SCIENCE CHINA
Information Sciences

• RESEARCH PAPER •

May 2024, Vol. 67, Iss. 5, 152106:1–152106:16
https://doi.org/10.1007/s11432-022-3852-0

Understanding and improving fairness in cognitive diagnosis

Zheng ZHANG[1,2], Le WU[3], Qi LIU[1,2*], Jiayu LIU[1,2], Zhenya HUANG[1,2], Yu YIN[1,2], Yan ZHUANG[1,2], Weibo GAO[1,2] & Enhong CHEN[1,2]

[1]*Anhui Province Key Laboratory of Big Data Analysis and Application, School of Computer Science and Technology & School of Data Science, University of Science and Technology of China, Hefei* 230026*, China;*
[2]*State Key Laboratory of Cognitive Intelligence, Hefei* 230088*, China;*
[3]*School of Computer Science and Information Engineering, Hefei University of Technology, Hefei* 230601*, China*

Received 8 December 2022/Revised 1 April 2023/Accepted 10 July 2023/Published online 25 April 2024

Abstract Intelligent education is a significant application of artificial intelligence. One of the key research topics in intelligence education is cognitive diagnosis, which aims to gauge the level of proficiency among students on specific knowledge concepts (e.g., Geometry). To the best of our knowledge, most of the existing cognitive models primarily focus on improving diagnostic accuracy while rarely considering fairness issues; for instance, the diagnosis of students may be affected by various sensitive attributes (e.g., region). In this paper, we aim to explore fairness in cognitive diagnosis and answer two questions: (1) Are the results of existing cognitive diagnosis models affected by sensitive attributes? (2) If yes, how can we mitigate the impact of sensitive attributes to ensure fair diagnosis results? To this end, we first empirically reveal that several well-known cognitive diagnosis methods usually lead to unfair performances, and the trend of unfairness varies among different cognitive diagnosis models. Then, we make a theoretical analysis to explain the reasons behind this phenomenon. To resolve the unfairness problem in existing cognitive diagnosis models, we propose a general fairness-aware cognitive diagnosis framework, FairCD. Our fundamental principle involves eliminating the effect of sensitive attributes on student proficiency. To achieve this, we divide student proficiency in existing cognitive diagnosis models into two components: bias proficiency and fair proficiency. We design two orthogonal tasks for each of them to ensure that fairness in proficiency remains independent of sensitive attributes and take it as the final diagnosed result. Extensive experiments on the Program for International Student Assessment (PISA) dataset clearly show the effectiveness of our framework.

Keywords fairness, intelligent education, cognitive diagnosis, psychometrics, adversarial learning

1 Introduction

Intelligent education, an important avenue for artificial intelligence (AI), employs AI technology to investigate the learning laws of students. As an interdisciplinary study, the research of intelligent education has attracted considerable attention from scholars in different fields [1–5], such as education, machine learning, and psychology. In intelligence education, one of the key research topics is cognitive diagnosis (CD), which aims to measure the proficiency level of students in terms of specific knowledge about certain concepts (e.g., Geometry). Figure 1(a) shows a toy example of CD. Generally, the student usually first chooses to practice a set of exercises (e.g., e_1, e_2, e_3) and provide responses (correct or wrong). Then, the CD can infer the degree of mastery the student possesses over the corresponding concepts (e.g., Geometry). With the comprehensive understanding of students, CD could be further applied to numerous applications, such as student assessment [6] and educational recommendation systems [7], which can lessen the burden of both teachers and students and offer effective learning experiences for students.

In the literature, massive research efforts have been undertaken from both psychometrics and machine learning to further their understanding through CD. With the help of psychometric theories, many researchers manually designed linear interaction functions that combine the multiplication of students' and

* Corresponding author (email: qiliuql@ustc.edu.cn)

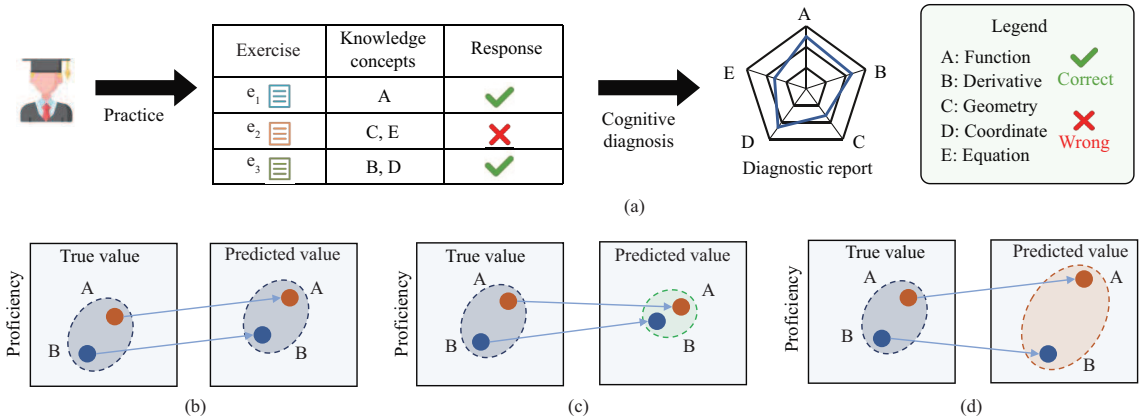

Figure 1 (Color online) Illustrative examples of (a) the cognitive diagnosis system, (b) the fairness definition, (c) the unfair performance—narrow the gap and (d) the unfair performance—widen the gap. A and B represent the groups divided by sensitive attributes (e.g., gender).

exercises' trait features linearly [1,3,8,9]. Among them, item response theory (IRT) [1] has been widely recognized as a classical framework in which interpretable parameters accompanied by item response functions are adopted to assess student performance. In terms of machine learning, the researchers have adopted deep neural networks to model high-order interaction functions. Representative methods such as NeuralCD (NCD) [10] have been proposed by leveraging neural networks to model more complex student-exercise interaction records. The result demonstrates that NeuralCD can achieve both accurate and interpretable diagnostic results.

Despite considerable efforts, we observe that most existing studies primarily focus on enhancing the accuracy of cognitive diagnostic performance while ignoring a fundamental but critical criterion in education — fairness. Here, fairness refers to the principle that sensitive attributes, which are characteristics prone to discrimination such as gender and race, should not be an obstacle to achieving educational potential [11]. In other words, various groups classified by sensitive attributes (e.g., gender, region) should be treated similarly [12]. In this paper, we explore the crucial issue of fairness in CD. Following the classical fairness definition of equal opportunity [13], we define fairness in CD as the proficiency gap between different groups divided by sensitive attributes should be maintained; i.e., it is unfair when CD models narrow or widen the gap of proficiency. In Figures 1(b)–(d), A and B depict two student groups separated by a sensitive attribute (e.g., gender); B possesses a lower level of proficiency compared to A, which can be attributed to the lack of study resources. It is deemed unfair that the predicted proficiency gap between these two groups is narrowed or widened, as shown in Figures 1(c) and (d), which may further lead to unfair education outcomes (e.g., college admission) in practical scenarios [11].

In this paper, we comprehensively analyze the fairness issue in CD and try to answer two questions: (1) Are the results of existing CD models affected by sensitive attributes? (2) If yes, how can we mitigate the impact of sensitive attributes to ensure fair diagnosis results? Specifically, to address the first question, we conduct experimental CD studies on the public Program for International Student Assessment (PISA) dataset. With rigorously controlled experiments, we can draw the conclusion that unfairness indeed exists in CD models, and they exhibit different unfair performances (i.e., they narrow or widen the gap). In order to explain this phenomenon, we conduct a theoretical analysis and reveal that model complexity results in different unfair performances.

After verifying that CD results are indeed affected by sensitive attributes, we attempt to answer the second question. In an attempt to address this question, several approaches have been proposed in the literature, such as adversarial learning [14–16] and regularization methods [17], among which adversarial learning-based methods reveal their theoretical elegance and have achieved widespread success in domains such as recommender system [17], healthcare [18]. For instance, Shao et al. [15] introduced an adversarial framework, FairCF, to analyze fair representations independent of sensitive attributes for the recommendation. Because of the success of adversary learning approaches, we appropriate and expand the application of this technology into CD to eliminate unfairness. However, due to the instability of the training process of adversarial learning [19] and the complex modeling process in cognitive diagnostics [20], the diagnosis results may still contain biased information about sensitive user attributes in practice. To address this challenge, we propose a decomposed adversarial learning-based CD framework, FairCD.

Specifically, we categorize student proficiency in existing CD models into two components: bias proficiency and fair proficiency, and design two orthogonal tasks for each of them to ensure fairness in proficiency independent of sensitive attributes. This is then taken as the final diagnosed result. More concretely, we adopt an adversarial learning task, ensuring fair proficiency to directly eliminate the effect of sensitive attributes. Moreover, an attribute prediction task is applied for bias proficiency to capture biases related to sensitive attributes, which further ensures fair proficiency independent of sensitive attributes. Finally, we conduct extensive experiments on PISA, and the results demonstrate the effectiveness of FairCD.

The main contributions of this work are as follows.

• **Fairness exploration.** To the best of our knowledge, fairness in CD has not been investigated in any prior research. We first realize the existence of unfairness in CD models and confirm the presence of various unfair performances. Then, we conduct a theoretical analysis to explain this phenomenon.

• **Fairness improvement.** We propose a fairness-aware cognitive diagnosis framework, FairCD, which can enhance the fairness of all CD models.

• **Fairness evaluations.** We perform comprehensive experimental evaluations on PISA to demonstrate the effectiveness of our work in maintaining both utility and fairness.

2 Related work

2.1 Cognitive diagnosis

The CD serves as a fundamental scientific topic in many real-world educational scenarios, such as student assessment [21–25] and educational recommender systems [7,26]. It is largely derived from psychometrics in the early years, with IRT [1] being one of the standard CD models commonly used in GRE [27]. IRT defined student-exercise interactions using manually constructed functions such as the logistic function and considered each student as a single latent characteristic. Subsequently, MIRT [8] was proposed by extending the single trait features in IRT into multiple dimensions. Although these models were designed effectively based on psychometric theories and the diagnostic results were well interpreted, these studies relied on handcrafted interaction functions and could only exploit users' numerical response records [28]. With the development of machine learning, researchers turned to develop several CD models to address these problems from a machine learning perspective, among which NeuralCD [29] was one of the typical models. NeuralCD used neural networks to learn the interactions between students and exercises and got satisfactory results. Based on NeuralCD, several methods were proposed. Ghosh et al. [30] extended current methods beyond the scope of accuracy prediction to accurately predict the options students choose in multi-choice problems, which could detect individual student errors. Gao et al. [20] proposed a relation map-driven CD to capture the relation between the student-exercise concept. Cheng et al. [31] proposed IK-NeuralCD to express the importance of knowledge points in student modeling, which improved the degree of fitting the complex relationship between students and exercises. However, although existing models can generate satisfactory reports, the fairness issue remains underexplored.

2.2 Fairness in machine learning

As machine learning continues to be widely used in modern society [32–36], researchers realize the significance of fairness [17,37–40]. Existing studies on fairness can be concluded from two perspectives: the definition of fairness and techniques for improving fairness. Fairness definition can be divided into two categories: (1) individual fairness, which requires similar individuals to be treated similarly [41], and (2) group fairness, which requires that the disadvantaged group be treated similarly to the advantaged group [12]. In this work, we focus on group fairness in CD. The most representative definitions of group fairness include demographic parity [41], equal opportunity [13], and equalized odds [13]. Demographic parity required that all subgroups received the same proportion of positive outcomes. However, it was limited as the base rates of subgroups differed. Equalized odds attempted to maintain consistency in both true positive and false positive rates across different subgroups, which might not directly address the issue of equitable access to educational opportunities. Equalized opportunities aimed to ensure that people with equivalent qualifications or abilities possessed equal chances of achieving positive outcomes, regardless of their sensitive characteristics. This metric is particularly relevant in the context of education, as it seeks to create a level playing field for all students to achieve their full potential. Thus, we extend the definition of equalized opportunities to our educational application. Approaches can be classified

Table 1 Statistics of the datasets

Dataset	Group	# Students	# Questions	# Records
PISA-OECD	OECD	159897	183	4668107
	Non-OECD	158593	183	4443408
PISA-GENDER	Female	202442	183	5746608
	Male	201167	183	5741626

based on the stage at which the mechanism operates [42, 43], which is divided into pre-processing [44], in-processing [45], and post-processing approaches [46]. Pre-processing approaches attempt to remove underlying bias from training data before the learning process. In-processing approaches attempt to revise the training of the models to achieve fairness. Post-processing methods directly change the predictive labels of trained models to mitigate unfairness. We hold the distinction of being the initial researchers to investigate the fairness of CD models and suggest an in-processing approach based on deconstructed adversarial learning to ensure the fairness of all CD models.

2.3 Educational fairness

Education is a major factor in determining how people spend their adult lives. A greater level of education translates into higher incomes, better health, and a longer life [47, 48]. It remains doubtful that any child can reasonably be expected to succeed in life if they are deprived of education opportunities [49]. Fairness in education implies ensuring that personal and social circumstances (e.g., gender) should not be an obstacle for an individual to fulfill their educational potential, which is a principle along the same vein as the concept of equal opportunity [11]. This definition is also followed in this paper. Extensive investigations have been conducted on educational fairness. For example, Hutt et al. [50] aimed at predicting on-time graduation from college applications. Hu et al. [51] investigated the fairness of identifying at-risk students. Yu et al. [52] analyzed fair prediction of college success from different sources of student data. Li et al. [53] proposed a fair logistic regression model to address the fairness of AI prediction in education. Gómez et al. [54] researched the understudied impact of recommender systems on instructors' exposure to online platforms and offered a method for promoting fairness in recommendation visibility and exposure across countries. However, limited work was conducted to investigate fairness in the fundamental educational research area of CD, which revealed important application potential in various educational contexts.

3 Preliminaries

3.1 Data description

PISA[1)] is one of the most famous worldwide testing programs, gaining honorary status as the Olympic Games in testing projects and attracting nearly one hundred regions or countries. Specifically, PISA measures 15-year-olds' ability on several topics including reading and science. Moreover, it provides questionnaires to collect students' contexts that contain some sensitive attributes (e.g., gender, region).

In this paper, we obtain a public dataset of PISA 2015[2)] for our study, which focuses on diagnosis assessment on the "science" topic. We select two sensitive attributes, i.e., gender and region, to explore the fairness issue of CD. We specifically categorize data into male and female gender groups, OECD groups, and Non-OECD groups based on whether the region is part of the Organization for Economic Co-operation and Development (OECD) countries. To avoid the influence of students answering different questions, we chose 57 locations where students answered the same questions, filtered out students with missing sensitive information, and created two datasets (i.e., PISA-GENDER, PISA-OECD). The basic statistics are shown in Table 1.

3.2 Cognitive diagnosis

In this subsection, we formally define the CD problem. Assume there are N students, M exercises, and K knowledge concepts, which are defined as $U = \{u_1, u_2, \ldots, u_N\}, E = \{e_1, e_2, \ldots, e_M\}$, and $K =$

1) https://www.oecd.org/pisa/.
2) https://www.oecd.org/pisa/data/2015database/.

$\{k_1, k_2, \ldots, k_K\}$. The response logs R are a set of triplets (u, e, y), where y is the score obtained by student u on exercise e. Given response logs R, the goal of CD is to mine students' proficiency.

There are two main categories of CD models: traditional methods that rely on manually designed functions and deep learning-based methods that model complex cognitive interactions. We choose two models to investigate fairness in CD: IRT from the first category, which has been widely implemented in GRE [27], and NeuralCD from the second category, which serves as the foundation for several advanced models [20, 30]. Please note that our investigation of these two most representative methodologies is broad enough to be applied to other CD models. The model details are as follows.

IRT [1, 55] models each student i as a proficiency variable θ_i, each exercise as a discriminating factor a_j and a difficulty factor b_j, and a logistic function is used to forecast the likelihood that student i will answer exercise j correctly based on a logistic function[3)]:

$$\hat{y}_{ij} = 1/(1 + \mathrm{e}^{a_j(\theta_i - b_j)}). \tag{1}$$

NeuralCD [10] is a novel deep CD model that generalizes the student's proficiency and exercise parameters into high dimensions and adds neural networks to learn their complicated interactions. Furthermore, NeuralCD requires a Q-matrix (often provided by experts) $Q \in \mathbb{R}^{M \times K}$, where $Q_{ij} = 1$ if exercise e_i is related to the knowledge concept k_j and $Q_{ij} = 0$ if not.

$$\hat{y}_{ij} = f(Q_j \circ (h_i^s - h_j^{\mathrm{diff}}) \times h_j^{\mathrm{disc}}), \tag{2}$$

where $h_i^s \in \mathbb{R}^K$ is the proficiency vector of student i, $h_j^{\mathrm{diff}} \in \mathbb{R}^K$ and $h_j^{\mathrm{disc}} \in \mathbb{R}^1$ are difficulty and discrimination factors of exercise j, respectively, $\circ$ is the element-wise product, Q_j is exercise j's factor that arises from the Q-matrix, and f represents multiple full connection layers.

3.3 Fairness in cognitive diagnosis

In this paper, we consider group fairness in CD. Here, we divide students into subgroups based on sensitive attributes. For simplicity, we consider binary-sensitive factors such as gender[4)]. The subgroups can be represented by A and B. Following the core idea of the classical fairness definition of equal opportunity [13], we propose the FairCD definition.

Definition 1 (FairCD)**.** A CD model is considered to be fair if the gap between true proficiency and predicted proficiency is identical across both groups (i.e., $F_{\mathrm{CD}} = 0$).

$$F_{\mathrm{CD}} = \left(\frac{1}{|A|}\sum_{i \in A}\hat{Y}_i - \frac{1}{|B|}\sum_{i \in B}\hat{Y}_i\right) - \left(\frac{1}{|A|}\sum_{i \in A}Y_i - \frac{1}{|B|}\sum_{i \in B}Y_i\right), \tag{3}$$

where Y_i, $\hat{Y}_i$ are the actual correct rate, the predicted correct probability by CD for the i-th student from groups A or B. For instance, suppose student i answers ten questions. He/she accurately answered five of them, although a CD model indicates that he/she will answer four questions. Then, Y_i equals 0.5 and $\hat{Y}_i$ equals 0.4. Owing to the fact that we cannot obtain the true proficiency of students, we use the correct rate and predicted correct probability to represent the true proficiency and predicted proficiency. This alternative approach is based on the assumption that these two groups of students answer the same questions as obtained by our dataset preprocessing in the data description. The same approach has been widely implemented in many other instances, such as the recommender system [17].

In Definition 1, the closer F_{CD} is to 0, the fairer this model is. Meanwhile, $F_{\mathrm{CD}} > 0$ indicates a wider gap, while $F_{\mathrm{CD}} < 0$ indicates a narrower gap. For example, in Table 2, the original gap between A and B is 0.1, and the cognitive diagnosis model predicted gap is 0.25; thus, F_{CD} is 0.15. As a result, the CD model widens the distance between A and B. The original gap between B and C is 0.1, but the cognitive diagnostic model projected a gap of 0; therefore, F_{CD} is -0.1, indicating that the cognitive diagnosis model closes the gap between B and C.

3) Here we adopt two-parameter logistic IRT model.

4) Our definition is easily extensible to different types of sensitive properties.

Table 2 Example of fairness in cognitive diagnosis

Group	Student	Actual correct rate		Predicted correct probability	
		Individual	Group	Individual	Group
A	a_1	0.8	0.7	0.9	0.8
	a_2	0.6		0.7	
B	b_1	0.7	0.6	0.6	0.55
	b_2	0.5		0.5	
C	c_1	0.5	0.5	0.6	0.55
	c_2	0.5		0.5	

Table 3 F_{CD} of NeuralCD and IRT[a)]

Group	NeuralCD	IRT
OECD/Non-OECD	0.0279 ↑	−0.0852 ↓
Male/Female	0.0059 ↑	−0.0115 ↓

a) ↑ means widening the gap; ↓ means narrowing the gap.

Table 4 Variances of predicted correct probability for IRT and NeuralCD

Model	PISA-OECD	PISA-GENDER
IRT	0.0088	0.0065
NeuralCD	0.0554	0.0561

4 Understanding fairness in cognitive diagnosis

In this section, we attempt to answer the first question: are the results of existing CD models affected by sensitive attributes? We train IRT and NeuralCD on PISA and compute the F_{CD} metric in (3) for each CD model; the results are shown in Table 3. We believe the metric reflects the fact that NeuralCD and IRT are unfair. Meanwhile, these two CD models exhibit distinct unfair phenomena (i.e., NeuralCD widens the gap, and IRT narrows the gap).

To better understand the fairness in cognitive diagnosis, we need to figure out why different models have different fairness performance. We discover that the outcomes predicted by IRT in different groups are essentially identical, while the results predicted by NeuralCD are vastly different. Therefore, we compute the variances of expected results of different models. The results are shown in Table 4. Here, let Var denote the variances of predicted correct probability. In each dataset, Var(NeuralCD) > Var(IRT). Inspired by this revelation, we intend to establish a link between variance and various unfair performance. Furthermore, we can explore the reasons behind different unfair performance. As such, we provide the following Lemma 1.

Lemma 1. Let $\hat{Y}_A, \hat{Y}_B$ indicate the projected accurate probability of groups A and B predicted by a CD model, and $\hat{g}$ represent the predicted correct probability gap (i.e., $\hat{Y}_A - \hat{Y}_B$), g represent the actual correct rate gap between A and B. Assume $\hat{Y}_A, \hat{Y}_B$ i.i.d. $\sim N(\mu, \sigma^2)$. Therefore, (1) the smaller σ, the greater the probability of closing the gap (i.e., $P(\hat{g} < g)$); (2) the larger σ, the greater the probability of increasing the gap (i.e., $P(\hat{g} > g)$).

Proof. Since $\hat{Y}_A, \hat{Y}_B$ i.i.d. $\sim N(\mu, \sigma^2)$, we have $\hat{g} \sim N(0, 2\sigma^2)$; then, we can get $\frac{\hat{g}}{\sqrt{2}\sigma} \sim N(0,1)$. $P(\hat{g} < g) = P(\frac{\hat{g}}{\sqrt{2}\sigma} < \frac{g}{\sqrt{2}\sigma}) = \int_{-\infty}^{\frac{g}{\sqrt{2}\sigma}} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{(-\frac{x^2}{2})} \mathrm{d}x$. Because g is calculated from training records, it can be considered a constant. Therefore, when σ decreases, $\int_{-\infty}^{\frac{g}{\sqrt{2}\sigma}} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{(-\frac{x^2}{2})} \mathrm{d}x$ increases; i.e., $P(\hat{g} < g)$ increases. $P(\hat{g} > g) = P(\frac{\hat{g}}{\sqrt{2}\sigma} > \frac{g}{\sqrt{2}\sigma}) = \int_{\frac{g}{\sqrt{2}\sigma}}^{\infty} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{(-\frac{x^2}{2})} \mathrm{d}x$. When σ increases, $\int_{\frac{g}{\sqrt{2}\sigma}}^{\infty} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{(-\frac{x^2}{2})} \mathrm{d}x$ gets larger. Thus, we have $P(\hat{g} > g)$ increases. This completes the proof.

Lemma 1 can be better understood through an extreme example. Assume the variance of a CD model's anticipated accurate probability is equal to zero; hence, we may infer that all predicted results of CD models are the same, and the difference between different groups is reduced to 0.

Using Lemma 1 and the results (i.e., Var(NeuralCD) > Var(IRT)) in Table 4, the reason behind NeuralCD and IRT possessing different unfair performance can be transformed into the reason behind the larger observed variance in NeuralCD compared to IRT. The following theoretical analysis is obtained.

Theorem 1. Assume the parameters in NeuralCD $h^s - h^{\mathrm{diff}} \sim N(0, \sigma^2 I)$; the model complexity leads

to Var(NeuralCD) > Var(IRT). Further, we conclude that model complexity leads to NeuralCD widening the gap and IRT narrowing the gap.

Proof. The interaction layer in NeuralCD is defined as $x = Q_j \circ (h_i^s - h_j^{\text{diff}}) \times h_j^{\text{disc}}$; for simplicity, we only consider NeuralCD with one connection layer and the sigmoid activation function $f_1 = \delta(w^{\mathrm{T}} x + b)$. We assume $h^s - h^{\text{diff}} \sim N(0, \sigma^2 I)$; thus, we have $x \sim N(0, (\sigma \cdot h^{\text{disc}})^2 \text{diag}(Q)^2)$. The distribution can therefore be estimated as $w^{\mathrm{T}} x + b \sim N(b, w^{\mathrm{T}} (\sigma \cdot h^{\text{disc}})^2 \text{diag}(Q)^2 w)$. A second order Taylor expansion is applied to δ at the origin $\delta(x) \approx \delta(0) + x\delta'(0) + \frac{x^2}{2}\delta''(0) = \frac{1}{2} + \frac{1}{4}x$. We can approximate that $\text{Var}(f_1) = \frac{1}{16} w^{\mathrm{T}} (\sigma \cdot h^{\text{disc}})^2 \text{diag}(Q)^2 w$ by combining these results. Finally, we obtain $\text{Var}(f_1) \approx \frac{1}{16} (\sigma \cdot h^{\text{disc}})^2 \sum_{i=1}^{K} (w_i \cdot q_i)^2$. When the dimension K is set to 1 and the parameters w, q are both set to 1, NeuralCD degenerates into IRT. As a result, the dimension and number of parameters in NeuralCD result in a higher variance as compared to IRT. In the meantime, these two factors reflect the model's complexity. Thus, we can attribute the inequality Var(NeuralCD) > Var(IRT) to model complexity. Combining it with Lemma 1, we can deduce that the model complexity causes NeuralCD to widen the gap and IRT to close it. This completes the proof.

Theorem 1 underscores a vital consideration for CD researchers: the complexity of the models they develop can considerably impact fairness. As a result, when developing novel CD models, it is essential to examine the potential effects of their complexity on performance in various settings and among diverse populations. This evaluation facilitates the improvement of more equitable and fair CD tools that can effectively address the needs of a broad range of individuals and circumstances.

5 Improving fairness in cognitive diagnosis

After confirming the existence of unfair phenomena in CD, we now address the second question: how can we mitigate the impact of sensitive attributes to ensure fair diagnosis results? There are two requirements: (1) the method should be model agnostic and can enhance the fairness of all the CD models; (2) the strategy should achieve fairness while keeping proficiency estimates accurate.

We summarize the commonalities of CD models to satisfy these two requirements. In general, CD models comprise three parts: student proficiency $\boldsymbol{\theta}^{\mathrm{u}}$, exercise factors $\boldsymbol{e}$, and the interaction function f [56]. Different CD models are distinguished by the different methods by which these three components are modeled. Taking IRT (introduced in preliminaries) as an example, student proficiency corresponds to a single dimension variable θ_i, exercise factors correspond to discrimination factor a_j and difficulty factor b_j, and the interaction function is (1). Based on these commonalities, the most direct method is to eliminate the impact of sensitive attributes on student proficiency $\boldsymbol{\theta}^{\mathrm{u}}$ based on adversarial learning, which consists of two components: (1) a trained filter module that filters the effect of sensitive attributes on student competence $\boldsymbol{\theta}^{\mathrm{u}}$; (2) a discriminator module that attempts to forecast the corresponding qualities based on the filtered student proficiency $\boldsymbol{\theta}^{\mathrm{u}}$. Through adversarial training, the effect of sensitive attributes can be removed. However, because the training process of adversarial learning can be unstable [19], the student proficiency $\boldsymbol{\theta}^{\mathrm{u}}$ may still contain biased knowledge about sensitive user qualities in practice.

To address this issue, we present FairCD, a deconstructed adversarial learning-based CD framework, FairCD, which can further ensure that $\boldsymbol{\theta}^{\mathrm{u}}$ does not contain information about sensitive attributes information. The architecture of FairCD is shown in Figure 2. We divide student proficiency $\boldsymbol{\theta}^{\mathrm{u}}$ in existing CD models into two components in this architecture: bias proficiency $\boldsymbol{\theta}^{\mathrm{b}}$ and fair proficiency $\boldsymbol{\theta}^{\mathrm{f}}$. Bias proficiency $\boldsymbol{\theta}^{\mathrm{b}}$ seeks to gather as much biased information regarding sensitive user attributes as possible, whereas $\boldsymbol{\theta}^{\mathrm{f}}$ aims to reduce the effect of sensitive user attributes as much as possible. Through these two components $\boldsymbol{\theta}^{\mathrm{f}}$, $\boldsymbol{\theta}^{\mathrm{b}}$, we can ensure $\boldsymbol{\theta}^{\mathrm{f}}$ independent of sensitive attributes and accept it as the ultimate student proficiency. To achieve this goal, we design two orthogonal tasks for $\boldsymbol{\theta}^{\mathrm{f}}$, $\boldsymbol{\theta}^{\mathrm{b}}$. We use an adversarial learning task for $\boldsymbol{\theta}^{\mathrm{f}}$ to directly remove the effect of sensitive attributes. Also, an attribute prediction task is applied for $\boldsymbol{\theta}^{\mathrm{b}}$ to capture biases related to sensitive attributes, which further ensures $\boldsymbol{\theta}^{\mathrm{f}}$ independent of sensitive attributes.

In the following section, we introduce these two tasks, respectively. Subsequently, we explain the approach behind integrating pre-existing CD models into the FairCD framework. Finally, we present the entire training algorithm of FairCD and provide the corresponding pseudocode.

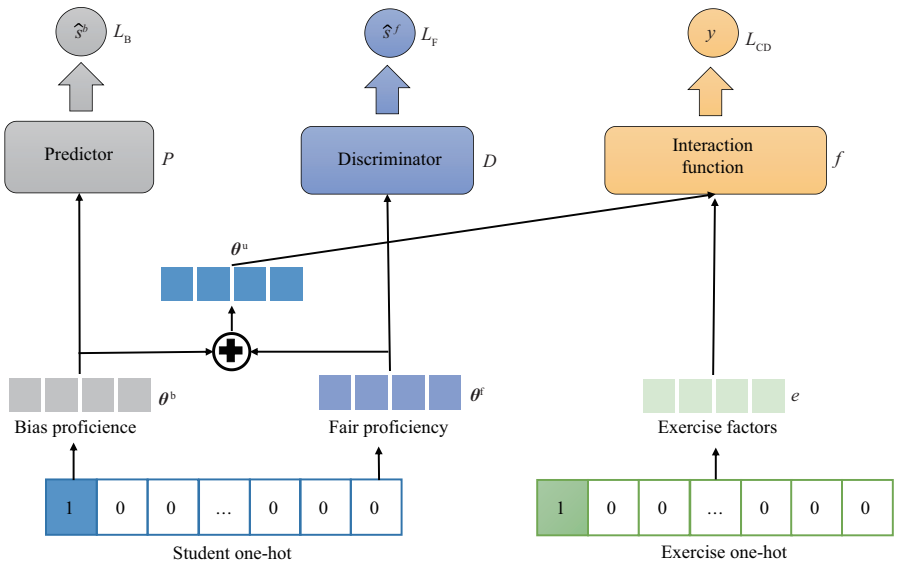

Figure 2 (Color online) Architecture of FairCD.

5.1 Adversarial learning task

Fair proficiency $\boldsymbol{\theta}^{\mathrm{f}}$ aims to eliminate the effect of sensitive user attributes as much as possible. First, we calculate each student's original fair proficiency by multiplying the student's one-hot representation vector $\boldsymbol{x}^{\mathrm{s}}$ by a trainable matrix $\boldsymbol{F}$:

$$\boldsymbol{\theta}^{\mathrm{f}} = \mathrm{sigmoid}(\boldsymbol{x}^{\mathrm{s}} \times \boldsymbol{F}). \tag{4}$$

We then use an adversarial learning task to reduce the effect of sensitive attributes for $\boldsymbol{\theta}^{\mathrm{f}}$. More specifically, we use an attribute discriminator which attempts to predict sensitive user attributes from fair proficiency. Our goal is to encourage the discriminator $\mathcal{D}$ to avoid predicting sensitive information. Thus, we can generate fair proficiency $\boldsymbol{\theta}^{\mathrm{f}}$, which does not contain sensitive information. To accomplish this, we maximize the cross entropy $\mathcal{L}_{\mathcal{F}}$:

$$\begin{aligned} \hat{s}^{\mathrm{f}} &= \mathrm{softmax}(\boldsymbol{W}^{\mathrm{f}}\boldsymbol{\theta}^{\mathrm{f}} + \boldsymbol{b}^{\mathrm{f}}), \\ \mathcal{L}_{\mathcal{F}} &= -\sum_i \sum_j s_{ij} \log \hat{s}^{\mathrm{f}}_{ij}, \end{aligned} \tag{5}$$

where $\boldsymbol{W}^{\mathrm{f}}$ and $\boldsymbol{b}^{\mathrm{f}}$ are parameters, $\hat{s}^{\mathrm{f}}$ is the predicted probability vector, s_{ij} and $\hat{s}^{\mathrm{f}}_{ij}$ represent ground truth and predicted probability of the i-th student's sensitive attribute in the j-th class.

5.2 Prediction task

Although an adversarial learning task was used to eliminate the effect of sensitive user attributes on $\boldsymbol{\theta}^{\mathrm{f}}$, biased information about sensitive user attributes may still be leaked into $\boldsymbol{\theta}^{\mathrm{f}}$ due to the unstable nature of the adversarial training process [19]. To solve this issue, we use an additional predictor $\mathcal{P}$ for bias proficiency $\boldsymbol{\theta}^{\mathrm{b}}$ to collect as much biased information as possible, ensuring that $\boldsymbol{\theta}^{\mathrm{f}}$ is independent of sensory qualities. Similar to $\boldsymbol{\theta}^{\mathrm{f}}$, $\boldsymbol{\theta}^{\mathrm{b}}$ is obtained by multiplying the student's one-hot representation vector $\boldsymbol{x}^{\mathrm{s}}$ with a trainable matrix $\boldsymbol{B}$. That is,

$$\boldsymbol{\theta}^{\mathrm{b}} = \mathrm{sigmoid}(\boldsymbol{x}^{\mathrm{s}} \times \boldsymbol{B}). \tag{6}$$

The loss function of the attribute prediction task is similar to the discriminator. The purpose of the attribute prediction task, unlike the adversarial learning task, is to capture more biased information linked to sensitive attributes; therefore, we directly minimize $\mathcal{L}_{\mathcal{B}}$ in model training.

$$\begin{aligned} \hat{s}^{\mathrm{b}} &= \mathrm{softmax}(\boldsymbol{W}^{\mathrm{b}}\boldsymbol{\theta}^{\mathrm{b}} + \boldsymbol{b}^{\mathrm{b}}), \\ \mathcal{L}_{\mathcal{B}} &= -\sum_i \sum_j s_{ij} \log \hat{s}^{\mathrm{b}}_{ij}, \end{aligned} \tag{7}$$

Algorithm 1 Detailed training procedures of FairCD

Require: Students U; exercises E; response logs $\boldsymbol{R}$; user sensitive labels S; training epochs M; discriminator training steps T;
Ensure:
Random initialize fair proficiency $\boldsymbol{\theta}^{\mathrm{f}}$ and bias proficiency $\boldsymbol{\theta}^{\mathrm{b}}$, exercise factors $\boldsymbol{e}$, discriminator's parameters $\boldsymbol{\theta}_{\mathrm{D}}$, predictor's parameters $\boldsymbol{\theta}_{\mathrm{P}}$;
for each epoch **do**
  Sample a batch of training data, including students-exercises pairs and corresponding responses and sensitive labels (u_i, e_j, y_{ij}, s);
  for each pair of input (u_i, e_j, y_{ij}, s) in the batch **do**
    Reconstruct the student proficiency $\boldsymbol{\theta}^{\mathrm{u}}$ through (8);
    Compute the cognitive diagnosis models loss $\mathcal{L}_{\mathrm{CD}}$ (9);
    Optimize $\boldsymbol{\theta}^{\mathrm{f}}$, $\boldsymbol{\theta}^{\mathrm{b}}$, $\boldsymbol{e}$ to minimize cognitive diagnosis loss $\mathcal{L}_{\mathrm{CD}}$ with $\boldsymbol{\theta}_{\mathrm{D}}$, $\boldsymbol{\theta}_{\mathrm{P}}$ fixed;
    for each discriminator training step **do**
      Compute the discriminator loss $\mathcal{L}_{\mathcal{F}}$ (5);
      Optimize $\boldsymbol{\theta}_{\mathrm{D}}$ to minimize discriminator loss $-\mathcal{L}_{\mathcal{F}}$ with $\boldsymbol{\theta}^{\mathrm{f}}$, $\boldsymbol{\theta}^{\mathrm{b}}$, $\boldsymbol{e}$ fixed;
      Compute the predictor loss $\mathcal{L}_{\mathcal{B}}$ (7);
      Optimize $\boldsymbol{\theta}_{\mathrm{P}}$ to minimize predictor loss $\mathcal{L}_{\mathcal{B}}$ with $\boldsymbol{\theta}^{\mathrm{f}}$, $\boldsymbol{\theta}^{\mathrm{b}}$, $\boldsymbol{e}$ fixed;
    end for
  end for
end for

where $\boldsymbol{W}^{\mathrm{b}}$ and $\boldsymbol{b}^{\mathrm{b}}$ are parameters, $\hat{s}^{\mathrm{b}}$ is the predicted probability vector, s_{ij} and $\hat{s}^{\mathrm{b}}_{ij}$ represent ground truth and predicted probability of the i-th student's sensitive attribute in the j-th class.

5.3 FairCD integration

Finally, we introduce a method to integrate existing CD models into the FairCD framework. After obtaining $\boldsymbol{\theta}^{\mathrm{f}}$ and $\boldsymbol{\theta}^{\mathrm{b}}$, we combine them to recreate the student proficiency $\boldsymbol{\theta}^{\mathrm{u}}$ in the original CD model:

$$\boldsymbol{\theta}^{\mathrm{u}} = \boldsymbol{\theta}^{\mathrm{f}} + \boldsymbol{\theta}^{\mathrm{b}}. \tag{8}$$

The likelihood of the student u_i successfully answering exercise e_j can be predicted by $\hat{y}_{ij} = f(\boldsymbol{\theta}^{\mathrm{u}}, \boldsymbol{e})$, where f and $\boldsymbol{e}$ are the interaction function and exercise factors inherited from the previous model. In the case of IRT, the interaction function is (1), and exercise factors are the discrimination factor a_j and difficulty factor b_j. Moreover, we need to maintain the accuracy of proficiency estimates. To accomplish this, we use the student performance prediction task to train cognitive diagnostic models. CD models are expected to minimize the difference between the anticipated probability $\hat{y}_{ij}$ and the true answer y_{ij}. The loss function of CD to maintain accuracy is as follows:

$$\mathcal{L}_{\mathrm{CD}} = -\sum_i \sum_j (y_{ij} \log \hat{y}_{ij} + (1 - y_{ij}) \log(1 - \hat{y}_{ij})). \tag{9}$$

FairCD strives toward fairness while maintaining the accuracy of proficiency estimations. Thus, combining (5), (7), and (9), our final loss function in FairCD can be formulated as follows:

$$\mathcal{L} = \mathcal{L}_{\mathrm{CD}} + \lambda_{\mathrm{B}} \mathcal{L}_{\mathcal{B}} - \lambda_{\mathrm{F}} \mathcal{L}_{\mathcal{F}}, \tag{10}$$

where adversarial coefficients λ_{B} and λ_{F} are hyperparameters that serve to control the tradeoff between diagnosis accuracy and fairness. Herein, we discuss the effect of λ_{B}, λ_{F} in Experiments (RQ3). Because all cognitive diagnostic models have three components $\boldsymbol{\theta}^{\mathrm{u}}$, $\boldsymbol{e}$, f, our FairCD is a general fairness-aware cognitive diagnosis that can improve the fairness of all CD models.

5.4 Training algorithm

In our implementation, we use mini-batch training for adversarial learning. Specifically, for each batch, we first feed the input to the cognitive diagnosis models to obtain $\mathcal{L}_{\mathrm{CD}}$, $\mathcal{L}_{\mathcal{B}}$, and $\mathcal{L}_{\mathcal{F}}$. The parameters of the discriminator and predictor are then fixed, and the cognitive diagnosis models are optimized by minimizing $\mathcal{L}_{\mathrm{CD}}$. Subsequently, $\mathcal{L}_{\mathcal{B}}$ and $-\mathcal{L}_{\mathcal{F}}$ are minimized for T steps while the parameters of CD models are fixed. Here, $T = 10$ in our implementation. The pseudocode for the entire training algorithm is given in Algorithm 1.

6 Experiments

In this section, we first introduce the experimental setup, following which we conduct experiments using the PISA dataset, comparing FairCD to other baselines to answer the following questions.

RQ1: Does FairCD outperform the fairness-aware baselines on both model utility and fairness?

RQ2: Can FairCD improve the fairness of CD applications (e.g., educational recommender system)?

RQ3: How will adversarial learning and prediction tasks influence FairCD?

RQ4: Does bias proficiency (i.e., $\boldsymbol{\theta}^{\mathrm{b}}$) contain useful information?

6.1 Experimental setup

Parameter settings and baselines. To establish a training procedure. The parameters are initialized using Xavier initialization [57], which fills the weights with random values sampled from $\mathcal{N}(0, \mathrm{std}^2)$, where $\mathrm{std} = \sqrt{2/(n_{\mathrm{in}} + n_{\mathrm{out}})}$. n_{in} is the number of neurons feeding into the weights, and n_{out} is the number of neurons fed the results. To evaluate the generalization of our method, we adopt IRT, NeuralCD, and MIRT models; MIRT is a multidimensional extension of IRT. For convenience, we use NCD to represent NeuralCD in the experiment. In terms of model parameter configuration, NCD adheres to the settings outlined in [10]. The fully connected layers have dimensions of 512, 256, and 1, respectively. The sigmoid function also serves as the activation function for all layers. MIRT's dimension of student proficiency parameters matches those in NeuralCD. For all datasets and models, we set the learning rate to 0.001 and the dropout rate to 0.2. We apply Adam as the optimization algorithm to update the model parameters. The discriminator $\mathcal{D}$ and predictor $\mathcal{P}$ in the deconstructed adversarial architecture are three-layer perceptrons with the activation function of LeakyReLU. We set the dropout rate to 0.1 and the slope of the negative section for LeakyReLU to 0.2 for them. The loss coefficients λ_{B}, λ_{F} in (10) are set to 0.2 and 0.4, respectively. We implement all models with PyTorch by Python and conduct our experiments on a Linux server with four 2.0 GHz Intel Xeon E5-2620 CPUs and a Tesla K20m GPU.

To the best of our knowledge, we hold the distinction of being the initial researchers delving into the field of fairness in cognitive diagnostics. In an effort to reveal the effectiveness of FairCD, we also compare FairCD with the following baselines.

- CD: original cognitive diagnosis models (i.e., IRT, MIRT, NeuralCD) that do not consider fairness.
- CD+REG: a well-known fairness improvement strategy that considers the fairness metric as a regularization for the loss and has been used in prior fairness studies [17, 46]. In our work, we use (3) as a regularization for (9).
- CD+DP: a fairness improving method regards demographic parity as a regularization [17].
- CD+AD: an adversarial learning method that serves to eliminate the effect of sensitive attributes. Specifically, a filter module that has been trained to filter the effect of sensitive attributes on student competence $\boldsymbol{\theta}^{\mathrm{u}}$ and a discriminator module that attempts to forecast corresponding attributes based on filtered student proficiency [14]. The adversarial architecture is the same as our experimental setting.
- FairCD ($\boldsymbol{\theta}^{\mathrm{b}}$): we use the bias proficiency ($\boldsymbol{\theta}^{\mathrm{b}}$) in our FairCD as input for students' performance prediction task.
- FairCD ($\boldsymbol{\theta}^{\mathrm{f}}$): we use the fair proficiency ($\boldsymbol{\theta}^{\mathrm{f}}$) in our FairCD as the final proficiency estimate for students' performance prediction task.

In summary, we compare FairCD with original CD models that do not consider fairness. CD+REG and CD+DP improve fairness in fairness-aware baselines by including alternative fairness regularization terms. CD+AD employs adversarial training methods to eliminate sensitive information.

Experimental evaluation. In terms of model evaluation, we employ $\boldsymbol{\theta}^{\mathrm{f}}$ as the final proficiency estimate. The evaluation can be divided into two parts. (1) Evaluation of accuracy. Because we cannot obtain the true proficiency of students, we utilize students' performance prediction tasks to demonstrate the utility of cognitive diagnostic, as in previous studies [10, 20, 58]. We adopt different metrics from the perspectives of regression and classification. From the regression perspective, we select MAE and RMSE to quantify the difference between predicted and actual scores. From the classification perspective, we consider that the incorrect and correct answers can be represented as 0 and 1, respectively, and we use AUC and ACC for model evaluation. (2) Fairness evaluation. We are interested in whether FairCD can promote fairness. For convenience, we use the $|F_{\mathrm{CD}}|$ metric to measure the fairness of cognitive diagnosis models. The closer this metric is to 0, the fairer the CD model is.

Table 5 Utility results in different datasets[a)]

	PISA-OECD				PISA-GENDER			
	AUC	ACC	MAE	RMSE	AUC	ACC	MAE	RMSE
IRT	0.715	0.661	0.415	0.462	0.716	0.660	0.417	0.461
IRT+REG	0.692	0.641	0.429	0.471	0.692	0.641	0.441	0.468
IRT+DP	0.691	0.640	0.437	0.467	0.683	0.633	0.451	0.475
IRT+AD	0.694	0.645	0.431	0.445	0.697	0.644	0.433	0.472
IRT+FairCD ($\boldsymbol{\theta}^{\mathrm{b}}$)	0.589	0.593	0.479	0.494	0.583	0.576	0.478	0.483
IRT+FairCD ($\boldsymbol{\theta}^{\mathrm{f}}$)	**0.704**	**0.651**	**0.428**	**0.465**	**0.713**	**0.659**	**0.420**	**0.463**
MIRT	0.737	0.668	0.3663	0.452	0.754	0.694	0.362	0.449
MIRT+REG	0.701	0.646	0.423	0.467	0.721	0.660	0.413	0.460
MIRT+DP	0.726	0.665	0.403	0.463	0.711	0.653	0.423	0.463
MIRT+AD	0.719	0.656	0.413	0.461	0.723	0.663	0.411	0.459
MIRT+FairCD ($\boldsymbol{\theta}^{\mathrm{b}}$)	0.598	0.613	0.456	0.478	0.596	0.606	0.448	0.487
MIRT+FairCD ($\boldsymbol{\theta}^{\mathrm{f}}$)	**0.727**	**0.667**	**0.394**	**0.462**	**0.739**	**0.674**	**0.384**	**0.458**
NCD	0.772	0.702	0.344	0.447	0.761	0.704	0.345	0.446
NCD+REG	0.723	0.660	0.462	0.471	0.718	0.660	0.449	0.467
NCD+DP	0.711	0.667	0.490	0.491	0.715	0.653	0.452	0.470
NCD+AD	0.725	0.671	0.460	0.474	0.721	0.663	0.446	0.466
NCD+FairCD ($\boldsymbol{\theta}^{\mathrm{b}}$)	0.612	0.621	0.497	0.499	0.607	0.613	0.471	0.493
NCD+FairCD ($\boldsymbol{\theta}^{\mathrm{f}}$)	**0.729**	**0.688**	**0.354**	**0.466**	**0.729**	**0.676**	**0.351**	**0.460**

a) Underline represents the best results, and bold represents the runner-up results.

Data partition. We conduct tests on two tasks: students' performance prediction and exercise recommendation. For each task, we perform an 80%/20% train/test split of each student's response log for each dataset (i.e., PISA-OECD, PISA-GENDER). For the exercise recommendation task, we observe students' actual performance on exercises that they have practiced in test sets.

6.2 Experimental results

Performance on utility and fairness (RQ1). In this section, we investigate FairCD efficacy by contrasting its performance with that of other baselines. For generalization purposes, we integrate FairCD with different CD benchmarks (i.e., IRT, MIRT, and NCD). The utility and fairness results are revealed in Table 5 and Figure 3. We can make the following observations from them.

• From the perspective of fairness (i.e., Figure 3), our findings reveal that all original CD models are unfair, suggesting the need to explore fairness in CD. We first discover that among fairness-aware approaches, CD+DP does not achieve adequate performance in some instances. This may be attributed to the fact that the CD+DP baseline improves the demographic parity fairness definition and does not improve $|F_{\mathrm{CD}}|$ directly. Second, in every situation, our framework surpasses all baseline techniques. This validates the effectiveness of our proposed framework's fairness promotion. Meanwhile, we notice that the fairness promotion on MIRT and NCD is higher than that on IRT. We hypothesize that this is due in part to the fact that the proficiency variable on MIRT and NCD is multidimensional, as opposed to being confined to a single dimension on IRT.

• From the perspective of model utility (i.e., Table 5), we first discover that all fairness-aware baselines diminish the original model utility. This phenomenon is the same as other domains, such as recommender systems [15, 59]. This is appropriate as fairness-aware algorithms tend to filter out knowledge of certain sensitive features from the proficiency of students, which will limit the information included and hence the utility performance to some extent. Second, we find that adversarial learning-based approaches (i.e., CD+AD, FairCD) outperform regularization-based methods in most circumstances, demonstrating the usefulness of adversarial learning-based methods. Most importantly, FairCD achieves the best performance among these fairness-aware methods.

• From the perspective of balancing the model utility and fairness, FairCD achieves both superior utility performance and fairness promotion in all cases. Based on such observations, we argue that our framework achieves superior performance in balancing the model utility and fairness.

Fairness improvement on CD application (RQ2). Numerous educational applications, such as educational recommender systems [60], have employed cognitive diagnostics. In this subsection, in

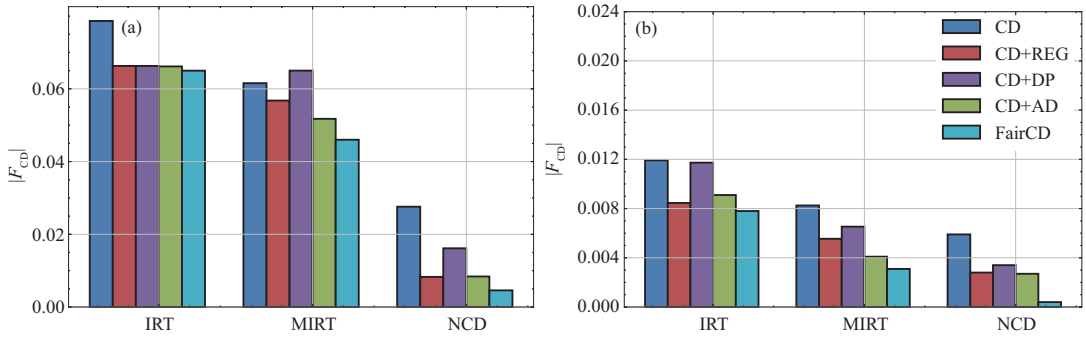

Figure 3 (Color online) Fairness results of different models in datasets (the lower, the better). (a) PISA-OECD; (b) PISA-GENDER.

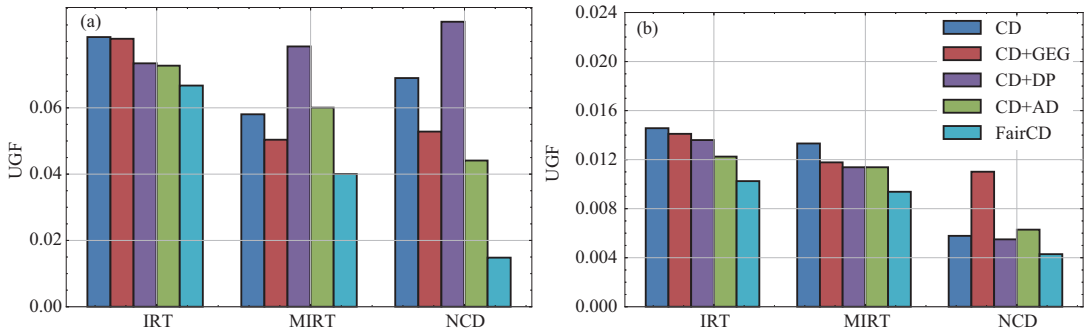

Figure 4 (Color online) UGF of different models in datasets (the lower, the better). (a) PISA-OECD; (b) PISA-GENDER.

an effort to validate a more realistic implication of FairCD, we explore the impact of FairCD on the fairness of educational recommender systems. Following [7], we focus on the most prevalent scenario: the system recommends non-mastered exercises to students based on the results of CD. As for fairness evaluation, following [46], we hope that the recommendation quality is identical across different groups. The user-oriented group fairness (UGF) statistic is as follows:

$$\mathrm{UGF} = |\mathcal{M}(A) - \mathcal{M}(B)|, \tag{11}$$

where $\mathcal{M}(A)$, $\mathcal{M}(B)$ represent recommendation qualities for groups A, B. Here we adopt HR@10. The result is shown in Figure 4. We first discover that the original recommender results based on CD models are unfair and that almost all FairCD methods can improve the fairness of educational recommendation systems, demonstrating the importance of enhancing cognitive diagnostic fairness. Then we can find the regularization-based methods (i.e., CD+REG, CD+AD) get unstable results compared with adversarial learning-based methods. We believe the reason is that regularization-based methods are proposed to optimize the specific fairness metric and do not achieve adequate performance in CD applications. Adversarial learning-based methods directly eliminate the effect of sensory qualities and can directly improve the fairness of cognitive diagnosis-based applications. Furthermore, among these methods, we demonstrate that FairCD exhibits superior results in all cases, demonstrating the effectiveness of FairCD.

Effectiveness of our proposed tasks (RQ3). To minimize sensitive attributes from student proficiency, we offer an adversarial learning task and a prediction task in FairCD. As discussed before, λ_{B} and λ_{F} defined in (10) control the effectiveness of these two tasks. Theoretically, the larger λ_{B} and λ_{F}, the greater the influence of the discriminator and predictor losses, implying that we observe a stricter demand for fairness and may have to sacrifice more CD utility performance to meet the requirement. We explore their usefulness in this part by altering the hyperparameters λ_{B} and λ_{F} on PISA-OECD. Because there are two hyperparameters, their influence is evaluated independently. We begin by varying the value of λ_{F} with $\lambda_{\mathrm{B}} = 0$. As revealed in Figure 5, we reveal that adversarial training tasks can help enhance fairness when λ_{F} increases from 0. When it is greater than 0.4, however, the fairness increase is minimal, and the utility performance drops more rapidly. Thus, when λ_{F} is approximately 0.4, it achieves the optimal utility-fairness balance. Subsequently, we vary the value of λ_{B} with $\lambda_{\mathrm{F}} = 0.4$. According to the data in Figure 6, fairness improves with the increase of λ_{B}, and the utility may decrease when λ_{B} is too

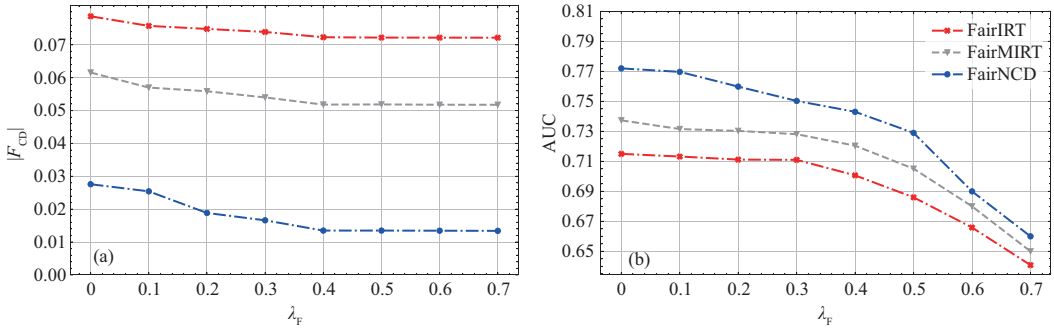

Figure 5 (Color online) Performance of FairCD with different λ_F. (a) Fairness; (b) utility.

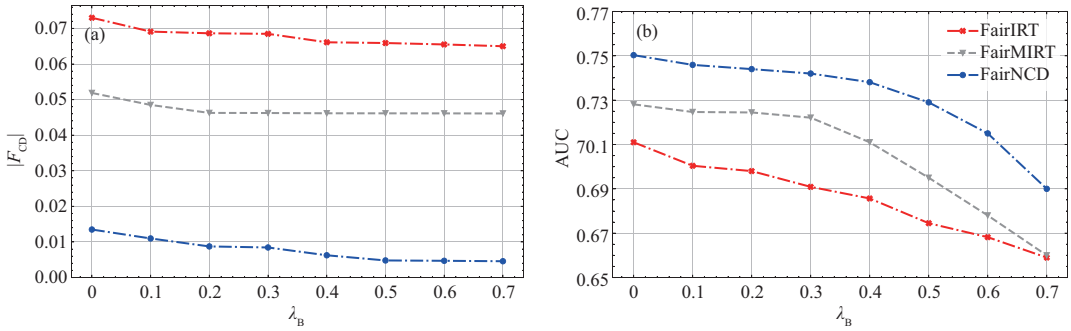

Figure 6 (Color online) Performance of FairCD with different λ_B. (a) Fairness; (b) utility.

large. Thus, a proper range of λ_B (0.1–0.3) can achieve an optimal tradeoff between fairness and utility.

Performance of bias proficiency (RQ4). In this paper, we divide student proficiency ($\boldsymbol{\theta}^u$) into bias proficiency ($\boldsymbol{\theta}^b$) and fair proficiency ($\boldsymbol{\theta}^f$). The goal of fair proficiency $\boldsymbol{\theta}^b$ is to acquire as much meaningful and unbiased student information as possible, while the goal of fair proficiency $\boldsymbol{\theta}^f$ is to capture as much useful and unbiased student information as possible, which we consider as the final proficiency estimate. However, $\boldsymbol{\theta}^b$ may also include insightful and objective student data. To evaluate the effectiveness of FairCD, we examine whether $\boldsymbol{\theta}^b$ includes useful information in this subsection. To do so, we use $\boldsymbol{\theta}^b$ as the input for the final students' performance prediction task. If the prediction task produces disappointing results, we can conclude that the bias proficiency has limited usable information, which also suggests that $\boldsymbol{\theta}^f$ captures as much useful student information as possible. FairCD ($\boldsymbol{\theta}^b$) consistently generates the worst outcomes across all CD benchmarks, according to the results shown in Table 5. This demonstrates that the bias proficiency contains minimal useful information, indicating the effectiveness of FairCD.

Case study. Further, we conduct a case study to better demonstrate the effectiveness of our method. We choose two students from each of the two categories (i.e., OECD/Non-OECD) who have answered the same questions and have answered more than 30 questions. In the meantime, they provide the same responses. The only distinction between them is their geographical location. We employ NCD to diagnose the proficiency of two students. In an ideal situation, the diagnosed difference between these two students would be 0. The case result is shown in Figure 7. As a result of NCD, the difference between them is expanded, demonstrating the unfairness of NCD. But after FairCD, the gap between these two students is closer to 0, showing that FairCD indeed can help CD maintain the gap.

7 Discussion

We now discuss the significance of fairness in cognitive diagnosis, which can be summarized from the following perspectives.

- Educational fairness: the results of CD serve as an essential reference for several high-stakes tests, such as the GRE and GMAT. Unfair cognitive diagnosing practices may unintentionally favor select groups, resulting in discrepancies in educational access and outcomes. We can establish a level playing field by promoting fairness, allowing every student to reach their full potential.
- Societal impact: advocating for fairness in cognitive diagnostics guarantees that kids receive equitable

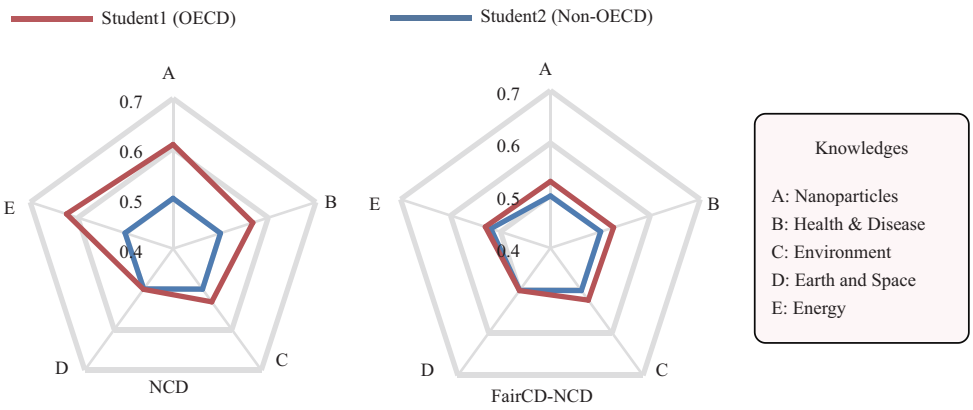

Figure 7 (Color online) Case study. The ideal knowledge proficiency gap between student1 and student2 should be 0. NCD widens the gap, and FairCD-NCD improves the fairness of NCD.

educational practices. This has the potential to break cycles and increase social mobility for historically marginalized groups, resulting in a more equal future for everybody.

In conclusion, the significance and societal impact of fairness in CD extends far and wide. We may contribute to the establishment of a more inclusive and just educational system that benefits all learners, ultimately leading to a more equitable society by encouraging equitable practices and addressing the potential harm caused by unfairness.

8 Conclusion and future work

We presented a concentrated investigation on the fairness issue in cognitive diagnostics in this paper and attempted to address two questions: (1) Are the results of existing CD models affected by sensitive attributes? (2) If so, how can we mitigate the impact of sensitive attributes to ensure fair diagnosis results? First, we discovered that unfairness exists in CD models, with varying degrees of unfairness. To explain this phenomenon, we conducted a theoretical analysis and found that model complexity leads to varying degrees of unfair performance. Then, we introduced FairCD, a framework for fairness-aware cognitive diagnostics that divides student performance into two components: bias proficiency and fair proficiency. We devised two orthogonal tasks for each of them to achieve fair proficiency regardless of sensitive traits, and we used this as the final diagnosed outcome. Finally, extensive experimental results on PISA clearly showed the effectiveness of our proposed framework.

In the future, we would like to analyze the fairness of more CD models and explore fairness in more educational tasks (e.g., knowledge tracing). Furthermore, we discovered that nearly all fairness-aware approaches were based on the well-known sensitive attribute labels situation. However, due to privacy concerns, students are not always willing to reveal sensitive information in real-world circumstances. Thus, the approaches via which fairness can be achieved without sensitive attributes are another crucial topic to be considered.

Acknowledgements This work was supported in part by National Key Research and Development Program of China (Grant No. 2021YFF0901003), National Natural Science Foundation of China (Grant Nos. 61922073, U20A20229), and University Synergy Innovation Program of Anhui Province (Grant No. GXXT-2022-042).

References

1 Lord F. A theory of test scores. Psychometric Monographs, 1952, 7: 84
2 Liu Q, Huang Z, Yin Y, et al. EKT: exercise-aware knowledge tracing for student performance prediction. IEEE Trans Knowl Data Eng, 2019, 33: 100–115
3 de la Torre J. DINA model and parameter estimation: a didactic. J Educational Behaval Stat, 2009, 34: 115–130
4 Templin J L, Henson R A. Measurement of psychological disorders using cognitive diagnosis models. Psychological Methods, 2006, 11: 287–305
5 Leighton J P, Gierl M J, Hunka S M. The attribute hierarchy method for cognitive assessment: a variation on Tatsuoka's rule-space approach. J Educational Measurement, 2004, 41: 205–237
6 Bi H, Ma H, Huang Z, et al. Quality meets diversity: a model-agnostic framework for computerized adaptive testing. In: Proceedings of IEEE International Conference on Data Mining (ICDM), 2020. 42–51
7 Huang Z, Liu Q, Zhai C, et al. Exploring multi-objective exercise recommendations in online education systems. In: Proceedings of the 28th ACM International Conference on Information and Knowledge Management, 2019. 1261–1270
8 Reckase M D. Multidimensional Item Response Theory Models. Berlin: Springer, 2009. 79–112

9 von Davier M. The DINA model as a constrained general diagnostic model: two variants of a model equivalency. Brit J Math Statis, 2014, 67: 49–71
10 Wang F, Liu Q, Chen E, et al. Neural cognitive diagnosis for intelligent education systems. In: Proceedings of the AAAI Conference on Artificial Intelligence, 2020. 6153–6161
11 Kizilcec R F, Lee H. Algorithmic fairness in education. 2020. ArXiv:2007.05443
12 Pedreshi D, Ruggieri S, Turini F. Discrimination-aware data mining. In: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2008. 560–568
13 Hardt M, Price E, Srebro N. Equality of opportunity in supervised learning. In: Proceedings of the 30th International Conference on Neural Information Processing Systems, 2016. 3323–3331
14 Bose A, Hamilton W. Compositional fairness constraints for graph embeddings. In: Proceedings of International Conference on Machine Learning, 2019. 715–724
15 Shao P Y, Wu L, Chen L, et al. FairCF: fairness-aware collaborative filtering. Sci China Inf Sci, 2022, 65: 222102
16 Wu L, Chen L, Shao P, et al. Learning fair representations for recommendation: a graph-based perspective. In: Proceedings of the Web Conference, 2021. 2198–2208
17 Yao S, Huang B. Beyond parity: fairness objectives for collaborative filtering. In: Proceedings of the 31st International Conference on Neural Information Processing Systems, 2017. 2925–2934
18 Lee M K, Rich K. Who is included in human perceptions of AI? Trust and perceived fairness around healthcare AI and cultural mistrust. In: Proceedings of the CHI Conference on Human Factors in Computing Systems, 2021. 1–14
19 Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial networks. Commun ACM, 2020, 63: 139–144
20 Gao W, Liu Q, Huang Z, et al. 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, 2021. 501–510
21 Rudner L M. Implementing the graduate management admission test computerized adaptive test. In: Proceedings of Elements of Adaptive Testing, 2009. 151–165
22 Mills C N. The GRE computer adaptive test: operational issues. In: Proceedings of Computerized Adaptive Testing: Theory and Practice, 2000
23 Zhuang Y, Liu Q, Huang Z, et al. A robust computerized adaptive testing approach in educational question retrieval. In: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2022. 416–426
24 Zhuang Y, Liu Q, Ning Y, et al. Efficiently measuring the cognitive ability of LLMs: an adaptive testing perspective. 2023. ArXiv:2306.10512
25 Gao W, Wang H, Liu Q, et al. 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, 2023. 983–992
26 Lin Z Q, Chen H X. Recommendation over time: a probabilistic model of time-aware recommender systems. Sci China Inf Sci, 2019, 62: 212105
27 McKinley R, Kingston N. Exploring the use of IRT equating for the GRE subject test in mathematics. ETS Res Report Ser, 1987, 1987: 1–35
28 Liu Q. Towards a new generation of cognitive diagnosis. In: Proceedings of the 30th International Joint Conference on Artificial Intelligence, 2021. 4961–4964
29 Wang F, Liu Q, Chen E, et al. NeuralCD: a general framework for cognitive diagnosis. IEEE Trans Knowl Data Eng, 2023, 35: 8312–8327
30 Ghosh A, Raspat J, Lan A. Option tracing: beyond correctness analysis in knowledge tracing. In: Proceedings of International Conference on Artificial Intelligence in Education, 2021. 137–149
31 Cheng Y, Li M, Chen H, et al. Neural cognitive modeling based on the importance of knowledge point for student performance prediction. In: Proceedings of the 16th International Conference on Computer Science & Education, 2021. 495–499
32 Wu J W, Shen L W, Guo W N, et al. Code recommendation for Android development: how does it work and what can be improved? Sci China Inf Sci, 2017, 60: 092111
33 Chen H H, Jin H, Cui X L. Hybrid followee recommendation in microblogging systems. Sci China Inf Sci, 2017, 60: 012102
34 He X, Liao L, Zhang H, et al. Neural collaborative filtering. In: Proceedings of the 26th International Conference on World Wide Web, 2017. 173–182
35 Huang H Y, Wang J Q, Fei C, et al. A probabilistic risk assessment framework considering lane-changing behavior interaction. Sci China Inf Sci, 2020, 63: 190203
36 Xiao S T, Shao Y X, Li Y W, et al. LECF: recommendation via learnable edge collaborative filtering. Sci China Inf Sci, 2022, 65: 112101
37 Hu H C, Guo Y F, Yi P, et al. Achieving fair service with a hybrid scheduling scheme for CICQ switches. Sci China Inf Sci, 2012, 55: 689–700
38 Ekstrand M D, Tian M, Azpiazu I M, et al. All the cool kids, how do they fit in? Popularity and demographic biases in recommender evaluation and effectiveness. In: Proceedings of Conference on Fairness, Accountability and Transparency, 2018. 172–186
39 Khademi A, Lee S, Foley D, et al. Fairness in algorithmic decision making: an excursion through the lens of causality. In: Proceedings of the Web Conference, 2019. 2907–2914
40 Liu J H, Yu Y, Bi H L, et al. Post quantum secure fair data trading with deterability based on machine learning. Sci China Inf Sci, 2022, 65: 170308
41 Dwork C, Hardt M, Pitassi T, et al. Fairness through awareness. In: Proceedings of the 3rd Innovations in Theoretical Computer Science Conference, 2012. 214–226
42 Zafar M B, Valera I, Rogriguez M G, et al. Fairness constraints: mechanisms for fair classification. In: Proceedings of Artificial Intelligence and Statistics, 2017. 962–970
43 Kang J, He J, Maciejewski R, et al. Inform: individual fairness on graph mining. In: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2020. 379–389
44 Calmon F, Wei D, Vinzamuri B, et al. Optimized pre-processing for discrimination prevention. In: Proceedings of Advances in Neural Information Processing Systems, 2017
45 Dong Y, Kang J, Tong H, et al. Individual fairness for graph neural networks: a ranking based approach. In: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, 2021. 300–310
46 Li Y, Chen H, Fu Z, et al. User-oriented fairness in recommendation. In: Proceedings of the Web Conference, 2021. 624–632

47 Simon F, Małgorzata K, Beatriz P. No more failures: ten steps to equity in education. OECD Publishing, 2007. doi: 10.1787/9789264032606-en

48 Rao Y S, Zhang J Z, Zou Y, et al. An advanced operating environment for mathematics education resources. Sci China Inf Sci, 2018, 61: 098102

49 Warren C J E. Brown v. Board of Education. United States Reports, 1954

50 Hutt S, Gardner M, Duckworth A L, et al. Evaluating fairness and generalizability in models predicting on-time graduation from college applications. In: Proceedings of the 12th International Conference on Educational Data Mining (EDM 2019), 2019

51 Hu Q, Rangwala H. Towards fair educational data mining: a case study on detecting at-risk students. In: Proceedings of the 13th International Conference on Educational Data Mining, 2020. 431–437

52 Yu R, Li Q, Fischer C, et al. Towards accurate and fair prediction of college success: evaluating different sources of student data. In: Proceedings of International Conference on Educational Data Mining (EDM 2020), 2020

53 Li C, Xing W, Leite W. Using fair AI to predict students' math learning outcomes in an online platform. Interactive Learn Environ, 2022. doi: 10.1080/10494820.2022.2115076

54 Gómez E, Zhang C S, Boratto L, et al. The winner takes it all: geographic imbalance and provider (un) fairness in educational recommender systems. In: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2021. 1808–1812

55 Embretson S E. Item Response Theory. London: Psychology Press, 2013

56 DiBello L V, Roussos L A, Stout W. 31a review of cognitively diagnostic assessment and a summary of psychometric models. Handbook of Statistics, 2006, 26: 979–1030

57 Glorot X, Bengio Y. Understanding the difficulty of training deep feedforward neural networks. In: Proceedings of the 13th International Conference on Artificial Intelligence and Statistics, 2010. 249–256

58 Liu Q, Wu R, Chen E, et al. Fuzzy cognitive diagnosis for modelling examinee performance. ACM Trans Intell Syst Technol, 2018, 9: 1–26

59 Wu C, Wu F, Wang X, et al. Fairness-aware news recommendation with decomposed adversarial learning. In: Proceedings of the AAAI Conference on Artificial Intelligence, 2021. 4462–4469

60 Chen J, Li H, Ding W, et al. An educational system for personalized teacher recommendation in K-12 online classrooms. In: Proceedings of International Conference on Artificial Intelligence in Education, 2021. 104–108