mathbf{r}_i) \,\|\, p(\boldsymbol{u}_i)\big)\,.\end{aligned} \quad (10)$$

Proposition 1 (Within-group dependence) *Under GLAD-SE, for two distinct students $i \neq i'$ in the same group g, since $\boldsymbol{\theta}_i = \boldsymbol{\mu}_\theta + B\boldsymbol{z}_g + \boldsymbol{u}_i$ with independent $\boldsymbol{u}_i$ and shared $\boldsymbol{z}_g \sim \mathcal{N}(\mathbf{0}, I_p)$, we have* $\mathrm{Cov}(\boldsymbol{\theta}_i, \boldsymbol{\theta}_{i'}) = \mathrm{Cov}(B\boldsymbol{z}_g, B\boldsymbol{z}_g) = B\,\mathrm{Cov}(\boldsymbol{z}_g)\,B^\top = BB^\top$.

Marginalizing $\boldsymbol{z}_g$ yields the group-marginal prior

$$p(\boldsymbol{\theta}_i \mid g(i) = g) = \mathcal{N}\Big(\boldsymbol{\mu}_\theta,\ BB^\top + \mathrm{diag}(\boldsymbol{\sigma}_u^2)\Big), \quad (11)$$

which exhibits a shared low-rank covariance $BB^\top$ among members of the same group.

Remark. In contrast to GLAD-SE and GLAD-GP, in GLAD-FE, $\Delta_g = B\boldsymbol{e}_g$ is deterministic, hence $\mathrm{Cov}(\boldsymbol{\theta}_i, \boldsymbol{\theta}_{i'}) = \mathbf{0}$ for $i \neq i'$ in the same group. Similarly, in GLAD-GP, student abilities are independent given the group-specific prior in Eq. (5), and the mean-field posterior preserves this factorization. Consequently, only GLAD-SE provides a probabilistic mechanism to model structured within-group dependence while retaining conditional independence given $\boldsymbol{z}_g$.

Permutation-invariant Amortized Inference

We implement amortized inference via encoders that map observed response sets to variational parameters. Amortized inference utilizes a neural network (encoder) to map varying lengths of student response sets into latent distribution parameters, and being permutation-invariant ensures the output is unaffected by the order of the items.

Student encoder. For each student i, we embed each observed interaction (j, r_{ij}) and pool over the set:

$$\boldsymbol{x}_{ij} = \mathrm{Emb}_e(j) + \mathrm{Emb}_r(r_{ij}), \quad \boldsymbol{h}_i = \mathrm{Pool}\big(\{\boldsymbol{x}_{ij}\}_{j\in\mathcal{J}(i)}\big), \quad (12)$$

then output $(\boldsymbol{\mu}_i^{(u)}, \boldsymbol{\sigma}_i^{(u)})$ for $q_\varphi(\boldsymbol{u}_i \mid \mathbf{r}_i)$,or $(\boldsymbol{\mu}_i^{(\theta)}, \boldsymbol{\sigma}_i^{(\theta)})$ for GLAD-GP.

Group encoder. We aggregate student representations within a group using a permutation-invariant set encoder:

$$\bar{\boldsymbol{h}}_g = \rho\left(\frac{1}{|\mathcal{S}_g|}\sum_{i\in\mathcal{S}_g}\phi(\boldsymbol{h}_i)\right), \qquad (\boldsymbol{\mu}_g^{(z)}, \boldsymbol{\sigma}_g^{(z)}) = \mathrm{Enc}_z(\bar{\boldsymbol{h}}_g), \quad (13)$$

which is invariant to the students ordering in the same group.

Training and Optimization

All model parameters, including the decoder parameters β, the global parameters $(\boldsymbol{\mu}_\theta, B)$, and the inference networks $q_\varphi(\boldsymbol{u}_i \mid \mathbf{r}_i)$ and $q_\varphi(\boldsymbol{z}_g \mid R_g)$, are trained jointly by maximizing the corresponding ELBO (Eqs. (6)–(10)). We employ the reparameterization trick for Gaussian latent variables to obtain low-variance stochastic gradients.

To scale to large grouped response logs, we perform mini-batch training by sampling a subset of groups and students within each group. All KL divergences are between diagonal Gaussians and admit closed-form computation.

Table 1: Statistics of Eedi dataset.

Statistic	Value	Statistic	Value
Users	4,451	Sparsity	0.6868
Items	916	Avg. Correctness	0.5302
Concepts	85	Groups	234
Responses	1,276,921	Avg. Student/Group	19.02

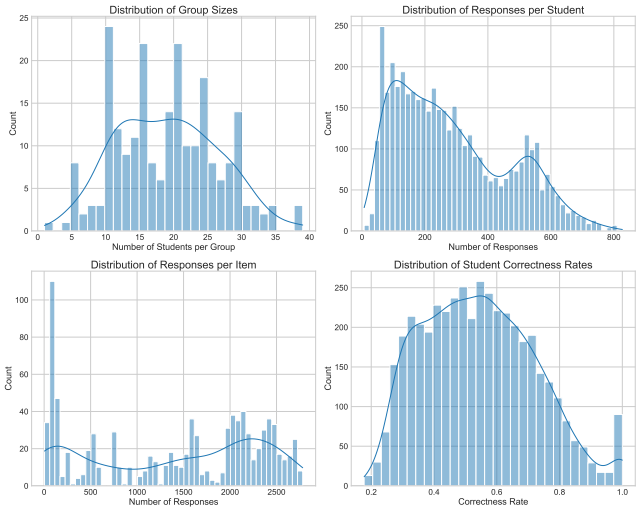

Figure 2: Statistical distributions of the Eedi dataset.

Experiments

We conduct experiments to answer the following questions: **RQ1:** Does leveraging group contexts improve the accuracy, reliability, and calibration of student performance prediction? **RQ2:** Is the proposed group modeling more robust to variations in group size, particularly in "cold-start" scenarios? **RQ3:** How does the model leverage group information internally? Specifically, does it capture group cohesiveness and reduce predictive uncertainty?

Dataset Description

We evaluate on large-scale grouped student responses from Eedi education platform[1].The dataset provides student responses of teaching classes, and knowledge concept mapping. The dataset is open-sourced in Kaggle[2].

The dataset reveals a relatively dense response matrix with a sparsity of 68.68%.The dataset comprises 4,451 students and 916 items covering 85 knowledge concepts. The students are organized into 234 distinct groups (classes), with an average group size of 19.02 students. Key data distributions are visualized in Figure 2.

We use an interaction-level split within each student. For every student, the attempted items are randomly shuffled and divided into training, validation, and test interactions with a ratio of 70%/10%/20%. Thus, the same student and group may appear in all three splits, but no student–item response interaction appears in more than one split. This setting evaluates held-out response prediction for partially observed students in known groups, rather than transfer to entirely unseen students or unseen groups. During training, the model only observes training interactions; validation interactions are used for model selection, and all reported metrics are computed on held-out test interactions.

[1]https://www.eedischool.com/

[2]https://www.kaggle.com/datasets/alejopaullier/eedi-external-dataset

Experimental Setting

All models are implemented in PyTorch. To ensure reproducibility, the random seeds are fixed from 42 to 51 for 10 runs. For training, we use the Adam optimizer with a learning rate of 0.01, a batch size of 32, and train all models for 20 epochs, which is sufficient for convergence. For models involving neural components, the hidden dimension is uniformly set to 64 to control model capacity and ensure fair comparison. We evaluate our proposed framework on four representative CDMs: IRT, MIRT, NCD, KaNCD. While specialized group-aware CDMs exist, our evaluation focuses on independent baselines to clearly ablate and demonstrate the plug-and-play efficacy of the GLAD framework across standard backbones.

Student Performance Prediction (RQ1)

Evaluation Metrics. We evaluate results of CDMs with metrics for discrimination, prediction quality, and calibration. For interactions with label and predicted probability, we report: **AUC** (%, ↑) for ranking ability; **ACC** (%, ↑) and **F1** (%, ↑) for prediction performance; **RMSE** (↓) for probability regression error; **NLL** (↓) for negative log-likelihood; and calibration errors **ECE** (↓) and **MCE** (↓), measuring the average and worst gap between predicted confidence and empirical accuracy across confidence bins.

Results. As shown in Table 2, all GLAD variants consistently outperform the Origin baseline across different CDMs, indicating that our method improves both predictive performance and reliability. For IRT, GLAD-GP achieves the best performance, while GLAD-FE and GLAD-SE further improve calibration. For MIRT, GLAD-FE attains the best overall results and achieves the lowest ECE and MCE, demonstrating clear gains in both accuracy and uncertainty quality. For NCD, GLAD-SE delivers the strongest and most stable improvements. On KaNCD, GLAD-GP achieves the best discrimination and error metrics, and importantly fixes the severe miscalibration of Origin, showing the effectiveness of GLAD in enhancing both performance and calibration.

Robustness to Group Size(RQ2)

Setting. To examine how sensitive each method is to the amount of group-level information, we stratify groups by their sizes and report performance in bucketed ranges (from large groups to small groups). We treat large groups (e.g.,

Table 2: Performance Comparison on Different CDMs. Best results are **bolded**.

CDMs	Model	AUC(%)↑	ACC(%)↑	F1(%)↑	RMSE↓	NLL↓	ECE↓	MCE↓
IRT	Origin	62.56 ± 0.35	59.44 ± 0.31	61.95 ± 1.73	0.5032 ± 0.0014	0.7310 ± 0.0075	0.1045 ± 0.0068	0.2693 ± 0.0268
	GLAD-GP	**71.18 ± 4.26**	**65.76 ± 3.33**	**67.27 ± 2.60**	**0.4663 ± 0.0137**	**0.6249 ± 0.0282**	0.0517 ± 0.0130	0.1140 ± 0.0265
	GLAD-FE	68.53 ± 0.56	63.64 ± 0.36	65.99 ± 0.70	0.4739 ± 0.0019	0.6403 ± 0.0040	**0.0316 ± 0.0084**	0.0846 ± 0.0168
	GLAD-SE	68.15 ± 0.08	63.40 ± 0.05	65.80 ± 0.34	0.4756 ± 0.0012	0.6439 ± 0.0025	0.0376 ± 0.0107	**0.0798 ± 0.0127**
MIRT	Origin	66.48 ± 1.87	62.21 ± 1.30	65.91 ± 1.69	0.4801 ± 0.0069	0.6548 ± 0.0169	0.0319 ± 0.0203	0.1032 ± 0.0737
	GLAD-GP	77.18 ± 0.21	70.88 ± 0.13	71.13 ± 0.47	0.4404 ± 0.0011	0.5745 ± 0.0032	0.0284 ± 0.0023	0.1813 ± 0.0157
	GLAD-FE	**77.84 ± 0.02**	**71.18 ± 0.04**	71.46 ± 0.21	**0.4360 ± 0.0002**	**0.5590 ± 0.0003**	**0.0129 ± 0.0017**	**0.0663 ± 0.0005**
	GLAD-SE	77.83 ± 0.02	71.13 ± 0.08	**71.34 ± 0.36**	0.4361 ± 0.0001	0.5592 ± 0.0000	0.0139 ± 0.0014	0.1471 ± 0.0365
NCD	Origin	62.41 ± 0.12	59.16 ± 0.09	61.41 ± 0.67	0.5076 ± 0.0018	0.7499 ± 0.0099	0.1215 ± 0.0055	0.2662 ± 0.0102
	GLAD-GP	77.60 ± 0.09	71.22 ± 0.07	71.24 ± 0.19	0.4373 ± 0.0007	0.5689 ± 0.0046	0.0193 ± 0.0065	0.0469 ± 0.0086
	GLAD-FE	77.72 ± 0.11	71.25 ± 0.09	71.42 ± 0.51	0.4366 ± 0.0005	0.5616 ± 0.0019	0.0182 ± 0.0061	0.0736 ± 0.0701
	GLAD-SE	**77.86 ± 0.12**	**71.29 ± 0.12**	**71.70 ± 0.47**	**0.4361 ± 0.0013**	**0.5609 ± 0.0040**	**0.0175 ± 0.0097**	**0.0398 ± 0.0236**
KaNCD	Origin	77.82 ± 0.06	71.13 ± 0.04	71.49 ± 0.54	0.4364 ± 0.0003	0.5604 ± 0.0009	0.0151 ± 0.0038	0.3041 ± 0.2336
	GLAD-GP	**77.97 ± 0.11**	**71.27 ± 0.11**	**71.50 ± 0.35**	**0.4354 ± 0.0007**	**0.5583 ± 0.0015**	0.0135 ± 0.0051	**0.0326 ± 0.0058**
	GLAD-FE	77.77 ± 0.05	71.08 ± 0.09	71.17 ± 0.62	0.4364 ± 0.0002	0.5605 ± 0.0007	0.0128 ± 0.0042	0.0362 ± 0.0055
	GLAD-SE	77.70 ± 0.08	71.08 ± 0.06	71.26 ± 0.47	0.4368 ± 0.0004	0.5602 ± 0.0011	**0.0127 ± 0.0044**	0.0426 ± 0.0078

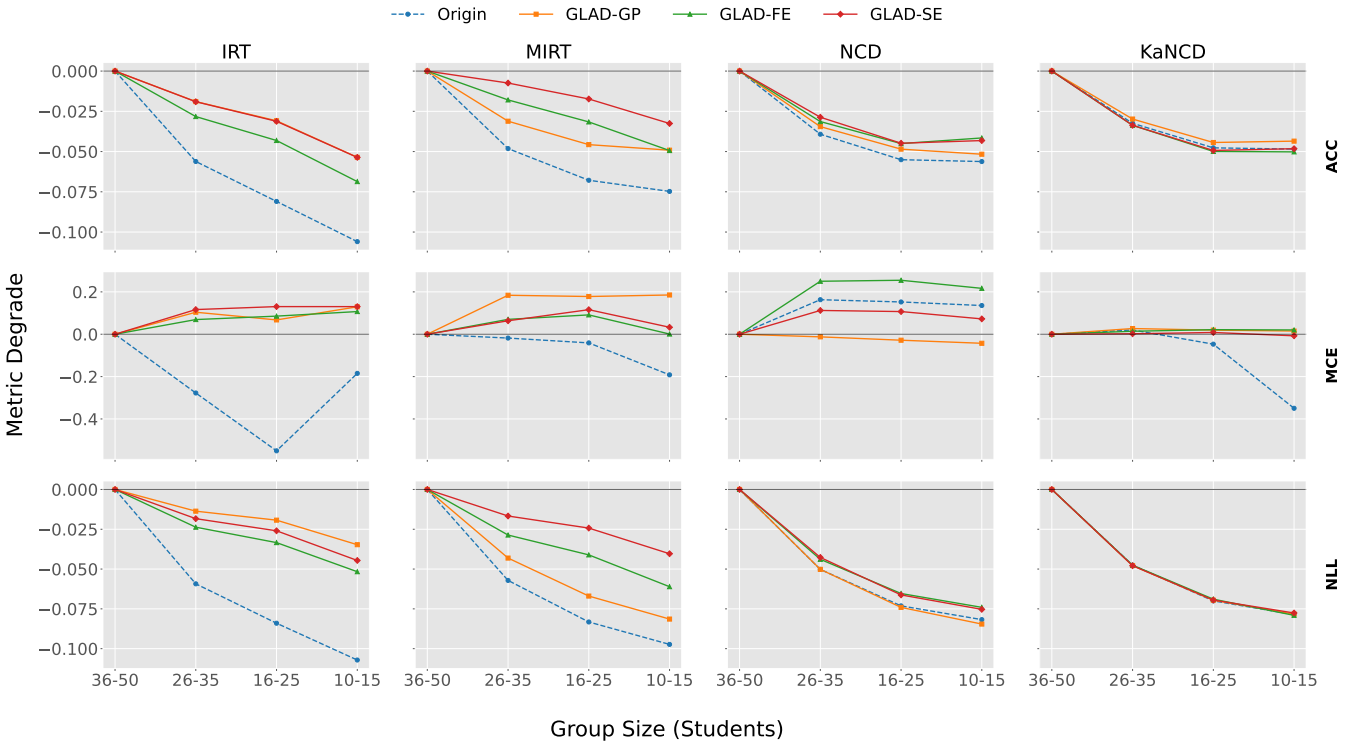

Figure 3: Bucket-wise Performance Degradation from Large to Small Groups.

36–50 students) as the reference point and measure the per-bucket performance change as group size decreases toward the smallest bucket (10–15), which approximates a colder setting with fewer within-group observations. We compare Origin (VIBO) with group-aware variants GLAD-GP, GLAD-FE, and GLAD-SE under four CDM backbones, and report metrics including AUC, NLL, and MCE. The change values are normalized by the reference bucket so that a larger magnitude of degradation indicates higher sensitivity to group size.

Results. Figure 3 shows that performance generally degrades as group size decreases, indicating that smaller groups provide less reliable group signals for modeling. Across backbones and metrics, GLAD-SE exhibits the smallest degradation in most cases, suggesting it is more robust under cold-start-like regimes where group information is sparse. In contrast, methods that rely primarily on fixed group representations (GLAD-FE) or a weak group prior (GLAD-GP) often show larger drops, implying higher dependence on group size and reduced generalization to small groups. Overall, these results highlight the advantage of stochastic group embeddings

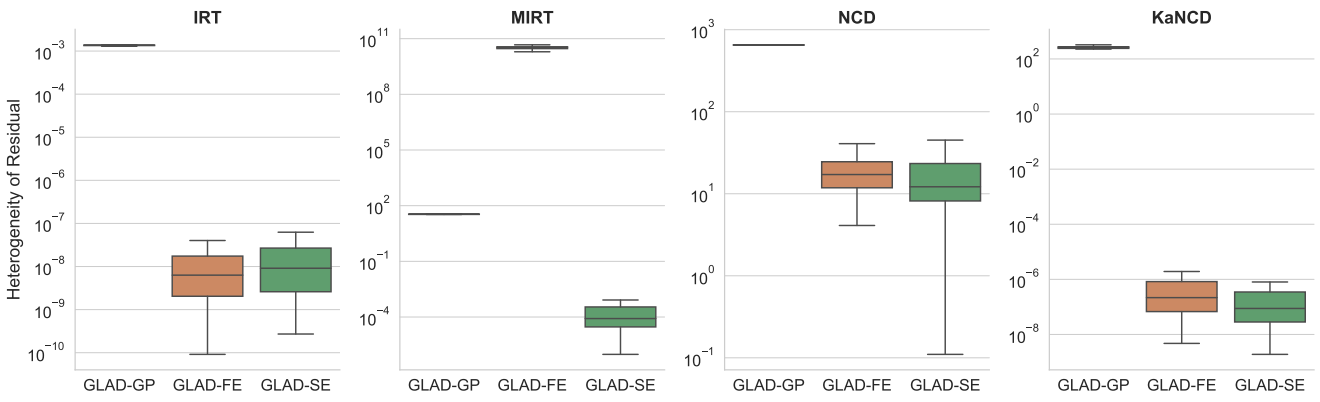

Figure 4: Comparison of student-group residual heterogeneity across different CDMs.

in maintaining stable performance when group size shrinks.

Group Cohesiveness via Heterogeneity(RQ3)

Setting. We study whether different group-aware variants learn cohesive group representations by measuring the *residual heterogeneity* within each student group. For a group g, we compute the dispersion of individual residual abilities (i.e., deviations from the group embedding) and aggregate it as a heterogeneity score; lower values indicate that students within the same group are more consistent.

Results. Figure 4 reports the distribution and averaged trends of group residual heterogeneity. Across most backbones, GLAD-SE yields the lowest heterogeneity, suggesting that stochastic group embeddings provide a more effective and robust inductive bias for capturing group-level commonality while suppressing idiosyncratic noise. In contrast, GLAD-GP typically produces much larger heterogeneity, indicating that a prior alone is insufficient to tightly align group members. Moreover, GLAD-FE can be sensitive in high-dimensional settings, where deterministic embeddings may lead to unstable residual scales, whereas GLAD-SE remains stable.

Uncertainty with Group Modeling(RQ3)

Setting. Beyond accuracy, we ask whether group-aware modeling can make predictions *more certain* by pooling evidence from group peers. For each student–item pair, we draw multiple Monte Carlo samples from the variational posterior and compute the standard deviation of the resulting predictive probabilities; we then report the mean ± std over the test set. We compare the Origin (no explicit group representation) with GLAD-SE (stochastic group embedding and individual residual) under four CDM backbones.

Results. Table 3 reveals a clear pattern: when the backbone is expressive (MIRT/NCD/KaNCD), GLAD-SE substantially *reduces* predictive uncertainty, suggesting that the stochastic group embedding provides a strong shared signal that shrinks the posterior and resolves ambiguity from sparse individual responses. The most striking case is MIRT, where GLAD-SE collapses uncertainty by nearly two orders of magnitude ($0.2363 \rightarrow 0.0060$), indicating that group-level information is especially valuable in high-dimensional ability spaces. Interestingly, IRT shows the opposite trend: GLAD-SE yields higher uncertainty than Origin. Rather than a simple "worse" outcome, this highlights a capacity mismatch: a 1D IRT backbone is too restrictive to absorb rich group structure, so the model expresses the remaining ambiguity as posterior dispersion instead of overconfidently collapsing—an effect that can be beneficial for risk-sensitive decision making and cold-start deployment. Overall, the uncertainty results complement performance metrics by showing when group information truly tightens beliefs and when it exposes under-modeling.

Table 3: Comparison of Uncertainty, which is measured by the standard deviation of the predicted posterior.

Model	Origin	GLAD-SE
IRT	**0.0361 ± 0.0292**	0.7512 ± 0.0336
MIRT	0.2363 ± 0.0392	**0.0060 ± 0.0001**
NCD	0.6877 ± 0.0292	**0.5235 ± 0.4049**
KaNCD	0.8673 ± 0.0230	**0.6042 ± 0.0004**

Conclusion

We propose GLAD, a hierarchical variational framework that captures within-group dependence via a shared latent context. Unlike standard independent approaches, our Stochastic Effect variant efficiently models group-level uncertainty, yielding robust ability estimates under sparsity. Experiments show consistent improvements across multiple baselines, particularly in cold-start scenarios.

Acknowledgments

This work was funded by the Central Guidance Fund for Local Science and Technology Development (No. 2023ZYDF078).

References

Bürkner, P.-C. (2021). Bayesian item response modeling in r with brms and stan. *Journal of Statistical Software*, *100*(5), 1–54. https://doi.org/10.18637/jss.v100.i05

Chen, Y., Silva Filho, T., Prudencio, R. B., Diethe, T., & Flach, P. (2019). β^3-irt: A new item response model and its applications. *The 22nd International Conference on Artificial Intelligence and Statistics*, 1013–1021.

De La Torre, J. (2009). Dina model and parameter estimation: A didactic. *Journal of educational and behavioral statistics*, *34*(1), 115–130.

Huang, J., Liu, Q., Wang, F., Huang, Z., Fang, S., Wu, R., Chen, E., Su, Y., & Wang, S. (2021). Group-level cognitive diagnosis: A multi-task learning perspective. *2021 IEEE International Conference on Data Mining (ICDM)*, 210–219.

Jiang, B., Wei, Y., Zhang, T., & Zhang, W. (2024). Improving the performance and explainability of knowledge tracing via markov blanket. *Information Processing & Management*, *61*(3), 103620.

Liu, Q., Wu, R., Chen, E., Xu, G., Su, Y., Chen, Z., & Hu, G. (2018). Fuzzy cognitive diagnosis for modelling examinee performance. *ACM Trans. Intell. Syst. Technol.*, *9*(4).

Liu, S., Yu, X., Ma, H., Wang, Z., Qin, C., & Zhang, X. (2023). Homogeneous cohort-aware group cognitive diagnosis: A multi-grained modeling perspective. *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, 4094–4098.

Lord, F. (1952). A theory of test scores. *Psychometric monographs*.

Wang, F., Huang, Z., Liu, Q., Chen, E., Yin, Y., Ma, J., & Wang, S. (2023). Dynamic cognitive diagnosis: An educational priors-enhanced deep knowledge tracing perspective. *IEEE Transactions on Learning Technologies*, *16*(3), 306–323. https://doi.org/10.1109/TLT.2023.3254544

Wang, F., Liu, Q., Chen, E., Huang, Z., Yin, Y., Wang, S., & Su, Y. (2023). Neuralcd: A general framework for cognitive diagnosis. *IEEE Transactions on Knowledge and Data Engineering*, *35*(8), 8312–8327. https://doi.org/10.1109/TKDE.2022.3201037

Wang, F., et al. (2025). Knowledge is power: Harnessing large language models for enhanced cognitive diagnosis. *Proceedings of the AAAI Conference on Artificial Intelligence*.

Wu, M., Davis, R., Domingue, B., Piech, C., & Goodman, N. (2020). Variational item response theory: Fast, accurate, and expressive. *Proceedings of the 13th International Conference on Educational Data Mining*, 257–268.

Yu, G., Xie, Z., Zhou, G., Zhao, Z., & Huang, J. X. (2025). Exploring long- and short-term knowledge state graph representations with adaptive fusion for knowledge tracing. *Information Processing & Management*, *62*(3), 104074.

Yu, X., Qin, C., Shen, D., Ma, H., Zhang, L., Zhang, X., Zhu, H., & Xiong, H. (2024). Rdgt: Enhancing group cognitive diagnosis with relation-guided dual-side graph transformer. *IEEE Transactions on Knowledge and Data Engineering*, *36*(7), 3429–3442.

Zhang, T., et al. (2024). Path-specific causal reasoning for fairness-aware cognitive diagnosis. *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*.