IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received March 9, 2021, accepted March 18, 2021, date of publication March 26, 2021, date of current version April 7, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3069055

Leveraging Cognitive Diagnosis to Improve Peer Assessment in MOOCs

JIA XU[1,2,3], (Member, IEEE), QIUYUN LI[1], JING LIU[1], PIN LV[1,2,3], (Member, IEEE), AND GE YU[4], (Senior Member, IEEE)

[1]College of Computer, Electronics and Information, Guangxi University, Nanning 530004, China
[2]Guangxi Colleges and Universities Key Laboratory of Parallel and Distributed Computing, Guangxi University, Nanning 530004, China
[3]Guangxi Key Laboratory of Multimedia Communications Network Technology, Guangxi University, Nanning 530004, China
[4]School of Computer Science and Engineering, Northeastern University, Shenyang 110004, China

Corresponding author: Pin Lv (lvpin@gxu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 62067001 and Grant U1811261, in part by the Projects of Higher Education Undergraduate Teaching Reform Project in Guangxi under Grant 2017JGZ103 and Grant 2020JGA116, in part by the Special funds for Guangxi Bagui Scholars, and in part by the Guangxi Natural Science Foundation under Grant 2019JJA170045.

ABSTRACT A major challenge faced by popular massive open online courses (MOOCs) is the assessment of large-scale open-ended assignments submitted by students. Recently, peer assessment has become a mainstream paradigm that helps grade open-ended assignments on a large scale. In peer assessment, students also become graders in grading a small number of their peers' assignments, and the peer grades are then aggregated to predict a true score for each assignment. The collected peer grades are usually inaccurate because graders have different reliabilities and biases. To improve accuracy, several probabilistic graph models have been proposed to model the reliability and bias of each grader. However, none of these models consider graders' competency information in the assignments to be graded, which has been found to be very effective. We propose two new probabilistic graph models to improve the accuracy of cardinal peer assessments based on the well-accepted cognitive diagnosis technique. Specifically, the cognitive diagnosis model DINA is applied to determine grader competency based on historical tests or assignments. Then, this information is used to optimize the modeling of grader reliability in each of the proposed models. Moreover, an effective model inference algorithm is proposed to infer true scores of assignments. Experimental results based on real world datasets show that the two proposed models outperform state-of-the-art models and that consideration of grader competency contributes to improved score estimation.

INDEX TERMS Peer assessment, cognitive diagnosis, probabilistic graph models, MOOCs.

I. INTRODUCTION

Peer assessment (or peer review), which is also known as peer grading, is an arrangement for peers to consider and specify the level, value, or quality of a product or performance of other equal-status peers [1]. Peer assessment is becoming increasingly important for Massive Open Online Courses (MOOCs) [2]–[4], since it provides a practical solution to the large-scale grading problem brought by large-scale participation of students in MOOCs. Automated grading software also offers help to solve the problem. However, there are less effective auto-grading solution for open-ended assignments (e.g., essays or problem-solving questions), since such assignments do not have standardized answers. Considering open-ended assignments are arguably important for evaluating learning outcomes of many MOOCs [5], many popular MOOC platforms, such as Coursera[1] and edX,[2] have already provided the function of peer assessment to help teachers assess massive submissions of students to open-ended assignments. Specifically, in these platforms, students play an additional role as graders of a small number of their peers' assignments, according to some rubrics or benchmarks given by the teacher. Grades offered by students (i.e., peer grades) are then aggregated by the MOOC platforms to produce an estimate to the true score for each assignment. Besides the benefit of reducing teachers' workloads of grading large-scale open-ended assignments, peer assessment is also believed to

The associate editor coordinating the review of this manuscript and approving it for publication was Eunil Park.

[1]https://www.coursera.org/
[2]https://www.edx.org/

bring other educational values, including learning from other students' solutions [6], inspiring of students' learning interests [7], enhancing of students' involvement in a course [8], and developing of students' higher-order thinking and metacognition skills [9].

Although peer assessment is very helpful for MOOCs, it is still a challenging problem to aggregate peer grades received for an open-ended assignment so as to derive more accurate estimates to the final scores of assignments. This paper considers the case of cardinal peer assessment, where the peer evaluations for assignments are in the form of numerical grades. The cardinal setting is currently the most common choice for popular MOOC platforms, e.g., Coursera and edX. These MOOC platforms simply use the median (or mean) of received peer grades as the final score. Considering the varied skill levels and attitudes of online students [10], such a simple aggregation strategy of peer grades used by existing MOOC platforms may be inaccurate due to the impact of two factors, namely grader bias and grader reliability [11]. Grader bias reflects any constant inflation or deflation of the peer grades given by the grader. The second impact factor, i.e., grader reliability, represents the variance in the difference between peer grades that the grader gives with respect to a group of assignments and the true scores of these assignments. If a grader gives grades randomly, the variance is large and the grader is deemed unreliable. If a grader carefully assigns grades based on the quality of assignments, the variance is small and the grader is deemed reliable. Bias and reliability of graders are two major factors concerned in peer assessment of MOOCs. On one hand, the evaluation of these two factors offers help in predicting the participation of students in the next assignment, based on the theory that a reliable grader would add a different dimension of information to a student's engagement in teaching activities [2]. On the other hand, recent study [2] have proven that the consideration of these two factors is very helpful in improving the estimates to final scores of assignments. Recently, in order to introduce these two factors into the procedure of peer assessment in MOOCs, different probabilistic graph model are presented in top conferences from education or computer science domain [2], [12]–[14]. Each of these proposed probabilistic graph models sets the true score of an assignment, the grader bias, and the grader reliability as latent random variables of a graphical network that follow certain distributions, and utilizes the conditional dependency network of these variables to infer their values by fitting the models on some observed values, such as the peer grades. These models are supposed to optimize the aggregation of peer grades in MOOCs.

Although these probabilistic graph models successfully improve the accuracy of estimates to the final scores of assignments, they build inadequate models for grader reliability. In specific, these models either model grader reliability based on a simple probability distribution [2], [13] or based on a probability distribution related to the true score of the grader to the graded assignment [12]–[14]. Cognitive diagnosis models (CDMs) [15], [16], which are famous

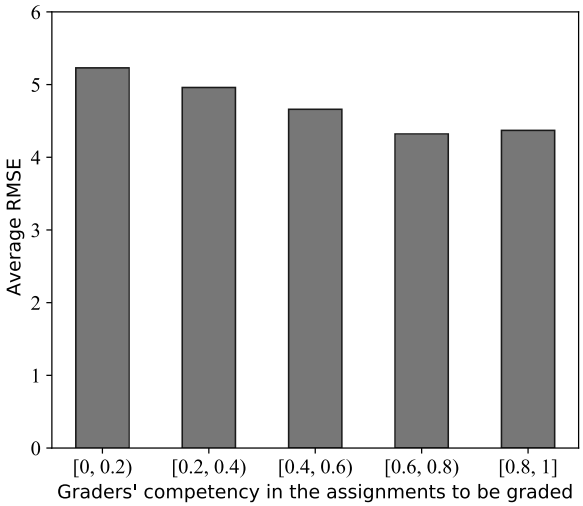


FIGURE 1. Correlation between competency and RMSE.

psychometric models used to quantify the competency of a student in a given question based on his/her performance gained in historical records [17], provide possibility to optimize the modeling of grader reliability. Fig. 1 illustrates the results of 2,109 peer-assessment records given by 260 graders regarding three open-ended assignments. Using historical test and assignment results of these graders as input, the deterministic input, noisy, ''and'' gate cognitive diagnosis model, i.e., DINA [18], was used to compute their competency values in each open-ended assignment that they need to grade. In Fig. 1, the X-axis displays different ranges of grader competency values for the three open-ended assignments, and the Y-axis shows the average RMSE between peer grades given by graders whose competency values fall into a certain range and the corresponding true score given by the teacher. As seen in the figure, grader reliability is affected by the grader's competency in the assignments: the lower the competency value, the larger the RMSE, and thus the lower the grader reliability; the higher the competency value, the smaller the RMSE, and therefore, the higher the grader reliability. Ignoring grader competencies leads to a suboptimal model of grader reliability, and, consequently, leads to a less accurate prediction to the true scores of assignments.

Recognizing the importance of grader competency in the assignments they grade, we developed two peer grading probabilistic graph models (named *CD-PG*$_1$ and *CD-PG*$_2$) based on graders' competency information derived by cognitive diagnosis. Specifically, the popular DINA model is used to compute grader competency values in open-ended assignments, based on their performances in previous tests and assignments. Then, the proposed probabilistic graph models estimate the true scores of the assignments from the peer grades and the relative peer grades (i.e., the difference in the grades given by the same grader to two different assignments) by modeling grader reliability based on the grader's competency and by modeling grader bias. Gaussian distributions are applied to model the true scores, peer grades, relative

peer grades, and grader biases separately in the probabilistic graph models. Two different probabilistic distributions are used to estimate grader reliability. In the $CD\text{-}PG_1$ model, the reliability of a grader follows a Gamma distribution, with the shape parameter determined by the grader's competency in the assignment to be graded, while in the $CD\text{-}PG_2$ model, it follows a Gaussian distribution with the mean equal to the grader's competency in the assignment. To evaluate our proposed probabilistic graph models, we conducted a group of peer assessment activities for open-ended assignments on the MOOC platform of Guangxi University[3] and collected a peer assessment dataset. Experimental results show that the proposed models improve the accuracy of the predicted scores of open-ended assignments by considering graders' competency information. The main contributions of this paper are summarized as follows:

(1) We find that a grader's competency in an assignment to be graded can help optimize the modeling of the grader's reliability for the assignment to improve the accuracy of cardinal peer-assessment prediction.
(2) We proposed two probabilistic graph models (i.e., $CD\text{-}PG_1$ and $CD\text{-}PG_2$), which make novel optimization to state-of-the-art models for cardinal peer assessment in MOOCs. The novelty of the proposed models comes from the utilization of a grader's competency information derived by cognitive diagnosis method to optimize the modeling of his/her grading reliability.
(3) Our proposals are evaluated using a real MOOC peer assessment dataset, and the experimental results show that our methods are more accurate than the state-of-the-art peer-assessment probabilistic graph models.

The rest of the paper is organized as follows. Section II reviews related works. Section III describes preliminary knowledge, including the DINA cognitive diagnosis model and definitions relevant to our problem. Section IV then discusses the proposed probabilistic graph models for peer assessment. The algorithm for inferring latent variables involved in the proposed probabilistic graph models is explained in Section IV-D. Then, experimental results are presented in Section V, followed by the conclusions and future work in Section VI.

II. RELATED WORK

Since the probabilistic graph models we presented in this paper can be viewed as part of a long tradition of models that have been proposed for the purposes of aggregating opinions from diverse people. In this section, we first review literature about opinion aggregation techniques from different application domains in Section II-A, and then we focus ourselves in analyzing related works of peer assessment in MOOCs in Section II-B.

A. OPINION AGGREGATION

In the rapidly growing domain called crowdsourcing [19], labels of an item provided by many workers are aggregated to render an estimation to the true label of the item. Probability models [20]–[22], neural networks [23]–[25], weighted sum [26], [27], or majority vote [28], [29] is used to achieve the aggregation. Among these aggregation methods, majority vote is the simplest aggregation way that chooses what the majority of workers agree on as the final label to the item, and thus is error-prone when there are many spammers. The weighted sum method weights a label given by a worker and then aggregates labels obtained from all workers by computing weighted sum based on their weights. The weight of a label given by a worker, take literature [26] as an example, is determined according to the worker's performance or confidence (generally defined as the worker's reliability). Though it is very direct to implement a weighted sum solution, it fails to consider some complex factor, such as a worker's bias, which have important impact on the aggregation results. Neural networks currently have been successfully applied to perform the aggregation of labels in a crowdsourcing task. In particular, in [25], Rodrigues *et al.* proposed to embed a crowd layer that considers the ability of individual worker in a neural network to fulfill the predication of true labels for every item. Lacking of interpretability is an eye-catching problem of neural network which is often criticized by researchers. Besides, a large amount of ground-truth data which are used to train the network are not available in most cases. For probability models, they provide a principled way to infer the true label of an item based on the observations (e.g., labels submitted by workers) and the conditional dependencies among latent variables. For example, in [20], Whitehill *et al.* proposed a probabilistic model, called GLAD, where the expertise of each worker and the difficulty of each image are set as latent variables of the model and the EM algorithm is adopted to infer true labels of items based on dependencies among those latent variables. Since the probability model can capture the dependencies between the true labels of items and other impact variables and is highly interpretable, it has become a hot research topic of recent years.

The peer review of journal/conference papers or funding proposals are also related to our problem. Restate that our most challenging problem is to improve the quality of aggregation of noisy grades given by peers. For the works of paper review, however, their major concerns are the policy that deal with anonymity [30], [31], and the optimization of the assignment of reviewer roles, based on some impact factors, such as expertise of peers, citation link structure, and conflicts of interest [32]–[34]. Only simple strategies are employed to aggregate comments or grades of reviewers. As to the review of funding proposals, the evidential reasoning (ER) approach [35], which is a generic probabilistic reasoning process, is widely used to combine multiple pieces of independent evidence with both weight and reliability of evidence considered [36]–[39]. However, many constraints, such as the requirement of professional knowledge background, are imposed to reviewers, which makes the ER approach inappropriate to solve the peer assessment problem in MOOCs.

[3]http://www.course.gxu.edu.cn/portal

To sum up, there exist many related works that aggregate opinions of diverse peers are proposed in non-educational domains. Among these works, due to the interpretability and the ability of modeling the impact factors to the estimation of true labels and the dependencies among those factors, methods built on probability inference procedure (i.e., probability models and ER approach) favored by most researchers. However, the proposed methods in these works are not suitable to solve the peer assessment problem in MOOCs. For one, in our problem setting, graders are also gradees, while there is a dichotomy between the workers and the items being labeled by the workers in crowdsourcing applications [2] and papers or proposals are usually assigned to different set of people to review. For another, when reviewing papers or proposals, the most consideration is the precision at the top after aggregating feedback of reviewers and mis-ranking items that are far from the top-k carries no real consequence. While in our peer assessment scenario, each evaluation towards an assignment carries the same importance, and we do not need to precisely rank students whose assignments have approximately the same quality.

B. PEER ASSESSMENT IN MOOCs

Next, we review related works under the context of MOOCs. Existing works on peer assessment aggregation can be divided into two categories, based on the data types: the *ordinal estimation* and the *cardinal estimation*.

In ordinal estimation, every grader is asked to rank a set of submissions of an assignment according to each submission's quality. Then, the goal is to aggregate the partial rankings of submissions (e.g., $x_1 \succ x_4 \succ x_2$) given by each grader to derive a full ranking of all submissions (e.g., $x_1 \succ \cdots \succ \ldots x_n$) [40]. For ordinal technologies, the Bradley-Terry model [41] is generalized to learn latent student scores by aggregating partial ranking information on assignments provided by the students [42]. In [43], several statistical ranking models for ordinal comparison, namely the Bradley-Terry model [41], the Plackett-Luce model [44], and the Mallows model [45], are used to learn the full ranking of all submissions from the collected individual student's partial ranking of submissions. To further improve the accuracy of the full ranking of submissions, on one hand, researchers introduce a variability parameter into the statistical ranking models to estimate the uncertainty inherent in the assessment process (i.e., the reliability of graders). On the other hand, the Bayesian approach is applied to estimate the uncertainty of each submission's position in the full ranking [46] and to allocate the ranking tasks for the students [47]. To reduce the sample complexity for partial ranking, Chan *et al.* proposed a multi-armed-bandit-style online algorithm, which optimizes both the allocation of ranking tasks among students and the aggregation of partial ranking submissions by taking students' reliabilities into consideration [48]. Mi *et al.* augmented ordinal models with cardinal predictions as priors and proved that such a combination may achieve further performance boosts in both cardinal and ordinal evaluations [12]. Recent years also witness the application of fuzzy set theory in ordinal peer assessment. For example, in [49], Capuano *et al.* presented a new model for ordinal peer assessment based on the principles of fuzzy group decision making, where the partial rankings provided by students are first transformed in fuzzy preference relations and then those relations are used to generate a global ranking between the submissions and to estimate their absolute grades. Although ordinal estimation does not require students to give specific scores for submissions, and thus reduces the difficulty, it has an important limitation [50]: by considering only partial ranking information, it is very difficult to quantify the quality difference between two submissions and assign reasonable scores.

Unlike ordinal estimation, cardinal estimation asks each grader to give a numerical grade (e.g., 96) for each assignment. Then, the target of cardinal peer estimation is to predict the true score of each assignment based on a group of peer grades of the assignment given by multiple graders. A major approach for cardinal estimation is the iterative algorithm, which updates the final scores of assignments and the weights of graders in an iterative manner. In [50], De Alfaro *et al.* proposed the Vancouver algorithm, which iteratively updates the grading accuracy of each grader and refines its prediction of the true score of every assignment based on these updated grading accuracies. In addition, Walsh proposed another iterative algorithm, named PeerRank [51], which is inspired by the idea of Google's well-known PageRank algorithm [52]. When estimating the true score for an assignment, the PeerRank algorithm weights a peer grade by its corresponding grader's ability to grade correctly, which is determined by the grader's performance in the course. In [53], a reputation-based algorithm was proposed that builds a trust graph over graders and uses that graph to compute weights for the aggregation of peer grades. Another major category of cardinal estimation methods is those based on probabilistic graph models. Such methods model the true score of an assignment and the bias and reliability of every grader as latent random variables following certain probabilistic distributions and infer the values of those variables by fitting the models based on observed peer grades. Our proposed methods belong to this category. Piech *et al.* first proposed three probabilistic graph models [2], namely PG_1 (which assumes that the true scores, observed peer grades, and grader bias follow Gaussian distributions, and the grader reliability follows a Gamma distribution), PG_2 (which extends PG_1 by linking the bias of a grader derived from historical grading tasks), and PG_3 (which extends PG_1 using the score of a student's assignment to estimate that student's reliability in grading peer assignments). Considering the assumption of PG_3, i.e., that the reliability of a grader fits a linear function of the grader's grade, is too strict, two extensions of PG_3, called PG_4 and PG_5, were proposed in [12]. PG_4 and PG_5 assume that the reliability of a grader follows a Gamma distribution where the shape parameter equals the score of the grader's own submission or follows a Gaussian distribution where the

mean is equal to the score of the grader's own submission. Recent studies show that the bias of a grader can be affected by the biases of the grader's friends [54], [55]. In view of this, Chan and King employ the social connections collected from a MOOC platform to optimize the modeling of the bias of each grader and thus extend the probabilistic graph models of PG_1, PG_4, and PG_5, separately [13]. However, all the above probabilistic graph models have a limitation. It stems from their assumption that the grades given by a grader to different submissions of an assignment are mutually independent, which is not in accordance with the actual situation. Therefore, Wang *et al.* introduced the observed relative peer grades of a grader (i.e., the difference between the peer grades given by the grader to two different submissions of an assignment) into the probabilistic graph model and proposed two novel models, referred to as PG_6 (built on PG_4) and PG_7 (built on PG_5), to estimate the true score of a submission [14]. PG_6 and PG_7 have obtained promising performance because the introduction of relative peer grades reduces the negative impact of data sparsity on parameter estimation. They are the probabilistic graph models most relevant to our work.

In conclusion, cardinal peer assessment has an advantage in quantifying the quality differences between two submissions of assignments, and thus there are more related models proposed in recent years. Similar to the conclusion drawn from related works from other domains, methods built on probability inference procedure, i.e., probabilistic graph models here, are the mainstream solutions for cardinal peer assessment in MOOCs. The effectiveness of these models [2], [12]–[14] are verified using the real-world peer assessment datasets provided by popular MOOC platforms, including Coursera, and XuetangX[4] from China. However, none of the existing probabilistic graph models consider the impact of grader competency for the assignment on grader reliability, which has been proven to be an important impact factor of grader reliability (see Fig. 1 for details). To overcome such limitations, a popular CDM – DINA – is applied to determine graders' competencies in open-ended assignments and optimized the PG_6 and PG_7 models by modeling graders' reliabilities based on their competency values. To help readers better get the differences between the state-of-the-art probabilistic graph models and the proposed models in this paper, a comparison are made in Table 1.

[4]www.xuetangx.com

III. PRELIMINARY

In this section, the popular CDM, i.e., DINA, is firstly introduced, which is applied to quantify graders' competencies in the open-ended assignments to be graded (Section III-A). Then, important concepts used throughout this paper and the peer assessment problem solved in this study are described (Section III-B).

A. DINA MODEL

Recently, there has been increasing interest in CDMs, which are psychometric models used to provide fine-grained information about students' strengths and weaknesses in learning [56]–[58]. Although many CDMs have been proposed, the DINA model [18] is highly preferred by researchers due to its easy interpretation and good model-data fit [59], [60]. Thus, the DINA model has been widely applied in recent years [61], [62]. In this paper, the DINA model is applied to quantify students' competencies in open-ended assignments by considering their performances in historical tests and assignments.

Let $E = \{e_1, \ldots, e_M\}$ represent a set of examinees and $T = \{t_1, \ldots, t_N\}$ be a set of questions from tests or assignments. Then, the $\boldsymbol{R}$-matrix (i.e., response matrix) that records the responses of each examinee in E to each question in T can be denoted by $\boldsymbol{R} = [r_{mn}]_{M\times N}$ with $r_{mn} \in [0, 1]$, where $r_{mn} = 1$ indicates that examinee e_m has given a correct answer to question t_n, and $r_{mn} = 0$ means the answer given by examinee e_m is wrong. The $\boldsymbol{R}$-matrix is set based on the examinees' historical test and assignment results. Let $\boldsymbol{KP} = \{kp_1, \ldots, kp_K\}$ be a set of knowledge points examined by the questions in E. The implementation of the DINA model requires the construction of a $\boldsymbol{Q}$-matrix [63] in the form of $\boldsymbol{Q} = [q_{nk}]_{N\times K}$ with $q_{nk} \in \{0, 1\}$, and the element on the n-th row and k-th column of the matrix (i.e., q_{nk}) indicates whether knowledge point kp_k is required to correctly answer question t_n. The $\boldsymbol{Q}$-matrix explicitly identifies the cognitive specification for every question in T. In the DINA model, a skill vector of examinee e_m in the form of $\boldsymbol{\alpha_m} = \{\alpha_{m1}, \ldots, \alpha_{mK}\}$ is used to represent the skill status of e_m. Specifically, the value of α_{mk} ($\alpha_{mk} \in [0, 1]$) is the competency of e_m relating to the k-th knowledge point in KP. $\alpha_{mk} = 1$ indicates that examinee e_m has fully mastered the k-th knowledge point, while $m_k = 0$ means that examinee e_m has not mastered the k-th knowledge point. In the DINA model, the skill vector of examinee e_m (i.e., $\boldsymbol{\alpha_m}$) and the $\boldsymbol{Q}$-matrix produce a latent response vector $\delta_m = \{\delta_{mn}\}$,

TABLE 1. Comparison of probabilistic graph models.

Models	Reliability	Bias	Peer grades	Relative peer grades	Social connections	Grader competency
PG_1, PG_2, PG_3 [2]	✔	✔	✔	✗	✗	✗
PG_4, PG_5 [12]	✔	✔	✔	✗	✗	✗
PG_6(2017), PG_7(2017), PG_8 [13]	✔	✔	✔	✗	✔	✗
PG_6(2019), PG_7(2019) [14]	✔	✔	✔	✔	✗	✗
$CD\text{-}PG_1$, $CD\text{-}PG_2$	✔	✔	✔	✔	✗	✔

where

$$\delta_{mn} = \prod_{k=1}^{K} \alpha_{mk}{}^{q_{nk}}. \quad (1)$$

The latent response in Equation (1) assumes a value of 1 if examinee e_m masters all the knowledge points required for question t_n and a value of 0 if examinee e_m completely fails to master at least one of the required knowledge points. Two parameters required for examinee e_m's response to question t_n are expressed by $\mathbb{g}_n = P(r_{mn} = 1|\delta_{mn} = 0)$ and $\mathbb{s}_n = P(r_{mn} = 0|\delta_{mn} = 1)$. $\eth_n$ represents the guessing probability that examinee e_m, who has not mastered all the required knowledge points for question t_n will randomly respond correctly to the question, while $\mathbb{s}_n$ is the slipping probability that examinee e_m who has mastered all the required knowledge points for question t_n will still answer wrongly. By introducing these two parameters, the item response function for examinee e_m on question t_n in the DINA model is given as follows.

$$P(\boldsymbol{\alpha_m}) = P\left(r_{mn} = 1 \mid \boldsymbol{\alpha_m}\right) = \mathbb{g}_n^{1-\delta_{mn}} \left(1 - \mathbb{s}_n\right)^{\delta_{mn}} \quad (2)$$

As shown in Equation (2), the DINA model is a conditional distribution of r_{mn} given a skill vector $\boldsymbol{\alpha_m}$. Then, the expectation-maximization (EM) algorithm is used to estimate the marginalized likelihood of Equation (2) and derive skill vector $\boldsymbol{\alpha_m}$ for examinee e_m [18].

In this work, it is assumed that all graders who participate in peer assessments for open-ended assignments have completed some objective-form tests or assignments before the peer assessment, and the objective-form tests or assignments examine some knowledge points that are also required by the open-ended assignments. This assumption is in line with the actual teaching situation because teachers tend to arrange objective-form tests or assignments in class to get timely feedback from students and assign open-ended assignments after class to help students consolidate the knowledge points learned in the class. Using student e_m's historical records of objective-form tests and assignments and the $\boldsymbol{Q}$-matrix as the input, the DINA model can diagnose skill vector $\boldsymbol{\alpha_m}$ for the student. Then, the competency of student e_m in open-ended assignment t_n can be defined by the latent response variable δ_{mn}, which is computed by multiplying the competency value (i.e., α_{mk}) of every knowledge point required by assignment t_n (see Equation (1)). The proposed probabilistic graph models of peer assessment in this study use the diagnosed competency values of graders in the open-ended assignment to be graded to optimize the modeling of the graders' reliabilities.

B. PROBLEM DEFINITION

Let U denote the set of students who have submitted their open-ended assignments on a MOOC platform and u_i denote an arbitrary student in U. Let V represent the set of students who act as graders for those open-ended assignments and v represent an arbitrary grader in V. Because the students who have submitted their open-ended assignments are usually required to grade their peers' submissions of those same assignments, U and V actually correspond to the same set of students, i.e., $|U| = |V|$. The following are definitions of important concepts that will be used in this paper.

True score: It is assumed that each submission of an open-ended assignment is associated with a true score. s_i represents the true score of student u_i's submission.

Peer grade: Peer grades are the observable peer scores given by graders to their peers' submissions. The notation z_i^v is used to denote the peer grade given by grader v to student u_i's submission. The set of all observed peer grades is denoted as $Z = \{z_i^v | u_i \in U, v \in V\}$.

Relative peer grade: The relative peer grade is denoted by d_{ij}^v, which measures the difference between two observed peer grades given by grader v for the submissions of students u_i and u_j (i.e., z_i^v and z_j^v, respectively). The set of all relative peer grades is denoted as $D = \{d_{ij}^v | u_i, u_j \in U, v \in V\}$.

Grader bias: Grader bias is denoted by b_v, which reflects grader v's tendency to either inflate or deflate peer grades. For example, given the true grade of student u_i's submission as $s_i = 10$, and grader v's bias as $b_v = -2$, the mean of the peer grade given by v to u_i's submission is $z_i^v = s_i + b_v = 10 + (-2) = 8$.

Grader reliability: Grader reliability is denoted by τ_v, which is defined as the precision of the peer grades given by grader v. Grader reliability, in fact, measures how close on average the peer grades given by grader v are to the corresponding true score of the submission after correcting for the bias of grader v. In this work, given an open-ended assignment, the reliability of grader v is modeled as a random variable following a Gamma distribution, with the shape parameter being set to v's competency value in the assignment in the proposed *CD-PG*$_1$ model, or a random variable following a Gaussian distribution with the mean being set to the v's competency value in the assignment in the proposed *CD-PG*$_2$ model. This means that we assume grader v with a higher competency in an open-ended assignment to be a more reliable grader for that assignment.

Unlike existing probabilistic graph models that estimate the true scores of open-ended assignments only by observed peer grades and relative grades [2], [12]–[14], this paper introduces new probabilistic graph models (i.e., *CD-PG*$_1$ and *CD-PG*$_2$) by exploiting graders' competency information (derived from graders' historical question-answering records) in the assignments they are to grade to improve the estimation accuracy. Our goal is to estimate the true score of each submission of an open-ended assignment effectively by modeling the relationships of the observed peer grades, grader reliabilities, grader biases, and true scores of submissions. More formally, our cardinal peer-assessment problem is defined as follows: given the set of students U, set of graders V, set of all observed peer grades Z, set of all observed relative peer grades D, and skill vectors of all graders represented as a matrix $A_{|V|\times|KP|} = [\ldots, \boldsymbol{\alpha}_m, \ldots]$, our goal is to train our probabilistic graph model to estimate the grader reliability τ_v, grader bias b_v for all graders $v \in V$, and true scores s_i for

TABLE 2. Notations.

Notation	Description
U	Set of all students who submitted their assignments; u_i is an arbitrary student in U
V	Set of all graders; v is an arbitrary grader in V
U_v	Set of students whose submissions are graded by grader v
V_{u_i}	Set of graders who give peer grades to the submission of student u_i
$\boldsymbol{R}$	The $\boldsymbol{R}$-matrix
$\boldsymbol{Q}$	The $\boldsymbol{Q}$-matrix
$\boldsymbol{\alpha_v}$	Skill vector of grader v, and α_{vk} denotes v's competency value in the k-th knowledge point
$\boldsymbol{A}$	Matrix containing skill vectors of all graders
δ_v	Competency value of grader v in an open-ended assignment
s_i	True score for student u_i's submission of an assignment
τ_v	Reliability of grader v
b_v	Bias of grader v
z_i^v	Peer grade given by grader v to student u_i's submission
Z	Set of all observed peer grades, i.e., $Z = \{z_i^v \mid u_i \in U, v \in V\}$
d_{ij}^v	Relative peer grade for the peer grades z_i^v and z_j^v
D	Set of all relative peer grades, i.e., $D = \{d_{ij}^v \mid u_i, u_j \in U, v \in V\}$

the submissions of all students $u_i \in U$. Table 2 summarizes all the notations used in our proposed probabilistic graph models.

IV. PROBABILISTIC GRAPH MODELS FOR PEER ASSESSMENT

In this work, two probabilistic graph models, namely *CD-PG*$_1$ and *CD-PG*$_2$, are proposed for cardinal peer assessment based on the cognitive diagnosis model – DINA. *CD-PG*$_1$ and *CD-PG*$_2$ both assume that a grader's grading reliability in an open-ended assignment is affected by the grader's competence in the assignment, which can be diagnosed by the DINA model based on the grader's performance in historical tests and assignments. Specifically, the *CD-PG*$_1$ model and the *CD-PG*$_1$ model are optimizations to the *PG*$_6$ model and the *PG*$_7$ model proposed in [14], respectively.

A. CD-PG$_1$ MODEL

The conditional dependence structure between the random variables in *CD-PG*$_1$ is illustrated by the graphical model [64] shown in Fig. 2. As shown in the figure, peer grade z_i^v, relative peer grade d_{ij}^v, and skill vector $\boldsymbol{\alpha_v}$ for grader v are the observed random variables in the model. The true score for student u_i's submission s_i, grader v's reliability τ_v, and bias b_v are the latent variables in the model to be estimated. The prior distribution of these latent variables is specified by the hyper-parameters μ_0, γ_0, η_0, and β_0. Formal definition of the *CD-PG*$_1$ model is given in Equation (3).

$$\begin{aligned}
\tau_v &\sim \Gamma\left(\prod_{k=1}^{K} {\alpha_{vk}}^{q_k}, \beta_0\right) \\
b_v &\sim \mathcal{N}\left(0, \frac{1}{\eta_0}\right) \\
s_i &\sim \mathcal{N}\left(\mu_0, \frac{1}{\gamma_0}\right) \\
z_i^v &\sim \mathcal{N}\left(s_i + b_v, \frac{1}{\tau_v}\right) \\
d_{ij}^v &\sim \mathcal{N}\left(s_i - s_j, \frac{2}{\tau_v}\right)
\end{aligned} \tag{3}$$

In the *CD-PG*$_1$ model, it is assumed that the true score, s_i, follows a Gaussian distribution with the mean equal to μ_0 and the variance equal to $1/\gamma_0$. Though different graders may have different biases in the peer assessment, we believe that the average bias of all graders is 0. Hence, the grader bias b_v is assumed to follow a zero-mean Gaussian distribution with the variance equal to $1/\eta_0$. Because grader reliability in an open-ended assignment is affected by his/her competence in the assignment, the reliability of grader v is modeled as a random variable that is assumed to follow a Gamma distribution where the shape parameter equals to v's competency value in the assignment (i.e., $\prod_{k=1}^{K} \alpha_{vk}{}^{q_k}$), and the rate parameter equals β_0. Based on the characteristics of Gamma distribution, the mean of grader v's reliability is δ_v/β_0. In the *CD-PG*$_1$ model, peer grade z_i^v, which is given by grader v to student u_i's submission, is assumed to follow a Gaussian distribution with the mean equal to the true score of submission s_i plus grader v's bias b_v, and the variance is inversely proportional to grader v's reliability (i.e., $1/\tau_v$). In the model, the relative peer grade d_{ij}^v, corresponding to grader v for grading student u_i's submission and student u_j's submission, is assumed to follow a Gaussian distribution with the mean equal to the difference between the true score for student u_i's submission and the

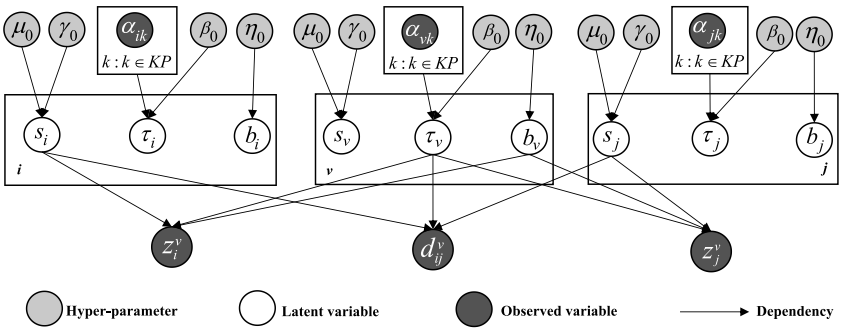


FIGURE 2. Graphical Model of *CD-PG*$_1$ and *CD-PG*$_2$.

true score for student u_j's submission (i.e., $s_i - s_j$), and the variance equals $2/\tau_v$.

B. CD-PG$_2$ MODEL

The formal definition of the *CD-PG*$_2$ model is shown in Equation 4, which shows the prior distribution information of different variables. Because the *CD-PG*$_2$ model also assumes that there is a dependency between the reliability of a grader and that grader's competency value in the open-ended assignment to be graded, the conditional dependence structure of *CD-PG*$_2$ is the same as that of *CD-PG*$_1$, which was shown in Fig. 2.

$$\begin{aligned}
\tau_v &\sim \mathcal{N}\left(\prod_{k=1}^{K} {\alpha_{vk}}^{q_k}, \frac{1}{\beta_0}\right) \\
b_v &\sim \mathcal{N}\left(0, \frac{1}{\eta_0}\right) \\
s_i &\sim \mathcal{N}\left(\mu_0, \frac{1}{\gamma_0}\right) \\
z_i^v &\sim \mathcal{N}\left(s_i + b_v, \frac{\lambda}{\tau_v}\right) \\
d_{ij}^v &\sim \mathcal{N}\left(s_i - s_j, \frac{2\lambda}{\tau_v}\right)
\end{aligned} \tag{4}$$

There are two differences between the *CD-PG*$_2$ model and the *CD-PG*$_1$ model. First, *CD-PG*$_2$ assumes that grader v's reliability τ_v follows a Gaussian distribution, while *CD-PG*$_1$ assumes that τ_v follows a Gamma distribution. Second, because *CD-PG*$_2$ assumes that grader v's reliability τ_v follows a Gaussian distribution, the scale of τ_v is determined by grader v's competency in the assignment to be graded by v (i.e., $\prod_{k=1}^{K} {\alpha_{vk}}^{q_k}$), which cannot be tuned. Under such circumstances, the variance of peer grade z_i^v and the variance of relative peer grade d_{ij}^v become non-tunable because they all depend on τ_v. To make the variance of z_i^v and the variance of d_{ij}^v tunable, *CD-PG*$_2$ introduces a hyper-parameter, λ, to specify the scale of the two variances. Then, z_i^v is assumed to follow a Gaussian distribution with a variance equal to λ/τ_v, and relative score d_{ij}^v is assumed to follow a Gaussian distribution with a variance equal to $2\lambda/\tau_v$.

C. MODEL INFERENCE

Given the above two formulated probabilistic graph models for cardinal peer assessment, the next phase is to infer the posterior distribution of every latent random variable (i.e., grader v's reliability τ_v, grader v's bias b_v, and the true score of student u_i's submission s_i), based on the values of observed random variables (i.e., peer grade z_i^v, relative peer grade d_{ij}^v, and grader v's skill vector $\boldsymbol{\alpha_v}$) in the model. Then, the estimated value of the true score for each student's submission can be derived. Thus, our inference problem can be formalized as $P(\{b_v|v \in V\}, \{\tau_v|v \in V\}, \{s_i|u_i \in U\}|Z, D, \boldsymbol{A})$.

One challenge in solving such inference problems comes from the correlations between latent variables in both models. For example, the dependency relationships among variables (see Fig. 2) show that the true score of student u_i's submission, s_i, can be accurately estimated only if the reliability, τ_v, of every grader who graded the submission can be accurately estimated. On the other hand, the dependency relationships among variables also indicate that to have a good estimation of grader v's reliability τ_v, a good estimation of the true scores of submissions graded by grader v are required. Therefore, our inference problem is a chicken-and-egg problem for inferring the posterior probability distributions of the latent variables based on the observable variables in the model [65]. To address this problem, the Gibbs sampling technique [66] is applied in this work. First, the Gibbs sampling is run for several iterations to draw a set of samples of a latent variable (e.g., $\{s_i^1, s_i^2, \ldots, s_i^{I_G}\}$, where I_G is the number of iterations) from an approximated posterior distribution. Then, the value of the latent variable is estimated based on the set of samples by empirical mean (e.g., $\bar{s}_i = \frac{1}{I_G}\sum_{t=1}^{I_G} s_i^t$), when the distribution of samples gradually tends to converge and stabilize. Considering that samples of latent variables generated in the burn-in iterations of Gibbs sampling are insufficiently accurate, samples generated in the burn-in iterations for each latent variable (generally the first n samples) are discarded.

The approximated posterior distributions for the latent variables in the *CD-PG*$_1$ model are as follows.

$$s \sim \mathcal{N}\left(\frac{Y}{R}, \frac{1}{R}\right),$$

where

$$\begin{aligned}
R &= \gamma_0 + \sum_{v \in V_{u_i}} \tau_v + \sum_{v \in V_{u_i}} \sum_{u_j \in U_v} \frac{\tau_v}{2}, \text{ and} \\
Y &= \mu_0\gamma_0 \\
&\quad + \tau_v \left(\sum_{v \in V_{u_i}} \left(z_i^v - b_v\right) + \sum_{v \in V_{u_i}} \sum_{u_j \in U_v} \frac{\left(d_{ij}^v + s_j\right)}{2} \right)
\end{aligned} \tag{5}$$

$$\begin{aligned}
\tau &\sim \Gamma\left(\prod_{k=1}^{K} {\alpha_{vk}}^{q_k} + \frac{|U_v|^2}{2}, \beta_0 \right. \\
&\quad \left. + \frac{\sum_{u_i \in U_v} \left(z_i^v - s_i - b_v\right)^2 + \sum_{u_i, u_j \in U_v} \frac{1}{2}\left(d_{ij}^v - s_i + s_j\right)}{2} \right)
\end{aligned} \tag{6}$$

$$b \sim \mathcal{N}\left(\frac{\sum_{u_i \in U_v} \tau_v \left(z_i^v - s_i\right)}{\eta_0 + |U_v|\,\tau_v}, \frac{1}{\eta_0 + |U_v|\,\tau_v} \right) \tag{7}$$

Following are the approximated posterior distributions for the latent variables in the *CD-PG*$_2$ model. Because, for latent variable τ_v in *CD-PG*$_2$, there is no closed-form distribution for the Gibbs samplings, a discrete approximation is performed to obtain the approximate posterior distribution of the variable.

$$s \sim \mathcal{N}\left(\frac{Y}{R}, \frac{1}{R}\right)$$

where

$$R = \gamma_0 + \sum_{v \in V_{u_i}} \frac{\tau_v}{\lambda} + \sum_{v \in V_{u_i}} \frac{\tau_v * (U_v \mid -1)}{2\lambda}, \text{ and}$$

$$Y = \gamma_0 \mu_0 + \frac{\tau_v}{\lambda} \left(\sum_{v \in V_{u_i}} \left(z_i^v - b_v \right) + \frac{\sum_{v \in V_{u_i}} \sum_{u_j \in U_v} \left(d_{ij}^v + s_j \right)}{2} \right) \tag{8}$$

$$\tau \propto \tau_v^{\frac{|U_v|}{2}} \times \exp\left(-\frac{\beta_0}{2} \left[\tau_v - \left(\prod_{k=1}^{K} \alpha_{vk}{}^{q_k} + \sum_{u_i \in U_v} \frac{\left(z_i^v - s_i - b_v \right)^2}{\lambda \beta_0} + \sum_{u_i, u_j \in U_v} \frac{\left(d_{ij}^v - s_i + s_j \right)^2}{2\lambda\beta_0} \right) \right]^2 \right) \tag{9}$$

$$b \sim \mathcal{N}\left(\frac{\sum_{u_i \in U_v} \frac{\tau_v}{\lambda} \left(z_i^v - s_i \right)}{\eta_0 + |U_v| \frac{\tau_v}{\lambda}}, \frac{1}{\eta_0 + |U_v| \frac{\tau_v}{\lambda}} \right) \tag{10}$$

The details of the inference process for the posterior distribution of every latent variable in the *CD-PG*$_1$ model and the *CD-PG*$_2$ model are provided in Appendix A.

D. MODEL INFERENCE ALGORITHM

In this section, the model inference algorithm for the two proposed probabilistic graph models is presented at first (Section IV-D1). Then, the complexity of the algorithm is analyzed (Section IV-D2).

1) ALGORITHM DESCRIPTION

Algorithm 1 presents the pseudocode of the model inference algorithm for models *CD-PG*$_1$ and *CD-PG*$_2$. As shown in Algorithm 1, the $\boldsymbol{R}$-matrix that records the responses of every grader to each question in the historical tests and assignments and the $\boldsymbol{Q}$-matrix that records the examined knowledge points of each question in the historical tests and assignments are two important inputs to the algorithm. Based on the $\boldsymbol{R}$-matrix and the $\boldsymbol{Q}$-matrix, the DINA model is applied to compute the skill vector of each grader and finally derive the $\boldsymbol{A}$-matrix, which consists of the skill vectors of all graders (Line 1). Then, based on one of the probabilistic graph models proposed in this paper, the prior probability distributions of latent variables s_i, τ_v, and b_v are assigned separately based on the setting of the model (Line 2). Next, Gibbs sampling is executed for I_G iterations to generate a sample set for every latent variable (Lines 3-17). For each iteration of the Gibbs sampling, Equations 5-7 are used to get samples for the latent variables if the *CD-PG*$_1$ model is applied, and Equations 8-10 are used to get samples for the latent variables if the *CD-PG*$_2$ model is applied. Let $\xi^{(t)}$ denote the set of sample sets of all latent variables for the t-th iteration. Then, each $\xi^{(t)}$ generated in the burn-in iteration of Gibbs sampling (i.e., $\xi^{(t)}$ with $t \leq \theta$) is discarded. Here, θ is the threshold of determining the burn-in iterations. Finally, the empirical mean of the remaining sample sets with respect to each latent variable is used as the final estimation for the latent variable (Line 18) and return the estimated values for the latent variables $\hat{s}_i$, $\hat{\tau}_v$, and $\hat{b}_v$ (Line 19).

Algorithm 1 Model Inference Algorithm

Input: the set of students U, set of graders V, set of knowledge points KP, $\boldsymbol{Q}$-matrix, $\boldsymbol{R}$-matrix, set of all observed peer grades Z, set of all relative peer grades D, number of iterations of Gibbs sampling I_G, threshold for determining the burn-in iterations θ, and probabilistic graph model for peer assessment *CD-PG*$_x$.

Output: $(\hat{s}_i, \hat{\tau}_v, \hat{b}_v)$ for all $u_i \in U$ and $v \in V$

1: $\boldsymbol{A}$ = DINA($\mathbb{s}_0$, $\mathbb{g}_0$, $\boldsymbol{Q}$, $\boldsymbol{R}$);
2: s_i, τ_v, b_v = setDistribution(*CD-PG*$_x$);
3: **for** each $t = 1 \rightarrow I_G$ **do**
4: **for** each s_i with $u_i \in U$ **do**
5: s'=gradeSampling($Z, D, \boldsymbol{A}$);
6: $s_{u_i} \leftarrow s'$;
7: **end for**
8: **for** each τ_v with $v \in V$ **do**
9: τ'=reliabilitySampling($Z, D, \boldsymbol{A}$);
10: $\tau_{v_i} \leftarrow \tau'$;
11: **end for**
12: **for** each b_v with $v \in V$ **do**
13: b'=biasSampling(Z);
14: $b_{v_i} \leftarrow b'$;
15: **end for**
16: $\xi^{(t)} \leftarrow (\{s_i \mid u_i \in U\}, \{\tau_v \mid v \in V\}, \{b_v \mid v \in V\})$;
17: **end for**
18: $(\{\hat{s}_i \mid u_i \in U\}, \{\hat{\tau}_v \mid v \in V\}, \{\hat{b}_v \mid v \in V\}) \leftarrow \frac{1}{I_G - \theta} \sum_{t=\theta+1}^{I_G} \xi^{(t)}$;
19: **return** $(\{\hat{s}_i \mid u_i \in U\}, \{\hat{\tau}_v \mid v \in V\}, \{\hat{b}_v \mid v \in V\})$;

2) COMPLEXITY ANALYSIS

The model inference algorithm described in Section IV-D1 has two functional modules: (1) the module that calculates the $\boldsymbol{A}$-matrix, which is composed of skill vectors of all graders derived from the cognitive diagnosis model – DINA (see Algorithm 1: Line 1); (2) the module that runs I_G iterations of Gibbs sampling to get I_G sample sets for every latent variable in the proposed probabilistic graph model (see Algorithm 1: Lines 3-17). Because the EM algorithm is used in the DINA model to compute the skill vectors of all graders, the time complexity of the first functional module in Algorithm 1 is equal to $O(|V| \times 2^{|KP|} \times I_{EM})$, where $|V|$ denotes the cardinality of graders, $|KP|$ denotes the cardinality of knowledge points, and I_{EM} represents the number of iterations in the EM algorithm. The time complexity of the second functional module of Algorithm 1 is $O(|U| \times |V_{u_i}| \times I_G + |V| \times |U_v| \times I_G)$, where $|U|$ is the cardinality of students who have submitted their open-ended assignments; $|V_{u_i}|$ denotes the cardinality of graders who assign grades to the submission of student u_i; $|U_v|$ denotes the cardinality of students whose submissions

are evaluated by grader v, and I_G is the number of iterations for Gibbs sampling. In real-world teaching practice, we have: (1) the students who make submissions to open-ended assignments are generally asked to participate in the peer assessment of the open-ended assignment, i.e., $|U| = |V|$; (2) the number of graders who give grades to the submission of a student generally equals the number of students whose submissions are evaluated by a grader, i.e., $|V_{u_i}| = |U_v|$. Therefore, the time complexity of the second functional module of Algorithm 1 can be simplified to $O(|U| \times |V_{u_i}| \times I_G)$. Integrating the time complexity of the two modules, the time complexity of Algorithm 1 is thus $O((|V| \times 2^{|KP|} \times I_{EM} + |U| \times |V_{u_i}| \times I_G)$.

Assuming that the cardinality of students who have submitted their open-ended assignments is far greater than both the number of knowledge points (i.e., $|U| \gg |KP|$) and the number of questions in historical tests and assignments (i.e., $|U| \gg |E|$), the memory consumption of Algorithm 1 mainly comes from storing every set of sample sets for latent variables generated in each iteration of Gibbs sampling (i.e., $\{\xi^{(t)}|1 \leq t \leq I_G\}$). Because the size of every $\xi^{(t)}$ is $I_G \times (|U| + 2|V|)$ and $|U| = |V|$, the storage space complexity of Algorithm 1 is $O(I_G \times |U|)$.

V. EXPERIMENTS

In this section, the experiments conducted to compare the performances of our proposals with other state-of-the-art methods for peer assessment are expatiated.

A. REAL DATASET

1) PEER GRADING DATASET

The real peer grading dataset was collected from a MOOC called "Database Principles" on the MOOC platform of Guangxi University by conducting a group of peer-assessment activities. This dataset contains the peer grades given by students and the grades given by two experienced teachers. The peer grades in the dataset are the results from 284 undergraduates majoring in computer science grading their peers' submissions corresponding to three assignments. Each assignment includes one open-ended question investigating the normalization theory of relational databases. Specifically, there are a total of 11 knowledge points about the normalization theory being investigated by the three assignments, and the ids and names of these knowledge points are as follows: (1) 1NF (the first normal form); (2) 2NF (the second normal form); (3) 3NF (the third normal form); (4) BCNF (the Boyce-Codd normal form); (5) primary attribute; (6) transitive functional dependency; (7) determinant; (8) functional dependency; (9) key; (10) partial functional dependency; (11) non-primary attribute. These 11 knowledge points are the more difficult knowledge points of the course because they are theoretical in nature. Compared with objective questions (e.g., single-choice questions), open-ended questions have proved to be more effective in helping students master these knowledge points. Fig. 3 shows the $\boldsymbol{Q}$-matrix that records the ids of the required knowledge points for correctly answering each assignment, as determined by the two experienced teachers.

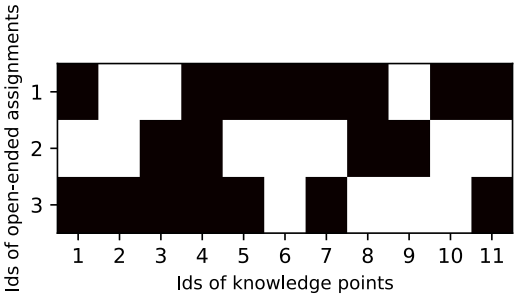


FIGURE 3. Q-matrix for the three open-ended assignments.

For the setting of each peer-assessment activity, every student who has submitted an open-ended assignment is required to act as a grader and give grades to three peers' submissions, according to rubrics specified by the teacher. The MOOC platform assigns submissions among graders randomly and ensures that each submission is graded by three graders. The identities of students who have submitted an open-ended assignment are concealed from the graders, and vice versa, throughout the entire process. At the end of the peer assessment, the MOOC platform uses the median of the peer grades given by the graders as the estimated final score of a submission.

Besides the peer grades given by the students, this dataset also contains grades given by two experienced teachers who have at least six years of teaching experience for the course. The average grade given by the two teachers for a student's submission to an assignment is considered as the ground-truth score for the submission. The summary statistics of the peer grading dataset are listed in Table 3.

TABLE 3. Summary of peer grading dataset for open-ended assignments.

	Assignment 1	Assignment 2	Assignment 3
Submissions	268	269	267
Teacher grades	536	538	534
Graders	254	260	254
Peer grades	694	725	690
Full grades	20	20	20

2) HISTORICAL ONLINE TEST DATASET

The two proposed probabilistic graph models (i.e., $CD\text{-}PG_1$ and $CD\text{-}PG_2$) make use of the competency information of a grader in an assignment to model that grader's reliability in grading the assignment. To calculate the competency values of graders in the three open-ended assignments in peer assessment, students were asked to complete an online test that contains 40 objective questions about the 11 knowledge points that are investigated by the three open-ended assignments mentioned above. The objective questions include single choice questions, multiple-choice questions, and judgment questions. The historical online test dataset contains the

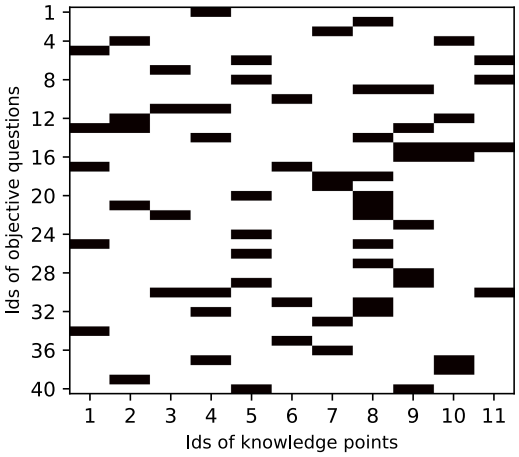


FIGURE 4. Q**-matrix of the 40 objective questions in the online test.**

$\boldsymbol{R}$-matrix, which stores the collected responses of students to the 40 objective questions, and the $\boldsymbol{Q}$-matrix, which is provided by the two teachers and records the ids of the knowledge points required to answer each question correctly. The $\boldsymbol{Q}$-matrix of the 40 objective questions in the online test is shown in Fig. 4. Then, the $\boldsymbol{R}$-matrix and $\boldsymbol{Q}$-matrix are used by the DINA model to compute the competency values of graders in the three open-ended assignments to be graded by them.

B. COMPARISON METHODS

To evaluate the effectiveness of our proposed probabilistic graph models, *CD-PG*$_1$ and *CD-PG*$_2$, our models are compred with the following state-of-the-art methods for peer assessment.

- **Median**: This method takes the median of peer grades given to a submission as the final score of the submission. This is the method used by most MOOC platforms to aggregate peer grades.
- **Mean**: This method simply assigns the mean of peer grades given to a submission as the final score of the submission.
- PG_6 **and** PG_7 [14]: PG_6 and PG_7 are state-of-the-art methods which solve the peer-assessment problem of open-ended assignments. These two models are probabilistic graph models that model the reliability and bias of each grader. Compared with previous research, the innovation of these two models stems from the introduction of relative peer grades into the model to obtain a more precise estimate of the true score of each submission. The difference between PG_6 and PG_7 is that PG_6 assumes that the prior distribution of grader reliability satisfies a Gamma distribution, while PG_7 assumes that it satisfies a Gaussian distribution.

Similar to PG_6 and PG_7, the proposed *CD-PG*$_1$ and *CD-PG*$_2$ models are also probabilistic graph models considering both the reliability and bias of each grader. Unlike PG_6 and PG_7, *CD-PG*$_1$ and *CD-PG*$_2$ apply graders' performance data in historical tests and assignments and use a CDM to derive graders' competencies in the open-ended assignments they grade, which are then used to improve the modeling of graders' reliabilities. *CD-PG*$_1$ corresponds to PG_6, both of which assume that the prior distribution of grader reliability follows a Gamma distribution. *CD-PG*$_2$ corresponds to PG_7, both of which assume that the prior distribution of grader reliability follows a Gaussian distribution.

C. EXPERIMENTAL SETUP

Hyper-parameters are used in the proposed probabilistic graph models (i.e., *CD-PG*$_1$ and *CD-PG*$_2$) and the related probabilistic graph models (i.e., PG_6 and PG_7), and it is important to set reasonable values for them. s_i, which represents the true score of student u_i's submission to an assignment, is the most important latent variable, and its prior distribution is assumed to be a Gaussian distribution by all models. The hyper-parameters of the Gaussian distribution, namely the mean (μ_0) and the variance ($1/\gamma_0$) of the distribution, are set as the mean and the variance, respectively, of the peer grades of all students' submissions to an assignment. Meanwhile, as claimed in the literature [14], β_0 in *CD-PG*$_1$ and PG_6, which decides the rate of the Gamma distribution for grader reliability, and λ in *CD-PG*$_2$ and PG_7, which determines the variance of the Gaussian distribution for peer grades, are the most critical hyper-parameters for these respective models. This is because the settings of these two hyper-parameters have a significant influence on the estimation accuracy of the true scores, while the impact of other hyper-parameters on the estimation accuracy is very small if they are set within a reasonable range. Therefore, β_0 in *CD-PG*$_1$ and PG_6 and λ in *CD-PG*$_2$ and PG_7 were the main tuned hyper-parameters in our experiments. Specifically, for *CD-PG*$_1$ and PG_6, with other variables being set to fixed values, the hyper-parameter β_0 is searched in the range of $[450, 700]$ with an interval of 50 to get the best performance, by following the tuning idea proposed in [12], [14]. As to *CD-PG*$_2$ and PG_7, the hyper-parameter λ is searched in the range of $[0.01, 0.25]$ with a step of 0.05, while the other variables were set to fixed values, and finally used the value of λ that obtained the best accuracy in estimating the true scores. Besides, the hyper-parameter η_0 is fine-tuned in every model in the range of $[0.04, 0.2]$ by following the tuning strategy proposed in [12]. The hyper-parameter β_0 involved in *CD-PG*$_2$ and PG_7 was set to 0.1, as proposed in [14]. For each model, the model inference algorithm that infers the values of latent variables in the model was executed 10 times, and the average estimated value of the 10 executions for latent variables was used in the experimental evaluation. During each execution of the model inference algorithm, every latent variable was sampled based on the Gibbs sampling method for 600 iterations, and the remaining samples after discarding the samples generated in the first 60 iterations (i.e., the burn-in iterations) were used to estimate the value of the latent variable.

All the peer-assessment methods involved in the comparison were implemented using Python (version 3.7) and were tested on a server running a 64-bit Windows 10 operating system equipped with an i5-8500 3 GHz CPU, 8 GB of memory, and a 1-TB hard disk drive. The RMSE is used to measure the deviations of the estimated scores from the ground-truth scores given by teachers. RMSE is a widely used metric for evaluating the effectiveness of cardinal peer-assessment methods [2], [13]. The formal definition of RMSE is given by Equation 11, where s_i represents the ground-truth score of student u_i's submission to an open-ended assignment; $\hat{s}_i$ denotes the estimated true score for the submission given by a peer-assessment method, and $|U|$ is the cardinality of students who have submitted the assignment.

$$\text{RMSE} = \sqrt{\frac{1}{|U|}\sum_{u_i \in U}\left(s_i - \hat{s}_i\right)^2} \tag{11}$$

D. PERFORMANCE ON A REAL DATASET

1) ACCURACY OF ESTIMATION

Table 4 compares the accuracy of the estimated true scores given by different peer-assessment methods. For the probabilistic-graph-model-based methods (i.e., PG_6, PG_7, $CD\text{-}PG_1$, and $CD\text{-}PG_2$), RMSE and STD refer to the *average* and the *standard deviation* of the 10 RMSE values over 10 executions of the model inference algorithm for each method, respectively. Table 4 shows that the *Median* method and the *Mean* method are the least accurate methods. This is because they fail to consider the reliability and the bias of graders, which are proven to be very important for enhancing the estimation accuracy. The proposed $CD\text{-}PG_1$ and $CD\text{-}PG_2$ methods based on cognitive diagnosis, are the most accurate methods compared with the other state-of-the-art solutions. In particular, the RMSE values of the $CD\text{-}PG_2$ method are on average 69% lower than those of the *Median* method, which is the mainstream peer-assessment method adopted by most popular MOOC platforms. It can also be seen from the table that in most cases (i.e., Assignment 1 and Assignment 2), the RMSE values of $CD\text{-}PG_2$ are lower than those of $CD\text{-}PG_1$. This shows that when the prior distribution of grader reliability is set to a Gaussian distribution, a probabilistic graph model better fits the peer grading dataset in most cases.

Because both $CD\text{-}PG_1$ and PG_6 assume that the prior distribution of grader reliability follows a Gamma distribution, and both $CD\text{-}PG_2$ and PG_7 assume that it follows a Gaussian distribution, these two pairs of models are compared as follows:

- ***CD-PG*$_1$ vs. *PG*$_6$:** From Table 4 we can see that $CD\text{-}PG_1$ and PG_6 have similar STDs of RMSE for all three assignments, and all their STD values are small. This indicates that both models act quite stably in predicting the true scores. It can also be observed from the table that the RMSE of $CD\text{-}PG_1$ is significantly lower than that of PG_6 under each assignment. In particular, the average RMSE of the true scores estimated by $CD\text{-}PG_1$ is on average 40% lower than that of PG_6 for all three assignments.
- ***CD-PG*$_2$ vs. *PG*$_7$:** It can be seen from Table 4 that the STDs of RMSE are the same for PG_7 and $CD\text{-}PG_2$ in all three assignments. The maximum STD of RMSE for these two models is only 0.01, which indicates that these two models also perform very stably in estimating the true scores. Moreover, the RMSE of $CD\text{-}PG_2$ is apparently lower than that of PG_7 for all settings, and the RMSE of $CD\text{-}PG_2$ is on average 41% lower than that of PG_7 for all three assignments.

In summary, by leveraging the diagnosed competency information of graders in the open-ended assignments and making use of such information to optimize the modeling of graders' reliabilities, the $CD\text{-}PG_1$ and $CD\text{-}PG_2$ methods successfully improved the accuracy of peer assessment, compared to the state-of-the-art methods.

TABLE 4. Experimental results.

	Assignment 1		Assignment 2		Assignment 3	
	RMSE	STD	RMSE	STD	RMSE	STD
Mean	4.61	0.00	4.16	0.00	4.53	0.00
Median	5.09	0.00	4.59	0.00	5.04	0.00
PG_6	3.32	0.01	2.67	0.01	3.32	0.02
$CD\text{-}PG_1$	2.37	0.01	1.78	0.02	**1.32**	**0.01**
PG_7	2.46	0.01	2.56	0.00	2.82	0.01
$CD\text{-}PG_2$	**1.63**	**0.01**	**1.31**	**0.00**	1.68	0.01

2) MAXIMUM GRADING DEVIATION

Table 5 compares the maximum grading deviations of different peer assessment methods by comparing them to the ground-truth scores of students' submissions. It shows that the maximum grading deviations of the *Mean* method and the *Median* method are both greater than that of the peer assessment method based on the probabilistic graph model. This is because the *Mean* method and the *Median* method estimate the true scores of students' submissions only by the peer grades; if the grading quality of graders with respect to a submission happens to be very low, these two methods will give very poor estimates for the score of the submission. In contrast, the probabilistic-graph-model-based peer assessment methods not only consider the peer grades, but also model the graders' reliabilities and biases to enhance

TABLE 5. Maximum deviation between an estimated grade and the ground truth for all students.

	Assignment 1	Assignment 2	Assignment 3
Mean	18	8	16
Median	18	8	16
PG_6	10.87	6.31	10.46
PG_7	10.74	6.36	10.81
$CD\text{-}PG_1$	6.26	5.51	**4.47**
$CD\text{-}PG_2$	**5.75**	**4.34**	5.24

the accuracy of estimation, and therefore they obtain more precise estimates for the scores of students' submissions. Meanwhile, it can also be observed from the table that the two proposed models have smaller maximum grading deviations than PG_6 and PG_7, which are also designed based on the probabilistic graph model. This observation shows that, by optimizing the model of grader reliability based on the grader's competency in the grading assignment, which is quantified by cognitive diagnosis, the proposed *CD-PG*$_1$ and *CD-PG*$_2$ models outperform the state-of-the-art peer assessment methods in terms of guaranteeing the accuracy of the estimated score for each submission.

3) SENSITIVITY OF HYPER-PARAMETERS

To show how hyper-parameter β_0 in the *CD-PG*$_1$ model and hyper-parameter λ in the *CD-PG*$_2$ model will influence the performance of the two models, experiments are conducted using different values for these two hyper-parameters, with all other parameters being fixed. Specifically, the value of β_0 in the *CD-PG*$_1$ model was set in the range of [450, 700] with an interval of 50, and the value of λ in the *CD-PG*$_2$ model was set in the range of $[0.01, 0.25]$ with an interval of 0.05. The results shown in Fig. 5 and Fig. 6 indicate that, within a reasonable range, these two models are robust to the settings of the hyper-parameters, and the RMSEs of their estimated scores for students' assignments are acceptable.

4) RUNNING TIME

Because the *Mean* and *Median* methods for peer assessment simply use the mean and the median of peer grades to predict the scores of students' assignments, their running time is very short, and these two methods are not compared in terms of their running time. Fig. 7 compares the running time of the model inference algorithm with respect to the probabilistic graph models compared in this paper. Specifically, the value of each running time in the figure is the average time consumption by executing the model inference algorithm of a probabilistic graph model 10 times with the same parameters.

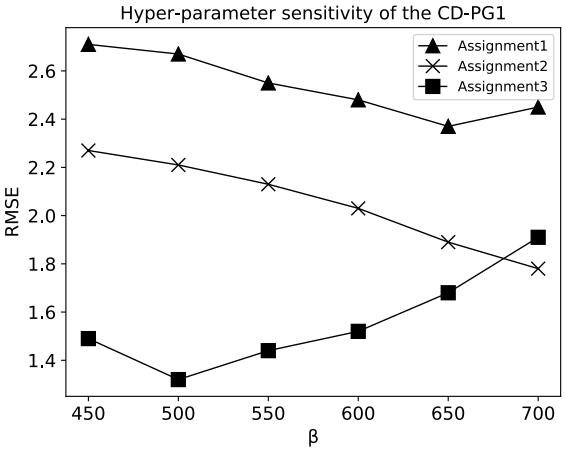


FIGURE 5. Sensitivity analysis of hyper-parameter β_0 for *CD-PG*$_1$.

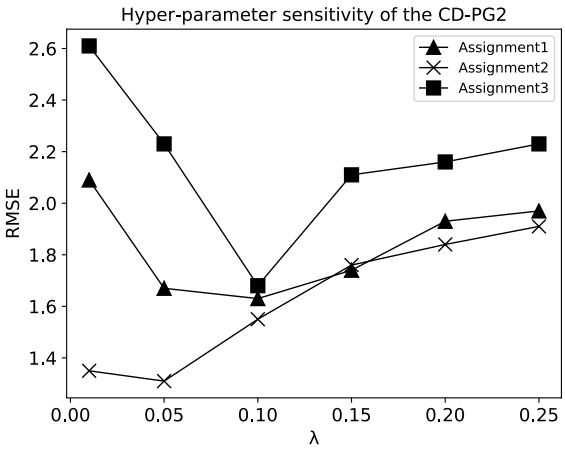


FIGURE 6. Sensitivity analysis of hyper-parameter λ for *CD-PG*$_2$.

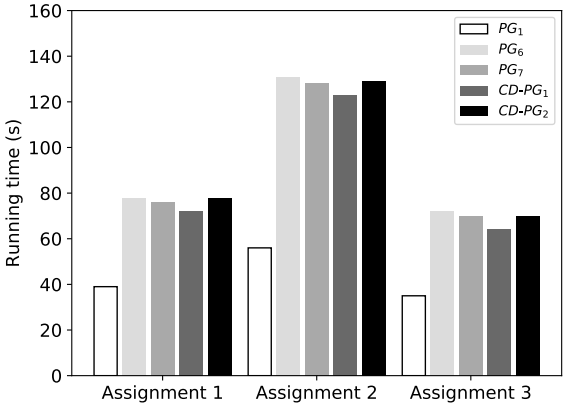


FIGURE 7. Running time comparison of probabilistic graph models.

As shown in Fig. 7, for different open-ended assignments, the running time of the model inference algorithm for each probabilistic graph model is greater than one minute. This is because Gibbs sampling is executed iteratively 600 times in each algorithm, which is the primary contributor to time consumption. It can also be observed from the figure that the time consumption in Assignment 2 is apparently greater than those in Assignment 1 and Assignment 3. This is because more peer grades were collected for Assignment 2 (see Table 3 for details), which increases the execution time. Another conclusion drawn from the figure is that the *CD-PG*$_1$ model costs less time in inferring the scores compared with other models because the posterior distribution of each latent variable in it has a closed form, which shortens the running time of its model inference algorithm. As for the other three models, a closed form of the posterior distribution for some of the latent variables in them cannot be found, so they consume more time than the *CD-PG*$_1$ model.

5) SUMMARY

Experimental results show that the proposed two cardinal estimation models for MOOCs gain on average 70%

reduction for the RMSE value compared with the simple aggregation strategies, i.e., computing the mean or the median of peer grades, which are widely used by mainstream MOOC platforms. Moreover, *CD-PG*$_1$ and *CD-PG*$_2$ beat the state-of-the-art probabilistic graph models, i.e., *PG*$_6$ and *PG*$_7$, for averagely reducing the RMSE value of compared models by 50%. Note that *CD-PG*$_1$ and *CD-PG*$_2$ differ from *PG*$_6$ and *PG*$_7$ due to the consideration of the diagnosed competency information of graders in modeling their grading reliability. Thus, applying cognitive diagnosis into the procedure of peer assessment is an effective way to get better estimates of the true scores of assignments. We are delighted to observe that some additional educational benefits can also be gained by employing the probabilistic graph models in peer assessment. For one, feedback collected from teachers who engaged in our peer assessment activities shows that the derived reliabilities and biases of graders computed by the models are important indicators to students' future performance in MOOCs, which once again explains the advantages of the proposed models from another point of view.

VI. CONCLUSION AND FUTURE WORK

With the proliferation of MOOCs, peer assessment has become the mainstream paradigm for large-scale grading of open-ended assignments. Because the biases and reliabilities of graders are unknown, it is a challenging problem to estimate the true score of a student's assignment based on peer grades given by multiple peer graders. Existing works contribute to the development of effective score-estimation methods by modeling grader bias and reliability. However, they ignore an important aspect in the modeling of grader reliability, which is the competency of the grader in the specific assignment to be graded. Real peer-assessment practices for open-ended assignments show that modeling the reliability of graders in terms of their competencies in the graded assignments can help improve the robustness of score estimation in peer assessment. In this paper, two probabilistic graph models are proposed that leverage graders' competency information in graded assignments to optimize the modeling of graders' reliabilities and achieve more accurate estimation of true scores. Such information about graders is determined using the cognitive diagnosis model, DINA, based on the performances of graders gained from historical tests or assignments. Moreover, an effective model inference algorithm is proposed to infer both model parameters and the true scores of students' assignments. Experimental results based on a real peer-grading dataset show that the two proposed models improve the accuracy of cardinal peer assessment.

Apart from the field of MOOCs, the proposed models can also be applied to crowdsourcing, if a crowdsourced task needs to predict a certain metric based on skill diagnosis information about crowdworkers. In the future, we will attempt to introduce other influencing factors with respect to the reliabilities and biases of graders, to further improve the models for peer assessment.

APPENDIX A
INFERENCE PROCESS

A. INFERENCE PROCESS FOR THE CD-PG$_1$ MODEL

The joint posterior distribution is

$$\begin{aligned} &P\left(Z, D \mid \{s_i\}_{u_i \in U}, \{b_v\}_{v \in V}, \{\tau_v\}_{v \in V}\right) \\ &\quad = \prod_i P\left(s_i \mid \mu_0, \gamma_0\right) \cdot \prod_v P\left(b_v \mid \eta_0\right) \\ &\qquad \cdot P\left(\tau_v \mid \prod_{k=1}^{K} {\alpha_{vk}}^{q_k}, \beta_0\right) \cdot \prod_{z_i^v} P\left(z_i^v \mid s_i, b_v, \tau_v\right) \\ &\qquad \cdot \prod_{d_{ij}^v} P\left(d_{ij}^v \mid s_i, s_j, \tau_v\right). \end{aligned} \tag{A.1}$$

Markov blanket (i.e., MB) denotes that a given feature is independent of all other feature conditions in the feature domain under its MB condition. For example, $MB(s_i)$ indicates that true score s_i is independent of other characteristic variables, namely, grader bias b_v and grader reliability τ_v. Therefore, when inferring s_i, variable s_i is fixed, while other variables are randomly initialized.

Consider now a fixed student u_i, who has submitted an assignment. The sampling step for s_i (i.e., the true score of student u_i's submission) is derived as follows:

$$\begin{aligned} s &\sim P\left(s_i \mid MB\left(s_i\right)\right), \\ &\propto P\left(s_i \mid \mu_0, \gamma_0\right) \cdot \prod_{v \in V_{u_i}} P\left(z_i^v \mid s_i, b_v, \tau_v\right) \\ &\quad \cdot \prod_{v \in V_{u_i}, u_j \in U_v} P\left(d_{ij}^v \mid s_i, s_j, \tau_v\right), \\ &\propto \exp\left(-\frac{1}{2}\gamma_0\left(s_i - \mu_0\right)^2\right. \\ &\quad + \sum_{v \in V_{u_i}} \left(-\frac{1}{2}\tau_v\left(z_i^v - \left(s_i + b_v\right)\right)^2\right) \\ &\quad \left. + \sum_{v \in V_{u_i}} \sum_{u_j \in U_v} \left(-\frac{1}{4}\tau_v\left(d_{ij}^v - \left(s_i - s_j\right)\right)^2\right)\right), \\ &\propto \exp\left(-\frac{1}{2}\left[\gamma_0\left(s_i - \mu_0\right)^2\right.\right. \\ &\quad \left.\left. + \sum_{v \in V_{u_i}} \left(\tau_v\left(z_i^v - \left(s_i + b_v\right)\right)^2\right)\right]\right) \\ &\quad \times \exp\left(\sum_{v \in V_{u_i}} \sum_{u_j \in U_v} \left(-\frac{1}{4}\tau_v\left(d_{ij}^v - \left(s_i - s_j\right)\right)^2\right)\right). \end{aligned} \tag{A.2}$$

The expression inside the exponent is quadratic; we thus complete the square, obtaining

$$\begin{aligned} &\gamma_0\left(s_i - \mu_0\right)^2 + \sum_{v \in V_{u_i}} \left(\tau_v\left(z_i^v - \left(s_i + b_v\right)\right)^2\right) \\ &\quad + \sum_{v \in V_{u_i}} \sum_{u_j \in U_v} \left(\frac{1}{2}\tau_v\left(d_{ij}^v - \left(s_i - s_j\right)\right)^2\right) \end{aligned}$$

$$
\begin{aligned}
&= \text{const.} + \gamma_0 \left(s_i^2 - 2\mu_0 s_i\right) \\
&\quad + \sum_{v \in V_{u_i}} \tau_v \left((s_i + b_v)^2 - 2\, z_i^v\, (s_i + b_v)\right) \\
&\quad + \sum_{v \in V_{u_i}} \sum_{u_j \in U_v} \left(\frac{1}{2}\tau_v \left((s_i - s_j)^2 - 2\, d_{ij}^v\, (s_i - s_j)\right)\right), \\
&= \text{const.} + \left(\gamma_0 + \sum_{v \in V_{u_i}} \tau_v + \sum_{v \in V_{u_i}} \sum_{u_j \in U_v} \frac{\tau_v}{2}\right) s_i^2 \\
&\quad - 2\, s_i \left(\gamma_0 \mu_0 + \sum_{v \in V_{u_i}} \tau_v \left(z_i^v - b_v\right)\right) \\
&\quad - \sum_{v \in V_{u_i}} \sum_{u_j \in U_v} s_i \tau_v \left(d_{ij}^v + s_j\right), \\
&= \text{const.} + R \left(s_i - \frac{Y}{R}\right)^2, \\
\text{where } R &= \gamma_0 + \sum_{v \in V_{u_i}} \tau_v + \sum_{v \in V_{u_i}} \sum_{u_j \in U_v} \frac{\tau_v}{2}, \\
\text{and } Y &= \mu_0 \gamma_0 + \sum_{v \in V_{u_i}} \tau_v \left(z_i^v - b_v\right) \\
&\quad + \sum_{v \in V_{u_i}} \sum_{u_j \in U_v} \frac{\tau_v \left(d_{ij}^v + s_j\right)}{2}. \qquad \text{(A.3)}
\end{aligned}
$$

Therefore, the sampling distribution is Gaussian:

$$
s \sim \mathcal{N}\left(\frac{Y}{R}, \frac{1}{R}\right). \qquad \text{(A.4)}
$$

Now, consider a fixed grader v. The sampling step for grader reliability τ_v is derived as follows:

$$
\begin{aligned}
\tau &\sim P\left(\tau_v \mid MB\left(\tau_v\right)\right), \\
&\propto P\left(\tau_v \mid \prod_{k=1}^{K} {\alpha_{vk}}^{q_k}, \beta_0\right) \cdot \prod_{u_i \in U_v} P\left(z_i^v \mid s_i, b_v, \tau_v\right) \\
&\quad \cdot \prod_{u_i, u_j \in U_v} P\left(d_{ij}^v \mid s_i, s_j, \tau_v\right), \\
&\propto {\tau_v}^{\prod_{k=1}^{K} {\alpha_{v_k}}^{q_k}} \\
&\quad \times \exp\left(-\beta_0 \tau_v + \sum_{u_i \in U_v} \sqrt{\frac{\tau_v}{2\pi}} \left(-\frac{\tau_v}{2}\left(z_i^v - (s_i + b_v)\right)^2\right)\right. \\
&\quad \left. + \sum_{u_i, u_j \in U_v} \sqrt{\frac{\tau_v}{4\pi}} \left(-\frac{\tau_v}{4}\left(d_{ij}^v - (s_i - s_j)\right)^2\right)\right), \\
&\propto {\tau_v}^{\prod_{k=1}^{K} {\alpha_{v_k}}^{q_k} + \frac{|U_v|^2}{2}} \\
&\quad \times \exp\left[-\beta_0 + \frac{1}{2}\left(\sum_{u_i \in U_v} \left(z_i^v - (s_i + b_v)\right)^2\right.\right. \\
&\quad \left.\left. + \sum_{u_i, u_j \in U_v} \left(d_{ij}^v - (s_i - s_j)\right)^2\right)\right] \tau_v. \qquad \text{(A.5)}
\end{aligned}
$$

From this, the sampling distribution can be recognized to be Gamma with

$$
\begin{aligned}
\tau \sim \Gamma\left(\prod_{k=1}^{K} {\alpha_{vk}}^{q_k} + \frac{|U_v|^2}{2}, \beta_0 + \frac{\sum_{u_i \in U_v} \left(z_i^v - s_i - b_v\right)^2}{2}\right. \\
\left. + \frac{\sum_{u_i, u_j \in U_v} \left(d_{ij}^v - s_i + s_j\right)}{4}\right). \qquad \text{(A.6)}
\end{aligned}
$$

Finally, the sampling set for grader bias b_v is derived as follows:

$$
\begin{aligned}
b &\sim P\left(b_v \mid MB\left(b_v\right)\right), \\
&\propto P\left(b_v \mid \eta_0\right) \cdot \prod_{u_i \in U_v} P\left(z_i^v \mid s_i, b_v, \tau_v\right), \\
&\propto \exp\left(-\frac{1}{2}\eta_0 b_v^2 - \frac{1}{2}\sum_{u_i \in U_v} \tau_v \left(z_i^v - (s_i + b_v)\right)^2\right), \\
&\propto \exp\left(-\frac{1}{2}\left[\eta_0 b_v^2\right.\right. \\
&\quad \left.\left. + \sum_{u_i \in U_v} \tau_v \left((s_i + b_v)^2 - 2 z_i^v\, (s_i + b_v)\right)\right]\right). \qquad \text{(A.7)}
\end{aligned}
$$

The expression inside the exponent is quadratic; we thus complete the square as follows:

$$
\begin{aligned}
&\eta_0 b_v^2 + \sum_{u_i \in U_v} \tau_v \left((s_i + b_v)^2 - 2 z_i^v\, (s_i + b_v)\right) \\
&\quad = \text{const.} + \left(\eta_0 + \sum_{u_i \in U_v} \tau_v\right) b_v^2 \\
&\qquad - 2\left(\sum_{u_i \in U_v} \tau_v \left(z_i^v - s_i\right)\right) b_v, \\
&\quad = \text{ const. } + R\left(b_v - \frac{Y}{R}\right)^2, \\
\text{where } R &= \eta_0 + \sum_{u_i \in U_v} \tau_v = \eta_0 + |U_v|\, \tau_v, \\
\text{and } Y &= \sum_{u_i \in U_v} \tau_v \left(z_i^v - s_i\right). \qquad \text{(A.8)}
\end{aligned}
$$

The sampling distribution for b is thus Gaussian, with

$$
b \sim \mathcal{N}\left(\frac{\sum_{u_i \in U_v} \tau_v \left(z_i^v - s_i\right)}{\eta_0 + |U_v|\, \tau_v}, \frac{1}{\eta_0 + |U_v|\, \tau_v}\right). \qquad \text{(A.9)}
$$

B. INFERENCE PROCESS FOR THE CD-PG$_2$ MODEL

Consider now a fixed student u_i who has submitted an assignment. The sampling step for s_i (i.e., the true score of student u_i's submission to the assignment) is derived as follows:

$$
\begin{aligned}
s &\sim P\left(s_i \mid MB\left(s_i\right)\right), \\
&\propto P\left(s_i \mid \mu_0, \gamma_0\right) \cdot \prod_{v \in V_{u_i}} P\left(z_i^v \mid s_i, b_v, \tau_v\right)
\end{aligned}
$$

$$
\begin{aligned}
&\cdot \prod_{v\in V_{u_i}} \prod_{u_j\in U_v} P\left(d_{ij}^v \mid s_i, s_j, \tau_v\right), \\
&\propto \exp\left(-\frac{1}{2}\gamma_0\left(s_i-\mu_0\right)^2\right) \\
&\quad\times \exp\left(\sum_{v\in V_{u_i}}\left(-\frac{\tau_v}{2\lambda}\left(z_i^v-\left(s_i+b_v\right)\right)^2\right)\right) \\
&\quad\times \exp\left(\sum_{v\in V_{u_i}}\sum_{u_j\in U_v}\left(-\frac{\tau_v}{4\lambda}\left(d_{ij}^v-\left(s_i-s_j\right)\right)^2\right)\right), \\
&\propto \exp\left(-\frac{1}{2}\gamma_0\left(s_i-\mu_0\right)^2\right) \\
&\quad\times \exp\left(\sum_{v\in V_{u_i}}\left(-\frac{\tau_v}{2\lambda}\left(z_i^v-\left(s_i+b_v\right)\right)^2\right)\right) \\
&\quad\times \exp\left(\sum_{v\in V_{u_i}}\sum_{u_j\in U_v}\left(-\frac{\tau_v}{4\lambda}\left(d_{ij}^v-\left(s_i-s_j\right)\right)^2\right)\right). \qquad \text{(B.1)}
\end{aligned}
$$

The expression inside the exponent is quadratic; we thus complete the square, obtaining

$$
\begin{aligned}
&\gamma_0\left(s_i-\mu_0\right)^2+\sum_{v\in V_{u_i}}\left(\frac{\tau_v}{\lambda}\left(z_i^v-\left(s_i+b_v\right)\right)^2\right) \\
&\quad+\sum_{v\in V_{u_i}}\sum_{u_j\in U_v}\left(\frac{\tau_v}{2\lambda}\left(d_{ij}^v-\left(s_i-s_j\right)\right)^2\right) \\
&\quad=\text{const.}+\gamma_0\left(s_i^2-2\mu_0 s_i\right) \\
&\quad\quad+\sum_{v\in V_{u_i}}\frac{\tau_v}{\lambda}\left(\left(s_i+b_v\right)^2-2\,z_i^v\left(s_i+b_v\right)\right) \\
&\quad\quad+\sum_{v\in V_{u_i},u_j\in U_v}\left(\frac{\tau_v}{2\lambda}\left(\left(s_i-s_j\right)^2-2\,d_{ij}^v\left(s_i-s_j\right)\right)\right), \\
&\quad=\text{const.}+\left(\gamma_0+\sum_{v\in V_{u_i}}\frac{\tau_v}{\lambda}+\sum_{v\in V_{u_i}}\sum_{u_j\in U_v}\frac{\tau_v}{2\lambda}\right)s_i^2 \\
&\quad\quad-2s_i\gamma_0\mu_0-\sum_{v\in V_{u_i}}\frac{2\tau_v}{\lambda}\left(z_i^v-b_v\right)s_i \\
&\quad\quad-\sum_{v\in V_{u_i}}\sum_{u_j\in U_v}\frac{\tau_v s_i\left(d_{ij}^v+s_j\right)}{\lambda}, \\
&\quad=\ \text{const.}\ +R\left(s_i-\frac{Y}{R}\right)^2, \\
&\text{where } R=\gamma_0+\sum_{v\in V_{u_i}}\frac{\tau_v}{\lambda}+\sum_{v\in V_{u_i}}\sum_{u_j\in U_v}\frac{\tau_v}{2\lambda}, \\
&\text{and } Y=\mu_0\gamma_0+\sum_{v\in V_{u_i}}\frac{\tau_v}{\lambda}\left(z_i^v-b_v\right) \\
&\quad\quad+\sum_{v\in V_{u_i}}\sum_{u_j\in U_v}\frac{\tau_v\left(d_{ij}^v+s_j\right)}{2\lambda}. \qquad \text{(B.2)}
\end{aligned}
$$

Therefore, the sampling distribution is Gaussian:

$$
s\sim\mathcal{N}\left(\frac{Y}{R},\frac{1}{R}\right). \qquad \text{(B.3)}
$$

Now, consider a fixed grader v. The sampling step for grader reliability τ_v is derived as follows:

$$
\begin{aligned}
\tau &\sim P\left(\tau_v \mid MB\left(\tau_v\right)\right), \\
&\propto P\left(\tau_v \mid \prod_{k=1}^{K}{\alpha_{vk}}^{q_k},\beta_0\right)\cdot\prod_{u_t\in U_v}P\left(z_i^v \mid s_i,b_v,\tau_v\right) \\
&\quad\cdot\prod_{u_t,u_j\in U_v}P\left(d_{ij}^v \mid s_i,s_j,\tau_v\right), \\
&\propto \exp\left(\left(-\frac{\beta_0}{2}\left(\tau_v-\prod_{k=1}^{K}{\alpha_{vk}}^{q_k}\right)^2\right)\right. \\
&\quad+\left(-\frac{1}{2}\sum_{u_t\in U_v}\sqrt{\frac{\tau_v}{2\pi}}\frac{\tau_v}{\lambda}\left(z_i^v-\left(s_i+b_v\right)\right)^2\right) \\
&\quad\left.+\left(-\frac{1}{2}\sum_{u_i,u_j\in U_v}\sqrt{\frac{\tau_v}{4\pi}}\frac{\tau_v}{2\lambda}\left(d_{ij}^v-\left(s_i-s_j\right)\right)^2\right)\right), \\
&\propto {\tau_v}^{\frac{|U_v|^2}{2}}\times\exp\left(-\frac{\beta_0}{2}\left(\tau_v-\prod_{k=1}^{K}{\alpha_{vk}}^{q_k}\right)^2\right) \\
&\quad\times\exp\left(\sum_{u_i\in U_v}-\frac{\tau_v}{2\lambda}\left(z_i^v-\left(s_i+b_v\right)\right)^2\right) \\
&\quad\times\exp\left(\sum_{u_i,u_j\in U_v}-\frac{\tau_v}{4\lambda}\left(d_{ij}^v-\left(s_i-s_j\right)\right)^2\right), \\
&\propto {\tau_v}^{\frac{|U_v|^2}{2}}\times\exp\left(-\frac{\beta_0\tau_v^2}{2}+\prod_{k=1}^{K}{\alpha_{vk}}^{q_k\beta_0\tau_v}\right) \\
&\quad\times\exp\left(\sum_{u_i\in U_v}\frac{\left(z_i^v-s_i-b_v\right)^2\tau_v}{\lambda}\right) \\
&\quad\times\exp\left(\sum_{u_i,u_j\in U_v}\frac{\left(d_{ij}^v-\left(s_i-s_j\right)\right)^2\tau_v}{2\lambda}\right), \\
&\propto {\tau_v}^{\frac{|U_v|^2}{2}}\times\exp\left(-\frac{\beta_0}{2}\left(\tau_v-Y\right)^2\right),
\end{aligned}
$$

where

$$
\begin{aligned}
Y=&\prod_{k=1}^{K}{\alpha_{vk}}^{q_k}+\sum_{u_i\in U_v}\frac{\left(z_i^v-s_i-b_v\right)^2}{\lambda\beta_0} \\
&+\sum_{u_i,u_j\in U_v}\frac{\left(d_{ij}^v-\left(s_i-s_j\right)\right)^2}{2\lambda\beta_0}. \qquad \text{(B.4)}
\end{aligned}
$$

Note that, unlike its analog from Model *CD-PG*$_1$, the sampling step for v in Model *CD-PG*$_2$ cannot be performed in

closed form. In our experiments, we instead sampled from a discretized approximation of the posterior distribution.

Finally, the sampling set for grader bias b_v is derived as follows:

$$\begin{aligned} b &\sim P\left(b_v \mid MB\left(b_v\right)\right) \\ &\propto P\left(b_v \mid \eta_0\right) \cdot \prod_{u_i \in U_v} P\left(z_i^v \mid s_i, b_v, \tau_v\right) \\ &\propto \exp\left(-\frac{1}{2}\eta_0 b_v^2 - \frac{1}{2}\sum_{u_i \in U_v}\frac{\tau_v}{\lambda}\left(z_i^v - (s_i + b_v)\right)^2\right) \\ &\propto \exp\left(-\frac{1}{2}\left[\eta_0 b_v^2 \right.\right. \\ &\quad \left.\left. + \sum_{u_i \in U_v}\frac{\tau_v}{\lambda}\left((s_i + b_v)^2 - 2z_i^v\left(s_i + b_v\right)\right)\right]\right) \end{aligned} \tag{B.5}$$

The expression inside the exponent is quadratic; we thus complete the square as follows:

$$\begin{aligned} &\eta_0 b_v^2 + \sum_{u_i \in U_v}\frac{\tau_v}{\lambda}\left((s_i + b_v)^2 - 2z_i^v\left(s_i + b_v\right)\right) \\ &\quad = \text{const.} + \left(\eta_0 + \sum_{u_i \in U_v}\frac{\tau_v}{\lambda}\right) b_v^2 \\ &\qquad - 2\left(\sum_{u_i \in U_v}\frac{\tau_v}{\lambda}\left(z_i^v - s_i\right)\right) b_v, \\ &\quad = \text{const.} + R\left(b_v - \frac{Y}{R}\right)^2, \\ &\text{where } R = \eta_0 + \sum_{u_i \in U_v}\frac{\tau_v}{\lambda} = \eta_0 + |U_v|\,\frac{\tau_v}{\lambda}, \\ &\text{and } Y = \sum_{u_i \in U_v}\frac{\tau_v}{\lambda}\left(z_i^v - s_i\right). \end{aligned} \tag{B.6}$$

The sampling distribution for b is thus Gaussian, with

$$b \sim \mathcal{N}\left(\frac{\sum_{u_i \in U_v}\frac{\tau_v}{\lambda}\left(z_i^v - s_i\right)}{\eta_0 + |U_v|\,\frac{\tau_v}{\lambda}}, \frac{1}{\eta_0 + |U_v|\,\frac{\tau_v}{\lambda}}\right). \tag{B.7}$$

ACKNOWLEDGMENT

The authors would like to thank Fei Mi from the Hong Kong University of Science and Technology and Hou Pong Chan from The Chinese University of Hong Kong for providing codes of related probabilistic graph models to them. They would also like to express our gratitude to Prof. Jimmy De La Torre at Rutgers Graduate School of Education for explaining details of the DINA model to them.

REFERENCES

[1] K. J. Topping, "Peer assessment," *Theory Pract.*, vol. 48, no. 1, pp. 20–27, 2009.
[2] C. Piech, J. Huang, Z. Chen, C. B. Do, A. Y. Ng, and D. Koller, "Tuned models of peer assessment in MOOCs," in *Proc. 6th Int. Conf. Educ. Data Mining.*, 2013, pp. 153–160.
[3] H. Luo, A. C. Robinson, and J.-Y. Park, "Peer grading in a MOOC: Reliability, validity, and perceived effects," *Online Learn.*, vol. 18, no. 2, pp. 1–14, Jun. 2014.
[4] M. Formanek, M. C. Wenger, S. R. Buxner, C. D. Impey, and T. Sonam, "Insights about large-scale online peer assessment from an analysis of an astronomy MOOC," *Comput. Educ.*, vol. 113, pp. 243–262, Oct. 2017.
[5] D. E. Paré and S. Joordens, "Peering into large lectures: Examining peer and expert mark agreement using peerScholar, an online peer assessment tool," *J. Comput. Assist. Learn.*, vol. 24, no. 6, pp. 526–540, Oct. 2008.
[6] L.-H. Hsia, I. Huang, and G.-J. Hwang, "Effects of different online peer-feedback approaches on students' performance skills, motivation and self-efficacy in a dance course," *Comput. Educ.*, vol. 96, pp. 55–71, May 2016.
[7] T. T. Vu and G. Dall'Alba, "Students' experience of peer assessment in a professional course," *Assessment Eval. Higher Educ.*, vol. 32, no. 5, pp. 541–556, 2007.
[8] T. Hovardas, O. E. Tsivitanidou, and Z. C. Zacharia, "Peer versus expert feedback: An investigation of the quality of peer feedback among secondary school students," *Comput. Educ.*, vol. 71, pp. 133–152, Feb. 2014.
[9] J. Mok, "A case study of students' perceptions of peer assessment in Hong Kong," *ELT J.*, vol. 65, no. 3, pp. 230–239, Jul. 2011.
[10] Y. Han, W. Wu, Y. Yan, and L. Zhang, "Human-machine hybrid peer grading in SPOCs," *IEEE Access*, vol. 8, pp. 220922–220934, 2020.
[11] F. Garcia-Loro, S. Martin, J. A. Ruipérez-Valiente, E. Sancristobal, and M. Castro, "Reviewing and analyzing peer review inter-rater reliability in a MOOC platform," *Comput. Educ.*, vol. 154, Sep. 2020, Art. no. 103894.
[12] F. Mi and D. Yeung, "Probabilistic graphical models for boosting cardinal and ordinal peer grading in MOOCs," in *Proc. 29th AAAI Conf. Artif. Intell.*, 2015, pp. 454–460.
[13] H. P. Chan and I. King, "Leveraging social connections to improve peer assessment in MOOCs," in *Proc. 26th Int. Conf. World Wide Web Companion (WWW Companion)*, 2017, pp. 341–349.
[14] T. Wang, Q. Li, J. Gao, X. Jing, and J. Tang, "Improving peer assessment accuracy by incorporating relative peer grades," in *Proc. 12th Int. Conf. Educ. Data Mining*, 2019, pp. 450–455.
[15] R. A. Henson, J. L. Templin, and J. T. Willse, "Defining a family of cognitive diagnosis models using log-linear models with latent variables," *Psychometrika*, vol. 74, no. 2, p. 191, 2009.
[16] J. Chen and J. de la Torre, "A general cognitive diagnosis model for expert-defined polytomous attributes," *Appl. Psychol. Meas.*, vol. 37, no. 6, pp. 419–437, Sep. 2013.
[17] H. Ravand and A. Robitzsch, "Cognitive diagnostic model of best choice: A study of reading comprehension," *Educ. Psychol.*, vol. 38, no. 10, pp. 1255–1277, Nov. 2018.
[18] J. de la Torre, "DINA model and parameter estimation: A didactic," *J. Educ. Behav. Statist.*, vol. 34, no. 1, pp. 115–130, Mar. 2009.
[19] J. Surowiecki and M. P. Silverman, "The Wisdom of crowds," *Amer. J. Phys.*, vol. 75, no. 2, pp. 190–192, 2005.
[20] J. Whitehill, P. Ruvolo, T. Wu, J. Bergsma, and J. R. Movellan, "Whose vote should count more: Optimal integration of labels from labelers of unknown expertise," in *Proc. Adv. Neural Inf. Process. Syst. 23rd Annu. Conf. Neural Inf. Process. Syst.*, Vancouver, BC, Canada, 2009, pp. 2035–2043.
[21] T. Wang, H. Xiao, F. Ma, and J. Gao, "IProWA: A novel probabilistic graphical model for crowdsourcing aggregation," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2019, pp. 677–682.
[22] A. R. Kurup and G. P. Sajeev, "Aggregating unstructured submissions for reliable answers in crowdsourcing systems," in *Proc. 9th Int. Symp. Embedded Comput. Syst. Design (ISED)*, Dec. 2019, pp. 1–7.
[23] M. Y. Guan, V. Gulshan, A. M. Dai, and G. E. Hinton, "Who said what: Modeling individual labelers improves classification," in *Proc. 32nd AAAI Conf. Artif. Intell., (AAAI), 30th Innov. Appl. Artif. Intell. (IAAI), 8th AAAI Symp. Educ. Adv. Artif. Intell. (EAAI)*. New Orleans, LA, USA: AAAI Press, Feb. 2018, pp. 3109–3118.
[24] O. Isupova, Y. Li, D. Kuzin, S. J. Roberts, K. J. Willis, and S. Reece, "BCCNet: Bayesian classifier combination neural network," 2018, *arXiv:1811.12258*. [Online]. Available: https://arxiv.org/abs/1811.12258
[25] F. Rodrigues and F. C. Pereira, "Deep learning from crowds," in *Proc. 32nd AAAI Conf. Artif. Intell., 30th Innov. Appl. Artif. Intell., 8th AAAI Symp. Educ. Adv. Artif. Intell. (EAAI)*, S. A. McIlraith and K. Q. Weinberger, Eds. New Orleans, LA, USA: AAAI Press, Feb. 2018, pp. 1611–1618. [Online]. Available: https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16102

[26] B. I. Aydin, Y. S. Yilmaz, Y. Li, Q. Li, J. Gao, and M. Demirbas, "Crowdsourcing for multiple-choice question answering," in *Proc. 28th AAAI Conf. Artif. Intell.*, Montreal, QC, Canada, Jul. 2014, pp. 2946–2953.

[27] D. Yue, G. Yu, D. Shen, and X. Yu, "A weighted aggregation rule in crowdsourcing systems for high result accuracy," in *Proc. IEEE 12th Int. Conf. Dependable, Autonomic Secure Comput.*, Dalian, China, Aug. 2014, pp. 265–270.

[28] H. Li and B. Yu, "Error rate bounds and iterative weighted majority voting for crowdsourcing," 2014, *arXiv:1411.4086*. [Online]. Available: https://arxiv.org/abs/1411.4086

[29] W. Li, M. Huhns, W. T. Tsai, and W. Wu, Eds., "Crowdsourcing," in *Progress in IS*. Berlin, Germany: Springer.

[30] S. Rooyen, F. Godlee, S. Evans, R. Smith, and N. Black, "Effect of blinding and unmasking on the quality of peer review," *J. Gen. Internal Med.*, vol. 14, no. 10, pp. 622–624, Oct. 1999.

[31] S. Nobarany and K. S. Booth, "Understanding and supporting anonymity policies in peer review," *J. Assoc. Inf. Sci. Technol.*, vol. 68, no. 4, pp. 957–971, Apr. 2017.

[32] L. Charlin, R. S. Zemel, and C. Boutilier, "A framework for optimizing paper matching," in *Proc. 27th Conf. Uncertainty Artif. Intell.*, Barcelona, Spain, Jul. 2011, pp. 86–95.

[33] B. Li and Y. T. Hou, "The new automated IEEE INFOCOM review assignment system," *IEEE Netw.*, vol. 30, no. 5, pp. 18–24, Sep./Oct. 2016.

[34] S. Price and P. A. Flach, "Computational support for academic peer review: A perspective from artificial intelligence," *Commun. ACM*, vol. 60, no. 3, pp. 70–79, 2017.

[35] Y. M. Wang, J. B. Yang, D. L. Xu, and K. S. Chin, "The evidential reasoning approach for multiple attribute decision analysis using interval belief degrees," *Eur. J. Oper. Res.*, vol. 175, no. 1, pp. 35–66, 2006.

[36] W.-D. Zhu, F. Liu, Y.-W. Chen, J.-B. Yang, D.-L. Xu, and D.-P. Wang, "Research project evaluation and selection: An evidential reasoning rule-based method for aggregating peer review information with reliabilities," *Scientometrics*, vol. 105, no. 3, pp. 1469–1490, Dec. 2015.

[37] F. Liu, W. D. Zhu, Y. W. Chen, D. L. Xu, and J. B. Yang, "Evaluation, ranking and selection of R&D projects by multiple experts: An evidential reasoning rule based approach," *Scientometrics*, vol. 111, no. 3, pp. 1501–1519, 2017.

[38] Y.-W. Du, N. Yang, and J. Ning, "IFS/ER-based large-scale multiattribute group decision-making method by considering expert knowledge structure," *Knowl.-Based Syst.*, vol. 162, pp. 124–135, Dec. 2018.

[39] W. Zhu, S. Li, Q. Ku, and C. Zhang, "Evaluation information fusion of scientific research project based on evidential reasoning approach under two-dimensional frames of discernment," *IEEE Access*, vol. 8, pp. 8087–8100, 2020.

[40] F. L. Wauthier, M. I. Jordan, and N. Jojic, "Efficient ranking from pairwise comparisons," in *Proc. 30th Int. Conf. Mach. Learn.*, vol. 28, 2013, pp. 109–117.

[41] R. A. Bradley and M. E. Terry, "Rank analysis of incomplete block designs: I. The method of paired comparisons," *Biometrika*, vol. 39, nos. 3–4, pp. 324–345, 1952.

[42] N. B. Shah, J. K. Bradley, A. Parekh, M. Wainwright, and K. Ramchandran, "A case for ordinal peer-evaluation in MOOCs," in *Proc. NIPS Workshop Data Driven Educ.*, 2013, pp. 1–8.

[43] K. Raman and T. Joachims, "Methods for ordinal peer grading," in *Proc. 20th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2014, pp. 1037–1046.

[44] D. R. Luce, "Individual choice behavior: A theoretical analysis," *J. Amer. Statal Assoc.*, vol. 67, no. 293, pp. 1–15, 2005.

[45] M. S. M. Sajjadi, M. Alamgir, and U. von Luxburg, "Peer grading in a course on algorithms and data structures: Machine learning algorithms do not improve over simple baselines," in *Proc. 3rd ACM Conf. Learn. @ Scale*, 2016, pp. 369–378.

[46] K. Raman and T. Joachims, "Bayesian ordinal peer grading," in *Proc. 2nd ACM Conf. Learn. @ Scale*, Mar. 2015, pp. 149–156.

[47] A. E. Waters, D. Tinapple, and R. G. Baraniuk, "BayesRank: A Bayesian approach to ranked peer grading," in *Proc. 2nd ACM Conf. Learn. @ Scale*, 2015, pp. 177–183.

[48] H. P. Chan, T. Zhao, and I. King, "Trust-aware peer assessment using multi-armed bandit algorithms," in *Proc. 25th Int. Conf. Companion World Wide Web (WWW Companion)*, 2016, pp. 899–903.

[49] N. Capuano, V. Loia, and F. Orciuoli, "A fuzzy group decision making model for ordinal peer assessment," *IEEE Trans. Learn. Technol.*, vol. 10, no. 2, pp. 247–259, Apr. 2017.

[50] L. D. Alfaro and M. Shavlovsky, "CrowdGrader: A tool for crowdsourcing the evaluation of homework assignments," in *Proc. 45th ACM Tech. Symp. Comput. Sci. Educ.*, 2014, pp. 415–420.

[51] T. Walsh, "The peerrank method for peer assessment," in *Proc. 21st Eur. Conf. Artif. Intell. (ECAI)*, vol. 263, 2014, pp. 909–914.

[52] L. Page. (1998). *The Pagerank Citation Ranking: Bringing Order to the web*. [Online]. Available: http://ilpubs.stanford.edu:8090/422/1/1999-66.pdf

[53] P. Gutierrez, N. Osman, and C. Sierra, "Collaborative assessment," in *Proc. 17th Int. Conf. Catalan Assoc. Artif. Intell.*, vol. 269, 2014, pp. 136–145.

[54] P. Singla and M. Richardson, "Yes, there is a correlation: -From social networks to personal behavior on the Web," in *Proc. 17th Int. Conf. World Wide Web*, 2008, pp. 655–664.

[55] S. Yang, B. Long, A. J. Smola, N. Sadagopan, Z. Zheng, and H. Zha, "Like like alike: Joint friendship and interest propagation in social networks," in *Proc. 20th Int. Conf. World Wide Web*, vol. 2011, pp. 537–546.

[56] R. Wu, Q. Liu, Y. Liu, E. Chen, Y. Su, Z. Chen, and G. Hu, "Cognitive modelling for predicting examinee performance," in *Proc. 24th Int. Joint Conf. Artif. Intell.*, 2015, pp. 1017–1024.

[57] T. Zhu, Q. Liu, Z. Huang, E. Chen, D. Lian, Y. Su, and G. Hu, "MT-MCD: A multi-task cognitive diagnosis framework for student assessment," in *Proc. 23rd Int. Conf. Database Syst. Adv. Appl.*, vol. 10828, 2018, pp. 318–335.

[58] S. Cheng, Q. Liu, E. Chen, Z. Huang, Z. Huang, Y. Chen, H. Ma, and G. Hu, "DIRT: Deep learning enhanced item response theory for cognitive diagnosis," in *Proc. 28th ACM Int. Conf. Inf. Knowl. Manage.*, 2019, pp. 2397–2400.

[59] L. T. DeCarlo, "On the analysis of fraction subtraction data: The DINA model, classification, latent class sizes, and the *Q*-matrix," *Appl. Psychol. Meas.*, vol. 35, no. 1, pp. 8–26, Jan. 2011.

[60] L. T. DeCarlo, "Recognizing uncertainty in the *Q*-matrix via a Bayesian extension of the DINA model," *Appl. Psychol. Meas.*, vol. 36, no. 6, pp. 447–468, Sep. 2012.

[61] T. Zhu, Z. Huang, E. Chen, Q. Liu, R. Wu, L. Wu, Y. Su, Z. Chen, and G. Hu, "Cognitive diagnosis based personalized question recommendation," *Chin. J. Comput.*, vol. 40, no. 1, pp. 176–191, 2017.

[62] C. Wang, Q. Liu, E. Chen, Z. Huang, T. Zhu, Y. Su, and G. Hu, "The rapid calculation method of DINA model for large scale cognitive diagnosis," *Chin. J. Electron.*, vol. 46, no. 5, pp. 1047–1055, 2018.

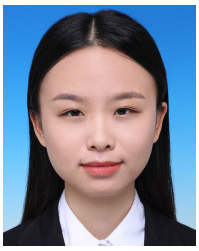
[63] S. Embretson (Whitely), "A general latent trait model for response processes," *Psychometrika*, vol. 49, no. 2, pp. 175–186, Jun. 1984.

[64] M. I. Jordan, "Graphical models," *Statist. Sci.*, vol. 19, no. 1, pp. 140–155, 2004.

[65] D. Christiansen, *A Chicken-and-Egg Problem*. Piscataway, NJ, USA: IEEE Press, 1991.

[66] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-6, no. 6, pp. 721–741, Nov. 1984.

JIA XU (Member, IEEE) received the Ph.D. degree in computer science and technology from Northeastern University, China, in 2013. From 2008 to 2009, she was an Intern Student with the School of Computing, National University of Singapore. She joined the Advanced Digital Science Center, Illinois at Singapore, as a Postdoctoral Research Fellow. She is currently an Associate Professor with the School of Computer, Electronics and Information, Guangxi University, China. She was awarded the position of Professor of Bagui Young Scholars of Guangxi Province, in 2019. Her research interests include educational data analysis, data management, distributed parallel data processing, and data privacy protection. She is a member of ACM, and a Senior Member of the China Computer Federation (CCF). She is a Committee Member of the CCF Technical Committee of Database (TCDB). She received the CCF Excellent Doctoral Dissertation Award, in 2014.



QIUYUN LI received the B.S. degree from the University of International Relations, China, in 2018. She is currently pursuing the M.S. degree in computer science and technology with Guangxi University, China. Her research interests include cognitive diagnosis in educational applications, peer assessment for MOOC platforms, and developing online pedagogical tools for colleges and universities.

JING LIU received the B.S. degree in information management and information systems from Dalian Maritime University, China, in 2017. She is currently pursuing the M.S. degree in computer science and technology with Guangxi University, China. Her research interests include educational data analysis, peer assessment for MOOC platforms, and artificial intelligence technologies.

PIN LV (Member, IEEE) received the B.S. degree from Northeastern University, China, in 2006, and the Ph.D. degree in computer science from NUDT, China, in 2012. He is currently an Associate Professor with the School of Computer, Electronics and Information, Guangxi University, China. He is also a Key Member of information with Guangxi University and the Guangxi Key Laboratory of Multimedia Communications and Network Technology. His research interests include wireless networks, mobile computing, and computer education. He is a member of ACM, and a Senior Member of the China Computer Federation (CCF). He serves as a Committee Member for the CCF Technical Committee of Cooperative Computing (TCCC).

GE YU (Senior Member, IEEE) received the Ph.D. degree in computer science from Kyushu University, Japan, in 1996. He is currently a Professor with Northeastern University, China. He has published more than 200 papers in refereed journals and conferences. His research interests include distributed and parallel database, OLAP and data warehousing, data integration, graph data management, and so on. He is a Fellow of the China Computer Federation (CCF). He is a member of ACM.