Unified Uncertainty Estimation for Cognitive Diagnosis Models

Fei Wang
School of Computer Science and Technology, University of Science and Technology of China & State Key Laboratory of Cognitive Intelligence
Hefei, China
wf314159@mail.ustc.edu.cn

Qi Liu
School of Computer Science and Technology, University of Science and Technology of China & State Key Laboratory of Cognitive Intelligence
Hefei, China
qiliuql@ustc.edu.cn

Enhong Chen*
Anhui Province Key Laboratory of Big Data Analysis and Application, University of Science and Technology of China & State Key Laboratory of Cognitive Intelligence
Hefei, China
cheneh@ustc.edu.cn

Chuanren Liu
The University of Tennessee Business Analytics and Statistics
Knoxville, United States
cliu89@utk.edu

Zhenya Huang
School of Computer Science and Technology, University of Science and Technology of China & State Key Laboratory of Cognitive Intelligence
Hefei, China
huangzhy@ustc.edu.cn

Jinze Wu
iFLYTEK Research & State Key Laboratory of Cognitive Intelligence
Hefei, China
hxwjz@mail.ustc.edu.cn

Shijin Wang
iFLYTEK AI Research (Central China) & State Key Laboratory of Cognitive Intelligence
Hefei, China
sjwang3@iflytek.com

ABSTRACT

Cognitive diagnosis models have been widely used in different areas, especially intelligent education, to measure users' proficiency levels on knowledge concepts, based on which users can get personalized instructions. As the measurement is not always reliable due to the weak links of the models and data, the uncertainty of measurement also offers important information for decisions. However, the research on the uncertainty estimation lags behind that on advanced model structures for cognitive diagnosis. Existing approaches have limited efficiency and leave an academic blank for sophisticated models which have interaction function parameters (e.g., deep learning-based models). To address these problems, we propose a unified uncertainty estimation approach for a wide range of cognitive diagnosis models. Specifically, based on the idea of estimating the posterior distributions of cognitive diagnosis model parameters, we first provide a unified objective function for mini-batch based optimization that can be more efficiently applied to a wide range of models and large datasets. Then, we modify the reparameterization approach in order to adapt to parameters defined on different domains. Furthermore, we decompose the uncertainty of diagnostic parameters into data aspect and model aspect, which better explains the source of uncertainty. Extensive experiments demonstrate that our method is effective and can provide useful insights into the uncertainty of cognitive diagnosis.

CCS CONCEPTS

• **Applied computing** → **Education**.

KEYWORDS

Intelligent Education, Cognitive Diagnosis, Uncertainty

ACM Reference Format:
Fei Wang, Qi Liu, Enhong Chen, Chuanren Liu, Zhenya Huang, Jinze Wu, and Shijin Wang. 2024. Unified Uncertainty Estimation for Cognitive Diagnosis Models. In *Proceedings of the ACM Web Conference 2024 (WWW '24), May 13–17, 2024, Singapore, Singapore.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3589334.3645488

1 INTRODUCTION

Cognitive diagnosis is a class of methods that have been widely studied in areas such as education [19], psychometric [28], medical diagnosis [31], and crowdsourcing [21, 36]. The main purpose of cognitive diagnosis is to obtain examinees' cognitive states from their activities. Particularly, in educational area, such as the online learning platforms, cognitive diagnosis obtains students' knowledge proficiencies from the their learning activities (e.g., question

*Corresponding author.

WWW '24, May 13–17, 2024, Singapore, Singapore

ACM ISBN 979-8-4007-0171-9/24/05…$15.00
https://doi.org/10.1145/3589334.3645488

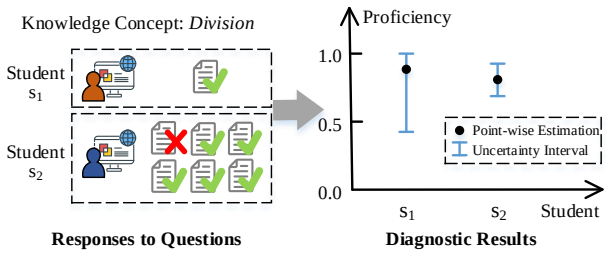


Figure 1: A toy example.

answering), as well as estimates the attributes of questions (e.g., question difficulty). A toy example is illustrated in Figure 1, where two students have answered questions that relate to the knowledge concept "*Division*". After diagnosis, we know that s_1 has mastered "*Division*" well while s_2 has a lower proficiency (black points). Cognitive diagnosis usually serves as the core of intelligent tutoring systems, which provide personalized support for learners.

In practice, however, the diagnostic results of students are not always highly reliable. In the example of Figure 1, although both students s_1 and s_2 are diagnosed to have high proficiency of "*Division*", the diagnostic result of s_1 is not as reliable as s_2. The reason is that s_1's proficiency of "*Division*" is inferred based on a single response related to "*Division*", which may cause severe bias. The uncertainty of diagnosis has important influence on personalized teaching. The system can assign less practice of "*Division*" to s_2; while for s_1, more questions or better cognitive diagnosis models are needed to obtain an exact proficiency assessment. Furthermore, in a recommender system, more diverse learning resources can be recommended to students with higher uncertainty [13]. In computerized adaptive testing, reducing uncertainty of diagnosis is an important target when selecting the next test question for an examinee [2]. However, most existing diagnosis models cannot tell how confident they are with their point-wise diagnosis.

In recent years, more sophisticated model structures have been proposed for better diagnosis, including deep learning-based models such as NeuralCD [33]. However, the research on the uncertainty estimation of cognitive diagnosis remains on several traditional non-deep learning-based models. For example, the Bayesian method is the most representative for item response theory (IRT) based models [9]. The application of existing methods is limited due to the following challenges. 1) Limited application range of training algorithms. The widely accepted training algorithms for existing methods, such as Expectation-Maximization (EM) based algorithms and Metropolis-Hasting (MH) sampling-based algorithms, are inefficient or even inapplicable to complex diagnosis models (e.g., deep learning-based models) having large-scale parameters and on large datasets. 2) Insufficient estimation of parameters. Generally, there are two types of parameters in cognitive diagnosis models, i.e., the diagnostic parameters that represent the features of students and questions, and function parameters that decide the interaction functions among diagnostic parameters. Existing methods only consider diagnostic parameters, because they are proposed based on traditional cognitive diagnosis models, where the interaction functions are fixed without extra parameters. However, in the state-of-the-art deep learning-based models, the interaction functions are modeled with neural networks, where additional uncertainty from neural network parameters should be considered.

Our Work. In this paper, we propose a unified **U**ncertainty estimation approach for **C**ognitive **D**iagnosis models (abbreviated as UCD), which can both be applied to traditional latent trait models and fill the vacancy for deep learning-based models. 1) Based on the idea of learning the posterior distributions of the parameters, we derive a unified objective function for mini-batch-based optimization, which can be applied to both deep and non-deep learning models. 2) We propose a derivative reparameterization approach, which not only facilitates the efficient gradient descending-based training but also conveniently adapts to parameters with different domains of definition. 3) By further consideration of the difference between diagnostic parameters and function parameters, we factorize the uncertainty of diagnostic parameters into data uncertainty and model uncertainty. Through extensive experiments on real-world datasets, we validate the effectiveness of UCD and provide some useful insights into the uncertainty of cognitive diagnosis models. The codes and public data are available at: https://github.com/LegionKing/UCD.

2 RELATED WORK

Cognitive Diagnosis. Existing cognitive diagnosis methods can be generally classified into non-deep learning models and deep learning-based models. Representative non-deep learning cognitive diagnosis models include continuous latent trait models, such as Item Response Theory (IRT) [9] and Multidimensional Item Response Theory (MIRT) [27]; and discrete classification models, such as Deterministic Input Noisy "And" Gate model [6], and Higher-order DINA [7]. By contrast, deep learning-based approaches achieve state-of-the-art and capture attentions in recent year. Wang et al. [32] proposed a NeuralCD framework that introduces neural networks to learn the interaction between students and questions while keeping interpretability. Several extensions based on NeuralCD have been proposed, such as [20, 24, 33, 35].

Uncertainty Quantification. Uncertainty quantification plays a critical role in the process of decision making and optimization in many fields [14, 22]. In cognitive diagnosis, the uncertainty of diagnostic parameters has been studied for traditional models. Fully Bayesian sampling-based methods [26]) and the multiple imputation method [37] characterize the uncertainty of IRT and MIRT by the variations of diagnostic results. Frequentist methods [25, 29] use standard error to reflect the uncertainty. Duck-Mayr et al. [8] proposed a Gaussian process based method for nonparametric IRT models. However, the estimation algorithm could be time consuming, and function parameters are not considered. In deep learning, Bayesian approximation and ensemble learning techniques are two widely-studied types of methods [1] that quantify the uncertainty. Bayesian approximation typically uses a probability distribution to characterize the uncertainty of parameters and model outputs. Representative methods include the Monte Carlo dropout [34], variational inference [30], and Bayesian neural network based models [3]. Ensemble learning approaches [10, 17] train the deep learning model multiple times and then average the model predictions. Although inspiring, these methods have not been applied to CDMs yet. It should be noted that in CDMs, the focus is the diagnostic results (i.e., the estimated parameters) instead of the model predictions,

which is opposite to deep learning models. Moreover, the difference between diagnostic parameters and function parameters are not recognized in existing methods.

3 PRELIMINARY

3.1 Task Overview

In the educational area, cognitive diagnosis is essentially a measurement of students' knowledge states. Through fitting students' response data by cognitive diagnosis models, the estimated values of student-related parameters are the diagnostic results, which represent the students' levels of knowledge mastery. Suppose there are students $S = \{s_1, s_2, \ldots, s_M\}$, questions $E = \{e_1, e_2, \ldots, e_N\}$, and the Q-matrix $Q \in \{0,1\}^{N \times K}$ which indicates the related knowledge concepts (KC) of the questions (i.e., $Q_{jk} = 1$ means that question e_j involves knowledge concept c_k). Then, the cognitive diagnosis task can be formalized as follows.

Problem Definition. The observed data includes students' response logs $R = \{r_{ij}\}$ and the Q-matrix Q, where $r_{ij} \in \{0, 1\}$ denotes the student s_i's response to question e_j (i.e., incorrect or correct). Our goal is to estimate the uncertainty of diagnostic results (e.g., students' proficiencies on knowledge concepts) provided by cognitive diagnosis models. Here, the probability distribution is adopted to depict the uncertainty.

3.2 Representative Cognitive Diagnosis Models

We briefly introduce the basic structure of cognitive diagnosis models (CDMs) and some representative methods. Generally, a CDM contains two parts: (1) the diagnostic parameters (Φ), indicating the proficiency levels of students (α_i) and properties of questions (β_j); (2) the interaction function about student and question parameters which outputs the probability of correctly answering the question, i.e., $p_{ij} = F(\alpha_i, \beta_j, \Omega)$, where Ω denotes the parameters of the interaction function. Figure 2 demonstrates the structures of two representative cognitive diagnosis models, i.e., IRT and NeuralCDM. After training the CDM to fit responses, the estimated diagnostic parameters α_i are diagnostic results.

As a representative traditional model, the IRT estimates the interaction function $p_{ij} = 1/\{1 + e^{-1.7 \times \beta_j^{\text{disc}}(\alpha_i - \beta_j^{\text{diff}})}\}$, where β_j^{disc} and β_j^{diff} indicate the discrimination and difficulty of question e_j respectively ($\beta_j = \{\beta_j^{\text{disc}}, \beta_j^{\text{diff}}\}$), and α_i indicates the ability of student s_i. IRT has been extended to Multidimensional IRT (MIRT) by using multidimensional vectors of student and question traits [27]. There is no extra functional parameters in these CDMs, i.e., $\Omega = \emptyset$.

As for deep learning-based cognitive diagnosis models, Wang et al. proposed a general framework as well as a model called NeuralCDM, where the interaction function is learned from data by neural networks [32]. The formulation is as follows:

$$\boldsymbol{x}_{ij} = \boldsymbol{Q}_j \circ (\boldsymbol{\alpha}_i - \boldsymbol{\beta}_j^{\text{diff}}) \times \beta_j^{\text{disc}}, \tag{1}$$

$$\boldsymbol{f}_1 = \text{Sigmoid}(\boldsymbol{W}_1 \times \boldsymbol{x}_{ij} + \boldsymbol{b}_1), \tag{2}$$

$$\boldsymbol{f}_2 = \text{Sigmoid}(\boldsymbol{W}_2 \times \boldsymbol{f}_1 + \boldsymbol{b}_2), \tag{3}$$

$$p_{ij} = \text{Sigmoid}(\boldsymbol{W}_3 \times \boldsymbol{f}_2 + b_3), \tag{4}$$

where $\boldsymbol{\alpha}_i$ indicates student s_i's proficiency on each knowledge concept; $\boldsymbol{\beta}_j^{\text{diff}}$ indicates the difficulty of each knowledge concept tested by question e_j; β_j^{disc} indicates the discrimination of question e_j; $\boldsymbol{Q}_j$ is the j-th row of Q-matrix. $\Omega = \{\boldsymbol{W}_1, \boldsymbol{W}_2, \boldsymbol{W}_3, \boldsymbol{b}_1, \boldsymbol{b}_2, b_3\}$ are network parameters, where each element in $\boldsymbol{W}_*(* = 1, 2, 3)$ is nonnegative.

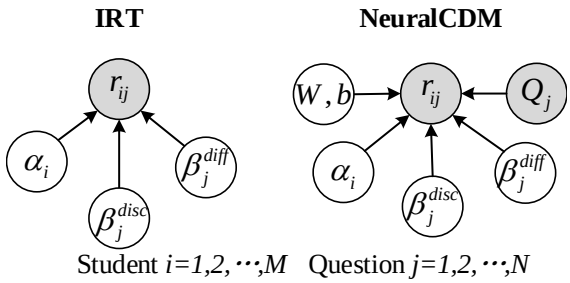


Figure 2: The model structures of IRT and NeuralCDM

4 UNCERTAINTY ESTIMATION FOR COGNITIVE DIAGNOSIS MODELS

We first introduce an overview of our approach. Then, we provide a unified objective function for mini-batch-based training which can be applied to different CDMs on large datasets, and the reparameterization trick that facilitates the gradient computation of different parameter distributions. Finally, we introduce the decomposition of the uncertainty to better estimate the parameters.

4.1 Overview

As most continuous latent trait CDMs and existing deep learning-based CDMs fall under the umbrella of the framework described in 3.2, we choose to make minor modifications to the framework so that our approach can be applied to a wider range of CDMs and avoid impairing the diagnosing ability of the original model structures. Furthermore, in order to obtain the uncertainty of parameters during model training, we change the point-wise estimations of parameters into estimating the posterior distributions. The variance of a posterior distribution directly depicts the uncertainty of the parameter. Uncertainty intervals can also be obtained as an indicator of uncertainty, which is adopted by some studies [5, 11]. Consequently, we propose a unified Bayesian approach called UCD.

For convenience, we treat the parameters as random variables and represent all the variables with $\Psi = \Phi \cup \Omega$, where Φ denotes the diagnostic variables, including student variables $\alpha = \{\alpha_i, i = 1, 2, \ldots, M\}$ and question variables $\beta = \{\beta_j, j = 1, 2, \ldots, N\}$. The overall generative process of the responses $R = \{r_{ij}\}$ modeled by UCD is depicted in Figure 3. To directly estimate the posterior distribution $p(\Psi|R)$ is intractable. Instead, we adopt a practical solution that approximates $p(\Psi|R)$ with a parametric distribution $q(\Psi|\theta)$ which has good statistical properties [3]. Furthermore, by assuming the independence among the variables, the distribution can be factorized to:

$$p(\Psi|R) \simeq q(\Psi|\theta) = q(\Phi|\theta_\Phi) q(\Omega|\theta_\Omega), \tag{5}$$

where θ_Φ and θ_Ω are learnable parameters (notations without circles in Figure 3) that define the distributions of Φ and Ω respectively. Therefore, the goal of model training changes to finding the optimal parameters $\theta = \theta_\Phi \cup \theta_\Omega$ that make $q(\Psi|\theta)$ closest to $p(\Psi|R)$. Along

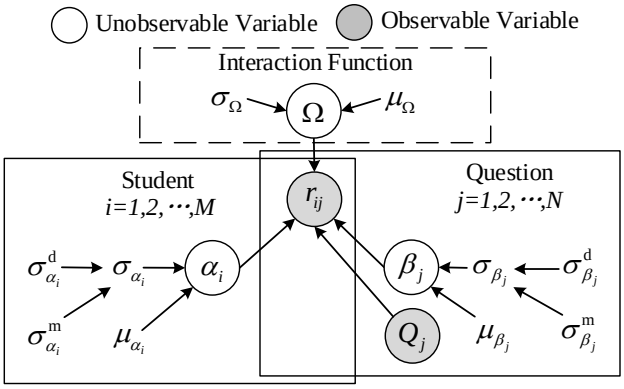


Figure 3: The graphic model of UCD

this way, we introduce the derivation of the objection function in the following subsection.

4.2 Objective Function

In this subsection, we derive the objective function for mini-batch-based optimization, which can be used for different CDMs. Primarily, we choose to minimize the Kullback-Leibler divergence (D_{KL}) [16], which is a widely accepted measurement of the distance between probability distributions. Therefore, the optimal θ^* can be calculated as:

$$\begin{aligned}\theta^* &= \arg\min_\theta D_{KL}[q(\Psi|\theta)||p(\Psi|R)] \\ &= \arg\min_\theta D_{KL}[q(\Psi|\theta)||p(\Psi)] - \mathbb{E}_{q(\Psi|\theta)} \log p(R|\Psi), \quad (6)\end{aligned}$$

where $p(\Psi)$ is the prior distribution of the variables. $D_{KL}[q(\Psi|\theta)||p(\Psi)] - \mathbb{E}_{q(\Psi|\theta)} \log p(R|\Psi)$ is not an ideal objective function yet, as there is the calculation of expectation. Based on the Monte Carlo approach [3], the expectation can be approximated with the average of samplings. In addition, we incorporate the mini-batch-based training strategy in order to facilitate complicated CDMs and large datasets. Specifically, assuming that there are M_b mini-batches, and for each data sample, we draw M_c variable samples from the distribution $q(\Psi|\theta)$. Then for i-th batch, let $F'_i(\theta) = \pi_i L_A - L_B$, where:

$$L_A = D_{KL}[q(\Psi|\theta)||p(\Psi)], \quad L_B = \sum_j \frac{1}{M_c} \sum_{m=1}^{M_C} \log p(R_j|\Psi_{jm}). \quad (7)$$

Here, R_j is the j-th response in the batch, Ψ_{jm} is the m-th sample from $P(\Psi|\theta)$ for R_j, and $\sum_{i=1}^{M_b} \pi_i = 1$. We can adopt $\pi_i = \frac{2^{M_b - i}}{2^{M_b} - 1}$ [3]. Furthermore, we place weights on the KL divergence of diagnostic variables and function variables to adjust their learning rates:

$$L'_A = \zeta_0 D_{KL}[q(\Phi|\theta_\Phi)||p(\Phi)] + \zeta_1 D_{KL}[q(\Omega|\theta_\Omega)||p(\Omega)], \quad (8)$$

where ζ_0 and ζ_1 are hyper-parameters. Finally, the objective function for the i-th mini-batch is:

$$F_i(\theta) = \pi_i L'_A - L_B. \quad (9)$$

Minimizing $F_i(\theta)$ means better approximating the prior distribution (lower L'_A) and higher probability of reconstructing the responses (higher L_B). Although the objective follows the conventional Bayesian methods, we first use it to unify the uncertainty estimation for both traditional latent-trait CDMs and deep-learning-based CDMs, and propose refinements specially designed for CDMs in the following subsections.

4.3 Reparameterization

We adopt the gradient descent algorithm to optimize the parameters, as gradient descent can be applied to both deep learning and non-deep learning models and is more efficient than EM-based or MH sampling-based algorithms in traditional approaches. However, there still exists a problem that, if we directly sample Ψ from the distribution $q(\Psi|\theta)$, the gradient of θ in L_B will not be able to be calculated. Therefore, the reparameterization trick is adopted. To facilitate variables defined on different domains and simplify the sampling process, we propose a theorem derived from the proposition in [3] as follows:

Theorem 4.1. *Suppose there is a function $h(x)$ and its inverse function $g(x)$. Let ϵ be a random variable having a probability density $\epsilon \sim N(0,1)$, and let $\Psi = g(\mu + \sigma\epsilon)$. Then we have $h(\Psi) \sim N(\mu, \sigma^2)$, and for a function $f(\Psi, \theta)$, we have:*

$$\frac{\partial}{\partial\theta}\mathbb{E}_{q(\Psi|\theta)}[f(\Psi,\theta)] = \mathbb{E}_{q(\epsilon)}[\frac{\partial f(\Psi,\theta)}{\partial\Psi}\frac{\partial\Psi}{\theta} + \frac{f(\Psi,\theta)}{\partial\theta}]. \quad (10)$$

The proof is provided in Appendix A. Based on Theorem 4.1, the partial derivative with respect to θ of an expectation can be calculated as the expectation of a partial derivative, and the expectation can be further approximated with MC sampling. If we select a distribution for Ψ that $h(\Psi) \sim N(\mu, \sigma^2)$, here $\theta = \{\mu, \sigma\}$, then an unbiased partial derivative with respect to θ of $\mathbb{E}_{q(\Psi|\theta)} \log p(R|\Psi)$ (in Eq. (6)) can be calculated with the following steps: (1) draw samples of ϵ from $N(0,1)$; (2) let $\Psi = \Psi(\theta, \epsilon) = g(\mu + \sigma\epsilon)$; (3) calculate $\partial\mathbb{E}_{q(\Psi|\theta)} \log p(R|\Psi)/\partial\theta = \partial L_B/\theta$.

According to the domain of definition of the Ψ, different distributions $q(\Psi|\theta)$ can be selected. Using ψ to denote any variable in Ψ, the corresponding distribution can be selected as shown in Table 1. Compared to the original reparameterization, the derived method simplified the implementation through Theorem 4.1 and Table 1, as it reduces the hassle of finding suitable sampling probability distributions for different parameter domains.

With the usage of the above probability distributions for each variable ψ, the corresponding parameters that need to be estimated during training are $\theta_\psi = \{\mu_\psi, \sigma_\psi\}$, where $\psi \in \Psi = \alpha \cup \beta \cup \Omega$. It should be noted that, with the assumption of variable independence, all variables are fully factorized, i.e., the covariance of a multidimensional variable is 0.

4.4 Decomposition of the Uncertainty

A significant difference between diagnostic variables and function variables is that: function variables are affected by all the responses in data, while the diagnostic variables are mainly affected by related responses. For example, in IRT, the distribution of α_i is estimated according to student s_i's responses; in NeuralCDM, the distribution of student s_i's proficiency on knowledge concept c_k (α_{ik}) is estimated according to s_i's responses to questions that involve c_k. Therefore, even if the responses to a student/question are highly consistent (illustrated as s_1 in Figure 1), there still exists relatively high uncertainty if related responses are too few, which we call data uncertainty. Another factor that matters is the characteristics

Table 1: Distributions selected for variables defined on different domains.

Domain of ψ	$h_\psi(x)$	$g_\psi(x)$	Examples
$(-\infty, +\infty)$	x	x	The student ability and question difficulty in IRT and MIRT; the network bias in NeuralCDM. We get $\psi \sim N(\mu, \sigma^2)$.
$(a, +\infty)$	$\ln(x-a)$	$e^x + a$	The discrimination in IRT and MIRT; the weights of neural networks in NeuralCDM. Here, $a = 0$, which means $\psi \sim log - norm(\mu, \sigma^2)$.
(a, b)	$\text{Logit}(\frac{x-a}{b-a})$	$\text{Sigmoid}(x)(b-a)+a$	the student ability, question difficulty and discrimination in NeuralCDM. Here, $a = 0, b = 1$, which means $\psi \sim logit - norm(\mu, \sigma^2)$.

of CDMs themselves, such as the fitting ability and the stability of parameter estimation. Such characteristics can make it difficult to estimate diagnostic diagnostic parameters as definite values, leading to model uncertainty.

To be specific, the distribution parameter σ_ϕ is decomposed into σ_ϕ^m and σ_ϕ^d ($\phi \in \alpha \cup \beta$), where σ_ϕ^m indicates the model uncertainty learned from the CDM, and σ_ϕ^d is monotonically decreasing with the number of related responses. In addition, considering that σ_ϕ^d should be positive and has a diminishing marginal utility when there is sufficiently large number of relevant responses, we formulate it as $\sigma_\phi^d = \lambda_0 e^{-\lambda_1 \tau}$, where τ is the number of responses related to ϕ; λ_0 and λ_1 are learnable weights that adjust the rate of decreasing (two sets of λ_0 and λ_1 can be used for questions and students respectively when the amount of responses related to a question differs too much from that to a student). Then, we use $\sigma_\phi = \sigma_\phi^m \times \sigma_\phi^d$.

The whole graphical model of UCD is illustrated in Figure 3, and the training algorithm is summarized in Appendix B.

4.5 Model Complexity

The space complexity of UCD-integrated CDMs depends on the $q(\Psi|\theta)$ we choose. In our case, although UCD doubles the number of parameters, the space complexity is still O(M + N + U), where M, N and U are the numbers of students, questions and function parameters.

The increase in the number of parameters does not affect the time cost much, as it does not change the gradient descent algorithm (we did not observe appreciably more epochs before convergence in our experiments). The extra time cost mainly comes from the sampling process, especially the sampling of neural network parameters. For example, in Eq. (2), $\boldsymbol{W}_1$ is sampled for each data sample in $\boldsymbol{x}_{ij}$, changing the matrix-matrix multiplication ($\boldsymbol{W}_1 \times \boldsymbol{x}_{ij}$) to multiple matrix-vector multiplications, which is difficult for parallel GPU computing. Nevertheless, this is an acceptable trade-off to obtain the uncertainty, especially in deep learning-based CDMs where traditional uncertainty estimation methods can not be applied.

5 EXPERIMENTS

We conduct comprehensive experiments to answer the following research questions:

- **RQ1** Can UCD provide reasonable uncertainty for different CDMs?
- **RQ2** Whether the captured uncertainty relevant to the decomposed sources?
- **RQ3** Can UCD more efficiently deal with sophisticated CDMs and large datasets?
- **RQ4** What personalized diagnostic information can UCD provide?
- **RQ5** Does UCD avoid impairing the diagnostic ability of the CDMs?

Table 2: The statistics of the datasets.

	FrcSub	Math	Eedi
number of students	536	7,756	17,740
number of questions	20	1,993	8,987
number of KCs	8	305	286
number of responses	10,720	637,798	610,032

5.1 Dataset Description

We use three real-world datasets, i.e., FrcSub, Math and Eedi, in the experiments. FrcSub is a widely used dataset in cognitive diagnosis modeling, which consists of students' responses to fraction-subtraction questions [23]. Math is a dataset collecting the test performances of senior high school students. Eedi is the dataset released by the NeurIPS 2020 education challenge (track 1), containing students' answers to mathematics questions from Eedi[1]. We use a subset of the original data starting from 04/01/2020 to 05/01/2020. Table 2 shows some basic statistics.

5.2 Experimental Setup

To evaluate the effectiveness of our method, we applied UCD to two representative non-deep learning CDMs, i.e., IRT [9] and MIRT [27], and two representative deep learning-based CDMs, i,e, NeuralCDM [32] and KaNCD [33]. In addition, we also compare our UCD with the fully Bayesian sampling-based method [26] (FB) on IRT and MIRT, multiple imputation [37] (MI) on IRT [2], and the nonparametric method GPIRT [8]. As ensemble-based method is the only available baseline that can be directly applied on NeuralCDM and KaNCD, we compare UCD with deep ensemble [17] (DE).

The fully Bayesian sampling-based approach was implemented using PyStan[3] of which the underlying implementation is in C language, and the number of warm-up samples is set to 500; GPIRT is implemented based on the R package provided by the authors[4]; the other approaches were implemented with Pytorch in Python. All experiments were run on a Linux server with Intel Xeon Gold 5218 CPU and Tesla V100 GPU.

The responses of each student in the datasets are divided into train:validate:test = 0.7:0.1:0.2. M_c is set to be 5. ζ_0 and ζ_1 are both

[1] https://competitions.codalab.org/competitions/25449
[2] Frequentist methods are not compared with because they use standard error instead of probability distribution (or uncertainty interval) to represent the uncertainty of student proficiency.
[3] https://pystan.readthedocs.io/
[4] https://github.com/duckmayr/gpirt/blob/main/

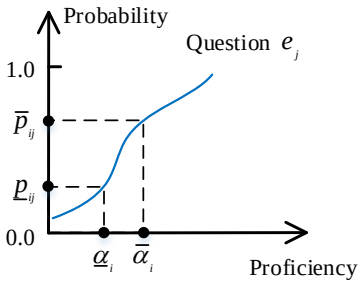


Figure 4: A unidimensional illustration of interval transformation.

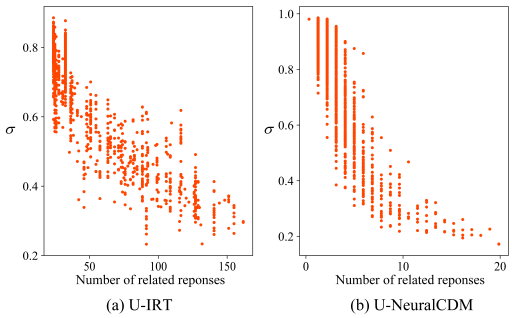


Figure 5: The σ_α of students estimated by U-IRT and U-NCDM.

selected from [0.01, 0.1, 1, 1.5]. For all the standard deviation parameters (σ_*), to ensure that they are positive, we instead make $\sigma_* = \text{Softplus}(\eta_*)$, and learn η_* through training. We select $N(0,1)$, log-norm(0,1) and logit-norm(0,1) as the prior distributions for variables defined on $(-\infty, +\infty)$, $(0, +\infty)$ and $(0, 1)$ respectively. To initialize the network variables, we initialize a matrix W with Kaiming initialization [12], and then let $\mu_W = \ln(|W|)$. The Adam algorithm [15] is used for optimization, and the learning rate is 0.002.

5.3 Evaluation of Uncertainty Intervals (RQ1)

The uncertainty of the diagnostic results (i.e., student variable α) is characterized by their estimated posterior distributions, and can be further concretized with the confidence intervals (uncertainty intervals) of the distributions. To facilitate the evaluation with observable responses, we project the intervals of students' knowledge proficiencies $[\underline{\boldsymbol{\alpha}}_i, \overline{\boldsymbol{\alpha}}_i]$ to the intervals of model predictions $[\underline{p}_{ij}, \overline{p}_{ij}]$. This is achieved by taking advantage of the monotonicity of CDMs. As the monotonicity assumption in CDMs indicates, the model prediction monotonically increases with any dimension of knowledge proficiency $\boldsymbol{\alpha}_i$ [27]. Figure 4 illustrates a unidimensional example, where the curve depicts the predicted probability (that a student can correctly answer the question e_j) with respect to the student's knowledge proficiency. Specifically, we first obtain the 95% confidence interval of the estimated knowledge proficiency $[\underline{\boldsymbol{\alpha}}_i, \overline{\boldsymbol{\alpha}}_i]$, where $\underline{\boldsymbol{\alpha}}_i = g(\boldsymbol{\mu}_{\alpha_i} - 1.96\boldsymbol{\sigma}_{\alpha_i})$ and $\overline{\boldsymbol{\alpha}}_i = g(\boldsymbol{\mu}_{\alpha_i} + 1.96\boldsymbol{\sigma}_{\alpha_i})$]. Here, $g(\cdot)$ is the function discussed in Table 1. Next, we sample question variables ($\boldsymbol{\beta}_j$) and network variables (Ω) 50 times and calculate their corresponding predictions with $\underline{\boldsymbol{\alpha}}_i$ and the corresponding interaction of the CDM. $\underline{p}_{ij} = \mathbb{E}_{q(\boldsymbol{\beta}_j, \Omega | \theta_{\beta_j}, \theta_\Omega)} p(r_{ij} = 1 | \underline{\boldsymbol{\alpha}}_i)$ is finally approximated with the average of these predictions. Similarly, $\overline{p}_{ij}$ can be obtained. DE is exceptional, for which $[\underline{\boldsymbol{\alpha}}_i, \overline{\boldsymbol{\alpha}}_i]$ is directly obtained from the predictions of multiple trained CDM instances.

In order to evaluate whether reasonable uncertainty intervals are obtained, Prediction Interval Coverage Probability (PICP) and Prediction Interval Average Width (PIAW) are widely accepted metrics [1]. PICP calculates the proportion of true values lying in the interval, while PIAW calculates the average widths of the intervals. To adapt to binary response labels (0 or 1) in our experiments, we adjust the formulation as follows:

$$PICP = \frac{1}{n}\sum_{i,j} c_{ij}, \quad PIAW = \frac{1}{n}\sum_{i,j}(\overline{p}_{ij} - \underline{p}_{ij}), \tag{11}$$

where n is the number of responses in the test set, and

$$c_{ij} = \begin{cases} 1, & [0.5r_{ij}, 0.5(1 + r_{ij})] \cap [\underline{p}_{ij}, \overline{p}_{ij}] \neq \emptyset, \\ 0, & \text{otherwise.} \end{cases} \tag{12}$$

Well estimated intervals should have a PICP close to the confidence level, and the same PICP with a smaller PIAW indicates a tighter interval. Furthermore, with a certain confidence level, a CDM having a smaller PIAW usually indicates more confident diagnostic results.

The results of the models are presented in Table 3.[5] We have the following observations. First, UCD achieves PICPs closer to 0.95, which indicates accurate uncertainty estimation. Second, on IRT, MI tends to underestimate the uncertainty. The uncertainty estimated by UCD is consistent with the traditional FB method, and UCD performs better than FB on FrcSub and Math. On MIRT, FB overestimates the uncertainty (the abnormally high PIAW), while UCD provides reasonable results. These validate the effectiveness of UCD on traditional CDMs. On NeuralCDM and KaNCD, UCD gets better results most time. Moreover, the comparability issue among the CDMs trained multiple times (i.e., scale linking [18]) is dismissed in DE.

5.4 Analysis of the Uncertainty Source (RQ2)

As stated in subsection 4.4, the uncertainty of diagnostic variables comes from both data aspect and model aspect. Better insights into the uncertainty source can be useful in applications, such as deciding the number of questions or repeats of knowledge concepts in an examination, and selecting suitable CDMs that have a better balance between diagnosis accuracy and model uncertainty on the data. For better understanding, we visualize the uncertainty parameters σ_α estimated on Math in Figure 5. For brevity, we use the prefix "U-" to identify the CDMs integrated with UCD.

For data aspect, as can be observed in Figure 5, there is a tendency that diagnostic variables with more related responses should have lower uncertainty. To fully validate whether this tendency is captured by UCD, we calculate the Spearman rank correlation coefficient [38] between the σ_α of students and the number of responses related to α in the training set. The results are presented in Table 4. As expected, we can observe strong negative correlations, which validate the tendency. Here we focus on the diagnosed proficiencies of students (α), which is the goal of CDMs, and the same results can be observed for question parameters (β).

[5]The results of GPIRT on Math and Eedi were not obtained because the iteration stop condition is too hard to meet for large datasets, causing unacceptable running time.

Table 3: Experimental results of student performance prediction (uncertainty interval).

Dataset	Metric	IRT				MIRT		NeuralCDM		KaNCD	
		FB	MI	GPIRT	UCD	FB	UCD	DE	UCD	DE	UCD
FrcSub	PICP	0.935 ± .001	0.885 ± .008	0.933 ± .001	**0.957 ± .001**	**1.000 ± .000**	0.922 ± .001	0.868 ± .003	**0.956 ± .003**	0.899 ± .004	**0.918 ± .001**
	PIAW	0.340 ± .001	0.277 ± .007	0.335 ± .001	0.342 ± .003	0.999 ± .000	0.264 ± .012	0.085 ± .012	0.472 ± .007	0.191 ± .008	0.247 ± .005
Math	PICP	0.867 ± .001	0.802 ± .006	-	**0.883 ± .001**	**1.000 ± .000**	0.940 ± .002	0.816 ± .003	**0.927 ± .004**	**0.875 ± .004**	0.837 ± .003
	PIAW	0.159 ± .001	0.269 ± .006	-	0.194 ± .002	0.990 ± .000	0.393 ± .007	0.084 ± .006	0.468 ± .004	0.176 ± .011	0.143 ± .004
Eedi	PICP	**0.898 ± .001**	0.830 ± .007	-	0.892 ± .001	**1.000 ± .000**	0.941 ± .002	0.831 ± .002	**0.946 ± .004**	**0.864 ± .005**	0.827 ± .003
	PIAW	0.266 ± .001	0.226 ± .006	-	0.247 ± .003	0.999 ± .000	0.493 ± .008	0.124 ± .005	0.471 ± .004	0.195 ± .020	0.107 ± .005

Table 4: The Spearman rank correlations between σ_α and the number of related questions. The results of U-IRT and U-MIRT cannot be calculated on FrcSub because all students answer the same number of questions.

Dataset	U-IRT	U-MIRT	U-NeuralCDM	U-KaNCD
FrcSub	-	-	-0.96	-0.92
Math	-0.91	-0.89	-0.94	-0.60
Eedi	-0.91	-0.69	-0.85	-0.42

Table 5: The Spearman rank correlations between σ_α^m and the fitting ability. The results of U-IRT and U-MIRT cannot be calculated on FrcSub because all students answer the same number of questions.

Dataset	U-IRT	U-MIRT	U-NeuralCDM	U-KaNCD
FrcSub	-	-	0.82	0.23
Math	0.73	0.53	0.90	0.05
Eedi	0.60	-0.63	0.99	-0.16

For model aspect, as we can observe from Figure 5, although there is a decreasing tendency with the number of responses, there are variances on a certain number of responses, which are caused by σ_ϕ^m. The estimated σ_ϕ^m has a more complicated relation with the properties of CDMs, which can be difficult to fully analyze. We here provide a viewpoint that we observed in experiments. In general, the distance between model predictions and the true response labels indicates the ability of the model to reconstruct the responses. Therefore, this distance can be an indicator of the model characteristic, which may be relevant to the model uncertainty. Along this way, we calculate the Spearman rank correlation between this distance and the estimated σ_ϕ^m of student variables.

For CDMs diagnosing latent abilities (no corresponding relationship with Q-matrix, e.g., U-IRT, U-MIRT), the overall distance of student s_i is:

$$\text{dist}(s_i) = \sum_{r_{ij} \in R_i} |\hat{p}_{ij} - r_{ij}|, \tag{13}$$

where R_i is the set of responses of s_i in data; $\hat{p}_{ij}$ is expected prediction of input s_i and e_j; r_{ij} is the true response. Then, the Spearman rank correlation between $\{\text{dist}(s_i), i = 1, 2, \ldots, M\}$ and $\{\sigma_{\alpha_i}^m, i = 1, 2, \ldots, M\}$ is calculated.

For CDMs diagnosing explicit knowledge proficiencies (having corresponding relationship with Q-matrix, e.g., U-NeuralCDM, U-KaNCD), the overall distance of student s_i's proficiency on knowledge concept c_k is:

$$\text{dist}(s_i^k) = \sum_{r_{ij} \in R_i^k} |\hat{p}_{ij} - r_{ij}|, \tag{14}$$

where R_i^k is the set of responses of s_i to the questions requiring c_k. Then, the Spearman rank correlation between $\{\text{dist}(s_i^k), i = 1, 2, \ldots, M, k = 1, 2, \ldots, K\}$ and $\{\sigma_{\alpha_{ik}}^m, i = 1, 2, \ldots, M, k = 1, 2, \ldots, K\}$ is calculated. The results are presented in Table 5, where we can observe obvious correlations on most models, which partially explains the differences of model uncertainty (σ_α^m). The relatively weak correlation presented by U-KaNCD should be caused by that KaNCD actually models the associations among knowledge concepts, which is not measured by Eq. (14). It should be noticed that the evaluation here provides a viewpoint to understand σ_α^m. The whole relation between σ_α^m and CDMs can be more complicated.

5.5 Comparison of the Efficiency (RQ3)

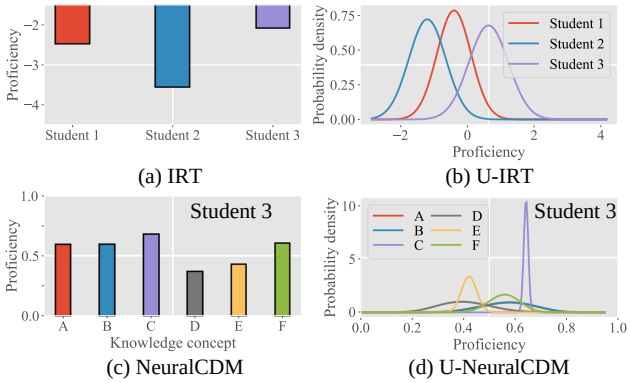
As stated in the Introduction, one of the limitations of traditional uncertainty estimation approaches is the limited application range of training methods, which are inefficient and even inapplicable to complex cognitive diagnosis models (CDMs) and large datasets. Here, we provide the training time costs (until convergence) of the fully Bayesian sampling-based approach, multiple imputation approach, and our UCD in Table 6. We can observe from the table that the model complexity and data size have a significant impact on the time cost of traditional approaches. Specifically, FB-MIRT requires much more time cost than FB-IRT, and their time cost increases dramatically on larger datasets, i.e., Math and Eedi. Similarly, GPIRT requires unacceptable time costs when applied to Math and Eedi. In contrast, the time cost increment of UCD is more moderate. Moreover, UCD can be applied to deep learning-based CDMs (e.g., NeuralCDM and KaNCD) where traditional approaches are not applicable. The ensemble-based uncertainty estimation approaches from deep learning academia are essentially not for CDMs, and the time cost is N times the original CDM, where N is the number of trials for CDM training. Larger N can provide more accurate estimations but leads to higher time costs.

5.6 Illustration of Diagnostic Information (RQ4)

Through integrating UCD, a CDM can provide more information about the diagnostic results. Here we present an example of diagnostic results provided by IRT, U-IRT, NeuralCDM and U-NeuralCDM, in Figure 6. We randomly select three students from FrcSub, and present their responses to three questions in the table, and the diagnostic reports in the subfigures. (For conciseness, we only present part of the responses and diagnositc reports.) From the figure, we can observe that, both IRT and NeuralCDM provides point-wise

	Concept	Student 1	Student 2	Student 3
Question 1	A, B, C	✗	✗	✓
Question 2	C, D, E	✓	✗	✗
Question 3	C, E, F	✗	✗	✓

A: Find a common denominator
B: Column borrow to subtract the second numerator from the first
C: Subtract numerators
D: Convert a whole number to a fraction
E: Separate a whole number from a fraction
F: Borrow from whole number part



(a) IRT (b) U-IRT (c) NeuralCDM (d) U-NeuralCDM

Figure 6: Differences of diagnostic results.

Table 6: Comparison of time cost.

Approach	CDM	FrcSub	Math	Eedi
FB	IRT	15s	1h 21min	2h 10min
	MIRT	3min 15s	>12h	>12h
MI	IRT	16min 30s	>12h	>12h
GPIRT	IRT	3min	-	-
UCD	IRT	90s	7min	9min 20s
	MIRT	1min 58s	7min 45s	19min 31s

proficiencies of students. For U-IRT, similar proficiencies are reported (i.e., Student 2 < Student 1 < Student 3); For U-NeuralCDM, the modes of the contributions are also close to the results of NeuralCDM (e.g., the proficiency on F is around 0.65). What's more, U-IRT and U-NeuralCDM provide the uncertainty of their diagnostic results. For example, in Figure 6(d), U-NeuralCDM is quite confident in C (having the most related responses), but more uncertain on B. Based on the uncertainty information, users (e.g., teachers) can decide whether to assign additional questions for better diagnosis; downstream applications, such as learning materials recommendation, can pay more attention to confident diagnostic results. Reducing uncertainty can also be considered in the next-question-selection process in computerized adaptive testing [2].

5.7 Impact on Diagnostic Ability (RQ5)

In algorithm designing, it is common to encounter situations where it is difficult to simultaneously satisfy different objectives, requiring a trade-off (e.g., accuracy and efficiency in recommender systems). Ideally, when estimating the uncertainty of CDMs, we do not expect negative impacts on the original diagnostic ability of the CDMs. Therefore, UCD is designed with mild modifications of the original CDM structures in order to smoothly conduct the uncertainty estimation. To validate it, we evaluate the diagnostic performances of CDMs before and after integrating UCD. Following [32], we use the diagnosed results to predict students' performances on questions in the test set, and use AUC and accuracy as metrics. The results of different models are presented in Table 7. Fortunately, we did not observe such degradation from our method. Moreover, for non-deep learning-based U-IRT and U-MIRT, there are considerable

Table 7: Experimental results of student performance prediction (point-wise/expectation).

Dataset	FrcSub		Math		Eedi	
Metric	AUC	Acc	AUC	Acc	AUC	Acc
IRT	0.829	0.778	0.809	0.779	0.796	0.758
U-IRT	0.881	0.805	0.815	0.781	0.808	0.766
MIRT	0.877	0.807	0.810	0.774	0.781	0.744
U-MIRT	0.894	0.822	0.822	0.782	0.806	0.764
NeuralCDM	0.894	0.824	0.808	0.772	0.811	0.768
U-NeuralCDM	0.899	0.826	0.806	0.775	0.810	0.765
KaNCD	0.900	0.835	0.824	0.783	0.809	0.765
U-KaNCD	0.903	0.838	0.822	0.783	0.811	0.764

improvements that might benefit from the regularization of prior distributions of diagnostic variables and the gradient descending algorithm.

6 CONCLUSION

In this paper, we proposed a unified solution to the uncertainty estimation of cognitive diagnosis models (UCD). Compared to traditional approaches, UCD follows the Bayesian strategy but provides better efficiency, and more sufficiently models the differences among parameters into the uncertainty from both data and model aspects. Therefore, UCD can not only be applied to traditional non-deep learning latent trait models but also fill the vacancy for deep learning-based models.

In UCD, we introduced a unified objective function and derived a reparameterization approach that can be applied to large-scale diagnosis model parameters defined on different domains. The current solution is based on the independence assumption among model parameters. In future studies, UCD can be further improved by considering the covariance among diagnostic parameters to better fit advanced cognitive diagnosis models (e.g., KaNCD).

ACKNOWLEDGMENTS

This research was supported by grants from the National Key R&D Program of China (2022ZD0117103) and the National Natural Science Foundation of China (62337001, U20A20229, 62106244).

REFERENCES

[1] Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mohammad Ghavamzadeh, Paul Fieguth, Xiaochun Cao, Abbas Khosravi, U Rajendra Acharya, et al. 2021. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion* 76 (2021), 243–297.

[2] Haoyang Bi, Haiping Ma, Zhenya Huang, Yu Yin, Qi Liu, Enhong Chen, Yu Su, and Shijin Wang. 2020. Quality meets diversity: A model-agnostic framework for computerized adaptive testing. In *2020 IEEE International Conference on Data Mining (ICDM)*. IEEE, 42–51.

[3] Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. 2015. Weight uncertainty in neural network. In *International conference on machine learning*. PMLR, 1613–1622.

[4] Léon Bottou. 2010. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*. Springer, 177–186.

[5] Yinghao Chu, Mengying Li, Hugo TC Pedro, and Carlos FM Coimbra. 2015. Real-time prediction intervals for intra-hour DNI forecasts. *Renewable energy* 83 (2015), 234–244.

[6] Jimmy De La Torre. 2009. DINA model and parameter estimation: A didactic. *Journal of educational and behavioral statistics* 34, 1 (2009), 115–130.

[7] Jimmy De La Torre and Jeffrey A Douglas. 2004. Higher-order latent trait models for cognitive diagnosis. *Psychometrika* 69, 3 (2004), 333–353.

[8] JBrandon Duck-Mayr, Roman Garnett, and Jacob Montgomery. 2020. Gpirt: A Gaussian process model for item response theory. In *Conference on uncertainty in artificial intelligence*. PMLR, 520–529.

[9] Susan E Embretson and Steven P Reise. 2013. *Item response theory*. Psychology Press.

[10] Elisabetta Fersini, Enza Messina, and Federico Alberto Pozzi. 2014. Sentiment analysis: Bayesian ensemble learning. *Decision support systems* 68 (2014), 26–38.

[11] Ethan Goan and Clinton Fookes. 2020. Bayesian neural networks: An introduction and survey. *Case Studies in Applied Bayesian Data Science: CIRM Jean-Morlet Chair, Fall 2018* (2020), 45–87.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*. 1026–1034.

[13] Junyang Jiang, Deqing Yang, Yanghua Xiao, and Chenlu Shen. 2019. Convolutional Gaussian Embeddings for Personalized Recommendation with Uncertainty. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence* (Macao, China) *(IJCAI'19)*. AAAI Press, 2642–2648.

[14] HM Dipu Kabir, Abbas Khosravi, Mohammad Anwar Hosen, and Saeid Nahavandi. 2018. Neural network-based uncertainty quantification: A survey of methodologies and applications. *IEEE access* 6 (2018), 36218–36234.

[15] Diederik P Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *International Conference on Learning Representations*.

[16] Solomon Kullback and Richard A Leibler. 1951. On information and sufficiency. *The annals of mathematical statistics* 22, 1 (1951), 79–86.

[17] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems* 30 (2017).

[18] Won-Chan Lee and Guemin Lee. 2018. IRT linking and equating. *The Wiley handbook of psychometric testing: A multidisciplinary reference on survey, scale and test development* (2018), 639–673.

[19] Jacqueline Leighton and Mark Gierl. 2007. *Cognitive diagnostic assessment for education: Theory and applications*. Cambridge University Press.

[20] Jiatong Li, Fei Wang, Qi Liu, Mengxiao Zhu, Wei Huang, Zhenya Huang, Enhong Chen, Yu Su, and Shijin Wang. 2022. HierCDF: A Bayesian Network-based Hierarchical Cognitive Diagnosis Framework. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 904–913.

[21] Jiaran Li, Richong Zhang, Samuel Mensah, Wenyi Qin, and Chunming Hu. 2023. Classification-oriented dawid skene model for transferring intelligence from crowds to machines. *Frontiers of Computer Science* 17, 5 (2023), 175332.

[22] I Lira and D Grientschnig. 2010. Bayesian assessment of uncertainty in metrology: a tutorial. *Metrologia* 47, 3 (2010), R1.

[23] Qi Liu, Runze Wu, Enhong Chen, Guandong Xu, Yu Su, Zhigang Chen, and Guoping Hu. 2018. Fuzzy cognitive diagnosis for modelling examinee performance. *ACM Transactions on Intelligent Systems and Technology (TIST)* 9, 4 (2018), 1–26.

[24] Haiping Ma, Manwei Li, Le Wu, Haifeng Zhang, Yunbo Cao, Xingyi Zhang, and Xuemin Zhao. 2022. Knowledge-Sensed Cognitive Diagnosis for Intelligent Education Platforms. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*. 1451–1460.

[25] Jeffrey M Patton, Ying Cheng, Ke-Hai Yuan, and Qi Diao. 2014. Bootstrap standard errors for maximum likelihood ability estimates when item parameters are unknown. *Educational and Psychological Measurement* 74, 4 (2014), 697–712.

[26] Richard J Patz and Brian W Junker. 1999. A straightforward approach to Markov chain Monte Carlo methods for item response models. *Journal of educational and behavioral Statistics* 24, 2 (1999), 146–178.

[27] Mark D Reckase. 2009. Multidimensional item response theory models. In *Multidimensional Item Response Theory*. Springer, 79–112.

[28] Christine A Reid, Stephanie A Kolakowsky-Hayner, Allen N Lewis, and Amy J Armstrong. 2007. Modern psychometric methodology: Applications of item response theory. *Rehabilitation Counseling Bulletin* 50, 3 (2007), 177–188.

[29] Jason D Rights, Sonya K Sterba, Sun-Joo Cho, and Kristopher J Preacher. 2018. Addressing model uncertainty in item response theory person scores through model averaging. *Behaviormetrika* 45, 2 (2018), 495–503.

[30] Jakub Swiatkowski, Kevin Roth, Bastiaan Veeling, Linh Tran, Joshua Dillon, Jasper Snoek, Stephan Mandt, Tim Salimans, Rodolphe Jenatton, and Sebastian Nowozin. 2020. The k-tied normal distribution: A compact parameterization of Gaussian mean field posteriors in Bayesian neural networks. In *International Conference on Machine Learning*. PMLR, 9289–9299.

[31] Michael L Thomas. 2011. The value of item response theory in clinical assessment: a review. *Assessment* 18, 3 (2011), 291–307.

[32] Fei Wang, Qi Liu, Enhong Chen, Zhenya Huang, Yuying Chen, Yu Yin, Zai Huang, and Shijin Wang. 2020. Neural cognitive diagnosis for intelligent education systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 6153–6161.

[33] Fei Wang, Qi Liu, Enhong Chen, Zhenya Huang, Yu Yin, Shijin Wang, and Yu Su. 2022. NeuralCD: A General Framework for Cognitive Diagnosis. *IEEE Transactions on Knowledge and Data Engineering* (2022).

[34] Guotai Wang, Wenqi Li, Michael Aertsen, Jan Deprest, Sébastien Ourselin, and Tom Vercauteren. 2019. Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. *Neurocomputing* 338 (2019), 34–45.

[35] Xinping Wang, Caidie Huang, Jinfang Cai, and Liangyu Chen. 2021. Using Knowledge Concept Aggregation towards Accurate Cognitive Diagnosis. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. 2010–2019.

[36] Jacob Whitehill, Ting-fan Wu, Jacob Bergsma, Javier Movellan, and Paul Ruvolo. 2009. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. *Advances in neural information processing systems* 22 (2009).

[37] Ji Seung Yang, Mark Hansen, and Li Cai. 2012. Characterizing sources of uncertainty in item response theory scale scores. *Educational and psychological measurement* 72, 2 (2012), 264–290.

[38] Jerrold H Zar. 2005. Spearman rank correlation. *Encyclopedia of biostatistics* 7 (2005).

A PROOF OF THEOREM 4.1

Proof. As $g(x)$ is the inverse function of $h(x)$, it is easy to get $h(\Psi) = (\mu + \sigma\epsilon) \sim N(\mu, \sigma^2)$.

Then, we prove that $q(\Psi|\theta)\mathrm{d}\Psi = q(\epsilon)\mathrm{d}\epsilon$.

$$\begin{aligned} q(\Psi|\theta)\mathrm{d}\Psi &= h'(\Psi)f(h(\Psi))\mathrm{d}\Psi \\ &= h'(\Psi)f(h(\Psi))\mathrm{d}g(h(\Psi)) \\ &= h'(\Psi)f(h(\Psi))g'(h(\Psi))\mathrm{d}h(\Psi) \\ &= f(\mu+\sigma\epsilon)\mathrm{d}(\mu+\sigma\epsilon) \\ &= \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(\mu+\sigma\epsilon-\mu)^2}{2\sigma^2}} \cdot \sigma\mathrm{d}\epsilon \\ &= \frac{1}{\sqrt{2\pi}}e^{-\frac{\epsilon^2}{2}}\mathrm{d}\epsilon \\ &= q(\epsilon)\mathrm{d}\epsilon. \end{aligned}$$

Therefore, we have:

$$\begin{aligned} \frac{\partial}{\partial\theta}\mathbb{E}_{q(\Psi|\theta)}[f(\Psi,\theta)] &= \frac{\partial}{\partial\theta}\int f(\Psi,\theta)q(\Psi|\theta)\mathrm{d}\Psi \\ &= \frac{\partial}{\partial\theta}\int f(\Psi,\theta)q(\epsilon)\mathrm{d}\epsilon \\ &= \mathbb{E}_{q(\epsilon)}[\frac{\partial f(\Psi,\theta)}{\partial\Psi}\frac{\partial\Psi}{\theta} + \frac{f(\Psi,\theta)}{\partial\theta}]. \end{aligned}$$

□

B UCD TRAINING

The overall training algorithm of UCD is summarized as follows, where l is the learning rate. Existing gradient descent algorithms,

such as SGD [4] and Adam [15], can be adopted to update the parameters (line 11-12). The training process does not make much difference with the original cognitive diagnosis, except that it introduces the sampling of variables and the modification of objective function.

As described in the experimental setup (5.2), we sample 5 samples for each variables in a mini-batch, i.e., M_c=5, which is enough for our experiments. Larger M_c will result in more computation. For both ζ_0 and ζ_1, we conducted grid search in the range of [0.01, 0.1, 1, 1.5]. 10% of each student's response data was used as the validation set. The best epoch for each choice of hyperparameters was chosen by the highest AUC and Acc on the validation set, because these are the basic metrics for cognitive diagnosis itself. Furthermore, the best choice of hyperparameters is decided by a high PICP (in our case, close to 0.95) with relatively small PIAW.

Algorithm 1 UCD training algorithm

Input: Responses R; Q-matrix Q
Parameter: Parameters of the approximated posterior distributions, i.e., $\theta_\Phi = \{\mu_\Phi, \sigma_\Phi^m, \lambda_0, \lambda_1\}, \theta_\Omega = \{\mu_\Omega, \sigma_\Omega\}$
Output: Approximated posterior distributions of diagnostic variables $q(\Phi|\theta_\Phi)$

1: **while** not converged **do**
2: **for** batch i in R **do**
3: **for** variable ψ in $\Phi \cup \Omega$ **do**
4: Draw M_c samples of ϵ from $N(0,1)$
5: **if** ψ is a diagnostic variable in Φ **then**
6: $\sigma_\psi^d = \lambda_0 e^{-\lambda_1 \boldsymbol{\tau}}$, $\sigma_\psi = \sigma_\psi^d \times \sigma_\psi^m$
7: **end if**
8: Let $\psi = g_\psi(\mu_\psi + \sigma_\psi \epsilon)$ (Table 1)
9: **end for**
10: Calculate the loss $F_i(\theta) = \pi_i L'_A - L_B$, where $L'_A = \zeta_0 D_{KL}[q(\Phi|\theta_\Phi)||p(\Phi)] + \zeta_1 D_{KL}[q(\Omega|\theta_\Omega)||p(\Omega)]$, $L_B = \sum_j \frac{1}{M_c} \sum_{m=1}^{M_C} \log p(R_j|\Psi_{jm})$. Eq. ((7)-(9))
11: **for** $\theta \in \theta_\Phi \cup \theta_\Omega$ **do**
12: $\theta \leftarrow \theta - l \nabla_\theta F_i(\theta)$
13: **end for**
14: **end for**
15: **end while**
16: **return** $q(\Phi|\theta_\Phi)$

C SUPPLEMENT OF RQ4

Here we provide more illustrations of the diagnostic information as the supplement of RQ4. Figure 7 lists the full responses of the randomly selected students in 5.6. Figure 8 illustrates the full diagnostic information of Student 3 provided by the original CDMs and the UCD models. The results provided by IRT and U-IRT remain the same with Figure 6 and thus are not presented again. We can observe that the variances of the distributions provided by U-NeuralCDM and U-KaNCD have similar relation with the number of responses relevant to the knowledge concepts. Strictly speaking, the estimation of MIRT does not involve knowledge concepts, therefore the diagnostic results θ is not associated with the predefined knowledge concepts. This makes U-MIRT produce similar uncertainty on different dimensions of θ.

	Correctly answered questions	Incorrectly answered questions
Student 1	2, 7, 8, 11, 15, 16	1, 3, 5, 9, 12, 17, 18, 20
Student 2	4, 6	1, 2, 3, 5, 9, 10, 13, 14, 15, 16, 18, 20
Student 3	1, 3, 4, 7, 8, 10, 11, 12, 16, 20	2, 9, 13, 17

(a) Responses of the three students

Knowledge concept	Knowledge concept name	Related question IDs
A	Find a common denominator	1, 7, 11, 5, 13
B	Column borrow to subtract the second numerator from the first	1, 18
C	Subtract numerators	1, 2, 3, 4, 5, 6, 7, 8, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20
D	Convert a whole number to a fraction	2, 15, 19
E	Separate a whole number from a fraction	4, 5, 2, 9, 10, 3, 13, 14, 16, 17, 18, 19, 20
F	Borrow from whole number part	4, 10, 3, 13, 17, 18, 19, 20
G	Simplify before subtracting	4, 19, 20
H	Reduce answers to simplest form	5, 10, 12

(b) Knowledge concepts and related questions

Figure 7: All responses of the randomly selected students, the knowledge concepts and questions.

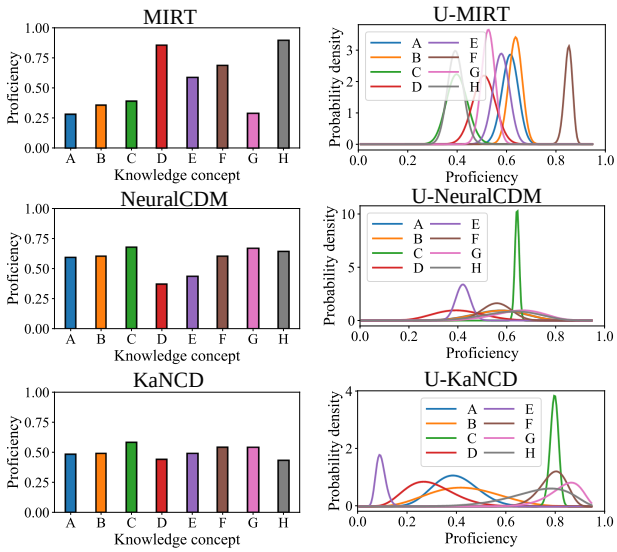


Figure 8: Diagnostic results of Student 3's proficiency on all knowledge concepts.