A Simple Yet Effective Interpretable Bayesian Personalized Ranking for Cognitive Diagnosis

Arthur Batel[a,*], **Idir Benouaret**[b], **Joan Fruitet**[c], **Marc Plantevit**[b] **and Celine Robardet**[a]

[a]INSA Lyon, CNRS, LIRIS UMR 5205, F-69621 Villeurbanne, France
[b]EPITA Research Laboratory (LRE), FR-94276, Le Kremlin-Bicêtre, France
[c]Humans Matter, FR-69002 Lyon, France

Abstract. In the field of education, the automatic assessment of student profiles has become a crucial objective, driven by the rapid expansion of online tutoring systems and computerized adaptive testing. These technologies aim to democratize education and enhance student assessment by providing detailed insights into student profiles, which are essential for accurately predicting the outcomes of exercises, such as solving various types of mathematical equations. We aim to develop a model capable of predicting responses to a large set of questions within the Multi-Target Prediction framework while ensuring that this model is explainable, allowing us to quantify student performance in specific knowledge areas. Existing cognitive diagnosis algorithms often struggle to meet the dual requirement of accurately predicting exercise outcomes and maintaining interpretability. To address this challenge, we propose an alternative to the complexity of current advanced machine learning models. Instead, we introduce a direct yet highly effective Bayesian Personalized Ranking algorithm, called CD-BPR, which incorporates interpretability as a core learning objective. Extensive experiments demonstrate that CD-BPR not only performs better in predicting exercise outcomes but also provides superior interpretability of estimated student profiles, thus fulfilling both key requirements.

1 Introduction

In computer-aided education, the accurate assessment of student proficiency is a main purpose, particularly with the advent of online tutoring systems and computerized adaptive testing. These technological advancements aim to democratize access to education and enhance the assessment process, necessitating a deep understanding of individual student capabilities to predict their performance on various educational tasks. Cognitive Diagnosis Models (CDMs) play a pivotal role in this landscape by providing estimates of student proficiency levels. CDMs are statistical frameworks used to assess and diagnose individuals' strengths and weaknesses in specific cognitive skills or knowledge domains. Unlike traditional testing models that primarily focus on providing a single global score, CDMs aim to offer a detailed profile of an individual's mastery of various attributes or skills. The motivation behind developing CDMs stems from the need for more nuanced diagnostic information that can inform targeted interventions and personalized instruction. By identifying specific areas where an individual may struggle or excel, educators, psychologists, and other professionals can tailor their approaches to better support learning and development.

Recent advancements in CDMs have significantly enhanced our understanding of individual cognitive processes across various domains, including psychiatric and psychological [13, 11, 4], and educational assessment [19, 14, 15]. However, existing models often face challenges in simultaneously achieving high predictive accuracy in such a multi-target prediction framework [16] and maintaining interpretability. While models like NCDM [17] harness neural networks to enhance the interaction between users and questions, they encounter challenges in discerning the cognitive processes that underlie individuals' responses to test items. This opacity contrasts sharply with the requirement for interpretability in educational contexts, where comprehending specific cognitive traits and processes is essential for improving the learning process and capturing strengths and weaknesses of learners. Consequently, there is an increasing acknowledgment of the necessity for interpretable models that shed light on the nuanced cognitive dimensions implicated in education, surpassing the predictions provided by current state-of-the-art methods. On the other hand, the attempt of the IRR [14] method to improve interpretability do not achieves fully satisfying results, but also suffer from a backlash on prediction performances, and come at high computational cost.

In response to this challenge, this article introduces CD-BPR, a Bayesian Personalized Ranking model designed specifically for cognitive diagnosis tasks. Traditional models often fall short by being either overly simplistic and underperforming or overly complex without achieving satisfactory results. In contrast, CD-BPR strikes a balance by embracing simplicity while effectively capturing the nuances of student profiles. By leveraging a Bayesian framework, CD-BPR not only achieves superior predictive performance but also enhances the interpretability of estimated student proficiency by explicitly incorporating the interpretability as a learning objective in a more direct way than what has been done in IRR. CD-BPR learns a representative vector to assess users' skills across multiple dimensions. This vector predicts user performance, ranks potential answers, and provides a self-explanatory skill profile. By adhering to the monotonicity assumption ensuring that the likelihood of correct responses consistently increases with the user's proficiency in the relevant dimension, CD-BPR enables personalized learning and targeted skill enhancement.

Our contributions are summarized as follows: (1) We define a novel, simple yet accurate model built upon the Bayesian Person-

* Corresponding Author. Email: arthur.batel@insa-lyon.fr

alized Ranking (BPR) framework, incorporating interpretability in a direct way to identify influential factors in educational assessment. Unlike traditional black-box models, our approach facilitates transparent decision-making, providing insights into the factors driving each prediction. (2) We report an extensive empirical study conducted on several Education datasets against numerous state-of-the-art models. Results indicate that CD-BPR outperforms other models on both classification metrics and interpretability. (3) For reproducibility and open science purpose, the source code and the experiments are made available on a public repository[1].

The remainder of the paper is structured as follows: Section 2 provides a summary of related work and positions our contribution. Section 3 elaborates on our model. Our extensive experiments are detailed in Section 4. Finally, we conclude in Section 5.

2 Related work

Cognitive Diagnosis Models (CDMs) are a class of psychometric models designed to provide detailed information about an individual's specific knowledge and skills. One of the earliest and most influential CDM is DINA model [3] that operates under the assumption that each item measures a combination of specific skills and that a correct response to an item requires mastery of all the necessary skills. If any of the required knowledge's domains are not mastered, the probability of a correct response is reduced. The model incorporates a deterministic component that specifies the relationship between skills and item responses, as well as a probabilistic component that accounts for the possibility of guessing and slipping (i.e., random errors). By analyzing students' response patterns in relation to skill requirements, the DINA model provides insights into individual students' strengths and weaknesses, allowing for targeted interventions and instructional support.

Multidimensional Item Response Theory (MIRT) [7] is a statistical framework that allows for the estimation of multiple latent traits simultaneously, enabling the assessment of complex constructs involving multiple dimensions or sub-skills. MIRT models provide insights into how individuals' performances on assessment items are influenced by various latent traits, allowing for a more nuanced understanding of their strengths and weaknesses across different skill domains. By capturing interactions between latent traits and item characteristics, MIRT facilitates the development of more accurate and informative assessments, leading to improved educational practices and student outcomes.

Matrix-factorization-based Cognitive Diagnosis (MCD) [9] leverages matrix factorization techniques to jointly model the interactions between students and items in a low-dimensional latent space. By decomposing the observed item-response matrix into student and item latent representations, MCD uncovers the underlying structure of students' mastery profiles and item characteristics. This approach allows MCD to capture complex patterns of skill mastery and item difficulty, providing rich insights into students' cognitive strengths and weaknesses.

Recently, deep neural network models have demonstrated state-of-the-art results in cognitive diagnosis models (CDMs). The Neural Cognitive Diagnosis Model (NCDM) [17] uses neural networks to model and predict students' cognitive proficiency. Unlike traditional CDMs, NCDM learns complex patterns and relationships from raw assessment data, extracting high-level representations of students' cognitive abilities and item characteristics. NCDM can also integrate additional sources of information, such as temporal dynamics or sequential dependencies in student responses. Through its data-driven approach and adaptability to diverse assessment contexts, NCDM offers promising avenues for advancing cognitive diagnosis research and enhancing educational assessment practices.

While efforts have been made to enhance the accuracy of models for predicting item responses and estimating users' cognitive traits, many recently developed models operate as "black boxes," offering limited interpretability. This lack of interpretability raises concerns when using CDMs. In [14], the authors propose a first attempt to deal with this problem. They extend previous work by introducing the Item Response Ranking (IRR) framework for cognitive diagnosis. They propose to use a loss function that measures how well a CDM model can correctly rank pairs of users based on their responses using a pairwise objective function to better ensure the monotonicity property. The experimentation shows a limited improvement of the interpretability, but also a backlash on response prediction performance.

Our work aims at further improving interpretability while mainting top response prediction performances by providing a framework based on Bayesian Personalized Ranking (BPR) that incorporates interpretability in a more direct way.

3 Cognitive Diagnosis Based on Bayesian Personalized Ranking

Our approach aims to assess and diagnose individuals' cognitive abilities and skills within specific cognitive domains by developing a new CDM. To formalize our problem, let U be a set of users and E a set of test questions. Each question e has a binary answer outcome, either a correct or a wrong answer. The response logs R consist of a set of triplets (u, e, y), where $u \in U$, $e \in E$, and $y \in \{0, 1\}$. We denote by $Y_{ue} \equiv y$ for $(u, e, y) \in R$, the answer outcome by user u to question e. Each test question corresponds to one or more cognitive concepts or dimensions evaluated by the test, which are designated by the set K of *knowledge concepts*. The Q-matrix [12] indicates the concepts associated with a question: $Q_{ek} = 1$ iff the question e relates to the concept k.

Our objective is to approximate users' skills across different knowledge concepts by learning a vector $\mathbf{H}_u$ for each user u, where the vector has a length of $\#K$, corresponding to the number of knowledge concepts. This vector must not only estimate a user's performance on unseen questions, facilitating the ranking of potential answers (see Section 3.1), but also be self-interpretable, reflecting the user's proficiency in various knowledge concepts (see Section 3.2). Each component of the vector assesses the user's ability in the corresponding knowledge concept. It must adhere to the monotonicity assumption [18], which posits that the probability of correct responses should increase monotonically with the user's vector component related to the knowledge concept of the test question. Thus, a user who excels in questions associated with a concept k will have a higher value $\mathbf{H}_u[k]$ for that concept.

3.1 Learning embeddings based on pairwise ranking distances

Our proposed CDM is founded upon the Bayesian Personalized Ranking (BPR) framework [8], which is specifically designed for pairwise ranking tasks. In such tasks, the goal is to establish a ranking among user responses rather than predicting precise values. This framework is particularly well-suited for our application because

[1] The code is available at the following link: https://github.com/arthur-batel/cd_bpr_ecai

question responses inherently possess an order, and our focus is on discerning the relative performance among students. Moreover, the BPR model accommodates scenarios where students respond to only a subset of questions, making it highly suitable for estimating cognitive profiles from partially answered questions.

BPR, rooted in matrix factorization techniques, excels at capturing latent factors within data. It achieves this by decomposing the user-item interaction matrix, commonly used in recommender systems, into low-rank matrices representing users and items within a latent space. This work extends the application of BPR beyond binary interactions, by introducing embeddings for each question-answer pair, enabling the capture of more complex relationships. This augmentation allows the model to uncover hidden patterns and similarities among users, questions, and answers. BPR's focus on learning the ranking of question answers aligns well with the goals of Cognitive Diagnostic models, where understanding user abilities on questions is essential. The adaptability of BPR to large datasets further enhances its suitability for this task.

In addition to the embedding vector $\mathbf{H}_u$ associated to each user u, we associate an embedding vector $\mathbf{H}_{(e,y)}$ with each question-answer pair. These two types of embeddings are of same dimension (#K) and make possible to compare diverse elements into a shared latent space. By employing the Euclidean metric within this space, effective generalization of relationships between users and questions/answers is achieved. Using Euclidean distances in our model ensures the preservation of the triangle inequality, which maintains consistent and interpretable relationships between users and questions/answers. This geometric property supports a robust relative representation, ensuring that the proximity in the embedding space accurately reflects users' performance and similarities with questions [2].

The probability of predicting an answer to a test question for a user is proportional to the opposite of the squared Euclidean distance between their embeddings:

$$\widehat{P_{u,e,y}} \propto -||\mathbf{H}_u - \mathbf{H}_{(e,y)}||^2$$

To increase the probability associated with a given triplet (u, e, y) of the response log R (the answer y of user u to question e), the model is trained to prioritize the user's answer y for question e over the opposite answer $\overline{y}$. This is expressed as maximizing the posterior probability of the embedding vector $\mathbf{H}$ given the order on the question/answer pairs: $P(\mathbf{H} \mid \widehat{P_{u,e,y}} > \widehat{P_{u,e,\overline{y}}})$. Using Bayes' rule, this probability is proportional to $P(\widehat{P_{u,e,y}} > \widehat{P_{u,e,\overline{y}}} \mid \mathbf{H})P(\mathbf{H})$. The model aims to determine the parameters $\mathbf{H}$ that most effectively maximize the likelihood of accurately predicting the user's answer, favoring a higher probability for answer y compared to $\overline{y}$. By abusively assuming independence of questions, answers, and users, BPR estimates the model parameters using the maximum a posteriori probability (MAP):

$$\arg\max_{\mathbf{H}} = \log \prod_{(u,e,y)\in R} \prod_{\overline{y}\neq y} P(\widehat{P_{u,e,y}} > \widehat{P_{u,e,\overline{y}}} \mid \mathbf{H})P(\mathbf{H})$$

The probability $P(\widehat{P_{u,e,y}} > \widehat{P_{u,e,\overline{y}}} \mid \mathbf{H})$ is approximated by $\sigma(\widehat{P_{u,e,y}} - \widehat{P_{u,e,\overline{y}}})$, where $\sigma(z) = \frac{1}{1+e^{-z}}$ is the logistic sigmoid function, which approaches 1 when $\widehat{P_{u,e,y}}$ is greater than $\widehat{P_{u,e,\overline{y}}}$. The comprehensive learning objective involves minimizing the loss:

$$\mathcal{L}_1(\mathbf{H}) = -\sum_{(u,e,y)\in R} \sum_{\overline{y}\neq y} \log\left(\sigma(\widehat{P_{u,e,y}} - \widehat{P_{u,e,\overline{y}}})\right) + \lambda||\mathbf{H}||^2$$

where λ is a regularization hyperparameter.

3.2 Self-interpretable embedding

A crucial consideration in developing our model is to enhance the model's confidence by incorporating self-interpretable embeddings that align with the expectations of the model's users. The overarching goal is for the embeddings to effectively capture and represent the inherent skills of the users. This implies that if one user consistently outperforms another, their corresponding value in the embedding space should be higher, reflecting their superior aptitude.

To achieve this, we employ two key strategies aimed at enforcing self-interpretability within the model. The first strategy centers around embedding initialization, ensuring that the initial values assigned to the embeddings are conducive to meaningful interpretations. This involves a thoughtful setup that considers the inherent skills and performance levels of users, providing a solid foundation for subsequent model training. The second strategy focuses on embedding updating during the model's training process. As the model learns from data and refines its understanding, the embeddings are dynamically adjusted to better align with the evolving user performance patterns. This continuous updating mechanism contributes to the model's adaptability and ensures that the embeddings accurately represent the users' skills over time. Both strategies are designed to enhance the alignment of embeddings with user performance across dimensions while preserving prediction accuracy (refer to the ablation study in Section 4.3).

3.2.1 Self-interpreting embedding initialization

We propose a simple yet effective initialization procedure that begins with a random initialization of $\mathbf{H}_u$ and $\mathbf{H}_{(e,y)}$ using a uniform distribution over the interval $[0, 1)$. Then, the embedding value at dimension k is adjusted based on the number of logs related to k with answers higher or lower than 0.5:

$$\begin{aligned}
&\forall (u, e, y) \in R \text{ s.t. } Q_{ek} = 1, \text{ and } (x = u \text{ or } x = (e, y)),\\
&\mathbf{H}_x[k] \leftarrow \mathbf{H}_x[k] + 1, \text{ if } y > 0.5\\
&\mathbf{H}_x[k] \leftarrow \mathbf{H}_x[k] - 1, \text{ if } y < 0.5,
\end{aligned}$$

Following the embedding initialization, the obtained vectors undergo normalization to ensure consistency and stability. The initial step of random uniform initialization plays a crucial role in preventing gradient vanishing during the learning process.

3.2.2 Embedding learning preserving self-interpretation

To enforce the monotonicity assumption during the learning process, we introduce an additional loss function that penalizes deviations from the desired order of user proficiency across different dimensions associated with a question. Specifically, when considering a log triplet (u, e, y) and another log triplet $(\overline{u}, e, y')$ where $y' \leq y$, indicating that user $\overline{u}$ exhibited less proficiency for the same question, our objective is to ensure that, for all dimensions k associated with question e, the value of $\mathbf{H}_u[k]$ is greater than $\mathbf{H}_{\overline{u}}[k]$. This regularization term promotes the desired monotonic relationship, steering the model toward learning embeddings that align with the desired interpretability, emphasizing the importance of dimension-wise performance distinctions between users. This leads to the following loss function:

$$\mathcal{L}_2(\mathbf{H}) = -\sum_{(u,e,y)\in R} \sum_{\substack{(\overline{u},e,y')\in R \\ y'\leq y \\ Q_{ek}=1}} \log\left(\sigma(\mathbf{H}_u[k] - \mathbf{H}_{\overline{u}}[k])\right)$$

Algorithm 1: CD-BPR: Cognitive Diagnostic Bayesian Personalized Ranking

Data: Response logs R, matrix Q , learning rate α, number of epochs max_epochs.
Result: $\mathbf{H}_u$, $\mathbf{H}_{(e,y)}$ of size $\#K$.
Initialize user and item embeddings (see Section 3.2.1);
/* Compute quadruplets used in the learning process*/;
for $(u, e, y) \in R$ **do**
  for k *such that* $Q_{ek} = 1$ **do**
    Take the opposite value $\overline{y}$;
    Take $\overline{u}$ such that $\exists(\overline{u}, e, y') \in R$ with $y' \leq y$, or u if $\overline{u}$ does not exist;
    $B[k] \leftarrow B[k] \bigcup [u, (e, y), (e, \overline{y}), \overline{u})]$;
/* Learning process*/;
for *epoch* $= 1$ **to** *max_epochs* **do**
  for $k = 1$ **to** $\#K$ **do**
    for $(u, (e, y), (e, \overline{y}), \overline{u}) \in B[k]$ **do**
      $\Delta_+ \leftarrow -||\mathbf{H}_u - \mathbf{H}_{(e,y)}||^2$;
      $\Delta_- \leftarrow -||\mathbf{H}_u - \mathbf{H}_{(e,\overline{y})}||^2$;
      $\Delta_u \leftarrow \mathbf{H}_u[k] - \mathbf{H}_{\overline{u}}[k]$;
      $\mathcal{L} \leftarrow -\log \sigma\left(\Delta_+ - \Delta_-\right) - \log\left(\sigma(\Delta_u)\right)$;
      /* Update user and item embeddings using stochastic gradient descent:*/;
      $\mathbf{H} \leftarrow \mathbf{H} - \alpha \times \frac{\partial \mathcal{L}}{\partial \mathbf{H}}$;
return $\{\mathbf{H}_u \mid u \in R\}$, $\{\mathbf{H}_{(e,y)} \mid (e, y) \in R\}$;

where the logistic sigmoid function enforces $\mathbf{H}_u[k] > \mathbf{H}_{\overline{u}}[k]$. It acts as a gatekeeper, allowing us to maintain the prescribed hierarchy among the dimension-wise values. It regularizes the learning process, ensuring that the embeddings adhere to the specified constraints, and fostering the model's interpretability in capturing nuanced distinctions between users' performances on the same dimension.

The pseudo-code of CD-BPR is presented in Algorithm 1. The embedding $\mathbf{H}$ are learned using a variant of Stochastic Gradient Descent (SGD) called Adam optimizer. Samples in batch are made of a user u, its answer y to a question e, the opposite answer $\overline{y}$ to question e, and a user $\overline{u}$ whose answer is $(e, \overline{y})$. The corresponding embeddings are updated by back propagation of the gradient of the following loss function

$$\mathcal{L}(\mathbf{H}) = \mathcal{L}_1(\mathbf{H}) + \mathcal{L}_2(\mathbf{H})$$

whose derivative is defined as: $\frac{\partial \mathcal{L}(\mathbf{H})}{\partial \mathbf{H}} = -\sum_{(u,e,y)\in R}\left(1 - \sigma(\widehat{P_{u,e,y}} - \widehat{P_{u,e,\overline{y}}})\right) \times \frac{\partial}{\partial \mathbf{H}}(\widehat{P_{u,e,y}} - \widehat{P_{u,e,\overline{y}}}) + 2\lambda||\mathbf{H}|| + \sum_{(u,e,y)\in R}\sum_{\substack{(\overline{u},e,y')\in R \\ y'\leq y Q_{ek}=1}}(1 - \sigma(\mathbf{H}_u[k] - \mathbf{H}_{\overline{u}}[k])) + \frac{\partial}{\partial \mathbf{H}}(\mathbf{H}_u[k] - \mathbf{H}_{\overline{u}}[k])$.

By combining these two strategies, CD-BPR not only strives to achieve high predictive accuracy but also prioritizes the development of embeddings that are inherently interpretable and align with the real-world expectations of users and stakeholders. This dual emphasis on predictive power and interpretability is pivotal for building a robust and user-friendly cognitive diagnostic system.

4 Empirical validation of the method

This section assesses CD-BPR's performance on classification and interpretability metrics against other established methods using standard benchmark datasets.

Dataset	Dimensions					correct	density
	# U	# E	# K	# K/E	# R	rate	
ASSIST09	2493	17671	123	1.19	267415	0.658	0.006
ASSIST17	1702	3162	102	1.17	390281	0.437	0.073
ALGEBRA	830	2365	136	1.38	616730	0.729	0.314
MATH1	4209	20	11	3.35	84180	0.492	1
MATH2	3911	20	16	3.20	78220	0.466	1

Table 1. Main characteristics of the datasets.

4.1 Experimental Setup

Datasets. We use five real-world datasets containing records of student performance in mathematics exercises. Each dataset consists of logs containing the outcome of a mathematical exercises attempted by students. Among these datasets, ASSIST09 and ASSIST17 are two sparse datasets gathered via the online tutoring systems ASSISTments [5]. ALGEBRA ("Algebra | 2006-2007") comes from the KDD Cup 2010 development challenge [10]. MATH1 and MATH2 are two datasets presenting high school students performances [6]. Basic statistics of the cleaned datasets are presented in Table 1. All datasets were sanitized: logs with missing user, exercise, response or skill were removed as well as duplicates. Attention has been paid not to lose any knowledge concept associated with exercises while removing the duplicates. ASSIST09 and ASSIST17 were pre-processed using the same methodology as [18]. We filtered out users with fewer than 15 logs to avoid excessively sparse datasets. For ALGEBRA, we iteratively discarded users and exercises with fewer than 100 logs. Finally, no pre-processing was performed on MATH1 and MATH2. This methodology ensures that we compare our approach to competitive ones, using datasets that exhibit different data sparsity levels.

Baselines. To assess the performance of CD-BPR, we compare it against four state-of-the-art methods: MIRT [7], DINA [3], MCD [9], and NCDM [17]. We use the code available in the EduCDM repository [1] for the experiments. All baseline parameters were optimized using the Adam optimization algorithm. For each method, we set the number of dimensions of the latent space equal to the number of knowledge concepts #K in the dataset (refer to Table 1). We also consider the IRR-learned model. Due to space constraints, we report only the results of the best model in terms of accuracy and DOA, specifically the model with the highest harmonic mean of precision and DOA, on the ASSIST09 dataset. Among the three methods – IRR-MIRT, IRR-NCDM, and IRR-DINA – it is the combination of IRR with MIRT that maximizes this value.

Training and hyperparameter setting. We used a 5-fold cross-validation approach, splitting student logs into training, validation, and test sets (60%, 20%, and 20% respectively). Before splitting, user logs were shuffled to reduce bias from response time variations. Hyperparameters were tuned using a grid search strategy with early stopping mechanism based on validation accuracy. The best configuration was chosen across folds, and models were trained on combined training and validation data. Evaluation was then conducted on the test set to gauge the model's generalization performance.

Evaluation Measures. We evaluate the performance of the models using standard classification metrics, including Accuracy, Root Mean Square Error (RMSE), Area Under the Receiver Operating Characteristic Curve (ROC-AUC), Precision, Recall, and F1 Score. Additionally, we assess the quality and interpretability of the users' embeddings using the Degree of Agreement (DOA) measure [17] and a metric based on the Pearson Correlation Coefficient (PC-ER).

Algorithms	Accuracy	Precision	Recall	F1	ROC-AUC	RMSE	DOA	PC-ER
				ASSIST09				
CD-BPR	**0.740 ± 0.003**	**0.765 ± 0.002**	**0.867 ± 0.003**	**0.813 ± 0.002**	**0.786 ± 0.003**	**0.430 ± 0.001**	**0.762 ± 0.005**	**0.536 ±0.066**
NCDM	0.718 ± 0.002	0.752 ± 0.008	0.849 ± 0.023	0.797 ± 0.006	0.740 ± 0.003	0.467 ± 0.004	0.561 ± 0.056	-0.019 ±0.005
MCD	0.657 ± 0.007	0.731 ± 0.010	0.750 ± 0.021	0.740 ± 0.008	0.667 ± 0.010	0.519 ± 0.005	0.463 ± 0.008	-0.010 ±0.018
MIRT	0.602 ± 0.002	0.717 ± 0.003	0.644 ± 0.003	0.679 ± 0.002	0.619 ± 0.003	0.589 ± 0.002	0.482 ± 0.006	0.007 ± 0.007
DINA	0.660 ± 0.005	0.754 ± 0.001	0.711 ± 0.011	0.732 ± 0.006	0.722 ± 0.003	0.481 ± 0.003	0.559 ± 0.050	-0.025±0.006
IRR-MIRT	0.620 ± 0.002	0.748 ± 0.002	0.773 ± 0.004	0.760 ± 0.002	0.693 ± 0.004	0.472 ± 0.002	0.526 ± 0.009	0.144 ± 0.007
				ASSIST17				
CD-BPR	**0.737 ± 0.003**	**0.763 ± 0.003**	**0.867 ± 0.003**	**0.812 ± 0.003**	**0.783 ± 0.004**	**0.430 ± 0.001**	**0.754 ± 0.012**	**0.199±0.213**
NCDM	0.696 ± 0.003	0.692 ± 0.006	0.549 ± 0.023	0.612 ± 0.013	0.742 ± 0.003	0.476 ± 0.002	0.596 ± 0.120	-0.019±0.001
MCD	0.660 ± 0.003	0.613 ± 0.005	0.600 ± 0.012	0.606 ± 0.007	0.713 ± 0.006	0.480 ± 0.003	0.458 ± 0.007	-0.006 ±0.006
MIRT	0.623 ± 0.002	0.569 ± 0.002	0.559 ± 0.003	0.564 ± 0.002	0.659 ± 0.002	0.552 ± 0.001	0.460 ± 0.009	-0.001±0.005
DINA	0.620 ± 0.001	0.565 ± 0.003	0.569 ± 0.004	0.567 ± 0.003	0.710 ± 0.002	0.501 ± 0.001	0.587 ± 0.113	-0.020±0.003
IRR-MIRT	0.675 ± 0.005	0.613 ± 0.003	0.672 ± 0.003	0.641 ± 0.001	0.730 ± 0.001	0.473 ± 0.001	0.534 ± 0.008	0.072 ± 0.009
				ALGEBRA				
CD-BPR	**0.806 ± 0.002**	0.828 ± 0.001	**0.926 ± 0.002**	**0.874 ± 0.001**	**0.826 ± 0.002**	0.377 ± 0.001	**0.846 ± 0.013**	**0.425±0.014**
NCDM	0.789 ± 0.001	**0.840 ± 0.005**	0.878 ± 0.007	0.858 ± 0.001	0.813 ± 0.002	0.389 ± 0.001	0.580 ± 0.013	0.043± 0.003
MCD	0.801 ± 0.003	0.831 ± 0.001	0.914 ± 0.001	0.870 ± 0.000	**0.826 ± 0.000**	**0.364 ± 0.000**	0.479 ± 0.008	0.000±0.007
MIRT	0.682 ± 0.002	0.803 ± 0.002	0.746 ± 0.004	0.774 ± 0.002	0.676 ± 0.002	0.541 ± 0.002	0.481 ± 0.006	-0.002±0.005
DINA	0.677 ± 0.002	0.801 ± 0.002	0.741 ± 0.005	0.770 ± 0.002	0.772 ± 0.001	0.741 ± 0.005	0.582 ± 0.011	0.066±0.007
IRR-MIRT	0.645 ± 0.002	0.803 ± 0.004	0.689 ± 0.045	0.741 ± 0.024	0.669 ± 0.008	0.483 ± 0.013	0.485 ± 0.007	0.140 ± 0.013
				MATH1				
CD-BPR	**0.776 ± 0.008**	**0.704 ± 0.011**	**0.727 ± 0.012**	**0.715 ± 0.009**	**0.853 ± 0.007**	0.407 ± 0.004	**0.803 ± 0.003**	**0.027±0.028**
NCDM	0.563 ± 0.024	0.664 ± 0.033	0.542 ± 0.040	0.597 ± 0.037	0.700 ± 0.040	0.510 ± 0.026	0.563 ± 0.021	-0.055±0.006
MCD	0.611 ± 0.004	0.693 ± 0.006	0.732 ± 0.009	**0.712 ± 0.004**	0.804 ± 0.006	**0.389 ± 0.004**	0.521 ± 0.003	-0.006±0.019
MIRT	0.606 ± 0.016	0.700 ± 0.020	0.713 ± 0.033	0.706 ± 0.020	0.781 ± 0.026	0.414 ± 0.020	0.496 ± 0.025	**0.027 ±0.050**
DINA	0.469 ± 0.005	0.587 ± 0.017	0.124 ± 0.004	0.204 ± 0.006	0.767 ± 0.005	0.469 ± 0.007	0.573 ± 0.002	-0.027±0.003
IRR-MIRT	0.598 ± 0.009	0.607 ± 0.014	0.610 ± 0.022	0.608 ± 0.017	0.773 ± 0.015	0.453 ± 0.010	0.575 ± 0.003	**0.017 ± 0.010**
				MATH2				
CD-BPR	**0.732 ± 0.008**	**0.689 ± 0.009**	0.683 ± 0.009	**0.686 ± 0.009**	**0.804 ± 0.008**	0.433 ± 0.002	**0.843 ± 0.003**	**0.221±0.023**
NCDM	0.585 ± 0.026	0.638 ± 0.034	0.549 ± 0.034	0.591 ± 0.034	0.662 ± 0.032	0.557 ± 0.024	0.588 ± 0.002	0.050±0.003
MCD	0.644 ± 0.005	0.668 ± 0.007	**0.700 ± 0.007**	0.684 ± 0.005	0.775 ± 0.005	**0.418 ± 0.003**	0.585 ± 0.004	-0.000±0.018
MIRT	0.625 ± 0.012	0.643 ± 0.013	0.671 ± 0.015	0.657 ± 0.013	0.737 ± 0.017	0.457 ± 0.013	0.499 ± 0.018	-0.017±0.021
DINA	0.492 ± 0.004	0.553 ± 0.011	0.160 ± 0.007	0.249 ± 0.010	0.706 ± 0.005	0.504 ± 0.003	0.547 ± 0.010	0.026±0.003
IRR-MIRT	0.582 ± 0.005	0.598 ± 0.010	0.625 ± 0.008	0.612 ± 0.008	0.723 ± 0.010	0.489 ± 0.007	0.505 ± 0.002	0.122 ± 0.009

Table 2. Classification performances based on accuracy, precision, recall, ROC-AUC, F1, and RMSE scores, and quality of the embeddings based on DOA and PC-ER. We highlight in bold the highest mean values and comparable ones with respect to the standard deviation, except for PC-ER on ASSIST17 whose standard deviations are too high.

The Degree of Agreement (DOA) evaluates the alignment between users' embeddings and users' responses in the response log. It reflects the consistency of the model with observed user behavior, providing insights into the interpretability and reliability of the embeddings. Specifically, it assesses the extent to which the embeddings align with the observed data. Higher DOA values ($DOA \in [0,1]$) indicate better alignment between the model and the data. Formally, for a given knowledge concept k, $DOA(k)$ evaluates pairs of users (u,v) where the model predicts that u performs better than v ($\delta(\mathbf{H}_{uk} > \mathbf{H}_{vk})$). It counts the proportion of questions related to knowledge concept k ($\delta(Q_{ek})$) that both users have answered ($J(e,u,v)$), and for which u provides a higher value than v. This measure captures how well the model's user embeddings reflect the actual performance hierarchy among users for specific knowledge concepts.

$$DOA(k) = \frac{1}{\alpha}\sum_{u\in U}\sum_{v\in U}\delta(\mathbf{H}_{uk} > \mathbf{H}_{vk}) \times \frac{\sum_{e\in E}\delta(Q_{ek}) \wedge J(e,u,v) \wedge \delta(Y_{ue} > Y_{ve})}{\beta(u,v)}$$

with $\alpha = \sum_{u\in U}\sum_{v\in U}\delta(\mathbf{H}_{uk} > \mathbf{H}_{vk})\delta(\beta(u,v) > 0)$, $\beta(u,v) = \sum_{e\in E}\delta(Q_{ek}) \wedge J(e,u,v) \wedge \delta(Y_{ue} \neq Y_{ve})$, and

$$\delta(x) = \begin{cases} 1 & \text{if } x \text{ is true} \\ 0 & \text{otherwise} \end{cases} \quad J(e,u,v) = \begin{cases} 1 & \text{if } u \text{ and } v \text{ answered } e \\ 0 & \text{otherwise} \end{cases}$$

The PC-ER measure quantifies the correlation between students' embedding and their average response on questions grouped by knowledge concepts. We use the Pearson correlation coefficient ($\rho_{A,B} = \frac{\text{cov}(A,B)}{\sigma_A\sigma_B}$) with $A = \mathbf{H}_u$ and $B = \mathbf{M}_u$ with

$$\mathbf{M}_u[k] = \frac{\sum_{\substack{Q_{ek}=1 \\ (u,e,y)\in R}} Y_{ue}}{\#\{e \mid (u,e,y) \in R \text{ and } Q_{ek} = 1\}}$$

.

We use the Pearson correlation coefficient to measure the linear relationship between users' embeddings and average user responses, providing a quantitative assessment of the consistency and reliability of user embeddings, ensuring that the model's embeddings align closely with actual observed data. The final measure is the average value over all users: PC-ER $= \frac{1}{\#U}\sum_{u\in U}\rho_{\mathbf{H}_u,\mathbf{M}_u}$

4.2 Classification performance and interpretability evaluation

Across all experimental datasets, CD-BPR achieves superior Accuracy and ROC-AUC performances compared to other methods except for algebra where MCD reaches a similar ROC-AUC (see Table 2). CD-BPR also excels on Precision, Recall and F1 scores, for which it holds the best performance in most of the cases. Notably, NCDM, the most recent method rooted in deep learning is outperformed in the majority of cases by CD-BPR which is based on a lighter architecture. Indeed, CD-BPR exhibits a reduction in parameters ranging from 7.9% fewer on ASSIST09 to as much as 74% fewer on MATH1. Because of its excessive parameter count, NCDM's performances are particularly poor on the two smaller datasets, MATH1 and MATH2—184268 and 203205 parameters, respectively, compared to fewer than 47000 and 63000 for all other methods on these datasets.

The performance contrast between CD-BPR and its competitors, observed through its lower RMSE but superior Accuracy and ROC-AUC scores, underscores a fundamental divergence in learning principles. On the one hand, the competitors learn to predict the outcome by closing the gap between the true binary response outcome and a continuous prediction —a proxy rounded for classification— that can be interpreted as the positive response outcome probability. On the other hand, CD-BPR learns to correctly order the two probabilities of negative and positive outcomes. The response outcome with the highest probability is the one predicted by CD-BPR. This approach, based on Bayesian Personalized Ranking (BPR), de-emphasizes closing the gap between class probabilities and true outcomes. Instead, CD-BPR predicts effectively even with minor differences in class probabilities. For the sake of comparison, RMSE for CD-BPR is also computed with the probability of the positive outcome. The lower RMSE measures of CD-BPR, together with its superior performances on the classification metrics, therefore indicate that trying to learn the absolute probability value of the positive outcome proves ineffective in binary classification tasks. Instead, prioritizing probability ranking emerges as a more appropriate strategy.

Knowing CD-BPR is the most efficient method to predict students' responses, we compared its capacity to provide students proficiency estimates on knowledge concepts (KC) through their embedding. The first condition for students embedding to be interpreted as proficiency estimates is the compliance with the monotonicity assumption, measured with DOA and PC-ER. In this regard, CD-BPR consistently achieves significantly higher performances on both scores (Table 2). It is worth noticing that IRR-MIRT, the second best method on DOA and PC-ER, is dominated by CD-BPR on all datasets with a respective DOA gap of 30%, 29 %, 42%, 28% and 40% on ASSIST09, ASSIST17, ALGEBRA, MATH1 and MATH2. Moreover, the IRR version decreases the accuracy of the original MIRT method. The NCDM method, which aims at optimizing both accuracy and DOA, reaches the second best prediction performances of three out of the five datasets but stay far bellow the DOA performance of CD-BPR. Indeed, the latter improves DOA measures on ASSIST09, ASSIST17, ALGEBRA, MATH1 and MATH2 respectively by 36%, 27%, 45%, 39% and 43%. These results suggest that CD-BPR significantly better constraints the embeddings in the KC space than other methods.

Two specific cases are worth detailing. Firstly, the standard deviation above the 5-fold cross validation of PC-ER measure for CD-BPR on ASSIST17 is higher than the mean value. Referencing the ablation study (Table 3), it is evident that while $\mathcal{L}_2$ term predominantly contributes to the enhancement of PC-ER across datasets, it also serves as the primary driver of PC-ER variability on ASSIST17, as standard deviation reduces only upon its removal. However, an analysis of PC-ER measures for CD-BPR on individual folds reveals consistently positive values, with two exceeding 0.4, indicative of CD-BPR's superior performance relative to competitors on this dataset. Secondly, PC-ER is surprisingly low for CD-BPR on MATH1 with regard to the high DOA. Looking once again at the ablation study, it becomes apparent that the initialization method adversely affects the measure for both the MATH1 and MATH2 datasets. However, the impact on MATH1 seems to exceed $\mathcal{L}_2$ improvement of PC-ER. Individually setting each dimension to zero, we observe that removing dimension one annihilate the negative effect of initialization on MATH1 and allow CD-BPR to reach a PC-ER of 0.522 ± 0.037 whereas the mean PC-ER plus standard deviation of the competitors in the same conditions are below 0.03. In fact, this dimension is linked to a single question out of the 20 in the dataset. Since PC-ER is calculated as an average across dimensions, this particular item exerts a disproportionate influence on the measure. The low PC-ER therefore is a specific effect of CD-BPR initialization failure on a specific question. Similarly, removing dimension 7 of MATH2 allows CD-BPR to reach a PC-ER of 0.419 ± 0.033 whereas the competitors mean PC-ER plus standard deviation stay bellow 0.04.

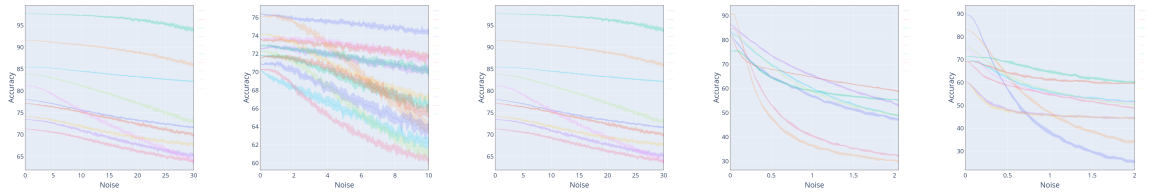
The second aspect of interpreting user embeddings as skill estimates is the independence of the dimensions from each other. We investigated whether individual components of a user's embedding can indicate their mastery of specific knowledge concepts (KC) or whether a combination of components is necessary. To do this, we measured the impact of each embedding component on the prediction accuracy for questions linked to the corresponding KC by introducing noise. Figure 1 shows the prediction accuracy, averaged over KCs, as subtractive noise is introduced into the user embedding component associated with the same KC. More precisely, we subtract from the k-th component of user embeddings the absolute value of a random variable following a normal distribution with a mean of 0 and x times the standard deviation of the component, where x is the value on the x-axis. Due to the limited number of data points in the test set for some dimensions, we exclude them from the graph, as they result in almost flat curves. To do this, we apply a minimum threshold value to the logs by KC, setting it at 2000 for ALGEBRA, MATH1 and MATH2, 1750 for ASSIST09 and 1500 for ASSIST17.

We notice a decline in predictions with the introduction of noise, signifying that altering the component associated with a KC affects predictions for questions related to the same concept. This observation underscores the strong correlation between the information associated with a KC and its corresponding component. Moreover, slower decrease of some dimensions can be explained by the fact that their corresponding questions are systematically related to other concepts. As a consequence, they are expected to be correlated to other dimensions. We can further hypothesize that the magnitude of decline on MATH1 and MATH2 datasets is greater because of the higher average number of logs per dimension compared to ASSIST09, ASSIST17 and ALGEBRA, which may allow CD-BPR to better learn to uncouple dimensions.

Overall, these findings indicate that CD-BPR excels in generating embeddings that are not only effective for classification tasks but also highly informative. Not only are the users embedding components coherent with their responses outcome, dimensionally wise, but there is also a direct consequence between an embedding component and the predicted outcome on a related knowledge concept. The elevated interpretability expands the potential utility of the model for stakeholders, suggesting additional avenues for exploration and analysis. In view of these results, it seems for example relevant to exploit CD-BPR embeddings to analyze student profiles.

4.3 Ablation study

To conclude our evaluation of CD-BPR, we aim to investigate the impact of the two strategies integrated into the BPR learning process on the classification performances and the interpretability. To achieve this, we conduct an ablation study, the results of which are presented in Table 3. The performances are evaluated through a 5-fold cross validation whose average and standard deviation on Accuracy, Precision, Recall, F1, ROC-AUC, RMSE, DOA and PC-ER is included in the table. Notably, the omission of $\mathcal{L}_2$ leads to a marginal reduction in classification performance across most datasets; however, two datasets are exceptions to this trend (i.e., MATH1 and MATH2). More significantly, the DOA and PC-ER experience a substantial decline



ASSIST09 ASSIST17 ALGEBRA MATH1 MATH2

Figure 1. Perturbation of the embeddings (subtracted noise) and its impact of the accuracy of the corresponding question answers (10 replications).

Datasets	Accuracy	Precision	Recall	F1	ROC-AUC	RMSE	DOA	PC-ER
				ASSIST09				
CD-BPR	**0.740 ± 0.003**	0.765 ± 0.002	**0.867 ± 0.003**	**0.813 ± 0.002**	**0.786 ± 0.003**	**0.430 ± 0.001**	**0.762 ± 0.005**	**0.536 ±0.066**
Without $\mathcal{L}_2$	0.737 ± 0.003	0.763 ± 0.002	0.866 ± 0.003	0.811 ± 0.002	0.782 ± 0.004	**0.430 ± 0.001**	0.534 ± 0.014	0.265 ±0.037
Without init	0.736 ± 0.003	**0.770 ± 0.002**	0.849 ± 0.002	0.807 ± 0.002	0.781 ± 0.003	0.433 ± 0.001	0.755 ± 0.004	0.456 ±0.051
Without $\mathcal{L}_2$ and init	0.737 ± 0.003	0.763 ± 0.003	**0.867 ± 0.003**	0.812 ± 0.003	0.783 ± 0.004	**0.430 ± 0.001**	0.478 ± 0.015	0.003 ±0.018
				ASSIST17				
CD-BPR	**0.737 ± 0.003**	**0.763 ± 0.003**	**0.867 ± 0.003**	**0.812 ± 0.003**	0.783 ± 0.004	**0.430 ± 0.001**	**0.754 ± 0.012**	**0.199±0.213**
Without $\mathcal{L}_2$	0.730 ± 0.001	0.711 ± 0.003	0.644 ± 0.002	0.676 ± 0.002	**0.798 ± 0.001**	0.431 ± 0.000	0.563 ± 0.011	0.046±0.011
Without init	0.730 ± 0.001	0.711 ± 0.003	0.643 ± 0.002	0.675 ± 0.002	0.796 ± 0.001	0.434 ± 0.000	0.572 ± 0.003	0.130 ±0.158
Without $\mathcal{L}_2$ and init	0.731 ± 0.001	0.712 ± 0.003	0.644 ± 0.002	0.676 ± 0.002	**0.798 ± 0.001**	0.431 ± 0.000	0.486 ± 0.020	-0.001 ± 0.025
				ALGEBRA				
CD-BPR	0.806 ± 0.002	0.828 ± 0.001	**0.926 ± 0.002**	**0.874 ± 0.001**	**0.826 ± 0.002**	0.377 ± 0.001	**0.846 ± 0.013**	**0.425±0.014**
Without $\mathcal{L}_2$	0.806 ± 0.002	0.828 ± 0.001	**0.926 ± 0.003**	**0.874 ± 0.001**	**0.826 ± 0.002**	0.376 ± 0.001	0.491 ± 0.017	0.088±0.009
Without init	**0.807 ± 0.001**	**0.830 ± 0.001**	0.923 ± 0.002	**0.874 ± 0.001**	0.822 ± 0.002	**0.380 ± 0.001**	0.735 ± 0.020	0.256±0.018
Without $\mathcal{L}_2$ and init	0.806 ± 0.002	0.829 ± 0.001	0.925 ± 0.002	**0.874 ± 0.001**	**0.826 ± 0.002**	0.376 ± 0.001	0.486 ± 0.016	0.001 ±0.031
				MATH1				
CD-BPR	0.776 ± 0.008	0.704 ± 0.011	0.727 ± 0.012	0.715 ± 0.009	0.853 ± 0.007	0.407 ± 0.004	0.803 ± 0.003	0.027±0.028
Without $\mathcal{L}_2$	**0.781 ± 0.007**	**0.709 ± 0.009**	0.729 ± 0.014	0.719 ± 0.009	**0.862 ± 0.007**	**0.397 ± 0.004**	0.467 ± 0.009	-0.298±0.03
Without init	0.778 ± 0.009	**0.709 ± 0.012**	0.719 ± 0.013	0.714 ± 0.010	0.856 ± 0.008	0.404 ± 0.004	**0.846 ± 0.007**	**0.454 ±0.091**
Without $\mathcal{L}_2$ and init	**0.781 ± 0.007**	**0.709 ± 0.009**	**0.731 ± 0.014**	**0.720 ± 0.008**	**0.862 ± 0.007**	**0.397 ± 0.004**	0.503 ± 0.076	0.036±0.080
				MATH2				
CD-BPR	0.732 ± 0.008	0.689 ± 0.009	0.683 ± 0.009	0.686 ± 0.009	0.804 ± 0.008	0.433 ± 0.002	0.843 ± 0.003	0.221±0.023
Without $\mathcal{L}_2$	**0.737 ± 0.011**	**0.693 ± 0.012**	0.692 ± 0.014	**0.693 ± 0.012**	**0.811 ± 0.010**	**0.428 ± 0.003**	0.523 ± 0.007	-0.004 ±0.006
Without init	0.730 ± 0.011	0.688 ± 0.012	0.676 ± 0.013	0.682 ± 0.012	0.802 ± 0.010	0.434 ± 0.003	**0.867 ± 0.006**	**0.547 ±0.064**
Without $\mathcal{L}_2$ and init	**0.737 ± 0.011**	**0.693 ± 0.012**	**0.693 ± 0.013**	**0.693 ± 0.012**	**0.811 ± 0.010**	**0.428 ± 0.003**	0.500 ± 0.057	0.061±0.035

Table 3. Ablation study: mean and standard deviation of the accuracy, precision, recall, f1, Area Under the ROC Curve (ROC-AUC), Root Mean Square Error (RMSE) and DOA scores for four ablation scenarios (without $\mathcal{L}_2$, without initialization, without both $\mathcal{L}_2$ and initialization, and vanilla BPR). Each experiment was replicated according to a 5-folds cross validation method. We highlight in bold the highest mean scores of every dataset.

when $\mathcal{L}_2$ is excluded, underscoring its importance in the model's predictive consistency. In scenarios where the model is deprived of initialization, the DOA and PC-ER exhibit a slight decrease, with the exception of MATH1 and MATH2. As detailed in section 4.2 it is explained by the failure of initialization on only one over-influential question in MATH1 and MATH2. Despite these alterations, the accuracy metric largely remains unaffected, indicating a resilience in the model's ability to correctly classify. However, when both $\mathcal{L}_2$ and initialization are absent, a noticeable decrement in accuracy is observed alongside a pronounced reduction in DOA and PC-ER, the latter almost reaching 0 correlation. This highlights the synergistic importance of these components in achieving optimal model performance and agreement.

The empirical validation shows CD-BPR surpasses current state-of-the-art models in Accuracy, AUC, and DOA across various datasets. Unlike most models, CD-BPR's efficient architecture ensures better classification performance and interpretability on both big and small datasets. An ablation study underscores the importance of $\mathcal{L}_2$ loss and initialization for high accuracy and DOA. These results underscore CD-BPR's value for students' profile analysis, thanks to its efficient parameterization and informative concept embeddings.

5 Conclusion

In this paper, we address the challenge of approximating students' skills in knowledge concepts using data on their success or failure in exercises. Our primary objectives are twofold: accurately predicting students' response outcomes and estimating interpretable proficiency scores. To achieve these goals, we introduce a novel Cognitive Diagnosis Model called CD-BPR. Leveraging the rank learning principle from Bayesian Personalized Ranking, our model incorporates a second learning objective to ensure interpretability in a more direct way, along with an astute parameter initialization method. We conduct extensive experiments on five datasets of mathematical exams to evaluate the performance of CD-BPR in both response outcome prediction and student profile interpretability. Our results demonstrate that CD-BPR outperforms existing approaches on all objectives, significantly improving the interpretability across all datasets as measured by the DOA and PC-ER metric. Additionally, a qualitative analysis of student proficiency approximations by CD-BPR confirms their interpretability, indicating potential for personalized education. Furthermore, an ablation study underscores the significance of our proposed contributions in CD-BPR, enhancing both classification performance and interpretability. By effectively combining these strengths, our model offers a nuanced understanding of student cognitive profiles, providing educators and educational platforms with a powerful tool for personalized teaching on a large scale. Future research could focus on characterizing the student profiles generated by our method to better target educational resources.

Acknowledgment

This research was funded by l'Agence Nationale de la Recherche (ANR), project PORTRAIT ANR-22-CE23-0006. For the purpose of open access, the author has applied a CC-BY public copyright license to any Author Accepted Manuscript (AAM) version arising from this submission.

This research has received funding from the European Union's H2020 European Innovation Council Pathfinder Pilot program under grant agreement No 964529.

References

[1] bigdata ustc. Educdm. https://github.com/bigdata-ustc/EduCDM, 2021.

[2] S. Chen, J. L. Moore, D. Turnbull, and T. Joachims. Playlist prediction via metric embedding. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 714–722, Beijing China, Aug. 2012. ACM. ISBN 978-1-4503-1462-6. doi: 10.1145/2339530.2339643. URL https://dl.acm.org/doi/10.1145/2339530.2339643.

[3] J. De La Torre. DINA Model and Parameter Estimation: A Didactic. *Journal of Educational and Behavioral Statistics*, 34(1):115–130, Mar. 2009. ISSN 1076-9986, 1935-1054. doi: 10.3102/1076998607309474. URL http://journals.sagepub.com/doi/10.3102/1076998607309474.

[4] J. De La Torre, L. A. Van Der Ark, and G. Rossi. Analysis of Clinical Data From Cognitive Diagnosis Modeling Framework. *Measurement and Evaluation in Counseling and Development*, 51(4):281–296, Feb. 2015. ISSN 0748-1756, 1947-6302. doi: 10.1177/0748175615569110. URL http://journals.sagepub.com/doi/10.1177/0748175615569110.

[5] M. Feng, N. Heffernan, and K. Koedinger. Addressing the assessment challenge with an online system that tutors as it assesses. *User Modeling and User-Adapted Interaction*, 19(3):243–266, Aug. 2009. ISSN 0924-1868, 1573-1391. doi: 10.1007/s11257-009-9063-7. URL http://link.springer.com/10.1007/s11257-009-9063-7.

[6] Q. Liu, R. Wu, E. Chen, G. Xu, Y. Su, Z. Chen, and G. Hu. Fuzzy Cognitive Diagnosis for Modelling Examinee Performance. *ACM Transactions on Intelligent Systems and Technology*, 9(4):1–26, July 2018. ISSN 2157-6904, 2157-6912. doi: 10.1145/3168361. URL https://dl.acm.org/doi/10.1145/3168361.

[7] M. Reckase. *Multidimensional Item Response Theory*. Statistics for Social and Behavioral Sciences. Springer New York, New York, NY, 2009. ISBN 978-0-387-89975-6 978-0-387-89976-3. doi: 10.1007/978-0-387-89976-3. URL http://link.springer.com/10.1007/978-0-387-89976-3.

[8] S. Rendle, C. Freudenthaler, Z. Gantner, and L. Schmidt-Thieme. BPR: Bayesian personalized ranking from implicit feedback. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, UAI '09, pages 452–461, Arlington, Virginia, USA, June 2009. AUAI Press. ISBN 978-0-9749039-5-8.

[9] T. Shiwei. Mcd implementation in educdm, Mar. 2021. URL https://github.com/bigdata-ustc/EduCDM/blob/main/EduCDM/MCD.

[10] J. Stamper, A. Niculescu-Mizil, S. Ritter, G. Gordon, and K. Koedinger. Data set from kdd cup 2010 educational data mining challenge, 2010. URL Finditathttp://pslcdatashop.web.cmu.edu/KDDCup/downloads.jsp.

[11] Z. Tan, J. de la Torre, W. Ma, D. Huh, M. E. Larimer, and E.-Y. Mun. A Tutorial on Cognitive Diagnosis Modeling for Characterizing Mental Health Symptom Profiles Using Existing Item Responses. *Prevention Science*, 24(3):480–492, Apr. 2023. ISSN 1573-6695. doi: 10.1007/s11121-022-01346-8. URL https://doi.org/10.1007/s11121-022-01346-8.

[12] K. K. Tatsuoka. Architecture of Knowledge Structures and Cognitive Diagnosis: A Statistical Pattern Recognition and Classification Approach. In *Cognitively Diagnostic Assessment*, pages 327–359. Routledge, New York, 1 edition, 1995. ISBN 978-0-203-05296-9. Num Pages: 33.

[13] J. L. Templin and R. A. Henson. Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods*, 11 (3):287–305, Sept. 2006. ISSN 1939-1463, 1082-989X. doi: 10.1037/1082-989X.11.3.287. URL http://doi.apa.org/getdoi.cfm?doi=10.1037/1082-989X.11.3.287.

[14] S. Tong, Q. Liu, R. Yu, W. Huang, Z. Huang, Z. A. Pardos, and W. Jiang. Item response ranking for cognitive diagnosis. In Z.-H. Zhou, editor, *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 1750–1756. International Joint Conferences on Artificial Intelligence Organization, 8 2021. doi: 10.24963/ijcai.2021/241. URL https://doi.org/10.24963/ijcai.2021/241. Main Track.

[15] S. Tong, J. Liu, Y. Hong, Z. Huang, L. Wu, Q. Liu, W. Huang, E. Chen, and D. Zhang. Incremental Cognitive Diagnosis for Intelligent Education. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1760–1770, Washington DC USA, Aug. 2022. ACM. ISBN 978-1-4503-9385-0. doi: 10.1145/3534678.3539399. URL https://dl.acm.org/doi/10.1145/3534678.3539399.

[16] W. Waegeman, K. Dembczynski, and E. Hüllermeier. Multi-target prediction: a unifying view on problems and methods. *Data Min. Knowl. Discov.*, 33(2):293–324, 2019. doi: 10.1007/S10618-018-0595-5. URL https://doi.org/10.1007/s10618-018-0595-5.

[17] F. Wang, Q. Liu, E. Chen, Z. Huang, Y. Chen, Y. Yin, Z. Huang, and S. Wang. Neural cognitive diagnosis for intelligent education systems. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 6153–6161, 2020.

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

[19] Y. Zhuang, Q. Liu, Z. Huang, Z. Li, S. Shen, and H. Ma. Fully adaptive framework: Neural computerized adaptive testing for online education. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 4734–4742, 2022.