# Path-Specific Causal Reasoning for Fairness-aware Cognitive Diagnosis

Dacao Zhang
zhdacao@gmail.com
School of Computer Science and Information Engineering, Hefei University of Technology
Hefei, Anhui, China

Kun Zhang*
zhang1028kun@gmail.com
School of Computer Science and Information Engineering, Hefei University of Technology
Hefei, Anhui, China

Le Wu
lewu.ustc@gmail.com
School of Computer Science and Information Engineering, Hefei University of Technology
Institute of Dataspace, Hefei Comprehensive National Science Center
Hefei, Anhui, China

Mi Tian
tianmixlb@gmail.com
School of Computer Science and Information Engineering, Hefei University of Technology
Hefei, Anhui, China

Richang Hong
hongrc.hfut@gmail.com
School of Computer Science and Information Engineering, Hefei University of Technology
Institute of Dataspace, Hefei Comprehensive National Science Center
Hefei, Anhui, China

Meng Wang
eric.mengwang@gmail.com
School of Computer Science and Information Engineering, Hefei University of Technology
Hefei, Anhui, China

## ABSTRACT

Cognitive Diagnosis (CD), which leverages students and exercise data to predict students' proficiency levels on different knowledge concepts, is one of fundamental components in Intelligent Education. Due to the scarcity of student-exercise interaction data, most existing methods focus on making the best use of available data, such as exercise content and student information (e.g., educational context). Despite the great progress, the abuse of student sensitive information has not been paid enough attention. Due to the important position of CD in Intelligent Education, employing sensitive information when making diagnosis predictions will cause serious social issues. Moreover, data-driven neural networks are easily misled by the shortcut between input data and output prediction, exacerbating this problem. Therefore, it is crucial to eliminate the negative impact of sensitive information in CD models. In response, we argue that sensitive attributes of students can also provide useful information, and only the shortcuts directly related to the sensitive information should be eliminated from the diagnosis process. Thus, we employ causal reasoning and design a novel *Path-Specific Causal Reasoning Framework* (*PSCRF*) to achieve this goal. Specifically, we first leverage an encoder to extract features and generate embeddings for general information and sensitive information of students. Then, we design a novel attribute-oriented predictor to decouple the sensitive attributes, in which fairness-related sensitive features will be eliminated and other useful information will be retained. Finally, we designed a multi-factor constraint to ensure the performance of fairness and diagnosis performance simultaneously. Extensive experiments over real-world datasets (e.g., PISA dataset) demonstrate the effectiveness of our proposed *PSCRF*.

## CCS CONCEPTS

- **Information systems → Personalization.**

## KEYWORDS

Cognitive Diagnosis, User modeling, Causal Reasoning, Sensitive Attribute, Fairness

**ACM Reference Format:**
Dacao Zhang, Kun Zhang, Le Wu, Mi Tian, Richang Hong, and Meng Wang. 2024. Path-Specific Causal Reasoning for Fairness-aware Cognitive Diagnosis. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '24), August 25–29, 2024, Barcelona, Spain.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3637528.3672049

## 1 INTRODUCTION

As a fundamental component in Intelligent Education, Cognitive Diagnosis (CD) requires an agent to mine student behavior data to access and identify the student's proficiency level in knowledge concepts [43]. It has been applied in various education scenarios, such as student performance prediction [14, 51], computerized adaptive testing [45], and exercise recommendation [25, 50, 53].

*Corresponding authors.

*KDD '24, August 25–29, 2024, Barcelona, Spain*

ACM ISBN 979-8-4007-0490-1/24/08
https://doi.org/10.1145/3637528.3672049

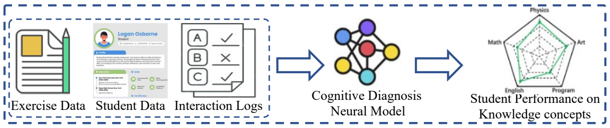

**Figure 1: A tiny example of general cognitive diagnosis.**

Literately, researchers have designed enormous neural network-based methods to realize accurate diagnosis and student modeling. As illustrated in Figure 1, existing models usually take multi-type information (e.g., student-exercise interaction logs, student's personal information, exercise text) as the input, and predict students' mastery of each concept. For example, Wang et al. [43] have built a deep full connection neural network to capture the complex student-exercise interaction records. Wang et al. [44] designed a novel KaNCD method to address the weak knowledge concepts coverage problem. Besides, there are many other methods for high-order student-exercise interaction modeling [1, 8, 26, 59] and graph-based modeling [15, 23, 27, 55].

Despite the great progress, these methods will inevitably introduce fairness issues while exploiting the full potential of student data. Taking Table 1 as an example, by conducting statistical analysis, we can observe that students who are from more affluent families or in more developed areas usually have better performance on exercise (e.g., 0.6434 points for rich boys v.s. 0.4736 points for poor boys). In fact, this phenomenon occurs more because those students receive better support or training (e.g., more books, computer access opportunities, etc), rather than better family circumstances. However, if we do not constrain the model to exploit all the data, it will easily learn the connections between sensitive information of students and student performance (e.g., using family wealth to predict the student proficiency level), which we name as spurious correlations. From the results in Table 1, we can observe this type of phenomenon. By using PISA data [32] to train NCD and KaNCD models directly, they will overestimate these advantaged students (e.g., 0.565 for Australia v.s. 0.3025 for Brazil), showing that they have taken advantage of sensitive information and made unfair predictions. If we apply the unfair results to real-world scenarios, it will exacerbate social prejudices and conflicts, bringing about bad social effects. More seriously, according to the results in Table 1, even if we do not use sensitive attributes as model inputs, NCD and KaNCD models still can infer sensitive attributes of students from the interaction logs and abuse them for better performance. Therefore, **it is crucial to exclude the abuse of student sensitive attributes while ensuring the diagnosis performance**.

Recently, plenty of fairness-aware methods have been proposed, such as data reweighting (resampling) [20, 36] and adversarial learning [4, 48, 57]. However, these strategies still have unavoidable shortcomings. For example, data resampling methods usually increase/decrease weights of certain student-exercise interactions to realize the fairness target. However, this strategy violates the principle in cognitive diagnosis that the same student should only respond to the same exercise once, and is also dependent on the sensitive attributes [21]. Meanwhile, adversarial learning uses an additional classifier to predict the sensitive attribute from user embeddings and eliminate corresponding information directly. This strategy is too coarse-grained to distinguish available information from sensitive information, leading to a decrease in model capability. To answer the above question, we propose that the fairness-related sensitive features from sensitive attributes should be eliminated as comprehensively as possible while diagnosis-related features from sensitive attributes should be retained as much as possible. For example, family wealth cannot be used as an influencing factor in determining the student proficiency level, while the quality of the learning environment can. For this goal, causal inference [34] is one promising direction. By distinguishing causation and correlations from biased real-world data, causal inference has made great progress in medicine [22], neuroscience [29], cognitive science [37], etc. It also has been proven useful in addressing bias issues in vision question answering [31], text classification tasks [35, 56], anomaly detection [47], user modeling [38], and so on.

**Table 1: The probability of students answering questions correctly (i.e., Data Statistics) and the predicted probability of students answering questions correctly by NCD and KaNCD among different groups. The groups are divided by the sensitive attributes of students (i.e., family wealth or country).**

| Model | Family Wealth | | | Country | |
|---|---|---|---|---|---|
| | Poor | Average | Wealth | Australia | Brazil |
| Data statistics | 0.4736 | 0.5448 | 0.6434 | 0.5516 | 0.3888 |
| NCD | 0.5140 | 0.5861 | 0.6789 | 0.5913 | 0.3293 |
| KaNCD | 0.4778 | 0.5589 | 0.6643 | 0.5650 | 0.3025 |
| NCD-*PSCRF* | 0.5545 | 0.5798 | 0.6155 | 0.5824 | 0.3321 |
| KaNCD-*PSCRF* | 0.5286 | 0.5581 | 0.6271 | 0.5680 | 0.3026 |

To this end, in this paper, we propose to employ causal inference and design a novel *Path-Specific Causal Reasoning Framework* (*PSCRF*) for fairness-aware CD modeling. Specifically, we leverage a causal graph to describe the correlations and causation between different factors and student proficiency levels. Based on the causal graph, we try to use *PSCRF* to calculate the path-specific effect of different inputs to the output. We first leverage an encoder to extract features from student-exercise interaction logs and generate embeddings for student IDs and sensitive attributes. Next, we design a novel attribute-oriented predictor (Decoupled Predictor (DP)) to realize the decoupling of sensitive attributes and useful information, in which fairness-related sensitive feature embeddings are used to predict the sensitive attributes and diagnosis-related feature embeddings are used to predict the useful information from sensitive attributes. Moreover, to ensure the quality of decoupling, we also design a multi-factor fairness constraint to restrict the distance of different embeddings. Then, the fairness-aware inference can be obtained by removing the fairness-related sensitive features from the diagnosis process. Finally, we conducted extensive experiments over real-world diagnosis data in various settings. Experimental results demonstrate that *PSCRF* can achieve impressive debiased performance while maintaining the accuracy of student proficiency level modeling. We also release the code to facilitate the community[1].

[1] https://github.com/NLPfreshman0/PSCRF

# 2 RELATED WORK

The related work can be summarized into two components: 1) *Cognitive Diagnosis*: giving a brief introduction of CD in intelligent education scenarios; 2) *Fairness-aware User Modeling*: focusing on fair user representation learning from biased real-world data.

## 2.1 Cognitive Diagnosis

Cognitive Diagnosis (CD) is a fundamental and pivotal task in many real-world intelligent education scenarios [9, 42]. It requires an agent to predict students' proficiency level of each knowledge concept through historical student-exercise interaction logs. DINA [10] and IRT [13] are two representative methods in this domain. DINA is a discrete CDM that assumes student mastery levels are binary (master the knowledge concept or not) [10]. IRT characterizes students' abilities as unidimensional and continuous latent traits and designs logistic-like interaction functions to model the probability of a student correctly answering an exercise [13].

To improve the diagnosis performance, various methods have been proposed to extend the capability of DINA and IRT [16, 19, 41, 46] and exploit the potential of student and exercise data. For example, Cheng and Liu [8] proposed a DIRT method to extract semantic features from the content of exercise texts for high-quality representation generation. Wang et al. [43] designed an NCD method to exploit student-exercise interactions for accurate student proficiency level modeling. Moreover, Zhou et al. [59] proposed to improve CD performance from the student perspective. They employed context and culture information of students to enrich the student proficiency representation, which is in favor of improving the diagnosis performance. Besides, other data issues in CD are also considered, such as weak knowledge concepts coverage problems [44] and non-interactive knowledge concepts problems [28].

## 2.2 Fairness-aware User Modeling

User modeling focuses on measuring user characteristics based on user-related data, which plays a crucial role in plenty of scenarios, such as user preference modeling in recommender system [49] and user proficiency level modeling in Intelligent Education [39]. Recent studies have demonstrated that user-related data may contain stereotypes or biased data, which will mislead models to learn the spurious correlations and make vulnerable and unfair decisions. To alleviate this problem, plenty of fairness-aware user modeling methods have been proposed, such as data reweighting [17, 20, 36], regularization [8, 52], and adversarial learning [2, 48, 57, 60]. Among all these methods, causal inference-based methods are one promising direction. For example, Zhao et al. [58] proposed a disentangled framework TIDE based on path-specific causal reasoning to deal with the popularity bias in user preference modeling in recommendations. Chen et al. [7] designed a novel data augmentation strategy to balance the training data, so that sensitive-related information will be inactivated when modeling user preference. Apart from this, other types of biases are also hot research topics, such as selection bias [5, 30], exposure bias [6], and unfairness [7, 12].

However, due to the sparsity characteristic of student and exercise data in education, existing methods mainly focus on exploiting the potential of data, ignoring the implicit sensitive information abuse problem. Since education plays a crucial role in influencing the trajectory of individuals' adult lives [40], it is urgent to focus more on this problem. Some works have made early attempts. For example, Yu et al. [54] conducted an analysis to explore the equitable prediction of short-term and long-term college success using various sources of student data. Li et al. [24] proposed a Fair-LR algorithm to achieve accurate and fair AI prediction to help to realize fair student modeling. Zhang et al. [57] divided student performance into bias proficiency and fair proficiency, then used only fair proficiency to make predictions. For fairness-aware CD modeling, enormous works remain unexplored, such as fair student representations, sensitive attribute utilization, and so on.

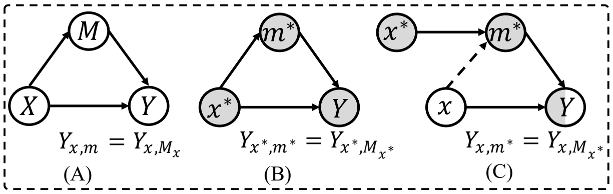

**Figure 2: The general causal graph and commonly used counterfactual notations.**

**Our Distinction.** We focus on a more impactful issue: *How to eliminate the abuse of student sensitive attributes from CD models while ensuring the diagnosis performance?* We argue that student sensitive attributes can also provide useful information, so directly removing them from CD models is not optimal. Thus, we design a novel *PSCRF* to realize the debiased CD learning while retaining the diagnosis performance. Specifically, *PSCRF* decouples student sensitive attributes into sensitive-related information that should not be used in diagnosis process, and sensitive-unrelated information that can be used to improve the diagnosis performance. Moreover, *PSCRF* leverages a multi-factor normalization to ensure the quality of debiased learning and diagnosis performance simultaneously.

# 3 PRELIMINARY

## 3.1 Prerequisite Knowledge

This section explains the causal graph and causal effect calculation mentioned, aiming to help readers understand the significance and importance of the causal graph.

**Causal Graph.** The causal graph is a Directed Acyclic Graph (DAG) $\mathcal{G} =< \mathcal{V}, \mathcal{E} >$, which describes the causal relationships between different variables. $\mathcal{V}$ is the node set and $\mathcal{E}$ is the edge set. As illustrated in Figure 2(A), the arrow indicates the direction of causality. For example, $X \longrightarrow Y$ denotes that variable $X$ has a direct effect on $Y$. $X \longrightarrow M \longrightarrow Y$ denotes that variable $X$ has indirect effect on $Y$ through mediator $M$. Following these notations, assume $X = x$, then the value of $Y$ can be calculated as follows:

$$Y_{x,m} = Y(X = x, M = m = M_x), \tag{1}$$

where the value of mediator $M$ can be calculated with $m = M_x = M(X = x)$.

**Causal Effect Calculation.** The causal effect is a comparison of the potential outcomes of giving two different interventions to the same variable. As shown in Figure 2(A)-(B), assume that $X = x$

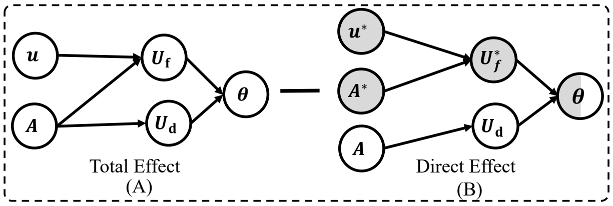

**Figure 3: The causal graph used in our proposed *PSCRF.***

is the treatment, then $Y_{x,m}$ denotes the potential outcome of the treatment. Similarly, $X = x^*$ is the no treatment, then $Y_{x^*,m^*}$ is the potential outcome of no treatment. Along this line, the causal effect can be calculated as follows:

$$Effect = Y_{x,m} - Y_{x^*,m^*} = P(y|do(X = x)) - P(y|do(X = x^*)), \quad (2)$$

where $do(X = x)$ and $do(X = x^*)$ is the intervention to the variable $X$. Note that $do(\cdot)$ operation requires that only the treatment variable is intervened, all other variables are not intervened. To satisfy this requirement, the counterfactual operation is proposed.

**Counterfactual Notation.** The counterfactual reason is to assume a scenario that all the other variables remain unchanged and only the treatment variable is changed. The causal effect of the treatment variable on the output variable can be calculated in this scenario. For example, in Figure 2(C), $Y_{x,M_{x^*}} = Y(X = x, M = M(X = x^*))$ denotes a typical counterfactual reasoning.

**Table 2: Educational Context examples from PISA dataset**

| Aspect | Question Examples |
|---|---|
| Home | Home Economic, Social and Cultural Status (ESCS) |
| | Highest education degree of parents |
| | Number of equipment, appliances, and rooms |
| Person | Whether students have a grade repetition experience |
| | how many days did students engage in out-school activities |

## 3.2 PISA Data Introduction

As mentioned above, the educational context of students can be used to obtain better student representations, which refer to the various features related to students' learning process [59]. OECD's Programme for International Student Assessment (PISA) focused on this topic and designed multiple questions to investigate and collect these educational contexts, such as family wealth, education degree of parents, and so on. For example, when investigating the highest education degree of students' parents, five options (e.g., 1-General senior, 2-Vocational senior, 3-Junior) are provided. Students can select one option based on their situations. This information can be used as student attributes. Table 2 lists some examples of these questions. Moreover, this organization has developed exercises to measure 15-year-olds' ability to employ reading, mathematics, and science knowledge and skills to meet real-life challenges [32]. They investigated students from different countries, and released the data and technical reports on a three-year cycle, which is suitable for student attribute-aware cognitive diagnosis.

With the guidance of technical report [32], we select ESCS index as the sensitive attribute example to tackle the problem in Section 1. Moreover, we leverage the Pearson Correlation Coefficient to select the useful but not sensitive attributes: *1) The number of books, 2) The number of tablet computers, 3) A link to the Internet, 4) A computer can be used for school work, 5) The number of E-book readers.* These selected attributes all exhibit strong correlations with ESCS index, which we have reported the results in Table 8 and Table 9 in the Appendix. However, they are not sensitive attributes and directly contribute to the development of students' abilities, which should be considered in the diagnosis process.

**Table 3: Notations and explanations in our proposed *PSCRF.***

| Notation | Explanation |
|---|---|
| $\boldsymbol{u}, \boldsymbol{u}*$ | Student ID embeddings and their counterfactuals |
| $A, A^*$ | sensitive attribute and counterfactual sensitive attribute |
| $\boldsymbol{U}_f, \boldsymbol{U}_f^*$ | The diagnosis-related features and their counterfactuals |
| $\boldsymbol{U}_d$ | The fairness-related feature |
| $\alpha$ | Learnable parameters for integrating $\boldsymbol{U}_f$ and $\boldsymbol{U}_d$ |
| $\phi_e$ | Parameters related to exercises(e.g., difficulty, discrimination) |
| $\boldsymbol{\theta}, \boldsymbol{\theta}^*$ | The student proficiency level and their counterfactuals |
| $\beta$ | Learnable parameters controlling the degree of debiasing |
| $\boldsymbol{\theta}_d$ | The fairness-aware student proficiency level |

# 4 TECHNICAL DETAILS OF *PSCRF*

## 4.1 Causal view of *PSCRF*

Based on the motivation in Section 1, we use the causal graph in Figure 3(A) to describe the causal relation among different paths. Specifically, $\boldsymbol{u}$ denotes the general representation of one student (i.e., ID embeddings). $\boldsymbol{A}$ is the corresponding sensitive attribute representation (e.g., ESCS embeddings). We argue that sensitive attributes contain fairness-related sensitive features and diagnosis-related features. The former should not be used in the diagnosis process while the latter should be used to improve the diagnosis performance. Therefore, we leverage the causal path $\boldsymbol{A} \longrightarrow \boldsymbol{U}_d \longrightarrow \boldsymbol{\theta}$ to denote the effect of fairness-related sensitive features, which should be removed from the entire graph. A toy example is that family wealth should not be considered in the diagnosis process since it will introduce unfairness to vulnerable groups.

Meanwhile, we use the causal path $(\boldsymbol{u}, \boldsymbol{A}) \longrightarrow \boldsymbol{U}_f \longrightarrow \boldsymbol{\theta}$ to denote the effect of diagnosis-related features from sensitive attributes and general information. one similar example is as follows: though family wealth cannot be used in the diagnosis process, we can exploit *the number of books* or *a link to the Internet* to better model student proficiency level since they directly contribute to the student's ability development. Based on this causal graph, we then introduce the corresponding implementation.

According to the principle of Average Treatment Effect (ATE), we can calculate the Total Effect (TE) of all input variables to the output prediction as follows:

$$TE = \boldsymbol{\theta}(u, A) - \boldsymbol{\theta}(u^*, A^*). \quad (3)$$

Next, we intend to calculate the effect of only fairness-aware sensitive features on the output prediction. For this target, we employ the Natural Direct Effect (NDE) as follows:

$$NDE = \boldsymbol{\theta}(u^*, A) - \boldsymbol{\theta}(u^*, A^*). \quad (4)$$

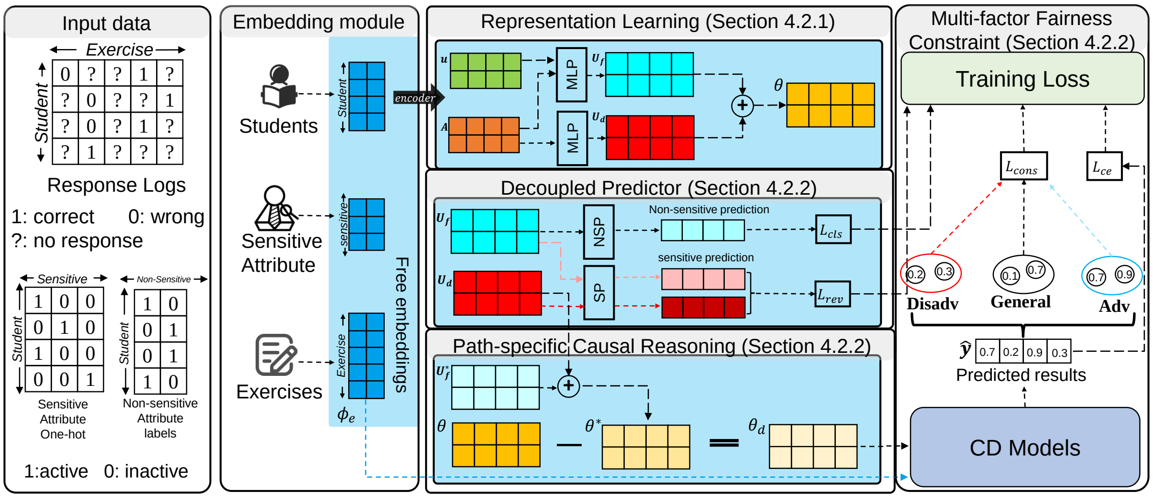

Figure 4: The overall framework of our proposed *PSCRF*.

Finally, we can obtain the debiased prediction by calculating the Total Indirect Effect (TIE) as follows:

$$TIE = TE - NDE = \boldsymbol{\theta}(u, A) - \boldsymbol{\theta}(u^*, A). \tag{5}$$

By maximizing the Total Indirect Effect (TIE) inference, we can achieve debiased learning in the diagnosis process.

## 4.2 Causal Implementation of *PSCRF*

Figure 4 illustrates the overall architecture of *PSCRF*, which consists of two main components: *Representation Learning Module* and *Decoupled and Constraint Module*. Next, we introduce each component in detail. Table 3 explains the notations in *PSCRF*.

*4.2.1 Representation Learning Module.* Similar to previous works [43, 59], we first embed input entities into latent embeddings, which including student embeddings $\boldsymbol{U} = [\boldsymbol{u}_1, \boldsymbol{u}_2, ..., \boldsymbol{u}_N]^\top \in \mathbb{R}^{N\times d}$, exercise embeddings $\boldsymbol{V} = [\boldsymbol{v}_1, \boldsymbol{v}_2, ..., \boldsymbol{v}_M]^\top \in \mathbb{R}^{M\times d}$, and sensitive attribute embeddings $\boldsymbol{A} = [\boldsymbol{a}_1, \boldsymbol{a}_2, ..., \boldsymbol{a}_T]^\top \in \mathbb{R}^{T\times d}$. $d$ is the embedding dimension. These embeddings are randomly initialized and will be updated during model learning.

After obtaining the embeddings of different entities, we intend to learn the representations of fairness-related sensitive features $\boldsymbol{U}_d$ and diagnosis-related features $\boldsymbol{U}_f$, which is realized by our designed proficiency modeling module. Specifically, we employ a fairness-related sensitive feature generator to obtain $\boldsymbol{U}_d$, which takes sensitive attribute embeddings as input and uses a Multi-Layer Perceptron (MLP) to generate representations as follows:

$$\boldsymbol{U}_d^i = \sigma(MLP_1(\boldsymbol{A}_{[i]})), \tag{6}$$

where $\sigma(\cdot)$ represents the sigmoid function, $\boldsymbol{A}_{[i]}$ is the sensitive attribute embedding set for the $i^{th}$ student. $\boldsymbol{U}_d^i$ is the sensitive-related feature representation of the $i^{th}$ student.

Similarly, a diagnosis-related feature extractor is used to generate $\boldsymbol{U}_f$ with student ID embeddings and sensitive attribute embeddings. Another MLP is used to obtain the representations:

$$\boldsymbol{U}_f^i = \sigma(MLP_2(concat(\boldsymbol{u}_i, \boldsymbol{U}_d^i))), \tag{7}$$

where $concat(\cdot)$ denotes the concatenation operation. $\boldsymbol{u}_i$ is the free ID embedding of the $i^{th}$ student.

After obtaining the results from two extractors, we leverage a learnable parameter $\alpha$ to fuse these two different features and generate student proficiency level $\boldsymbol{\theta}_i$ as:

$$\boldsymbol{\theta}_i = \sigma((1-\alpha)\boldsymbol{U}_f^i + \alpha\boldsymbol{U}_d^i). \tag{8}$$

*4.2.2 Decoupled and Constraint Module.* In the above module, we aim to extract fairness-related sensitive features and diagnosis-related features from input data. However, there are no explicit supervised signals. which poses a big challenge. In response, we design the following three modules. Next, we omit the student index $i$ for simplicity and introduce how we construct each module.

**(1) Decoupled Predictor (DP).** First, we intend diagnosis-related feature embedding $\boldsymbol{U}_f$ should include useful information from sensitive attributes and maintain general information of students. Therefore, we leverage the educational context as the guidance, and identify top-k educational context questions most correlated with the sensitive attribute by computing the Pearson correlation coefficient between them. These educational contexts often relate to students' learning environments and do not involve sensitive attributes (e.g., the number of books), which can be used to enhance the student proficiency level modeling. Thus, we leverage selected educational contexts as the prediction targets and formulate the optimization target as follows:

$$\mathcal{L}_{\text{cls}} = \frac{1}{K}\sum_{k=1}^{K} CE(MLP(\boldsymbol{U}_f), Label_k), \tag{9}$$

where $K$ represents the total number of non-sensitive attributes, $Label_k$ represents the label of the $k^{\text{th}}$ non-sensitive attribute, and CE denotes the cross-entropy function.

Meanwhile, to ensure $\boldsymbol{U}_d$ to focus only on the fairness-related sensitive features, we develop a novel sensitive attribute enhancement module. We first send $\boldsymbol{U}_d$ to an MLP to predict sensitive attributes, so that $\boldsymbol{U}_d$ can be better learned to represent sensitive attributes. Meanwhile, $\boldsymbol{U}_f$ should not contain these sensitive attributes. Thus, we send $\boldsymbol{U_f}$ to the same MLP to predict the counterfactual sensitive attributes. Therefore, $\boldsymbol{U_d}$ can only encode fairness-related sensitive features that should not be used in the diagnosis process. $\boldsymbol{U_f}$ will not include these sensitive features, which is in favor of fairness-aware diagnosis. This process can be formulated as follows:

$$\mathcal{L}_{\text{rev}} = \mathcal{L}(SMLP(\boldsymbol{U_d}), A) + \mathcal{L}(SMLP(\boldsymbol{U_f}), A^*), \tag{10}$$

where $\mathcal{L}$ represents the loss function, which can be mean squared error (MSE) for continuous values or cross-entropy for discrete values. $SMLP(\cdot)$ denotes the shared MLP. $A^*$ represents the counterfactual sensitive attribute label, which we will give a detailed explanation in the next section.

**(2) Path-specific Causal Reasoning.** According to Figure 3 and Section 4.1, we need to remove the path $\boldsymbol{A} \rightarrow \boldsymbol{U_d} \rightarrow \boldsymbol{\theta}$ to realize the fairness-aware diagnosis. Following the principle of Average Treatment Effect (ATE), we need to imagine a counterfactual world, which is shown in Figure 3(B). In the counterfactual world, only the path $\boldsymbol{A} \rightarrow \boldsymbol{U_d} \rightarrow \boldsymbol{\theta}$ remains unchanged. We need to block the effect of the path $(\boldsymbol{u}, \boldsymbol{A}) \rightarrow \boldsymbol{U_f} \rightarrow \boldsymbol{\theta}$. Thus, we use the counterfactual sensitive attributes to modify Eq.(7) and Eq.(8) as follows:

$$\begin{aligned} \boldsymbol{U}_f^* &= \sigma(MLP(concat(\boldsymbol{u}^*, \boldsymbol{U}_d^*))), \\ \boldsymbol{\theta}^* &= \sigma((1-\alpha)\boldsymbol{U}_f^* + \alpha \boldsymbol{U}_d), \end{aligned} \tag{11}$$

where $\{\boldsymbol{u}^*, \boldsymbol{U}_d^*\}$ are the counterfactual student representation and counterfactual fairness-related sensitive feature representation. For implementation, we use the mean representation of all student representations to realize $\boldsymbol{u}^*$, and use the mean representation of corresponding sensitive attributes to calculate $\boldsymbol{U}_d^*$. According to the causal inference, this intervention can help *PSCRF* to calculate the accurate effect of path $\boldsymbol{A} \rightarrow \boldsymbol{U_d} \rightarrow \boldsymbol{\theta}$. Then, we can obtain fairness-aware student proficiency level as follows:

$$\boldsymbol{\theta_d} = \sigma(\boldsymbol{\theta} - \beta\boldsymbol{\theta}^*), \tag{12}$$

where $\beta$ is a learnable parameter to control the degree of debiasing. $\boldsymbol{\theta_d}$ is used to realize the fairness-aware CD modeling.

**(3) Multi-factor Fairness Constraint.** To facilitate better fairness-aware diagnosis, we introduce a Multi-factor Fairness Constraint. Specifically, we partition students into disadvantaged, general, and advantaged groups based on the value of sensitive attributes. Since $\boldsymbol{\theta_d}$ is the fairness-aware student proficiency level representation, we intend the variance of the predicted means for different groups to be as low as possible. Meanwhile, since $\boldsymbol{U_d}$ should incorporate unwanted sensitive information as much as possible, we maximize the variance of the predicted means for different groups. Therefore, this constrain can be realized as follows:

$$\mathcal{L}_{\text{cons}} = std\left(\overline{y}_{\text{dis}}, \overline{y}_{\text{gene}}, \overline{y}_{\text{adv}}\right)_{\theta_d} - \text{std}\left(\overline{y}_{\text{dis}}, \overline{y}_{\text{gene}}, \overline{y}_{\text{adv}}\right)_{U_d}, \tag{13}$$

where $\overline{y}_{\text{dis}}$, $\overline{y}_{\text{gene}}$ and $\overline{y}_{\text{adv}}$ respectively denote the predicted means of three groups, while $std(\cdot)$ represents the variance.

### 4.3 Model Training

To ensure prediction accuracy, we also add traditional cross-entropy constraints to $\boldsymbol{U_f}$, $\boldsymbol{U_d}$, $\boldsymbol{\theta}$ and $\boldsymbol{\theta_d}$. First, we input them into the CD model to obtain prediction results:

$$Y_{\hat{\theta}} = CDM(\hat{\boldsymbol{\theta}}, \boldsymbol{\phi_e}), \tag{14}$$

where $CDM(\cdot)$ represents the CD models such as NCD or KaNCD. $\hat{\boldsymbol{\theta}}$ can be $\boldsymbol{U_f}$, $\boldsymbol{U_d}$, $\boldsymbol{\theta}$ or $\boldsymbol{\theta_d}$. $\boldsymbol{\phi_e}$ represents the parameters related to exercises(e.g., difficulty, discrimination). Then we minimize the cross-entropy loss between the predictions and labels:

$$\mathcal{L}_{\text{ce}} = \sum_{\theta_i \in \Theta} \text{CE}(Y_{\theta_i}, y), \tag{15}$$

where $\Theta = \{\boldsymbol{U_f}, \boldsymbol{U_d}, \boldsymbol{\theta}, \boldsymbol{\theta_d}\}$. $y$ is the true label. Finally, total loss is:

$$\mathcal{L}_{\text{total}} = w_1\mathcal{L}_{\text{ce}} + w_2\mathcal{L}_{\text{cls}} + w_3\mathcal{L}_{\text{rev}} + w_4\mathcal{L}_{\text{cons}}, \tag{16}$$

where $w_1$,$w_2$, $w_3$, and $w_4$ represent hyperparameters that balance the weights of each part of the loss.

**Table 4: The Statistics of datasets**

| Dataset | Students | Exercises | Exercise Records |
|---|---|---|---|
| Australia | 8,485 | 184 | 249,727 |
| Brazil | 5,777 | 183 | 143,314 |

## 5 EXPERIMENTS

### 5.1 Experimental Setup

**Data preprocessing.** We procured two prototypical datasets from the PISA-2015, representing Australia and Brazil, respectively, meticulously arranged in descending order based on their developmental status and the mean scholastic attainment of students[33]. In each dataset, there are 28 different self-acquired features [3], such as learning interests and self-efficacy. We used two representative sensitive attributes, namely ESCS (Index of Economic, Social, and Cultural Status) and the father's education level [32] to evaluate our method. Specifically, based on the data provided by PISA [32], we categorized students into three groups - *disadvantaged, general, and advantaged* - according to their sensitive attributes, for fairness-aware diagnosis. We filtered out students with fewer than 10 exercise records to ensure sufficient data for training. The Basic statistics of datasets are shown in Table 4. For each dataset, we performed a 70%/10%/20% training/validation/testing split.

**Evaluation Metrics.** Based on the target, we select two types of metrics. For diagnosis performance, following previous works [15, 44], we used widely used metrics: Area Under Curve (AUC) and Accuracy (ACC). Meanwhile, following the work [43], we also use the Degree of Agreement (DOA) for validation.

For fairness performance, since the abuse of sensitive attributes will mislead models to underestimate or overestimate the students from different groups, commonly used fairness metrics are used. We first employ Equal opportunity (EO) [18]:

$$EO = Std(TPR_{disadv}, TPR_{gene}, TPR_{adv}, ). \tag{17}$$

$TPR$ refers to True Positive Rates. $Std(\cdot)$ is the standard deviation. Since there are only two situations (i.e., correct and incorrect) for students answering questions, CD models should have equal capability of predicting the probability of students answering exercises correctly across different groups. Along this line, sensitive attributes can be proved to be not used in the diagnosis process. Moreover, since predictions of CD models have a big social influence in real-world scenarios, we argue that the rights of vulnerable groups

**Table 5: Evaluating accuracy and fairness performance associated with sensitive attribute ESCS**

| Model | | Australia | | | | | | Brazil | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | EO↓ | $D^{under}_{disadv}$ | IR↑ | AUC↑ | ACC↑ | DOA↑ | EO↓ | $D^{under}_{disadv}$ | IR↑ | AUC↑ | ACC↑ | DOA↑ |
| **IRT** | Base | 0.0338 | 0.0826 | 0.7353 | 0.7979 | 0.7266 | - | 0.0582 | 0.1407 | 0.5018 | 0.7794 | 0.7269 | - |
| | Base$^{\dagger}$ | 0.0604 | 0.1473 | 0.7025 | **0.8080** | **0.7322** | - | 0.1025 | 0.2510 | 0.4700 | **0.7958** | **0.7324** | - |
| | Reg | 0.0110 | 0.0270 | **0.7544** | 0.7961 | 0.7249 | - | 0.0277 | 0.0665 | 0.5301 | 0.7769 | 0.7250 | - |
| | Adv | 0.0286 | 0.0697 | 0.7449 | 0.7969 | 0.7264 | - | 0.0669 | 0.1609 | 0.4935 | 0.7797 | 0.7268 | - |
| | *PSCRF* | **0.0051** | **0.0002** | 0.7339 | 0.8022 | 0.7249 | - | **0.0162** | **0.0357** | **0.5760** | 0.7893 | 0.7255 | - |
| **MIRT** | Base | 0.0575 | 0.1408 | 0.7013 | 0.8027 | 0.7299 | - | 0.0913 | 0.2227 | 0.5109 | 0.7836 | 0.7280 | - |
| | Base$^{\dagger}$ | 0.0645 | 0.1523 | 0.6973 | **0.8088** | **0.7339** | - | 0.1251 | 0.3053 | 0.4663 | **0.7950** | **0.7316** | - |
| | Reg | 0.0284 | 0.0694 | 0.7279 | 0.8010 | 0.7278 | - | 0.0512 | 0.1246 | **0.5539** | 0.7813 | 0.7258 | - |
| | Adv | 0.0554 | 0.1357 | 0.7009 | 0.8030 | 0.7288 | - | 0.0956 | 0.2335 | 0.5036 | 0.7840 | 0.7283 | - |
| | *PSCRF* | **0.0098** | **0.0227** | **0.7520** | 0.7983 | 0.7237 | - | **0.0279** | **0.0403** | 0.5248 | 0.7804 | 0.7205 | - |
| **NCD** | Base | 0.0425 | 0.1040 | 0.7183 | 0.7868 | 0.7170 | 0.6248 | 0.0669 | 0.1588 | 0.5220 | 0.7675 | 0.7140 | 0.5972 |
| | Base$^{\dagger}$ | 0.0857 | 0.2039 | 0.6615 | 0.7911 | 0.7199 | 0.6384 | 0.1274 | 0.3108 | 0.4491 | 0.7718 | 0.7166 | 0.6394 |
| | Reg | 0.0331 | 0.0811 | 0.7277 | 0.7863 | 0.7172 | 0.6245 | 0.0522 | 0.1229 | 0.5370 | 0.7669 | 0.7131 | 0.5965 |
| | Adv | 0.0528 | 0.1292 | 0.6644 | 0.7801 | 0.7111 | 0.5715 | 0.0506 | 0.1234 | 0.5388 | 0.7601 | 0.7112 | 0.5648 |
| | *PSCRF* | **0.0029** | **0.0010** | **0.7538** | **0.7997** | **0.7234** | **0.7040** | **0.0030** | **0.0028** | **0.5599** | **0.7788** | **0.7209** | **0.6806** |
| **KaNCD** | Base | 0.0464 | 0.1133 | 0.7113 | 0.8017 | 0.7273 | 0.6584 | 0.0742 | 0.1792 | 0.4877 | 0.7793 | 0.7221 | 0.6046 |
| | Base$^{\dagger}$ | 0.0770 | 0.1878 | 0.6957 | **0.8076** | **0.7310** | 0.6917 | 0.1210 | 0.2963 | 0.5103 | **0.7910** | **0.7284** | **0.6848** |
| | Reg | 0.0255 | 0.0622 | 0.7299 | 0.8004 | 0.7260 | 0.6552 | 0.0464 | 0.1115 | 0.5138 | 0.7775 | 0.7207 | 0.6015 |
| | Adv | 0.0532 | 0.1303 | 0.7075 | 0.8009 | 0.7282 | 0.6615 | 0.0686 | 0.1664 | 0.5388 | 0.7802 | 0.7244 | 0.6357 |
| | *PSCRF* | **0.0110** | **0.0252** | **0.7484** | 0.8045 | 0.7299 | **0.7013** | **0.0363** | **0.0888** | **0.5145** | 0.7892 | 0.7267 | 0.6840 |

should be guaranteed. We should not be prejudiced against disadvantaged groups and assume that they will perform less well. Based on the principle of Equalized Odds [18], we propose the following evaluation metric to evaluate the fairness performance of models:

$$D^{under}_{disadv} = FNR_{disadv} - FNR_{adv}, \tag{18}$$

where $\{FNR_{disadv}, FNR_{adv}\}$ denote the False Negative Rates (FNRs) of disadvantaged and advantaged groups. The closer the value of $D^{under}_{disadv}$ is to 0, the better fairness performance the model is.

Meanwhile, for disadvantaged groups, we should also identify the top students as accurately as possible, so that they have opportunities to access higher levels of education. Thus, we select the absolute metric F2-score [11] to assess the proportion of high-achieving students of disadvantaged groups, which we name as Identified Rate (IR) and formulate as follows:

$$IR = \frac{5 \times precision_{disadv} \times recall_{disadv}}{(4 \times precision_{disadv}) + recall_{disadv}}, \tag{19}$$

where $precision_{disadv}$ and $recall_{disadv}$ denote the Precision and Recall of disadvantaged group. Note that the larger the value of $IR$ is, the better performance the model has.

**Implementation Details.** As our proposed method is model-agnostic, we apply *PSCRF* to four advanced CDMs to show its effectiveness and flexibility. Moreover, we compare with several recent approaches, including regularization-based methods and adversarial-based methods: *1) Base*: Basic cognitive diagnosis models (i.e., IRT, MIRT, NCD, and KaNCD) that do not consider bias; *2) Base*$^{\dagger}$: Basic Models with Sensitive Attributes; *3) Reg*: Regularization-based models, by adding Equal opportunity as a regularization to CDMs [18, 24]; *4) Adv*: Adversarial learning methods [4] to reduce relevance between sensitive attributes and student representation

For implementation, we set the learning rate to 0.001 and batch size to 512. We apply Adam as the optimization algorithm to update the model parameters. To obtain the best performance, we tune hyper-parameters on validation sets to select the best. The balance parameters in Eq.(16) are set to 1.0, 0.1, 0.5, and 1.0, respectively.

## 5.2 Overall Experiments

Table 5 and Table 6 report overall results under different sensitive attributes. We observe all basic CDMs suffered from unfair outcome issues. Besides, *Base*$^{\dagger}$ incorporates sensitive attributes into model learning, achieving better diagnosis performance. Meanwhile, it also exhibits more severe unfairness. All these suggest the urgency to explore fairness-aware learning in CD models. Therefore, we evaluate the model performance from the following three aspects:

For *fairness performance* (i.e., EO and $D^{under}_{disadv}$), we observe that *PSCRF* outperforms most baselines, proving its effectiveness. Moreover, Adversarial-based methods (Adv) show worse performance than regularization-based (Reg) methods in most cases. One possible reason is that the latter directly utilizes sensitive attribute group labels to optimize corresponding metrics. In contrast, *PSCRF* decouples sensitive attributes and only eliminates the fairness-related sensitive features that should not be used in diagnosis process.

Table 6: Evaluating accuracy and fairness performance associated with sensitive attribute Father's education level.

| Model | | Australia | | | | | | Brazil | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | EO↓ | $D^{under}_{disadv}$ | IR↑ | AUC↑ | ACC↑ | DOA↑ | EO↓ | $D^{under}_{disadv}$ | IR↑ | AUC↑ | ACC↑ | DOA↑ |
| IRT | Base | 0.0293 | 0.0705 | 0.7346 | 0.7979 | 0.7266 | - | 0.0366 | 0.0896 | 0.5209 | 0.7794 | 0.7269 | - |
| | Base† | 0.0440 | 0.1069 | 0.6980 | **0.8087** | **0.7312** | - | 0.0720 | 0.1723 | 0.5379 | **0.7959** | **0.7314** | - |
| | Reg | **0.0132** | 0.0303 | **0.7515** | 0.7969 | 0.7255 | - | 0.0190 | 0.0465 | 0.5468 | 0.7781 | 0.7262 | - |
| | Adv | 0.0231 | 0.0546 | 0.7319 | 0.7919 | 0.7200 | - | 0.0314 | 0.0716 | 0.5404 | 0.7752 | 0.7242 | - |
| | *PSCRF* | 0.0162 | **0.0015** | 0.7342 | 0.8034 | 0.7277 | - | **0.0021** | **-0.0049** | **0.5640** | 0.7911 | 0.7274 | - |
| MIRT | Base | 0.0437 | 0.1051 | 0.7084 | 0.8027 | 0.7299 | - | 0.0572 | 0.1398 | 0.5472 | 0.7836 | 0.7280 | - |
| | Base† | 0.0554 | 0.1326 | 0.7195 | **0.8106** | **0.7347** | - | 0.0849 | 0.1820 | 0.4931 | **0.7896** | 0.7279 | - |
| | Reg | **0.0194** | **0.0449** | 0.7383 | 0.8025 | 0.7297 | - | 0.0300 | 0.0735 | 0.5787 | 0.7812 | 0.7272 | - |
| | Adv | 0.0389 | 0.0947 | 0.7110 | 0.8043 | 0.7316 | - | 0.0528 | 0.1292 | 0.5548 | 0.7821 | **0.7282** | - |
| | *PSCRF* | **0.0194** | 0.0474 | 0.7095 | 0.8073 | 0.7285 | - | **0.0247** | **0.0422** | **0.5879** | 0.7827 | 0.7214 | - |
| NCD | Base | 0.0313 | 0.0747 | 0.7265 | 0.7868 | 0.7170 | 0.6248 | 0.0428 | 0.1042 | 0.5409 | 0.7675 | 0.7140 | 0.5972 |
| | Base† | 0.0477 | 0.1147 | 0.6981 | **0.8021** | 0.7263 | 0.6478 | 0.0855 | 0.1745 | **0.6095** | 0.7785 | 0.7026 | 0.6518 |
| | Reg | 0.0293 | 0.0679 | 0.6940 | 0.7834 | 0.7119 | 0.6167 | 0.0324 | 0.0794 | 0.5396 | 0.7689 | 0.7145 | 0.6010 |
| | Adv | 0.0323 | 0.0789 | 0.6791 | 0.7825 | 0.7130 | 0.5918 | 0.0484 | 0.1177 | 0.5470 | 0.7635 | 0.7150 | 0.5722 |
| | *PSCRF* | **0.0227** | **-0.0116** | **0.7362** | 0.8003 | **0.7276** | **0.7096** | **0.0280** | **0.0465** | 0.5745 | **0.7879** | **0.7293** | **0.6889** |
| KaNCD | Base | 0.0370 | 0.0887 | 0.7146 | 0.8017 | 0.7273 | 0.6584 | 0.0433 | 0.1058 | 0.5189 | 0.7793 | 0.7221 | 0.6046 |
| | Base† | 0.051 | 0.1207 | 0.6938 | **0.8084** | **0.7310** | **0.7183** | 0.0555 | 0.1302 | 0.5434 | 0.7862 | 0.7256 | 0.6731 |
| | Reg | 0.0251 | 0.0578 | 0.7364 | 0.8010 | 0.7274 | 0.6642 | **0.0288** | **0.0699** | **0.5826** | 0.7799 | 0.7242 | 0.6347 |
| | Adv | 0.0405 | 0.0972 | 0.7144 | 0.8006 | 0.7278 | 0.6618 | 0.0419 | 0.1020 | 0.5609 | 0.7802 | 0.7239 | 0.6352 |
| | *PSCRF* | **0.0114** | **0.0275** | **0.7768** | 0.8066 | 0.7269 | 0.7097 | 0.0340 | 0.0746 | 0.5130 | **0.7930** | **0.7278** | **0.6847** |

For the *trade-off between fairness and diagnosis performance*, all debiased baselines perform worse on diagnosis performance than *PSCRF*, proving that they cannot retain diagnosis-related features from sensitive attributes, which causes a larger decrease in diagnosis accuracy. *PSCRF* uses the newly designed DP module and multi-factor constraint to retain diagnosis-related features as much as possible, thus outperforming baselines.

For the *Identification capability* (i.e., IR) of high-achieving students from disadvantaged groups, *PSCRF* still achieves impressive performance, which proves that *PSCRF* can effectively ensure the rights of vulnerable groups. Moreover, by considering the results on $D^{under}_{disadv}$ metric, we can conclude that for different CD backbones, *PSCRF* can effectively balance fairness and diagnosis performance, demonstrating the flexibility and effectiveness of *PSCRF*.

## 5.3 Ablation Study

To verify the effectiveness of each component, we conduct an ablation study with ESCS and report results in Table 7. From the results, when only using $\mathcal{L}_{\text{ce}}$, $\boldsymbol{U_f}$ and $\boldsymbol{U_d}$ cannot decouple sensitive attributes effectively. Instead, they would introduce more bias, resulting in better diagnosis and worse fairness performances. When separately introducing $\mathcal{L}_{\text{cls}}$ and $\mathcal{L}_{\text{rev}}$, we observe improvements in fairness performance with minimal impact on accuracy, proving the effectiveness of learned $\boldsymbol{U_f}$ and $\boldsymbol{U_d}$. When using only $\mathcal{L}_{\text{cons}}$, *PSCRF* can learn how to remove fairness-aware sensitive features, achieving a substantial improvement in fairness. However, since $\boldsymbol{U_f}$ still contains some factors affecting fairness, this component can only generate a suboptimal outcome. Moreover, when removing each component, we observe varying degrees of performance degradation. Among all components, removing $\mathcal{L}_{\text{cons}}$ causes the most significant decrease in fairness performance and increase in diagnosis performance. This phenomenon not only proves the importance of multi-factor fairness constraint, but also shows *PSCRF* will abuse sensitive attributes to improve diagnosis performance. Moreover, $\mathcal{L}^{*}_{\text{cons}}$ denotes the removal of constraints on $\boldsymbol{U_d}$ for direct comparison with regularization methods. The result proves that the absence of constraints on $\boldsymbol{U_d}$ severely impacts fairness performance, aligning its performance comparably with regularization methods. Furthermore, we observe that $\mathcal{L}_{\text{cls}}$ and $\mathcal{L}_{\text{rev}}$ have a positive impact on fairness performance, removing them will cause a decrease in fairness performance. In conclusion, these components are all necessary for the superiority of *PSCRF*.

## 5.4 Parameter Sensitive Test

To conduct a deeper analysis on *PSCRF*, we also conduct parameter sensitive test on the weights $w_1$, $w_2$, $w_3$, $w_4$ in Eq.(16), whose results are summarized in Figure 5. According to the results, we can observe that with the increase of $w_4$, the diagnosis performance of *PSCRF* decreases slightly while the fairness performance increases rapidly. This phenomenon is consistent with the results in Table 7, proving the importance of $\mathcal{L}_{\text{cons}}$. Moreover, with the weight $w_1$ increasing, the diagnosis performance of *PSCRF* increases while the fairness performance decreases rapidly and is unstable. This is intuitive since a large $w_1$ will focus more on student data and impose

**Table 7: Ablation Study of IRT on the Australia Dataset**

| Conditions | EO↓ | $D^{under}_{disadv}$ | IR↑ | AUC↑ | ACC↑ |
|---|---|---|---|---|---|
| Base | 0.0338 | 0.0826 | 0.7353 | 0.7979 | 0.7266 |
| *PSCRF* | **0.0051** | **0.0002** | 0.7339 | 0.8022 | 0.7249 |
| w $\mathcal{L}_{\text{ce}}$ | 0.0545 | 0.1335 | 0.6885 | 0.8089 | 0.7329 |
| w $\mathcal{L}_{\text{cls}}$ | 0.0503 | 0.1231 | 0.6982 | 0.8088 | 0.7328 |
| w $\mathcal{L}_{\text{rev}}$ | 0.0525 | 0.1287 | 0.6919 | 0.8090 | 0.7332 |
| w $\mathcal{L}_{\text{cons}}$ | 0.0088 | 0.0069 | 0.7392 | 0.8016 | 0.7250 |
| w/o $\mathcal{L}_{\text{cls}}$ | 0.0137 | 0.0318 | 0.7206 | 0.8057 | 0.7279 |
| w/o $\mathcal{L}_{\text{rev}}$ | 0.0112 | -0.0057 | 0.7277 | 0.8022 | 0.7258 |
| w/o $\mathcal{L}_{\text{cons}}$ | 0.0609 | 0.1493 | 0.7127 | **0.8092** | **0.7337** |
| w/o $\mathcal{L}^*_{\text{cons}}$ | 0.0132 | -0.0322 | **0.7565** | 0.8021 | 0.7257 |

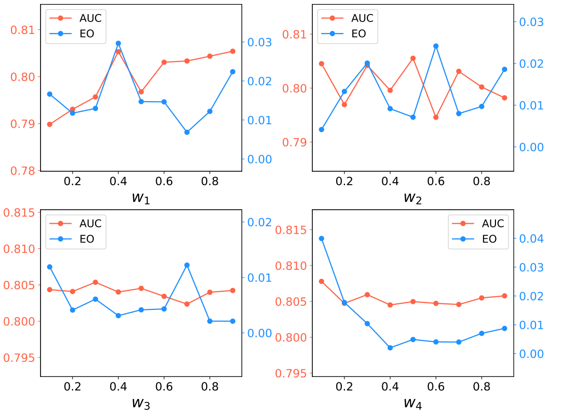

**Figure 5: *PSCRF* based on IRT with varying $w$**

more biased information. With the increase of $w_2$, both diagnosis performance and fairness performance fluctuate considerably. We speculate one possible reason is that $\mathcal{L}_{\text{cls}}$ is relatively simple and its impact on *PSCRF* is relatively small. Thus, a large weight will lead *PSCRF* unable to learn useful information. For $w_3$, a larger value will improve the fairness performance while retaining the diagnosis performance, proving the effectiveness of $\mathcal{L}_{\text{rev}}$.

## 5.5 Case Study

To better illustrate the effectiveness of *PSCRF*, we visualize the prediction distributions of different models using FNR (underestimate rate) and FPR (overestimate rate). As shown in Figure 6, we initially observe a conspicuous prediction bias in the *Base* model, characterized by underestimation for the disadvantaged group and overestimation for the advantaged group, displaying a distinct stepwise distribution between different groups. When introducing sensitive attributes, this unfairness is exacerbated (e.g., *Base*$^\dagger$ model). *Reg* and *Adv* methods can alleviate this unfairness to some extent. However, their prediction distributions still exhibit a stepwise pattern. As a comparison, *PSCRF* shows nearly consistent levels of underestimation and overestimation across different groups, achieving the lowest overestimation rates among the various models. A comparison with *Base*$^\dagger$ reveals that *PSCRF* has effectively mitigated the unfairness imposed on the disadvantaged group by a single sensitive attribute. The only shortfall is that our approach does not reduce the underestimation rates. We speculate that there might exist other sensitive attributes that are not considered.

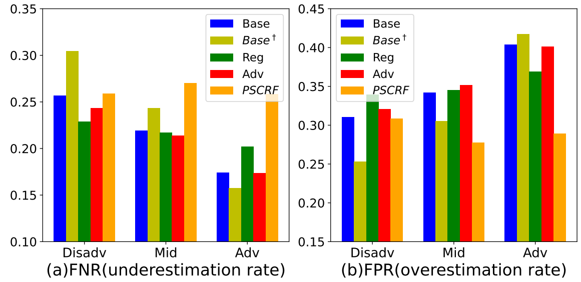

**Figure 6: Visualization of the prediction distributions.**

## 6 CONCLUSION

In this paper, we argued that current CD models concentrated more on the full exploitation of student-exercise interaction data, ignoring the potential risk of abuse of sensitive attributes. Moreover, we argued that student sensitive attributes could also provide useful information, so directly eliminating them was not optimal. To achieve fairness-aware CD learning, we proposed to incorporate casual inference and designed a novel *PSCRF*. By leveraging a newly designed attribute-oriented predictor to deal with the information from sensitive attributes. *PSCRF* decoupled sensitive attributes into fairness-related sensitive features that should not be used in the diagnosis process, and diagnosis-related features that should be used to enrich the student proficiency level modeling. Thus, *PSCRF* could achieve impressive fairness performance while retaining the diagnosis performance. Moreover, we designed a multi-factor fairness constraint to ensure the fairness performance and diagnosis performance simultaneously. Finally, we conducted extensive experiments over real-world datasets and multiple advanced CD models to demonstrate the effectiveness of *PSCRF*.

## 7 ACKNOWLEDGMENTS

This work was supported in part by grants from the National Science and Technology Major Project (Grant No. 2021ZD0111802), Joint Funds of the National Natural Science Foundation of China (Grant No. U22A2094), and the National Natural Science Foundation of China (Grant No. 62376086, U23B2031, 721881011).

## REFERENCES

[1] Hengyao Bao, Xihua Li, Xuemin Zhao, and Yunbo Cao. 2021. Exploring Student Representation For Neural Cognitive Diagnosis. *arXiv preprint arXiv:2111.08951* (2021).

[2] Ghazaleh Beigi, Ahmadreza Mosallanezhad, Ruocheng Guo, Hamidreza Alvari, Alexander Nou, and Huan Liu. 2020. Privacy-aware recommendation with private-attribute protection using adversarial learning. In *Proceedings of the 13th International Conference on Web Search and Data Mining*. 34–42.

[3] Peter M Blau and Otis Dudley Duncan. 1967. The American occupational structure. (1967).

[4] Avishek Bose and William Hamilton. 2019. Compositional fairness constraints for graph embeddings. In *International Conference on Machine Learning*. PMLR, 715–724.

[5] Jiawei Chen, Hande Dong, Xiang Wang, Fuli Feng, Meng Wang, and Xiangnan He. 2023. Bias and debias in recommender system: A survey and future directions. *ACM Transactions on Information Systems* 41, 3 (2023), 1–39.

[6] Jiawei Chen, Yan Feng, Martin Ester, Sheng Zhou, Chun Chen, and Can Wang. 2018. Modeling Users' Exposure with Social Knowledge Influence and Consumption Influence for Recommendation. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management.* 953–962.

[7] Lei Chen, Le Wu, Kun Zhang, Richang Hong, Defu Lian, Zhiqiang Zhang, Jun Zhou, and Meng Wang. 2023. Improving Recommendation Fairness via Data Augmentation. In *Proceedings of the ACM Web Conference 2023.* 1012–1020.

[8] Song Cheng and Qi Liu. 2019. Enhancing item response theory for cognitive diagnosis. *arXiv preprint arXiv:1905.10957* (2019).

[9] Ali Daud, Naif Radi Aljohani, Rabeeh Ayaz Abbasi, Miltiadis D Lytras, Farhat Abbas, and Jalal S Alowibdi. 2017. Predicting student performance using advanced learning analytics. In *Proceedings of the 26th international conference on world wide web companion.* 415–421.

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

[11] Divyaansh Devarriya, Cairo Gulati, Vidhi Mansharamani, Aditi Sakalle, and Arpit Bhardwaj. 2020. Unbalanced breast cancer data classification using novel fitness functions in genetic programming. *Expert Systems with Applications* 140 (2020), 112866.

[12] Michael D Ekstrand, Mucun Tian, Mohammed R Imran Kazi, Hoda Mehrpouyan, and Daniel Kluver. 2018. Exploring author gender in book rating and recommendation. In *Proceedings of the 12th ACM conference on recommender systems.* 242–250.

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

[14] Lina Gao, Zhongying Zhao, Chao Li, Jianli Zhao, and Qingtian Zeng. 2022. Deep cognitive diagnosis model for predicting students' performance. *Future Generation Computer Systems* 126 (2022), 252–262.

[15] Weibo Gao, Qi Liu, Zhenya Huang, Yu Yin, Haoyang Bi, Mu-Chun Wang, Jianhui Ma, Shijin Wang, and Yu Su. 2021. RCD: Relation map driven cognitive diagnosis for intelligent education systems. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval.* 501–510.

[16] Weibo Gao, Hao Wang, Qi Liu, Fei Wang, Xin Lin, Linan Yue, Zheng Zhang, Rui Lv, and Shijin Wang. 2023. Leveraging Transferable Knowledge Concept Graph Embedding for Cold-Start Cognitive Diagnosis. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval.* 983–992.

[17] Sahin Cem Geyik, Stuart Ambler, and Krishnaram Kenthapadi. 2019. Fairness-aware ranking in search & recommendation systems with application to linkedin talent search. In *Proceedings of the 25th acm sigkdd international conference on knowledge discovery & data mining.* 2221–2231.

[18] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. *Advances in neural information processing systems* 29 (2016).

[19] Jie Huang, Qi Liu, Fei Wang, Zhenya Huang, Songtao Fang, Runze Wu, Enhong Chen, Yu Su, and Shijin Wang. 2021. Group-level cognitive diagnosis: A multi-task learning perspective. In *2021 IEEE International Conference on Data Mining (ICDM).* IEEE, 210–219.

[20] Vasileios Iosifidis, Besnik Fetahu, and Eirini Ntoutsi. 2019. Fae: A fairness-aware ensemble framework. In *2019 IEEE International Conference on Big Data (Big Data).* IEEE, 1375–1380.

[21] Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, and Jonathan Taylor. 2023. Resampling methods. In *An Introduction to Statistical Learning: with Applications in Python.* Springer, 201–228.

[22] Stefan Konigorski. 2021. Causal inference in developmental medicine and neurology. *Developmental Medicine & Child Neurology* 63 (2021).

[23] 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.

[24] Yunqi Li, Hanxiong Chen, Zuohui Fu, Yingqiang Ge, and Yongfeng Zhang. 2021. User-oriented fairness in recommendation. In *Proceedings of the Web Conference 2021.* 624–632.

[25] Xin Lin, Zhenya Huang, Hongke Zhao, Enhong Chen, Qi Liu, Defu Lian, Xin Li, and Hao Wang. 2023. Learning Relation-Enhanced Hierarchical Solver for Math Word Problems. *IEEE Transactions on Neural Networks and Learning Systems* (2023).

[26] Qi Liu. 2021. Towards a New Generation of Cognitive Diagnosis.. In *IJCAI.* 4961–4964.

[27] Shuhuan Liu, Xiaoshan Yu, Haiping Ma, Ziwen Wang, Chuan Qin, and Xingyi Zhang. 2023. Homogeneous Cohort-Aware Group Cognitive Diagnosis: A Multi-grained Modeling Perspective. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management.* 4094–4098.

[28] 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.

[29] Ioana Marinescu, Patrick N. Lawlor, and Konrad Paul Kording. 2018. Quasi-experimental causality in neuroscience and behavioural research. *Nature Human Behaviour* 2 (2018), 891–898.

[30] Benjamin M Marlin and Richard S Zemel. 2009. Collaborative prediction and ranking with non-random missing data. In *Proceedings of the third ACM conference on Recommender systems.* 5–12.

[31] Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xian-Sheng Hua, and Ji-Rong Wen. 2021. Counterfactual vqa: A cause-effect look at language bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 12700–12710.

[32] OECD. 2017. PISA 2015 Assessment and Analytical Framework. https://doi.org/10.1787/9789264281820-en. , 260 pages.

[33] OECD. 2018. PISA 2015 Results (Volume I). https://doi.org/10.1787/19963777. , 492 pages.

[34] Judea Pearl, Madelyn Glymour, and Nicholas P Jewell. 2016. *Causal inference in statistics: A primer.* John Wiley & Sons.

[35] Chen Qian, Fuli Feng, Lijie Wen, Chunping Ma, and Pengjun Xie. 2021. Counterfactual inference for text classification debiasing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers).* 5434–5445.

[36] Bashir Rastegarpanah, Krishna P Gummadi, and Mark Crovella. 2019. Fighting fire with fire: Using antidote data to improve polarization and fairness of recommender systems. In *Proceedings of the twelfth ACM international conference on web search and data mining.* 231–239.

[37] Ladan Shams and Ulrik R Beierholm. 2010. Causal inference in perception. *Trends in Cognitive Sciences* 14 (2010), 425–432.

[38] Pengyang Shao, Le Wu, Kun Zhang, Defu Lian, Richang Hong, Yong Li, and Meng Wang. 2024. Average User-Side Counterfactual Fairness for Collaborative Filtering. *ACM Transactions on Information Systems* 42, 5 (2024).

[39] Shuanghong Shen, Zhenya Huang, Qi Liu, Yu Su, Shijin Wang, and Enhong Chen. 2022. Assessing Student's Dynamic Knowledge State by Exploring the Question Difficulty Effect. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval.* 427–437.

[40] Field Simon, Kuczera Małgorzata, and PONT Beatriz. 2007. *Education and training policy no more failures ten steps to equity in education: Ten steps to equity in education.* oecd Publishing.

[41] Shiwei Tong, Qi Liu, Runlong Yu, Wei Huang, Zhenya Huang, Zachary A Pardos, and Weijie Jiang. 2021. Item Response Ranking for Cognitive Diagnosis.. In *IJCAI.* 1750–1756.

[42] Wim J Van der Linden and Cees AW Glas. 2000. *Computerized adaptive testing: Theory and practice.* Springer.

[43] 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.

[44] 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).

[45] Hangyu Wang, Ting Long, Liang Yin, Weinan Zhang, Wei Xia, Qichen Hong, Dingyin Xia, Ruiming Tang, and Yong Yu. 2023. GMOCAT: A Graph-Enhanced Multi-Objective Method for Computerized Adaptive Testing. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining.* 2279–2289.

[46] Shanshan Wang, Zhen Zeng, Xun Yang, and Xingyi Zhang. 2023. Self-supervised Graph Learning for Long-tailed Cognitive Diagnosis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 110–118.

[47] Junfei Wu, Qiang Liu, Weizhi Xu, and Shu Wu. 2022. Bias mitigation for evidence-aware fake news detection by causal intervention. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval.* 2308–2313.

[48] Le Wu, Lei Chen, Pengyang Shao, Richang Hong, Xiting Wang, and Meng Wang. 2021. Learning fair representations for recommendation: A graph-based perspective. In *Proceedings of the Web Conference 2021.* 2198–2208.

[49] Le Wu, Xiangnan He, Xiang Wang, Kun Zhang, and Meng Wang. 2022. A survey on accuracy-oriented neural recommendation: From collaborative filtering to information-rich recommendation. *IEEE Transactions on Knowledge and Data Engineering* 35, 5 (2022), 4425–4445.

[50] Shangshang Yang, Haoyu Wei, Haiping Ma, Ye Tian, Xingyi Zhang, Yunbo Cao, and Yaochu Jin. 2023. Cognitive diagnosis-based personalized exercise group assembly via a multi-objective evolutionary algorithm. *IEEE Transactions on Emerging Topics in Computational Intelligence* (2023).

[51] Mengfan Yao, Siqian Zhao, Shaghayegh Sahebi, and Reza Feyzi Behnagh. 2021. Stimuli-sensitive Hawkes processes for personalized student procrastination modeling. In *Proceedings of the Web Conference 2021.* 1562–1573.

[52] Sirui Yao and Bert Huang. 2017. Beyond parity: Fairness objectives for collaborative filtering. *Advances in neural information processing systems* 30 (2017).

[53] Shengjun Yin, Kailai Yang, and Hongzhi Wang. 2020. A mooc courses recommendation system based on learning behaviours. In *Proceedings of the ACM Turing Celebration Conference-China*. 133–137.

[54] Renzhe Yu, Qiujie Li, Christian Fischer, Shayan Doroudi, and Di Xu. 2020. Towards Accurate and Fair Prediction of College Success: Evaluating Different Sources of Student Data. *International educational data mining society* (2020).

[55] Xiaoshan Yu, Chuan Qin, Dazhong Shen, Haiping Ma, Le Zhang, Xingyi Zhang, Hengshu Zhu, and Hui Xiong. 2024. RDGT: Enhancing Group Cognitive Diagnosis with Relation-Guided Dual-Side Graph Transformer. *IEEE Transactions on Knowledge and Data Engineering* (2024).

[56] Kun Zhang, Dacao Zhang, Le Wu, Richang Hong, Ye Zhao, and Meng Wang. 2024. Label-aware debiased causal reasoning for Natural Language Inference. *AI Open* 5 (2024), 70–78.

[57] Zheng Zhang, Le Wu, Qi Liu, Jiayu Liu, Zhenya Huang, Yu Yin, Yan Zhuang, Weibo Gao, and Enhong Chen. 2024. Understanding and improving fairness in cognitive diagnosis. *Science China Information Sciences* 67, 5 (2024), 152106.

[58] Zihao Zhao, Jiawei Chen, Sheng Zhou, Xiangnan He, Xuezhi Cao, Fuzheng Zhang, and Wei Wu. 2022. Popularity bias is not always evil: Disentangling benign and harmful bias for recommendation. *IEEE Transactions on Knowledge and Data Engineering* (2022).

[59] Yuqiang Zhou, Qi Liu, Jinze Wu, Fei Wang, Zhenya Huang, Wei Tong, Hui Xiong, Enhong Chen, and Jianhui Ma. 2021. Modeling context-aware features for cognitive diagnosis in student learning. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 2420–2428.

[60] Ziwei Zhu, Jianling Wang, and James Caverlee. 2020. Measuring and mitigating item under-recommendation bias in personalized ranking systems. In *Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval*. 449–458.

**Table 8: The statistics of useful but not sensitive attributes associated with ESCS**

| id | name | correlation | category |
|---|---|---|---|
| ST013Q01TA | How many books are there in your home? | 0.416 | 0: 0-100 books, 1: More than 100 books |
| ST012Q07NA | Tablet computers | 0.402 | 0: Zero or one 1: More than one |
| ST011Q06TA | A link to the Internet | 0.308 | 0: Yes 1: No |
| STo11Q04TA | A computer you can use for school work | 0.301 | 0: Yes 1: No |
| ST012Q08NA | E-book readers | 0.266 | 0: NaN 1: More than one |

**Table 9: The statistics of useful but not sensitive attributes associated with Father's education level**

| id | name | correlation | category |
|---|---|---|---|
| ST013Q01TA | How many books are there in your home? | 0.286 | 0: 0-100 books, 1: More than 100 books |
| ST012Q07NA | Tablet computers | 0.166 | 0: Zero or one 1: More than one |
| ST011Q09TA | Works of art (e.g. paintings) | 0.165 | 0: Yes 1: No |
| ST011Q16NA | Books on art, music or design | 0.146 | 0: Yes 1: No |
| ST011Q07TA | Classic literature (e.g. <Shakespeare>) | 0.142 | 0: Yes 1: No |

**Table 10: The Performance of *PSCRF* on Graph-Based RCD Model**

| Method | EO↓ | $\mathbf{D}^{under}_{disadv}$ | IR↑ | AUC↑ | ACC↑ | DOA↑ |
|---|---|---|---|---|---|---|
| Base | 0.0394 | 0.0965 | 0.7132 | 0.7971 | 0.7265 | 0.6209 |
| Base$^\dagger$ | 0.0622 | 0.1523 | 0.7000 | 0.8005 | 0.7269 | 0.6744 |
| *PSCRF* | 0.0032 | 0.0004 | 0.7433 | 0.8009 | 0.7265 | 0.6327 |

**Table 11: The statistical data for alpha and beta.**

| param | mean | variance |
|---|---|---|
| $\alpha$ | 0.2043 | 0.0826 |
| $\beta$ | 0.4041 | 0.1304 |

# A APPENDIX

## A.1 The selection of Educational Context

We present the most relevant useful attributes associated with different sensitive attributes in Table 8 and Table 9. We employ the Pearson correlation coefficient to calculate the correlation, and select the Top-k attributes as the educational context used in Eq.(9) in Section 4.2.2:

$$\rho_{X,Y} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}. \tag{20}$$

## A.2 The Performance of *PSCRF* on Graph-Based RCD Model

Since our *PSCRF* focuses more on fairness-aware cognitive diagnosis, we select fundamental and general CD models in the main text. To further demonstrate the generality of *PSCRF* , we applied PSCRF to the graph-based model RCD[15] and report the results on Australia dataset regarding the sensitive attribute ESCS in Table 10. It can be observed that *PSCRF* achieves similarly good performance on the RCD model as well.

## A.3 The relevant information about the learnable parameters $\alpha$ and $\beta$.

We want to know the values of $\alpha$ and $\beta$ after they have been learned, so we computed the mean and variance of their values after training on the Australia dataset using IRT model, as shown in Table 11. From the results, we can conclude that our proposed *PSCRF* realizes user-specific debiasing according to the input user's situation.