# A Cognitive Diagnosis Framework Based on Peer Assessment

Yu He, Xinying Hu, and Guangzhong Sun*
School of Computer Science and Technology, University of Science and Technology of China, Hefei, China
{heyu3761,hxiny}@mail.ustc.edu.cn,gzsun@ustc.edu.cn

## ABSTRACT

Given examinees' performance (i. e., scores) on each problem, cognitive diagnosis models can discover the latent characteristics of examinees. Traditional cognitive diagnosis models require teachers to provide scores in time. Thus we can hardly apply traditional models in large-scale scenarios, such as Massive Open Online Courses (MOOC). Peer assessment refers to a teaching activity in which students evaluate each other's assignments. The scores given by students could replace the teacher's assessments to a certain extent. In this paper, we propose a novel cognitive diagnosis model named Peer-Assessment Cognitive Diagnosis Framework (PACDF). This model combines peer assessments with cognitive diagnosis, aiming at reduce the burden of teachers. Specifically, we propose a novel probabilistic graphic model at first. This model characterizes not only the relationships between real scores and scores given by peer assessment, but also the relationship between examinees' skill proficiency and problem mastery. Then we adopt Monte Carol Markov Chain (MCMC) sampling algorithm to estimate the parameters of the model. Lastly, we use the model to predict examinees' performance. The experimental results show that PACDF could quantitatively explain and analyze skill proficiencies of examinees, thus perform better in predicting examinees' performances.

## CCS CONCEPTS

• **Computing methodologies** → **Latent variable models**; • **Applied computing** → **Computer-assisted instruction**.

## KEYWORDS

peer assessment, cognitive diagnosis, automated assessment, qualitative feedback, peer grading

**ACM Reference Format:**
Yu He, Xinying Hu, and Guangzhong Sun. 2019. A Cognitive Diagnosis Framework Based on Peer Assessment. In *ACM Turing Celebration Conference - China (ACM TURC 2019) (ACM TURC 2019), May 17–19, 2019, Chengdu, China.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3321408.3322850

*The corresponding author


*ACM TURC 2019, May 17–19, 2019, Chengdu, China*

ACM ISBN 978-1-4503-7158-2/19/05…$15.00
https://doi.org/10.1145/3321408.3322850

## 1 INTRODUCTION

Giving accurate and rapid feedback to examinees is very important in the field of education. It has been proved that rapid feedback cpoule improve students' performance: in a controlled experiment, students' final grades had been improved when feedback was delivered quickly, but not if delayed by 24 hours [1]. The shortcoming of traditional examinations is that the feedback to examinees is inaccurate or slow. Traditional examinations could only provide scores. But students with the same scores have different knowledge states, which cannot be explained. With the rapid development of education, the number of teachers is insufficient compared with the rapidly growing number of students. The teacher was heavily burdened with correcting the examination papers, especially in correcting the subjective problems.

Peer assessment refers to the teaching activities in which the students evaluate each other's assignments, and the results given by students can replace the teacher's assessment to a certain extent. Many studies have proved that there is a strong positive correlation between the scores given by students and teachers [2–5]. Relevant statistical results show that students can give consistent and reliable scores [3, 6–9].

Peer assessment is an application of crowdsourcing in education. At present, there are some studies about how to improve the accuracy of the results. Such as many-facet rasch measurement(MFRM) [10–13], hierarchical rater model(HRM) [14], partial credit model [15], signal detection rater model [16], rater bundle model [17]. Among them, three-facet rasch rating scale model is used widely. And the three faces are: the peer reviewers, the examinees and the problems. The model is widely used to estimate the true level of examinees. Another commonly used model is hierarchical rater model, which is a two-layer model. The first layer is to estimate the true scores of the assignments by analyzing the scores provided by the peer reviewers. And the second layer is to estimate the true level of the examinees through the true scores of the assignments which is calculated in the first layer.

Traditional examinations that only provide a single total score can no longer meet the needs of current teaching. The emergence of cognitive diagnosis theory compensates for the defect of traditional examinations. And it provides more abundant measurement information, that is, it could provide information about students' skills proficiency and problems mastery. By knowing specific diagnostic information in time, teachers could mastery the cognitive structure of students comprehensively [9, 18]. Figure 1 shows the contrast of the feedback provided by traditional examinations and cognitive diagnosis. Cognitive diagnosis could infer the students' mastery of the knowledge involved in the examination by analyzing their test paper. It aims at discovering the latent factors/characteristics of examinees. And it could be used to model a group of examinees to predict their possible scores for each problem, i.e. predicting examinees' performance (PEP). According to

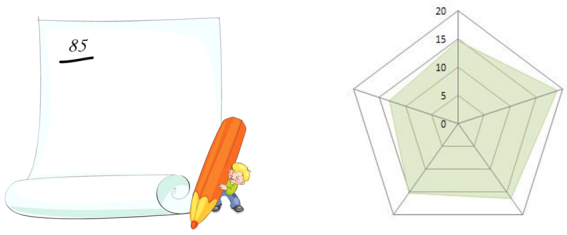


**Figure 1: The contrast of feedback provided by traditional examinations and cognitive diagnosis.**

the cognitive diagnosis model [18], examinees are characterized by proficiency in specific skills (problem solving skills, such as computational ability). The cognitive diagnosis model could provide personalized guidance for teachers' teaching activities. At present, the cognitive diagnosis model has achieved good results in the application of student-correction plan and early warning of dropout [19].

But existing methods could be improved, for example, previous models rely on scores provided by teachers too much, which lead to the difficulty of applying in large-scale scenarios such as massive open online courses(MOOC). It is difficult to give feedback and evaluation to the students due to severe mismatches in the number of students enrolled and the number of experts available in MOOC [20].

In this paper, we propose a peer-assessment cognitive diagnosis framework (PACDF) that combines peer assessment with cognitive diagnosis to predict examinee performance (PEP). Specifically, our model defines not only the relationship between real scores and scores given by peer assessment, but also the relationship between examinees' skill proficiency and problem mastery. We propose a Monte Carol Markov chain (MCMC) sampling algorithm to estimate the parameters and predict examinees' performance. Figure 2 shows a toy process of PACDF.

The main contributions of this paper are as follows:

As far as we know, this is the first attempt to combine peer assessment with cognitive diagnosis. Firstly, it could lighten the burden of teachers. Secondly, it could make more accurate and explanatory cognitive analysis of predicting examinee performance.

This model combines the theory of peer assessment and cognitive diagnosis to redefine the examinees' skill proficiency and problem mastery, aiming at predicting examinees' performance.

We design a simple but effective MCMC sampling algorithm for parameter estimation and conduct extensive experiments on synthetic datasets to demonstrate the effectiveness of PACDF.

## 2 RELATED WORK

In this section, we will introduce related work from two aspects: peer assessment and cognitive diagnosis model.

### 2.1 Peer Assessment

Peer assessment is an application of crowdsourcing in the fields of education. There are some studies about the existing challenges and solutions of peer assessment [21], as well as the impact of peer

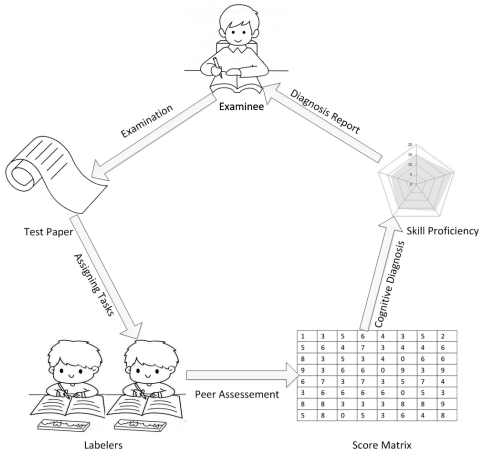


**Figure 2: A toy process of PACDF.**

assessment on the final review [22, 23]. Kulkarni demonstrated the feasibility of combining machine and peer assessment, which results in more detailed student feedback, and can be leveraged to provide early feedback [24].

A group of researchers, including educational researchers and computer scientists, proposed some methods to improve the accuracy and reliability of the crowdsourcing results. These methods can be roughly divided into pre-correction and post-correction. Pre-correction is to correct or estimate the error of the students before peer assessment. Using teacher's assessment as standard, students are trained to approach the standard, thereby reducing the possible scoring error of students [23]. Kulkarni proposed some techniques to improve the accuracy, such as giving students feedback about their grading bias, introducing short and customizable feedback snippets. Besides, rubrics that use a parallel sentence structure, unambiguous wording, and well-specified dimensions perform better [24]. Another method is to determine students' weight in peer assessment by preliminary verification [25]. Post-correction is to estimate the true score of the student's assignment by analyzing the existing data and computing students' scoring bias. Piech proposed a statistical model which can be used to estimate the true scores of students' homework in MOOC [4].It uses Bayesian methods to estimate the parameters. Goldin collected the results of peer assessment and teachers' scores, which were used to predict the true scores of students' homework and students' scoring bias. Unlike Piech's model, Goldin's model incorporates the difficulty factor as a parameter.

### 2.2 Cognitive Diagnosis Model

In educational psychology, many cognitive diagnostic models [18] were used to discover examinees' skill proficiency and predict examinees' performance. Common cognitive diagnosis models include Item response theory (IRT) model and DINA model. Item response theory (IRT) model describes students as one-dimensional ability variables, combining the potential features of examinations (such as difficulty, discrimination, etc.) [26–29] . Another basic approach is to determine the input, noisy "and" gate model (DINA) [30–33].

DINA model assume that the problems are related to a set of explicit skills, which is expressed by Q matrix. And Q matrix could be used to explain the diagnostic results. A student is described as a multidimensional vector, which could indicate whether she or he grasps the knowledge needed by the problem. Students' performance on the problem is affected by students' mastery of the knowledge examined by the problem [32]. At present, the cognitive diagnosis model has achieved good results in the application of student-correction plan and early warning of dropout [19].

## 3 COGNITIVE DIAGNOSIS MODELING

In this section, we will introduce our peer-assessment cognitive diagnosis framework (PACDF). The model is to determine the examinees' skill proficiency from the potential characteristics of the examinees. It could calculate the examinees' problem mastery. After that, the model considers the strictness of each student to generate the score they gave in peer assessment. We propose a MCMC sampling algorithm to infer the unobserved parameters of PACDF. For easy understanding, Table 1 shows some mathematical notation.

### 3.1 Model Description

We assume that the examinees' skill proficiency is a continuous variable with a value of $[0, 1]$, which has been used in the literature [34]. Besides, we assume that the examinees' skill proficiency is related to the examinees' basic ability and the degree of effort when learning this skill [35].

**Assumption 1** Skill proficiency is positively related to the examinees' basic ability and the level of effort in learning the skill.

So $\alpha_{j,k}$ is defined as:

$$\alpha_{j,k} = \frac{1}{1 + \exp(\theta_j + a_{j,k})} \tag{1}$$

CDMs assume that problem mastery is a result of interaction of the proficiency of the skills required by the problem [36]. There are two common approaches to describing the skills' interaction on problems: a conjunctive approach, which assumes that all skills must be known or a compensatory approach which assumes that the strength of one skill can compensate for the weakness of another skill. In this paper we assume that the skill's interaction on subjective problems is conjunctive, and the proportion of skills required by the problem will also have an impact on it. Considering the different scoring scales of subjective problems, we standardize the scores of subjective problems and divide the full score into a continuous variable with the value of [0, 1]. Then we assume that the score of examinees on subjective problems follow a Gaussian distribution, which is widely used in the literature.

In conclusion, we assume that examinees' problem mastery is related to the product of examinee skill proficiency and the ratio of skill to problem. The ratio of skill to problem is given by the teacher in advance. The problem mastery of examinee is redefined by the following assumptions:

**Assumption 2** A real score of an assignment submitted by an examinee(i.e. the problem mastery of this examinee), is the sum of the product of all skill proficiency of this examinee and skill ratios to this problem.

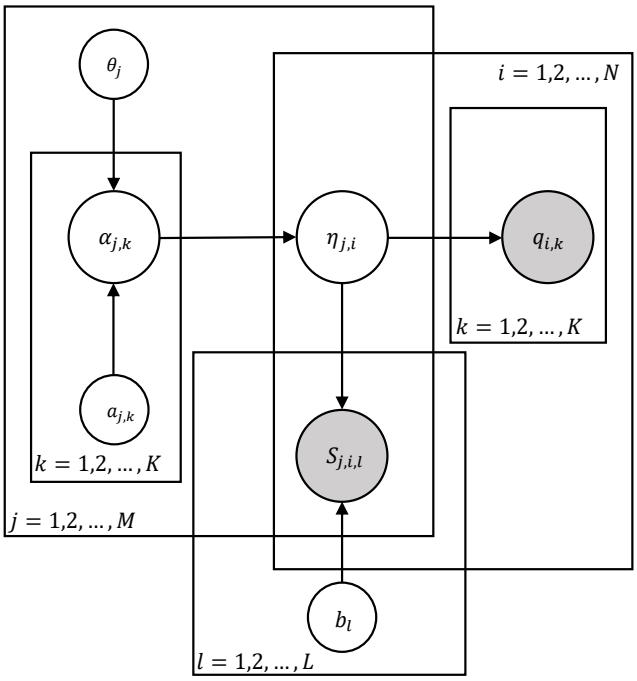


**Figure 3: The probability graph model of PACDF**

So $\eta_{j,i}$ is defined as:

$$\eta_{j,i} = \sum_{k=1}^{K} \alpha_{j,k} q_{i,k} \tag{2}$$

Then we define the strictness of the peer reviewers. The lower the value, the stricter the peer reviewer. We assume that the scores given by peer reviewer are related to the assignment's real score and peer reviewer's strictness [4]. The scores given by peer reviewers are defined by the following assumptions:

**Assumption 3** The scores for a certain assignment given by peer reviewers are subject to Gaussian distribution, and the mean is the product of the assignment's real score and peer reviewer's strictness.

So $S_{j,i,l}$ is defined as:

$$S_{j,i,l} \sim \mathcal{N}(b_l \eta_{ji}, \sigma_s^2, min_s, max_s) \tag{3}$$

Where $\mathcal{N}(\mu, \sigma^2, min, max)$ is a four-parameter Gaussian distribution in which parameter $\mu$ represents mean and $\sigma^2$ represents standard deviation. And it is supported on the range [min, max].

**Summary.** To better understand our proposed PACDF, we use a probability graph model shown in Figure 3 to represent it. Here, we can observe the scoring matrix, Q-matrix (the ratio of skill $k$ to problem $i$, if problem $i$ doesn't require skill $k$, then $q_{i,k} = 0$). The skill proficiency of examinee $a_{j,k}$ depends on the basic ability $\theta_j$ and the level of effort $\alpha_{j,k}$. The real score $\eta_{j,i}$ (i.e. the examinee's problem mastery ) depends on skill proficiency $a_{j,k}$ and the ratio of skill to problem $q_{i,k}$. The scores given by peer reviewers $S_{j,i,l}$ is affected by the assignment's real score $\eta_{j,i}$ and peer reviewer's strictness $b_l$.

**Table 1: Some improtant notations**

| Notation | Description |
|---|---|
| $\theta_j$ | the basic ability of examinee $j$ |
| $\alpha_{j,k}$ | the proficiency of examinee $j$ on skill $k$ |
| $b_l$ | the strictness of student $x$ in peer assessment |
| $a_{j,k}$ | the level of effort of examinee $j$ in learning skill $k$ |
| $q_{i,k}$ | the ratio of skill $k$ to problem $i$, if problem $i$ doesn't require skill $k$, then $q_{i,k} = 0$ |
| $\eta_{j,i}$ | the real score of examinee $j$ on problem $i$ (the mastery of examinee $j$ on problem $i$) |
| $S_{j,i,l}$ | the score given by student $x$ when assessing assignment of examinee $j$ on problem $i$ |

## 3.2 MCMC Training Algorithm

In this section, we will introduce an algorithm for effectively training the proposed PACDF model using MCMC Training Algorithm, which aims to infer the unshaded variables in Figure 3. Specifically, we assume that the prior distributions of parameters in PACDF are as follows:

$$\begin{aligned} \theta_j &\sim \mathcal{N}(0, \sigma_\theta^2) \\ a_{j,k} &\sim \mathcal{N}(0, \sigma_a^2) \\ b_l &\sim Beta(u_b, v_b) \end{aligned} \tag{4}$$

Let $\boldsymbol{\theta} = \{\theta_j\}$, $\mathbf{a} = \{a_{j,k}\}$, $\mathbf{b} = \{b_l\}$.

The functional forms of the prior distributions are chosen out of convenience, and the associated hyperparameters are selected to be reasonably vague within the range of realistic parameters. Then,the joint posterior distribution of $\boldsymbol{\theta}$,$\mathbf{a}$,$\mathbf{b}$ given the peer-assessment matrix $\mathbf{S}$ is as follows:

$$p(\boldsymbol{\theta}, \mathbf{a}, \mathbf{b}|\mathbf{S}) \propto L(\boldsymbol{\theta}, \mathbf{a}, \mathbf{b}) p(\boldsymbol{\theta}) p(\mathbf{a}) p(\mathbf{b}) \tag{5}$$

Where $L$ is the joint likelihood function of PACDF :

$$L(\boldsymbol{\theta}, \mathbf{a}, \mathbf{b}) = \prod_{i=1}^{N} \prod_{j=1}^{M} \prod_{x=1}^{X} N_{b_l \eta_{ji}, \sigma_s}(S_{j,i,l}) \tag{6}$$

Then, the posterior distribution of $\boldsymbol{\theta}$,$\mathbf{a}$,$\mathbf{b}$are as follows:

$$p(\mathbf{a}|\boldsymbol{\theta}, \mathbf{b}) \propto L(\boldsymbol{\theta}, \mathbf{a}, \mathbf{b}) p(\mathbf{a}) \tag{7}$$

$$p(\mathbf{b}|\boldsymbol{\theta}, \mathbf{a}) \propto L(\boldsymbol{\theta}, \mathbf{a}, \mathbf{b}) p(\mathbf{b}) \tag{8}$$

$$p(\boldsymbol{\theta}|\mathbf{a}, \mathbf{b}) \propto L(\boldsymbol{\theta}, \mathbf{a}, \mathbf{b}) p(\boldsymbol{\theta}) \tag{9}$$

Finally, we propose an MCMC algorithm to estimate parameter of Algorithm 1. Specifically, we first randomize all parameters as initial values. Then we compute the probability of the basic ability $\theta_j$, the level of effort in learning the skill $a_{j,k}$, the problem mastery $\eta_{j,i}$, and the strictness of student $b_l$ by using the peer-assessment matrix $\mathbf{S}$ and $\mathbf{Q}$-matrix. Next, the acceptance probability of the samples can also be calculated according to the algorithm 1 [1] . In this way, we could estimate the parameters with the MCMC formed through sampling.

**Predicting Examinee Performance.** After the training phase, we can easily obtain examinees' skill proficiency based on the basic ability and the level of effort in learning the skill. Then, according to the $\mathbf{Q}$-matrix, examinees' mastery of each problem is further calculated through the equation. So the PEP task is completed, that is, we could predict examinees' performance (i.e. score) on each problem.

[1]proposal distribution $q_{\mu,\sigma}$ in Algorithm 1 is defined as follow:

$$q_{\mu,\sigma}(\mathbf{x}) = \prod_{i=1}^{n} I[0 \le x_i \le 1] \times \frac{\mathcal{N}(\mathbf{x}|\mu, \sigma)}{\int_0^1 \dots \int_0^1 \mathcal{N}(\mathbf{x}|\mu, \sigma) dx_1 \dots dx_n};$$

**Algorithm 1:** Sampling algorithm for PACDF

**Input:** peer-assessment matrix $\mathbf{S}$,$\mathbf{Q}$-matrix
**Output:** samples of $\boldsymbol{\theta}, \mathbf{a}, \mathbf{b}$:$\{\boldsymbol{\theta}^{(t)}\}_{t=1}^{T}, \{\mathbf{a}^{(t)}\}_{t=1}^{T}, \{\mathbf{b}^{(t)}\}_{t=1}^{T}$
**begin**
  Initialize $\boldsymbol{\theta}^{(0)}, \mathbf{a}^{(0)}, \mathbf{b}^{(0)}$ with random values;
  **for** $i = 1$ **to** $T$ **do**
    Generate random samples from $\boldsymbol{\theta}^* \sim q_{\boldsymbol{\theta}^{(t-1)}, \sigma_q}$ ,
    and accept $\boldsymbol{\theta}^{(t)}$ with the following probability;

$$\min\{1, \frac{L(\boldsymbol{\theta}^*, \mathbf{a}^{(t-1)}, \mathbf{b}^{(t-1)}) p(\boldsymbol{\theta}^*)}{L(\boldsymbol{\theta}^{(t-1)}, \mathbf{a}^{(t-1)}, \mathbf{b}^{(t-1)}) p(\boldsymbol{\theta}^{(t-1)})}\}$$

    Generate random samples from $\mathbf{a}^* \sim q_{\mathbf{a}^{(t-1)}, \sigma_q}$, and
    accept $\mathbf{a}^{(t)}$ with the following probability;

$$\min\{1, \frac{L(\boldsymbol{\theta}^{(t)}, \mathbf{a}^*, \mathbf{b}^{(t-1)}) p(\mathbf{a}^*)}{L(\boldsymbol{\theta}^{(t)}, \mathbf{a}^{(t-1)}, \mathbf{b}^{(t-1)}) p(\mathbf{a}^{(t-1)})}\}$$

    Generate random samples from $\mathbf{b}^* \sim q_{\mathbf{b}^{(t-1)}, \sigma_q}$,
    and accept $\mathbf{b}^{(t)}$ with the following probability;

$$\min\{1, \frac{L(\boldsymbol{\theta}^{(t)}, \mathbf{a}^{(t)}, \mathbf{b}^*) p(\mathbf{b}^*)}{L(\boldsymbol{\theta}^{(t)}, \mathbf{a}^{(t)}, \mathbf{b}^{(t-1)}) p(\mathbf{b}^{(t-1)})}\}$$

# 4 EXPERIMENTS

In this section, we will introduce our experiment. At first, we choose two public datasets, which comprises of scores from two final mathematical exams for high school students including both objective and subjective problems [34]. Then we choose the part of subjective problems. According to the examinees' real scores, the scores given in peer assessment are simulated and generated. We denote these two datasets as M1 and M2.The brief description of these datasets is shown in Table 2.

In terms of the hyperparameters, we selected 80% as training sets for tuning. The experimental results show that $u_b, v_b$ have the greatest impact on the result. Fixed other parameters, we change

**Table 2: Datasets Description**

| | M1 | M2 |
|---|---|---|
| **Examinee** | 4209 | 3911 |
| **Skill** | 11 | 16 |
| **Problem** | 5 | 4 |
| **PeerReviewer** | 15 | 15 |

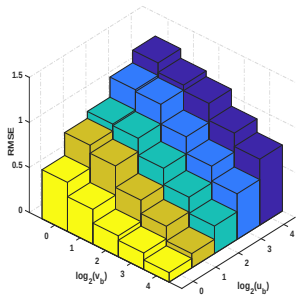


**Figure 4: Tuning Parameters**

ub, vb to tune the parameters, the root mean square error of the model are shown in the Figure 4.

For the prior distributions of parameters in PACDF, we set the hyperparameters as follows:

$$\sigma_\theta = 10, \sigma_a = 10$$
$$\sigma_s = 0.1, \sigma_q = 10$$
$$u_b = 1, v_b = 16$$

In the experiments, we set the number of iterations of Algorithm 1 to 5,000 and estimate the parameters based on the last 2,500 samples to guarantee the convergency of the Markov chain. Both our PACDF and other baseline approaches are implemented by using python on a Core i5 3.2Ghz machine with Windows 7 and 8 GB memory.

**PEP Task.** In order to verify the validity of PACDF, we conduct experiments on PEP task, which is to predict the score of examinees for each problem. In order to observe the behavior of the method under different sparsity levels, we built training sets of different size. We selected 20%, 40%, 60%, and 80% as training sets, and the rest were used for testing, respectively. We use root mean square error (RMSE) and mean absolute error (MAE) as the evaluation metrics. Then, we consider baseline approaches as follows:

- **FuzzyCDF** [34]: A fuzzy cognitive diagnosis framework which fuzzified the skill proficiency of examinees based on a fuzzy set assumption. It modelled the generation of the two kinds of problems by considering slip and guess factors. We choose the student with the highest accuracy in peer assessment of our model, then we use his score as the input of Fuzzy CDF.
- **PG1** [4]:A probabilistic model which assumes that the scores given by students are normally distributed. The mean is the sum of the true score and the bias, and the variance is the reciprocal of the student's reliability. The model converts the scores into normalized score (z-score) and estimates the parameters by using the Bayesian method.

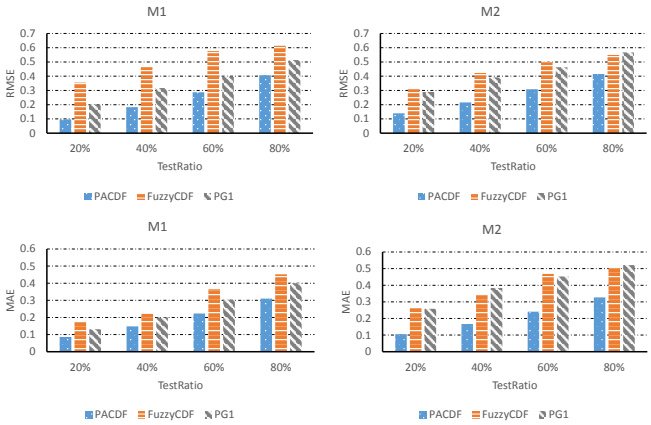


**Figure 5: PEP task Performance**

For comparison, we tuned the parameters to record the best performance of each algorithm. Figure 5 shows the PEP results of our PACDF and baseline methods on datasets.

From this figure, we could observe that PACDF performs best over all the datasets. Specifically, from the perspective of cognitive diagnosis, it defeated FuzzyCDF, and from the perspective of peer assessment, it defeated PG1. More importantly, as the sparsity of the training data increases (training data ratio drops from 80% to 20%), the superiority of our approach is becoming more and more apparent. For example, when the training data is 20%, the improvement of PACDF can reach 22%, 24%, respectively, compared to the optimal baseline method at the MAE metric. In summary, PACDF performs better in predicting examinees' performance, and is more suitable for data sparseness and examinees/problems are cold starts.

Fixing the training data ratio equals to 80%, Figure 6 shows the predictive performance for each problem in the datasets. From each graph, we can observe that PACDF outperforms almost all baselines on all problems.

**Discussion.** It can be seen from the experimental results that PACDF outperforms the baseline on predicting examinee performance. So PACDF could be used in providing accurate diagnosis report. More importantly, it could be used in large-scale scenarios (such as MOOC) to reduce the burden of teachers.

On the other hand, PACDF could be improved. First of all, the computational complexity of PACDF is too high at present. In the future, we will try to design an efficient sampling algorithm. Second, in addition to subjective questions, we need to consider objective problems in the cognitive diagnosis model. Third, we did not take how to assigning task into consideration. Last but not least, the model does not analyze the relationship between examinees' skill proficiency and their strictness.

## 5 CONCLUSION AND FUTURE WORKS

In this paper, we propose a peer-assessment cognitive diagnosis framework (PACDF) that combines peer assessment with cognitive

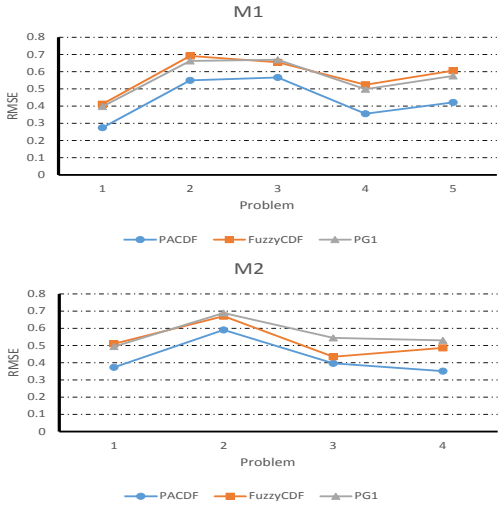


**Figure 6: PEP task Performance for each problem**

diagnosis to predict examinee performance. Specifically, our model defines the relationship between examinees' skill proficiency and problem mastery. We propose a Monte Carol Markov chain sampling algorithm to estimate the parameters and predict examinee performance. The experimental results show that the model could quantitatively explain and analyze the skill proficiency of each examinee, thus perform better in predicting examinee performance.

However, there is still room for improvement. For example, we could try to design an efficient sampling algorithm, model about objective problems, take task assignment into consideration and analyze the relationship between examinees' skill proficiency and their strictness in the future.In future, we will apply this framework to programming education, which is one of the most important and basic parts in Computer Education.

## ACKNOWLEDGMENTS

This work is supported by Youth Innovation Promotion Association of CAS, Anhui Provincial Major Teaching Reform Research Project (No. 2015zdjy004) and National Natural Science Foundation of China (No. 61432016).

## REFERENCES

[1] C. Kulkarni, M. S. Bernstein, and S. Klemmer. Peerstudio: Rapid peer feedback emphasizes revision and improves performance. 2015.
[2] L'Hadi Bouzidi and Alain Jaillet. Can online peer assessment be trusted? *Journal of Educational Technology & Society*, 12(4):257–268, 2009.
[3] Philip M. Sadler and Eddie Good. The impact of self- and peer-grading on student learning. *Educational Assessment*, 11(1):1–31, 2006.
[4] Chris Piech, Jonathan Huang, Zhenghao Chen, Chuong Do, Andrew Ng, and Daphne Koller. Tuned models of peer assessment in moocs. *Computer Science*, 2013.
[5] Nancy Falchikov and Judy Goldfinch. Student peer assessment in higher education: A meta-analysis comparing peer and teacher marks. *Review of Educational Research*, 70(3):287–322, 2000.
[6] David A. F. Haaga. Peer review of term papers in graduate psychology courses. *Teaching of Psychology*, 20(1):28–32, 1993.
[7] George A. Marcoulides and Mark G. Simkin. The consistency of peer review in student writing projects. *Journal of Education for Business*, 70(4):220–223, 1995.
[8] Kwangsu Cho, Christian D. Schunn, and Roy W. Wilson. Validity and reliability of scaffolded peer assessment of writing from instructor and student perspectives. *Journal of Educational Psychology*, 98(4):891–901, 2006.
[9] McGarr, Olliver, and Clifford. 'just enough to make you take it seriously': exploring students';attitudes towards peer assessment. *Higher Education*, 65(6):677–693, 2013.
[10] Thomas Eckes. Introduction to many-facet rasch measurement. *Frankfurt am*, 2011.
[11] Carol M Myford and Edward W Wolfe. Detecting and measuring rater effects using many-facet rasch measurement: Part i. *Journal of applied measurement*, 4(4):386–422, 2003.
[12] Farahman Farrokhi and Rajab Esfandiari. A many-facet rasch model to detect halo effect in three types of raters. *Theory & Practice in Language Studies*, 1(11), 2011.
[13] Everett V Smith Jr and Jonna M Kulikowich. An application of generalizability theory and many-facet rasch measurement using a complex problem-solving skills assessment. *Educational and Psychological Measurement*, 64(4):617–639, 2004.
[14] Richard J. Patz, Brian W. Junker, Matthew S. Johnson, and Louis T. Mariano. The hierarchical rater model for rated test items and its application to large-scale educational assessment data. *Journal of Educational & Behavioral Statistics*, 27(4):341–384, 2002.
[15] Geoff N. Masters. A rasch model for partial credit scoring. *Psychometrika*, 47(2):149–174, 1982.
[16] Lawrence T. Decarlo and Matthew S. Johnson. A hierarchical rater model for constructed responses, with a signal detection rater model. *Journal of Educational Measurement*, 48(3):333–356, 2011.
[17] Mark Wilson and Machteld Hoskens. The rater bundle model. *Journal of Educational & Behavioral Statistics*, 26(3):283–306, 2001.
[18] Louis V. Dibello, Louis A. Roussos, and William Stout. 31a review of cognitively diagnostic assessment and a summary of psychometric models 1 2. *Handbook of Statistics*, 26(06):979–1030, 2006.
[19] J. P. Leighton and M. J. Gierl. Cognitive diagnostic assessment for education: Theory and applications. *Journal of Qingdao Technical College*, 45(4):407–411, 2007.
[20] Nihar B. Shah and Joseph K. Bradley. A case for ordinal peer-evaluation in moocs.
[21] Y. Kotturi, C. Kulkarni, M. S. Bernstein, and S. Klemmer. Structure and messaging techniques for online peer learning systems that increase stickiness. In *Acm Conference on Learning*, 2015.
[22] Catherine M Hicks, Vineet Pandey, C Ailie Fraser, and Scott Klemmer. Framing feedback: Choosing review environment features that support high quality peer assessment. 2016.
[23] Hui Tzu Min. The effects of trained peer review on efl students' revision types and writing quality. *Journal of Second Language Writing*, 15(2):118–141, 2006.
[24] Chinmay Kulkarni, Koh Pang Wei, Huy Le, Daniel Chia, Kathryn Papadopoulos, Justin Cheng, Daphne Koller, and Scott R. Klemmer. Peer and self assessment in massive online classes. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 20(6):1–31, 2013.
[25] Stephen Balfour. Assessing writing in moocs: Automated essay scoring and calibrated peer review. *Research & Practice in Assessment*, 8(1):40–48, 2013.
[26] G. Rasch. "on general laws and the meaning of measurement in psychology,". In *Berkeley Symposium on Mathematical Statistics*, 1961.
[27] Susan E Embretson and Steven P Reise. Item response theory for psychologists. *Quality of Life Research*, 13(3):715–716, 2004.
[28] Xitao Fan. Item response theory and classical test theory: An empirical comparison of their item/person statistics. *Educational and psychological measurement*, 58(3):357–381, 1998.
[29] ALord Birnbaum. Some latent trait models and their use in inferring an examinee's ability. *Statistical theories of mental test scores*, 1968.
[30] Brian W Junker and Klaas Sijtsma. Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement*, 25(3):258–272, 2001.
[31] Jimmy De La Torre. The generalized dina model framework. *Psychometrika*, 76(2):179–199, 2011.
[32] Jimmy De La Torre. Dina model and parameter estimation: A didactic:. *Journal of Educational & Behavioral Statistics*, 34(1):115–130, 2009.
[33] Edward Haertel. An application of latent class models to assessment data. *Applied Psychological Measurement*, 8(3):333–346, 1984.
[34] Runze Wu, Qi Liu, Yuping Liu, Enhong Chen, Yu Su, Zhigang Chen, and Guoping Hu. Cognitive modelling for predicting examinee performance. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2015.
[35] Dornyei and Zoltan. Motivation strategies in the language classroom. *Elt Journal*, 57(3):308–310, 2001.
[36] Zachary A. Pardos, Neil T. Heffernan, Carolina Ruiz, and Joseph E. Beck. The composition effect: Conjuntive or compensatory? an analysis of multi-skill math questions in its. In *International Conference on Educational Data Mining*, 2008.