SC-PA: A Spot-checking Model Based on Stackelberg Game Theory for Improving Peer Assessment

Jia Xu[a,b,*], Panyuan Yang[a], Teng Xiao[a], Pin Lv[a,b], Minghe Yu[c] and Ge Yu[d]

[a] *School of Computer Electronics and Information, Guangxi University, Nanning 530004, China*
[b] *Guangxi Key Laboratory of Multimedia Communications and Network Technology, Nanning 530004, China*
[c] *Software College, Northeastern University, Shenyang 110004, China*
[d] *School of Computer Science and Engineering, Northeastern University, Shenyang 110004, China*

ARTICLE INFO

Keywords:
Peer assessment
Cooperative learning
Spot-checking
Stackelberg equilibrium
Game theory
Review resource allocation

ABSTRACT

Peer assessment can effectively solve the challenge of grading large-scale open-ended assignments on massive open online courses (MOOC) platforms. However, when students are asked to spend their spare time reviewing their peers' submissions, they may not have sufficient motivation which makes their assessment results questionable. To this end, this paper proposes a novel spot-checking peer assessment model, named SC-PA, to improve students' motivation in peer assessment. In SC-PA, a peer assessment activity is modeled as a Stackelberg game, where the teacher acts as the leader and the students are followers. Submissions spot-checked and graded by the teacher are treated as the review resources, which are allocated among students based on their review reliabilities. In particular, to ensure the utilities of both of the teacher and the students, an algorithm that computes the optimal allocation plan for review resources is carefully designed based on the concept of Stackelberg equilibrium. Then, unlike the classical spot-checking model, the spot-checking probability status of each student which is determined based on the computed plan is shown to the student, so as to give full play to the role of spot-checking in enhancing students' motivation for peer grading. Note that, to better evaluate the review reliability of a student, the review reliability of the student in SC-PA is quantified by using the entire historical review records of the student, while previous works only employ the review performance of the student in the latest peer assessment activity. Extensive empirical studies are conducted in the real settings of teaching environment to verify the effectiveness of the proposed SC-PA model in improving the peer assessment. The results indicated that the SC-PA model successfully enhanced students' motivation for peer assessment, and improved their performance in terms of the assessment accuracy, self-determination, and self-efficacy.

1. Introduction

Nowadays, massive open online courses (MOOCs) are drawing increasing attention because they provide millions of learners with open access to high quality courses via the Internet. The enrollment of a popular MOOC can reach tens of thousands of students. Thus, it is very challenging for the teachers of such popular MOOCs to grade the massive submissions of assignments generated by the large-scale participation of students. Although some automated grading techniques (Noorbehbahani and Kardan, 2011; Rico-Juan, Gallego and Calvo-Zaragoza, 2019; Lan, Vats, Waters and Baraniuk, 2015) have been proposed to address large-scale grading problems, there are less effective auto-grading methods for open-ended assignments. This is because the open-ended questions (e.g., essays or problem-solving questions) included in the assignments do not have standardized answers. Considering open-ended assignments are arguably critical for verifying the learning effectiveness of many MOOCs (Archibald, 1938), most popular MOOC platforms, including Coursera[1], edX[2], and ICourse[3], propose employing the idea of peer assessment (or peer grading) to help teachers assess the massive submissions of students towards open-ended assignments. Specifically, in these MOOC platforms, students play an additional role as graders of a small proportion of their peers' submissions, based on rubrics (or benchmarks) set by the teachers. Scores given by a group of students for a certain submission (also known

✉ xujia@gxu.edu.cn (J. Xu); panyuan@st.gxu.edu.cn (P. Yang); xiaoteng@st.gxu.edu.cn (T. Xiao); lvpin@gxu.edu.cn (P. Lv); yuminghe@mail.neu.edu.cn (M. Yu); yuge@mail.neu.edu.cn (G. Yu)
ORCID(s):

[1] https://www.coursera.org/
[2] https://www.edx.org/
[3] https://www.icourse163.org/

as peer scores) were then aggregated to generate an estimate of the true score of the submission. Besides reducing the workload of teachers for grading large-scale open-ended assignments, peer assessment is also believed to bring other educational values, including helping students learn from others (Hsia and Hwang, 2016; Lin, 2018), inspiring students' learning interests (Vu and Dall'Alba, 2007; Chang and Hsu, 2020), strengthening students' participation in a course (Hovardas, Tsivitanidou and Zacharia, 2014), and improving students' sense of responsibility (Li, Liu and Steckelberg, 2010).

Owing to the importance of peer assessment, researchers have proposed various incentive methods to motivate students in a peer assessment activity to give more accurate grades to their peers' submissions, which can be generally divided into two categories, namely gamification incentive methods and spot-checking incentive methods. Gamification incentive methods enhance the motivation of students in peer assessment by using various game elements, such as scores, achievement badges, levels, storylines, leaderboards, and other self-elements and social elements (Moccozet, Tardy, Opprecht and Léonard, 2013; Wu, Daskalakis, Kaashoek, Tzamos and Weinberg, 2015; Tenorio, Bittencourt, Isotani, Pedro and Ospina, 2016). Although gamification incentive methods are naturally interesting and very effective in stimulating the motivation of students in peer assessment, the game elements in them are generally designed based on specific learning contents. For example (Henderson, Kumaran, Min, Mott, Wu, Boulden, Lord, Reichsman, Dorsey, Wiebe et al., 2020), a game was designed to stimulate students' interest in biology by fusing the genes to create their own species, which limits the application of these methods in other peer assessment applications. In recent years, spot-checking incentive methods have attracted increasing increasing attention (Zarkoob, Fu and Leyton-Brown, 2019; Wright, Thornton and Leyton-Brown, 2015; Carbonara, Datta, Sinha and Zick, 2015; Wang, An and Jiang, 2018). The spot-checking incentive methods, which can support any type of peer assessment application, select a small number of student submissions (also known as spot-checked submissions). The true scores of these spot-checked submissions are generally obtained by asking teachers to evaluate them, after which they are rewarded or punished based on the comparison of scores given by the students and the true scores given by the teachers with respect to these submissions. Although these proposed spot-checking incentive methods for peer assessment have yielded promising results, they all fail to give full play to the role of spot-checking in enhancing students' motivation for peer assessment, since: 1) limited review resources (or called as spot-checking resources) of teachers are not reasonably allocated among students, which may lead to the waste of some review resources; 2) students are not aware of their probability of being spot-checked, which makes the spot-checking less effective in motivating them to give accurate scores.

To address this limitations of recent spot-checking incentive methods, **SC-PA**, a novel **S**pot-**C**hecking model for **P**eer **A**ssessment, is proposed on the Stackelberg game theory (Von Stackelberg, 2010). In SC-PA model, a peer assessment activity is treated as a Stackelberg game, where the teacher is the leader and the students act as followers. Only a small amount of submissions are spot-checked and graded by the teacher in the SC-PA model. The graded submissions having true scores given by the teacher, which are deemed as review resources, are allocated among students based on their review reliabilities. In particular, to ensure the utilities of both of the teacher and the students in the peer assessment activity are maximized and guarantee review resources are assigned to students with relatively low review reliabilities, an algorithm that computes the optimal allocation plan for review resources among students is carefully designed in SC-PA model based on the concept of Stackelberg equilibrium. Besides the employment of Stackelberg game theory to model the peer assessment activity, another innovation of SC-PA compared with recent spot-checking models is the exhibition of spot-checking probability status of every student to the student which is determined based on the derived optimal allocation plan. The exhibition of spot-checking probability status to students is a very effective incentive mechanism which can give full play to the role of spot-checking in motivating the students to give more accurate scores to their peers' submissions. To evaluate the effectiveness of the proposed SC-PA model for improving peer assessment, an online peer assessment system that implements both the SC-PA model and a classical spot-checking model was designed and deployed in Guangxi University. By applying the peer assessment system, empirical studies have been conducted for different spot-checking peer assessment models under the real classroom settings. Specifically, the RMSEs of student-giving scores with respect to the true scores of submissions were employed to evaluate the incentive effects of different peer assessment models. Questionnaires were also designed and used to explore the effects of different peer assessment models on students' intrinsic motivation, self-determination, self-efficacy, and career motivation. Empirical studies show that the proposed SC-PA model is more effective in improving the accuracy of student-giving scores than the classical spot-checking model. Moreover, by analyzing the feedback from the questionnaires, we found that the proposed SC-PA model is also very helpful in improving students' self-determination and self-efficacy.

2. Related work

2.1. Peer assessment

Peer assessment is the mainstream solution that addresses the massive evaluation problem of open-ended assignments submitted by large-scale students on MOOC platforms (Xu, Li, Liu, Lv and Yu, 2021). To be more specific, in a peer assessment activity, each student plays an additional role as a grader (Chang and Hsu, 2020) to give evaluation results (e.g., scores or textual feedback) (Topping, 1998) for a small part of their peers' submission of assignments, based on some rubrics given by the teachers (Li et al., 2010). Scores given by a group of students towards a certain submission were then aggregated to generate a prediction of the true score of the submission.

With the proliferation of online education platforms, more empirical studies about peer assessment have been performed by both academia and industry, which has had a positive impact on modern education. For example, Hwang and Hung (2014) proposed a peer assessment based game development approach in a an elementary school science course. Their empirical results demonstrate that students using the peer assessment based game development approach have better learning achievement, learning motivation, and problem-solving skills than students who learned with the conventional game development approach. Hsu (2016) focuses on the computer skills training and developed a peer assessment system using the idea of grid-based knowledge classification. Their experimental results showed that students using the peer assessment system performed significantly better than students who did not participate in the peer grading activity. With respect to the astronomy course provided by Coursera, Formanek, Wenger, Buxner and Impey (2017) organized and analyzed peer assessment activities. According to their analysis results, learners who performed well in grading their peers' assignments showed better engagement and performed better during the course. Most recently, Chang and Hsu (2020) develop a peer assessment approach that incorporates virtual reality activities for a natural science course. They found that the proposed approach not only improved the learning achievement of students but also enhances their self-efficacy and critical thinking tendencies.

Although peer assessment is helpful, obtaining accurate predictions of the true scores of submissions is challenging. Hence, many researchers have devoted themselves to improving the quality of peer assessment, which can be categorized into grade aggregation methods or incentive methods. Grade aggregation methods are designed to improve the predicted score of each submission by refining the aggregation of peer scores given by students with different reliabilities or biases. According to the different forms of peer scores, the grade aggregation methods can be divided into ordinal aggregation (Capuano, Caballé and Percannella, 2020; Luaces, Díez, Alonso-Betanzos, Troncoso and Bahamonde, 2015) and cardinal aggregation (Walsh, 2014; Wang, Jing, Li, Gao and Tang, 2019). Although the optimization of aggregation strategies of peer scores can improve the quality of peer assessment, it cannot increase the accuracy of peer scores given by the student, which means that the quality of peer assessment still has more room for improvement. To this end, incentive methods focus on enhancing the engagement of participants and fundamentally improving the accuracy of their grades given by them.

2.2. Incentive method

The incentive methods used to improve the accuracy of peer scores can generally be divided into two categories: gamification incentives and spot-checking incentives.

In recent years, many researchers have demonstrated the effectiveness of the gamification incentives under the Context of Peer Assessment. For example, Tenorio et al. (2016) proposed a gamified peer assessment model that was implemented in an intelligent tutoring system and aimed to monitor the peer review of students in a personalized manner. Their experimental results show that the average grade of a submission given by a group of students is equivalent to that given by an expert, and students grade their peers' submissions more carefully with gamification incentives. Moccozet et al. (2013) described an assessment framework that applies gamification components to encourage and moderate students' contributions to peer assessment work. Their study showed that the gamification mechanism is eye-catching for students and encourages them to be more serious in their grading works.

In contrast to gamification incentives, spot-checking incentives have attracted considerable attention because of their generalization ability to motivate students to give high-quality grades to their peers' submissions. To be more specific, the spot-checking incentives propose to utilizing the expert grades that are usually provided by the teachers to evaluate the accuracy of peer scores given by the students and, as a result, motivate the students to give more precise grades. In a spot-checking peer assessment activity, expert scores are treated as a review resources that are assigned to students. Students who are assigned to review resources and grade lazily will then be punished. However, according to Wang, An and Jiang (2020), a simple spot-checking strategy, for example, randomly selecting students to be checked, may lead to the waste of review resources and thus cannot give full play to stimulate students' participation in the peer grading activity. This is because the spot-checking resources in a simple spot-checking strategy are not reasonably allocated amongst the students, where many unreliable students may not be checked, while reliable students may be assigned many spot-checking resources. Therefore, many studies have recently been proposed to improve the spot-checking procedure. Wright et al. (2015) design a system named "Mechanical TA," which divided students into two groups, i.e., reliable students and unreliable students. Then, students in different groups are spot-checked with different probabilities, which successfully improves the utilization of review resources. However, giving students in the same group the same spot-checking probability is not reasonable, since their grading reliabilities still have relatively large differences. Considering this limitation, Carbonara et al. (2015) proposed treating each student as a single group and calculating the spot-checking probability of the student according to his/her performance in peer review. However, their model assumes that students' performance on their submissions is inversely proportional to their performance in peer review, which hardly holds true in an actual classroom setting. Wang et al. (2018) also define different checking probabilities for each student in a spot-checking model that determines the spot-checking probability of a student based on the grading performance of the student in the latest peer assessment activity and ignores the historical performance of the student. Although existing works on peer assessment have gained promising results in motivating students to give more accurate grades to their peers' submissions, students cannot perceive the probability states of being spot-checked, which reduces the incentive effect of these works in peer assessment. In this work, we propose to show students their spot-checking probability status, which computed based on the optimal review resources allocation plan derived by applying the Stackelberg equilibrium.

2.3. Game theory in education

Game theory is mainly used to analyze all the strategies that the decision-makers can adopt in competition or cooperation under specific rules and the benefits achieved (Fudenberg and Tirole, 1991). In recent years, game theory has been used to enhance students' learning in the education field. For example, Elbeck and DeLong (2016) designed an experiment based on the prisoner's dilemma game-theoretic model to make ultimate payoffs more explicit during "Principles of Marketing courses," which shows that the game mechanism successfully improves students' learning motivation. Burguillo (2010) proposed a framework for competition-based learning (CnBL) using the idea of game theory based tournaments to enhance students' learning motivation. In contrast to the above-mentioned two works, some scholars have proposed the use of game theory as an analytical tool to explain some negative phenomena in education to find positive solutions. For example, Chiong (2012) explained why most students are unwilling to cooperate in collaborative learning based on the evolutionary game theory and solved this problem by adjusting the Nash equilibrium. Noorani, Manshaei, Montazeri and Zhu (2018) proposed a game theory method based on the explanations and competition in learning. The results indicate that game theory drives learners to participate actively in different learning stages and helps them improve their knowledge more efficiently.

Stackelberg game theory (Leitmann, 1978) is a strategic game in economics in which the leader moves first, followed by a follower who moves sequentially. In recent years, the Stackelberg game has also been applied to educational scenarios to better motivate students in their learning process. As an example, Vallam, Bhatt, Mandal and Narahari (2021) formulate a mixed-integer linear program that views an Online Education Forum (OEF) as a single-leader-multiple-followers Stackelberg game. Their model studies the effect of instructor bias and budget on student participation levels and recommends an optimal plan for instructors to maximize student participation in OEFs.

In summary, existing spot-checking incentives never let the students perceive their probability statuses of being spot-checked, which fails to give full play to the role of spot-checking in enhancing students' grading motivation. Meanwhile, limited review resources of the teacher are not reasonably allocated among students, which may lead to the waste of some review resources and thus will also reduce the effect of spot-checking on stimulating students' to give more precise score to their peers' submission in the peer assessment activity. To solve the above limitations, a novel spot-checking model, named as SC-PA, is proposed and evaluated in this study. The research questions of this study are summarized as follows.

- **Research Question 1:** Do students under the setting of the proposed SC-PA model review their peers' submissions more carefully than students under the classical spot-checking peer assessment model?

- **Research Question 2:** Do students who observe a high probability of being spot-checked perform better than those who observe a low probability of being spot-checked in peer assessment activities?

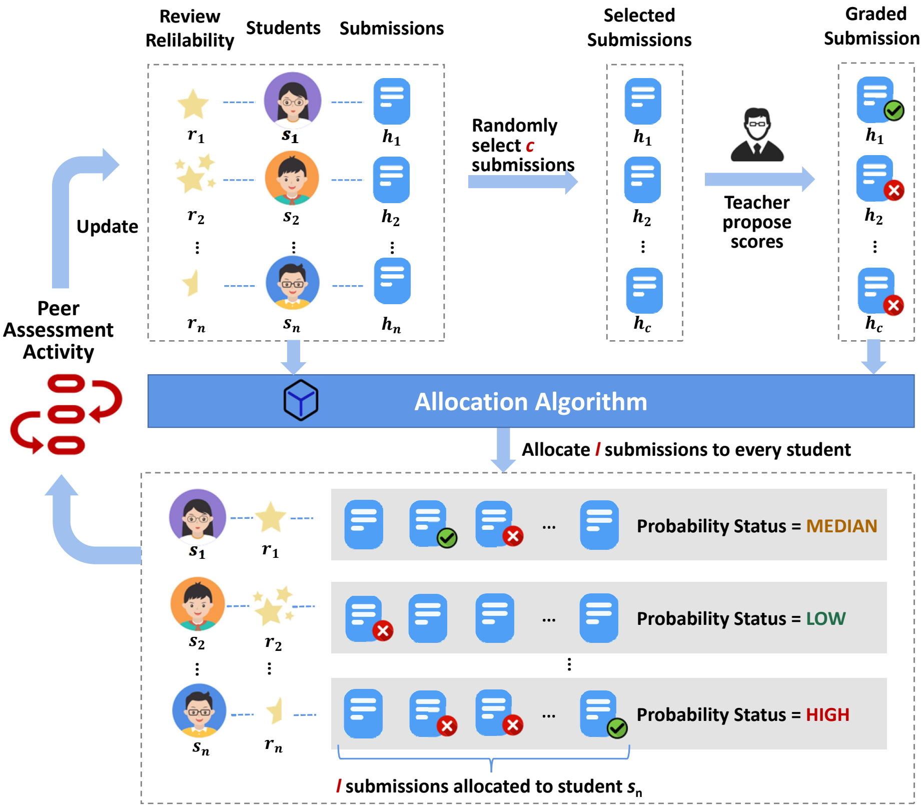


Figure 1: The Workflow of the SC-PA Model

- **Research Question 3:** Do students under the SC-PA model outperform those students under the classical spot-checking peer assessment model in terms of learning perceptions?

3. SC-PA model

3.1. Workflow of SC-PA model

Similar to the classical spot-checking workflow for peer assessment (Gamage, Whiting, Rajapakshe, Thilakarathne, Perera and Fernando, 2017), the workflow of the proposed SC-PA model also consists of four stages: submission, spot-checking, peer review, and settlement, as depicted in Figure 1.

As shown in Figure 1, in the submission stage of the SC-PA model, each student s_i with a review reliability r_i is asked to hand in his/her submission h_i for the same assignment. It is worth noting that a student who does not hand in the submission will not be allowed to participate in the subsequent stages of peer assessment and get a zero score for his/her assignment. Meanwhile, the student's review reliability value will not be updated after this peer assessment activity. For the spot-checking stage, the teacher is required to review c randomly selected submissions of students, denoted by $\{h_1, \dots h_c\}$, and the scores given by the teacher are regarded as the ground-truth scores of these submissions. Since each submission is allocated to l different students in the spot-checking stage, the c submissions graded by the teacher lead to $c \times l$ review resources, which are then allocated to students by the allocation algorithm. More specifically, the allocation algorithm first computes the optimal allocation plan of review resources among students, that is, determining the number of submissions that are graded by the teacher assigned to every student, by maximizing the utility of both the teacher and the students through the Stackelberg equilibrium. After that, l submissions are allocated to each student to evaluate the allocation algorithm of following the derived optimal allocation plan of review resources. The spot-checking stage is followed by the peer review stage, where students can observe their spot-checking probability statuses and are asked to review l peers' submissions assigned to each of them. In particular, the spot-checking probability status of a student is obtained by dividing the number of teacher-graded submissions assigned to the student by the number of submissions allocated to the student, i.e., l. It is worth noting that the submissions of students which have not been graded by any student will be graded by the teacher, to make sure each submission is evaluated in the peer assessment procedure. Finally, in the settlement stage of the SC-PA model, the final score of a student's submission is determined by comprehensively considering the scores of the submission given by the teacher or his/her peers and the student's performance in the peer review stage. Besides, the review reliability of each student will be updated in the settlement stage based on the student's performance in the peer review stage, and a student is penalized if the deviation between a peer score given by the student and the true score of a submission is too large.

In contrast to classical spot-checking solutions for peer assessment, spot-checking probability statuses are shown to students in the peer review stage of the SC-PA model, which is an effective incentive method giving full play to the role of spot-checking in enhancing students' reviewing motivation. Meanwhile, making the spot-checking probability statuses observable to the students makes the SC-PA model a non-cooperative game between students and teachers, which can be solved by the Stackelberg game theory (Leitmann, 1978).

Stackelberg game theory is a game with two players (i.e., leader and follower), where the leader acts first, and then the followers observe the leader's strategy and act accordingly. Unlike the Nash equilibrium, Stackelberg equilibrium refers to the optimal strategy of the leader in the case that followers acting according to the leader's strategy, and the leader's optimal strategy is composed of the optimal strategies of the followers. In the SC-PA model, the teacher acts as the leader, and the students perform as followers. In the game, the teacher's action is to propose an allocation of review resources among students, while the students are asked to review their peers' submissions in the peer review stage. Note that both the students and teachers have their own utility: Students want to spend as little effort as possible in the peer review stage while not being penalized, and teachers want to maximize the overall accuracy of the peer assessment results. Under such circumstances, the teacher is supposed to ensure that the review resources are optimally allocated among students so that students with different review reliabilities can be well motivated in the peer review stage.

In the following subsections, we elaborate on the implementation of the proposed SP-PA model.

3.2. Preliminaries

Let $S = \{s_1, s_2, \dots s_n\}$ denote the set of n students who have handed in their submissions for an assignment on a MOOC platform and s_i denote an arbitrary student in S. In a peer assessment activity, each student will be assigned with l peers' submissions of the same assignment for scoring, which indicates every submission will receive l peer scores. Among the set of l allocated submissions of a student s_i, we use the notation m_i ($m_i \leq l$) to represent the cardinality of submissions in the set that have been evaluated by the teachers in the spot-checking stage. Then, the notation $\boldsymbol{M} =< m_1, m_2, \dots m_n >$ is used to denote allocation plan of review resources among all students. Note that in the spot-checking stage we assume c submissions of students are spot-checked by the teacher, and the total number of review resources is thus $c \times l$. The following Equation 1 holds true.

$$\sum_{i=1}^{n} m_i = c \times l. \tag{1}$$

The following are definitions of important concepts that will be used in this paper.

True score: It was assumed that each submission was associated with a true score. Let t_i denote the true score of each student s_i's submission. In this paper, scores given by the teacher in the spot-checking stage are deemed as the true scores of those submissions being spot-checked.

Peer score: Peer scores are observable peer scores given by students to their peers' submissions. The notation u_i^j is used to represents the peer score of student s_j's submission which is given by student s_i.

Spot-checking probability: The spot-checking probability of a student s_i denoted by p_i, is the probability that s_i is spot-checked by the teacher, which is computed as $p_i = m_i/l$. The set of spot-checking probabilities for all students is denoted by $\boldsymbol{P} = \{p_1, p_2, \dots p_n\}$.

Given the spot-checking probability of a student, its spot-checking probability status is its discrete representation, which takes the value from the set {'LOW', 'MEDIAN', 'HIGH'}, where 'LOW' indicates the spot-checking probability is low and 'HIGH' means the probability is high. The spot-checking probability status of each student is shown to the student in the peer review stage of the SC-PA model, which is considered more intuitive than its corresponding probability value to enhance the motivation of students to review other peers' submissions.

Review Reliability: Review reliability represents the reliability of a student in evaluating his/her peers' submissions. Let r_i denotes the review reliability of student s_i. The set of review reliability values for all students is represented as $R = \{r_1, r_2, \dots r_n\}$.

According to the recent works on peer assessment, the review reliability of a student in terms of an assignment is directly reflected by the accuracy of peer scores given by the student to his/her peers' submissions (Wang, Fang, Jin and Ma, 2022), and can be inferred by the proficiency of the student to the assignment if the accuracy of peer scores is unknown (Piech, Huang, Chen, Do, Ng and Koller, 2013; Wang et al., 2019). Considering only c with ($c << n$) submissions of students are spot-checked and graded by the teacher in the spot-checking stage of the SC-PA model, there are a large number of submissions that do not gain their ground-truth scores. Under such conditions, we propose Equations 2 and 3 following similar ideas in Fang, Wang and Jin (2017) to quantify the review reliabilities of students under two different scenarios, respectively.

In the first scenario, the l submissions assigned to a student s_i contain some submissions with the teacher giving true scores. The review reliability of s_i after the k^{th} peer assessment activity, denoted as r_i^k, is computed by:

$$r_i^k = 0.5 \times r_i^{k-1} + (1 - 0.5) \times exp(-|\frac{\sum_{j \in H_i} (t_j - u_i^j)}{\varphi |H_i|}|), \tag{2}$$

where r_i^{k-1} is the review reliability of student s_i in the $(k-1)^{th}$ peer assessment activity, H_i is the set of submissions scored by both s_i and the teacher and $|H_i|$ is its cardinality, t_j is the true score of student s_j's submission, u_i^j is the peer score given by student s_i to student s_j's submission, φ is the full credit of the assignment, and $exp(x)$ is a mapping function that maps x into a range of $[0, 1]$. According to Equation 2, the review reliability of student s_i is determined by his/her review reliability in the previous peer assessment activity (i.e., r_i^{k-1}) and the quality of peer scores given by s_i in the current peer assessment activity (i.e., $exp(-|\frac{\sum_{j \in H_i} (t_j - u_i^j)}{\varphi |H_i|}|)$).

In the second scenario, the l submissions assigned to student s_i do not contain any submissions with a true score. The review reliability of s_i after the k^{th} peer assessment activity, i.e., r_i^k, is calculated as:

$$r_i^k = 0.5 \times r_i^{k-1} + (1 - 0.5) \times \frac{\left(\sum_{j \in N_i} (u_j^i \times r_j^{k-1})\right) / \left(\sum_{j \in N_i} r_j^{k-1}\right)}{\varphi}, \tag{3}$$

where N_i is the set of students who grade student s_i's submission in the current peer assessment activity and r_j^{k-1} is the review reliability of student s_j in the previous peer assessment activity. According to Equation 3, the review reliability of student s_i is determined by both his/her review reliability in the previous peer assessment activity (i.e., r_i^{k-1}) and the proficiency of the student to the assignment quantified by his/her submission score in the current peer assessment activity (i.e., $\frac{\left(\sum_{j \in N_i} (u_j^i \times r_j^{k-1})\right) / \left(\sum_{j \in N_i} r_j^{k-1}\right)}{\varphi}$).

Equations 2 and 3 also indicate that the review reliability r_i of a student s_i takes values from $[0, 1]$, and the closer the r_i is to 1, the higher the s_i's review reliability.

Score deviation: Score deviation is denoted as d_i^j, which is the deviation between the peer score u_i^j of student s_j's submission given by student s_i and the true score of the submission. To improve the peer assessment is to decrease the score deviations of all students involved in the peer assessment activity.

Table 1 summarizes the main notations used and their descriptions.

3.3. Stackelberg game defined by SC-PA model

The workflow of the SC-PA model is designed based on the Stackelberg game theory. In the game, the ***teacher*** acts as the ***leader*** while a set of n ***students*** $S = \{s_1, \dots s_n\}$ act as the ***followers***. The teacher has a total number of $c \times l$ review resources, which are allocated to students in S in the spot-checking stage. The allocation plan of review resources is a vector $\boldsymbol{M} = \{m_1, \dots m_n\} \in \mathbb{R}^n_{\geq 0}$ such that $\sum_{i=1}^n m_i = c \times l$, where m_i represents the review resources (i.e., the number of submissions with the teacher giving true scores) allocated to student s_i. The Stackelberg game defined by the SC-PA model proceeds as follows. First, the teacher determines a vector $\boldsymbol{M}$ such that $\sum_{i=1}^n m_i = c \times l$. Each student $s_i \in S$ observes m_i and tries to spend as little effort as possible in the peer review stage while not being penalized. We assume that the utility of each student is entirely identifiable with his/her score on the assignment. Then,

Table 1
Frequently-used Notations

Notation	Description
$S = \{s_1, \ldots s_n\}$	Set of n students, where $s_i \in S$ represents the i^{th} student.
$H = \{h_1, \ldots h_n\}$	Set of submissions of students in S for an assignment, where $h_i \in H$ is student s_i's submission.
c	Number of submissions for an assignment graded by the teacher during the spot-checking stage.
l	Number of submissions for an assignment graded by every student in the peer review stage.
$\boldsymbol{M} =< m_1, \ldots m_n >$	The allocation plan of review resources among all students. $m_i \in M$ $(m_i \leq l)$ denotes the number of submissions reviewed by student s_i which also have been graded by the teacher in the spot-checking stage.
t_i	The true score of student s_i's submission given by the teacher.
u_i^j	The peer score of student s_j's submission h_i given by student s_i.
$R = \{r_1, \ldots r_n\}$	Set of review reliability values of all students, where $r_i \in R$ is the review reliability of student s_i.
d_i^j	The deviation between the peer score u_i^j and the true score t_j of student s_j's submission.
p_i	The probability that the submissions reviewed by student s_i are spot-checked by the teacher is computed as $p_i = m_i/l$.
o_i^j	The probability that student s_i is punished in the settlement stage, due to his/her low accuracy (i.e., large d_i^j) when grading student s_j's submission.
φ	Full credit of the assignment.
θ	Penalty threshold.

we get that the expected utility of a student $s_i \in S$ under the allocation plan $\boldsymbol{M}$, denoted as $\mathcal{U}_s$, is

$$\mathcal{U}_{s_i}(m_i) = g_i - \tau \times \sum_{j=1}^{m_i} o_i^j, \tag{4}$$

where g_i is the score of student s_i's submission h_i on the assignment, τ is a parameter corresponding to the scale of punishment defined by the teacher, and o_i^j represents the probability of s_i being punished for misgrading student s_j's submission h_j on the same assignment. Note that the second part (i.e., $\tau \times \sum_{j=1}^{m_i} o_i^j$) in Equation 4 corresponds to the amount of punishment imposed on the score of student s_i's submission g_i.

Next, we discuss the utility of the teachers. To improve the accuracy of peer assessment, the teacher is interested in making the sum of deviation scores (i.e., $\sum d_i^j$) as small as possible, with d_i^j measuring the difference between the peer score u_i^j and the true score t_j of the student s_j's submission. The target of the teacher can be converted to another problem that minimizes the sum of the probability of each student being penalized due to the misgrading of every submission assigned to him/her, that is, o_i^j. The teacher's utility function, denoted as $\mathcal{U}_t$, is given by

$$\mathcal{U}_t(\boldsymbol{M}) = -\sum_{i=1}^{n} \sum_{j=1}^{m_i} o_i^j. \tag{5}$$

The objective of the SC-PA model is to find an allocation plan $\boldsymbol{M}$ of review resources that maximizes the teacher's utility, assuming that every student $s_i \in S$ observes $m_i \in \boldsymbol{M}$ and then plays a response accordingly in the peer review stage. The optimal allocation plan $\boldsymbol{M}$ of review resources is determined by the Stackelberg equilibrium which is defined in Definition 1.

Definition 1. ***Teacher-Optimal Stackelberg Equilibrium (TeaOptStlEq for short).***

- *A set of n students is $S = \{s_1, \ldots s_n\}$*
- *A set of students' review reliabilities is $R = \{r_1, \ldots r_n\}$*
- *The number of submissions on an assignment graded by the teacher in the spot-checking stage is c.*
- *The number of submissions allocated and graded by each student in the peer review stage is l.*

Equilibrium is a situation in the game theory. When both players' strategies affect each other's utility, each player chooses an invariant strategy such that their utility functions reach their maximum. In the SC-PA model, the teacher's strategy is to determine the allocation plan of review resources $\boldsymbol{M} = \{m_1, \dots m_n\}$, which has an impact on students' review quality. Student s_i's strategy is to maximize his/her utility function by adjusting each o_i^j and minimizing $\sum_{j=1}^{m_i} o_i^j$ in Equation 4 after perceiving the value of m_i via the probability status related to the value of m_i/l. In turn, the adjustment of o_i^j by the student affects the teacher's utility. The teacher-optimal Stackelberg equilibrium is an allocation plan M^* that maximizes the utility functions of both the students and teachers. The optimization variant of Definition 1 is thus rewritten as

$$\underset{M=\{m_1,\dots,m_n\}}{\boldsymbol{Maximize}} - \sum_{i=1}^{n}\sum_{j=1}^{m_i} o_i^j$$
$$s.t. \begin{cases} \sum_{j=1}^{m_i} o_i^j \in \underset{o_i^j \in [0,1]}{\boldsymbol{argmax}}\, \mathcal{U}_{s_i}(m_i) \\ \forall m_i \in \boldsymbol{M},\ \ 0 \le m_i \le l, \\ \sum_{i=1}^{n} m_i = c \times l. \end{cases} \tag{6}$$

In the proposed SC-PA model, for each student $s_i \in S$, after perceiving the value of m_i with respect to the student's spot-checking probability p_i, we assume that s_i will try to spend as little effort as possible in the peer review stage while maximizing his/her utility $\mathcal{U}_{s_i}(m_i)$. Under such circumstances, to determine the best utility of student s_i, the deviation score d_i^j between the peer score u_i^j given by s_i to student s_j's submission and the true score of the submission is assumed to follow the Gaussian distribution below:

$$d_i^j \sim N\left(0, \frac{1}{r_i p_i + 1}\right), \tag{7}$$

where r_i is the review reliability of student s_i and p_i is the spot-checking probability of s_i. Apparently, d_i^j is inversely proportional to s_i's review reliability r_i and the spot-checking probability p_i is related to m_i. The involvement of p_i in Equation 7 echoes the assumption that s_i will try to spend as little effort as possible in the peer review stage while maximizing his/her utility $\mathcal{U}_{s_i}(m_i)$. This is because m_i (or p_i) is undoubtedly an important factor considered by student s_i when the student tries to determine a strategy for reviewing his/her peers' submissions. Hence, the utility function of every student can be maximized by quantifying every o_i^j based on the aforementioned distribution information of the score deviation variable d_i^j. Consider that a student s_i is punished because of the misgrading of his/her peer s_j's submission h_j when the condition $d_i^j \geq \theta$ is met, where θ is a predefined threshold. Then, given the distribution of d_i^j as shown in Equation 7, o_i^j, that is, the probability of s_i being punished for misgrading student s_j's submission h_j after he/she senses the spot-checking probability, is computed using Equation 8.

$$o_i^j = 2 - 2\sqrt{\frac{r_i p_i + 1}{2\pi}} exp(-\frac{\theta^2(r_i p_i + 1)}{2}), \tag{8}$$

where all notations in the equation follow the descriptions in Table 1.

The $\sum_{j=1}^{m_i} o_i^j \in \underset{o_i^j \in [0,1]}{\mathbf{argmax}}\, \mathcal{U}_{s_i}(m_i)$ in Equation 6 can be determined by computing each o_i^j based on Equation 8.

3.4. Computing optimal allocation plan of review resources

As mentioned in Subsection 3.3, computing the optimal allocation plan of review resources, i.e., $\boldsymbol{M}$, is to solve the TeaOptStlEq problem. In this subsection, an efficient and effective dynamic programming algorithm is proposed to address the TeaOptStlEq problem.

Algorithm 1 presents the details of the proposed dynamic programming algorithm. The algorithm includes solving sub-problems in which a subset of students with a subset of review resources is considered. Let $x_i(j)$ be the maximum possible utility that the teacher can achieve from the set of students with indices not greater than i, and the number of review resources equaling j. $f(i, k)$ is the utility the teacher obtains from student s_i when k review resources are allocated to the student. Note that the computation of $f(i, k)$ depends on the utility function of the student s_i'. Because

Equations 4 and 5 show that both student s_i's utility and the teacher's utility are inversely proportional to $\sum_{j=1}^{m_i} o_i^j$, we set each $f(i,k)$ with $0 \le i \le n$ and $0 \le k \le l$ to the smallest $\sum_{j=1}^{m_i} o_i^j$, to maximize the utilities of both the students and teacher. Furthermore, we define $a_i(k)$ as the number of review resources allocated to student s_i in the case of $x_i(j)$ and m_i^* as the number of review resources allocated to student s_i according to the outputted optimal allocation plan of review resources $\boldsymbol{M}^*$.

Next, the accuracy of Algorithm 1 is discussed. Suppose for the current iteration of the loop in the algorithm, the indices are i and j. We have solved the subproblems for all $i' < i$ and all j' in the range $[0, c \times l]$ (Lines 6-9 in Algorithm 1). We find the optimal value (i.e., the smallest value) for $x_{i-1}(j-1) + f(i,k)$ by trying every k in the range of $[0,l]$. Since the total utility of the current iteration (i.e., $x_i(j)$) is equal to the sum of the utility from each target (i.e., $f(*,*)$), the maximum value of $x_i(j)$ is the maximum possible utility for $x_{i-1}(j-1)$ added by the value of $f(i,k)$ (Line 7 in Algorithm 1), after which the optimal number of review resources k for the target $f(i,k)$ is stored in the variable $a_i(j)$ (Line 8 in Algorithm 1). Finally, at the end of the algorithm, we arrive at the maximum utility for the teacher (i.e., $x_n(c \times l)$), and can derive the optimal allocation plan of review resources $\boldsymbol{M}^* =< m_1, \dots, m_n >$ by assessing the values stored in a (Lines 12-15 in Algorithm 1). The time complexity and the space complexity of the algorithm are $O(n \times c \times l^2)$ and $O(n \times c \times l)$, respectively.

Algorithm 1: Computing Optimal Allocation Plan of Review Resources.

```
   Input: $S = \{s_1, \dots, s_n\}, R = \{r_1, \dots, r_n\}, l, c$
   Output: $\boldsymbol{M}^* =< m_1^*, \dots, m_n^* >$
1  initialization
2  $\forall_{i,j} : x_i(j) = +\infty$
3  begin
4      for $i = 0;\ i \le n;\ i++$ do
5          for $j = 0;\ j \le c \times l;\ j++$ do
6              for $k = 0;\ k \le l;\ k++$ do
                   /* When student $s_i$ gets $k$ review resources, the smallest value of $\sum_{j=1}^{k} o_k^j$
                      is $f(i,k)$, which is computed based on $r_i \in R$.                */
7                  $x_i(j) = min(x_{i-1}(j-1) + f(i,k), x_i(j))$;
8                  $a_i(j) = i$;
9              end
10         end
11     end
12     for $q = n, z = c \times l;\ q \ge 0;\ q--$ do
13         $m_q^* = a_q(z);\ z = z - m_q^*$;
14     end
15     return $\boldsymbol{M}^*$
16 end
```

Our algorithm is based on the premise that students who are not assigned review resources can maintain high-quality reviews. To ensure that the premise holds, we provided feedback functionality in our self-developed peer assessment system, allowing students to give feedback on the scores of their peers' submissions. If an evaluated student gives a feedback that the peer score for his/her submission is unreasonable, then the teacher will grade the submission and decide whether the evaluator should be punished, according to Equation 2.

4. Experiments

To test the effectiveness of the proposed SC-PA model, we developed an online peer assessment system in which both the proposed SC-PA model and a classical spot-checking model were implemented to facilitate the organization of peer assessment activities to make a comparison between these two models. Specifically, the classical spot-checking model compared in the experiment randomly determines the students to be spot-checked without informing them of their spot-checking probabilities and punishes students who misgrade their peers' submissions as the SC-PA model does in the settlement stage of peer assessment.

Table 2
Statistics of Experiments 1 and 2

Properties	Experiment 1	Experiment 2
The number of assignments (or peer assessment activities)	4	8
The number of submissions of students	751	475
The number of peer score records	2,226	1,326
The number of students involved	207	64

4.1. Participants

All students participating in our experiment were undergraduates majoring in computer science at Guangxi University. Two experiments were conducted:

- **The experiment for the course "Data Structure" (Exp. 1)**: 207 students from three parallel classes participated in Exp. 1. In particular, the experimental group comprised 72 students from one of these classes who were asked to review their peers' submissions under the setting of our proposed SC-PA model. On the other hand, the control group consisted of 135 students from the remaining two classes who are required to evaluate their peers' submissions under the setting of the classical spot-checking model. We organized four peer assessment activities for both the experimental and control groups.

- **Experiment for the course "Database Principles" (Exp. 2)**: There were 64 students from the same class taking part in the Exp. 2. Unlike Exp. 1, we organized eight peer assessment activities in Exp. 2, where the previous four activities are set with the SC-PA model, and the remaining half of the activities were associated with the classical spot-checking model. Under such circumstances, the first four peer assessment activities were deemed as the experimental group, while the last four activities were treated as the control group.

4.2. Experimental settings

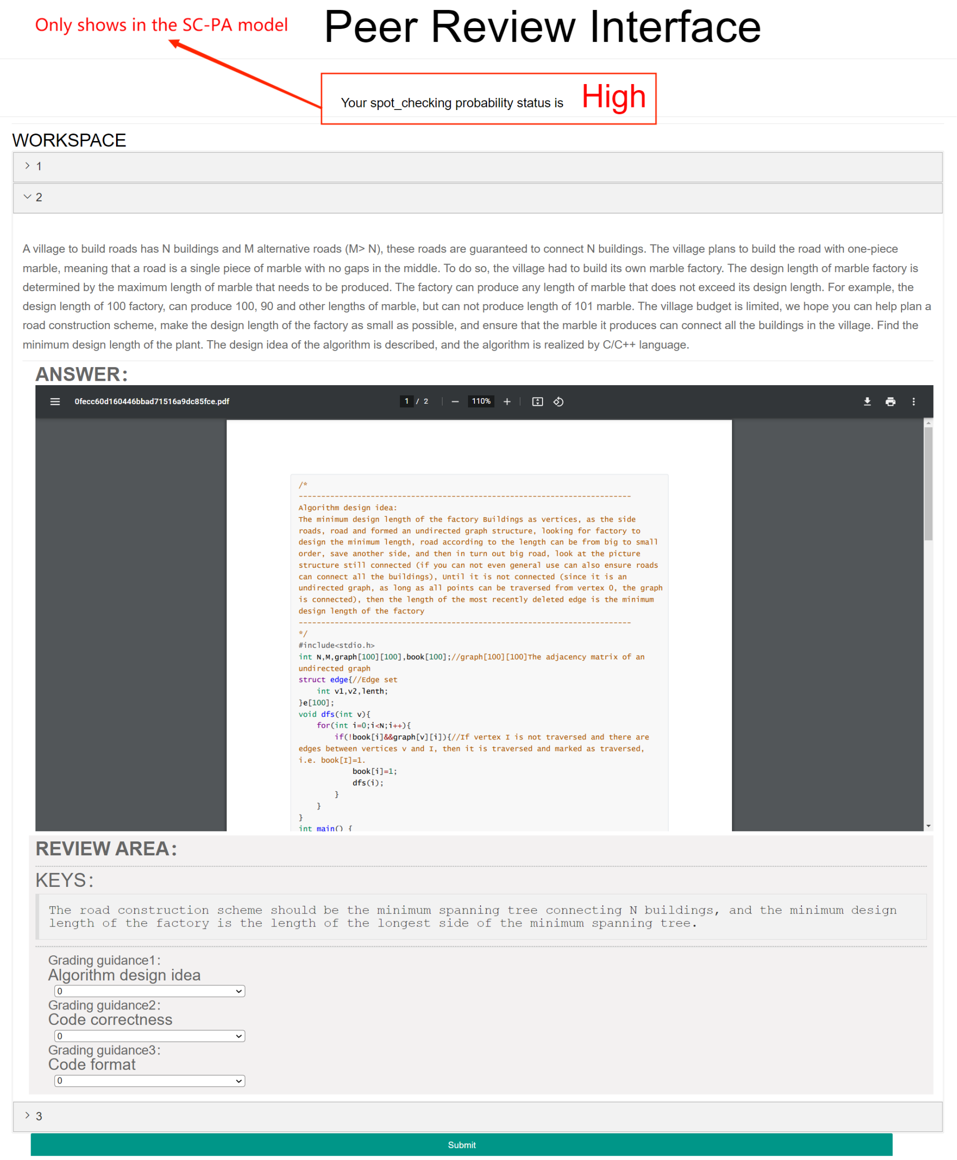
For both Exp. 1 and Exp. 2, each peer assessment activity is corresponding to one open-ended assignment with full credit equaling to ten (i.e., $\varphi = 10$). In the spot-check stage of each peer assessment activity, the teacher randomly spot-checks and grades ten submissions of students on the same assignment (i.e., $c = 10$). In the peer review stage, every student is asked to review three submissions of their peers (i.e., $l = 3$). To ensure fairness, the whole process of peer assessment is double-blind. Figure 2 displays the peer review interface in our developed peer assessment system, where students review their peers' submissions. Restate that one important innovation of the proposed SC-PA model is the exhibition of spot-checking probability status having values in {'LOW','MEDIUM','HIGH'} to every student in the peer review interface. A status of 'LOW' indicates 0 or 1 review resources are allocated to the student, while a status of 'MEDIUM' and 'HIGH' means 2 and 3 review resources are assigned to the student respectively. In the settlement stage, the final score of a spot-checked submission is the true score given by the teacher. Otherwise, the final score of a submission that has not been graded by the teacher is calculated by weighing each of its peer scores by the review reliability of the peer evaluator. Meanwhile, if the deviation between a submission's peer score given by a student and its true score is more than 3 (i.e., $\theta = 3$), the student will receive punishment on the final score of his/her submission. Note that the review reliabilities of all students in the first peer assessment activity are set to 0.5 to handle the cold-start problem. Since all students in Exp. 2 act as members of both the experimental and control groups, they are asked to fill out a questionnaire designed by us. This process is not mandatory at the end of the last peer assessment activity. Table 2 shows other important statistics of the two experiments, while Figure 3 displays scenarios of students participating in the peer assessment activities in two core courses of computer science major.

4.3. Experimental results

In this subsection, we analyze the experimental results to answer the three research questions proposed at the end of Section 2.

4.3.1. Analysis of accuracy

The analysis of accuracy is to answer the ***first research question:*** *Do students under the setting of the proposed SC-PA model review their peers' submissions more carefully than the students in the classical spot-checking peer*



Only shows in the SC-PA model
Peer Review Interface
Your spot_checking probability status is High
WORKSPACE
1
2
ANSWER:
0fecc60d160446bbad71516a9dc85fce.pdf
1 / 2
110%
REVIEW AREA:
KEYS:
3
Submit

A village to build roads has N buildings and M alternative roads (M> N), these roads are guaranteed to connect N buildings. The village plans to build the road with one-piece marble, meaning that a road is a single piece of marble with no gaps in the middle. To do so, the village had to build its own marble factory. The design length of marble factory is determined by the maximum length of marble that needs to be produced. The factory can produce any length of marble that does not exceed its design length. For example, the design length of 100 factory, can produce 100, 90 and other lengths of marble, but can not produce length of 101 marble. The village budget is limited, we hope you can help plan a road construction scheme, make the design length of the factory as small as possible, and ensure that the marble it produces can connect all the buildings in the village. Find the minimum design length of the plant. The design idea of the algorithm is described, and the algorithm is realized by C/C++ language.

```
/*
---------------------------------------------------------------------------
Algorithm design idea:
The minimum design length of the factory Buildings as vertices, as the side
roads, road and formed an undirected graph structure, looking for factory to
design the minimum length, road according to the length can be from big to small
order, save another side, and then in turn out big road, look at the picture
structure still connected (if you can not even general use can also ensure roads
can connect all the buildings), Until it is not connected (since it is an
undirected graph, as long as all points can be traversed from vertex 0, the graph
is connected), then the length of the most recently deleted edge is the minimum
design length of the factory
---------------------------------------------------------------------------
*/
#include<stdio.h>
int N,M,graph[100][100],book[100];//graph[100][100]The adjacency matrix of an
undirected graph
struct edge{//Edge set
    int v1,v2,lenth;
}e[100];
void dfs(int v){
    for(int i=0;i<N;i++){
        if(!book[i]&&graph[v][i]){//If vertex I is not traversed and there are
edges between vertices v and I, then it is traversed and marked as traversed,
i.e. book[I]=1.
            book[i]=1;
            dfs(i);
        }
    }
}
int main() {
```

The road construction scheme should be the minimum spanning tree connecting N buildings, and the minimum design length of the factory is the length of the longest side of the minimum spanning tree.

Grading guidance1: Algorithm design idea: 0

Grading guidance2: Code correctness: 0

Grading guidance3: Code format: 0

Figure 2: The Review Interface in Our Self-developed System

assessment model? In this section, we first adopt the root mean square error (RMSE) as the metric to compare the accuracy of peer scores under the setting of the two spot-checking peer assessment models. Then, we analyze the percentage of adversarial strategies frequently employed by students in peer assessment to test whether the SC-PA model inhibits students from adopting such a strategy.

RMSE is a metric that is widely used to evaluate the accuracy of peer assessment. It measures the deviations of the peer scores from the true scores of submissions, which are computed based on the following equation:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(t_i - \hat{t}_i)^2}, \tag{9}$$

where t_i is the true score for student s_i's submission h_i, $\hat{t}_i$ indicates the average value of peer scores of h_i, and n is the cardinality of submissions in the recent peer assessment activity. Note that true scores of submissions are provided by teachers who have at least 5 years of teaching experience for the corresponding course.

(a) In the 'Data Structure' Course (b) In the 'Database Principles' Course

Figure 3: Scenarios of Peer Assessment Activities

Table 3
Different RMSE in Exp.1 (Best values are in bold.)

Assignments	Experimental Group	Control Group
Assignment 1	**2.12**	2.67
Assignment 2	**2.29**	2.63
Assignment 3	**1.46**	1.92
Assignment 4	**1.77**	2.23

Table 4
Different RMSE in Exp.2 (Best values are in bold.)

Assignments	Experimental Group	Control Group
Assignment 1	**1.40**	1.81
Assignment 2	**1.29**	2.14
Assignment 3	**1.26**	2.49
Assignment 4	**1.28**	2.89

Tables 3 and 4 list the RMSE of peer scores gained in every peer assessment activity corresponding to a certain open-ended assignment in the context of both Exp. 1 and Exp. 2. For each assignment in both the experiments, the RMSE of the experimental group was lower than that of the control group. This is beacuse the SC-PA model shows students the spot-checking probability status, effectively motivating them to review the assignments carefully. The motivational effects of the SC-PA model are discussed in detail in Section 4.3.2.

In Exp. 1, the experimental group showed a mean decrease in RMSE of 24.57% compared with the control group. However, in Exp. 2, the RMSE of the experimental group was significantly smaller than that of the control group. While this result is exciting, we also considered whether other factors affected our experiment. In Exp. 2, students participated in a total of eight peer assessment activities. We were concerned that too many peer assessment activities would reduce the students' patience, and the control group's method would hardly motivate the students. Therefore, we asked the students in Exp. 2, through a questionnaire, whether excessive peer assessment activities would reduce motivation. The conjecture proved to be incorrect, following our research results in Section 4.3.3. The number of peer assessment activities was not enough annoy the students. This demonstrates that the SC-PA model does improves the accuracy of peer assessment.

Next, we analyzed an adversarial strategy that students often use in classical peer assessment to examine whether the SC-PA model inhibits students from utilizing this strategy.

Alfaro and Shavlovsky (2016) found that some students adopted the strategy of countering peer assessment by giving a high score to all submissions. We call this strategy "all correct". We observed this behavior in the classroom,

Table 5
Percentage of "all correct" Records in Exp.1 (Best values are in bold.)

Assignments	Experimental Group	Control Group
Assignment 1	**9.31%**	18.89%
Assignment 2	**5.88%**	19.95%
Assignment 3	**4.37%**	5.88%
Assignment 4	**4.46%**	26.76%

Table 6
Percentage of "all correct" Records in Exp.2 (Best values are in bold.)

Assignments	Experimental Group	Control Group
Assignment 1	**4.71%**	10.90%
Assignment 2	**5.88%**	17.26%
Assignment 3	**2.19%**	27.33%
Assignment 4	**2.19%**	50.70%

Table 7
Mean of Individual RMSE in Exp.1 (Best values are in bold.)

Assignments	High Check Group	Low Check Group
Assignment 1	**1.65**	2.10
Assignment 2	**1.50**	1.95
Assignment 3	**0.69**	1.35
Assignment 4	**1.34**	1.52

which significantly threatened the accuracy of peer assessment. To further confirm the accuracy of the SC-PA model, we examined the proportion of students using the "all correct" strategies in the experiment. In our experiment, the full credit of each assignment was ten points, and we believe that giving a review record of nine or more points for an assignment with a true score of less than six employs the "all correct" strategy.

In Tables 5 and 6, we observe that the proportion of "all correct" records in the experimental group is smaller compared than that in the control group, which proves that the SC-PA model has an improving effect on inhibiting students from using "all correct". In Assignment 3 of Exp.1, the proportion of "all correct" records in every group is much less than in the other assessments. We found that this was due to the low difficulty of Assignment 3, which resulted in over 70% of submissions in the control group receiving a score of nine or higher. However, for the other assignments, it did not exceed 40 percent. Therefore, in Assignment 3, "all correct" is less likely to be found because it is correct in the face of a high-scoring assignment result.

4.3.2. Analysis of incentive

The analysis of incentives is to answer the ***second research question:*** *Do students who observe a high probability of being spot-checked perform better than those who observe a low probability of being spot-checked in peer assessment activities?* To test the impact of different spot-checking probability statuses on students, we grouped students in the experimental group because only the SC-PA model shows students' spot-checking probability statuses. In specific, students with spot-checking probability status equaling to 'MEDIUM' or 'HIGH' are assigned to a group named *High Check Group*, while the remaining students whose status is 'LOW' are allocated to a group named *Low Check Group*. We employed the individual RMSE as the evaluation metric to evaluate the motivational effects of different spot-checking probability statuses on students.

Table 8
Mean of Individual RMSE in Exp.2 (Best values are in bold.)

Assignments	High Check Group	Low Check Group
Assignment 1	**0.85**	1.54
Assignment 2	**0.80**	1.20
Assignment 3	**0.61**	1.28
Assignment 4	**0.59**	1.10

The Individual RMSE measures the performance of individual students in peer assessments, with smaller numbers indicating better performance on the peer assessment. The calculation of individual RMSE was calculated as follows:

$$Individual\ RMSE = \sqrt{\frac{1}{l}\sum_{j=1}^{l}(t_j - u_i^j)^2} \tag{10}$$

where t_j indicates the true score of student s_j's submission, u_i^j indicates the peer score of student s_j's submission given by student s_i, and l is the number of submissions for an assignment graded by s_i in the peer review stage.

Tables 7 and 8 show the mean of individual RMSE for both the experiments. We observed that the individual RMSE of the High Check Group was smaller than the individual RMSE of the Low Check Group for all four assignments in both experiments, demonstrating that showing students a high spot-checking probability status produces more effective motivation. We can also noticed that the individual RMSE of the Low Check Group decreased as the peer assessment activity progressed. This proves that the SC-PA model also has a motivating effect on those students in the Low Check Group with high review reliabilities.

4.3.3. Analysis of learning perception

The analysis of learning perception is to addresses the ***third research question:*** *Do students under the setting of the SC-PA model outperform students under the setting of the classical spot-checking peer assessment model in terms of learning perceptions?* Before conducting the experiment, two questionnaires were designed for Exp. 2 to analyze the perceived different learning effects between the SC-PA model and the classical spot-checking peer assessment. We did not conduct a questionnaire survey in Exp. 1, for the students in Exp. 1 did not simultaneously experience the two spot-checking peer assessment models. As the questionnaire survey was not mandatory, 53 students who participated in the peer assessment activity provided feedback. Table 9 lists the items in the two questionnaires.

As shown in Table 9, Questionnaire 1 focused on studying the motivation of students under the settings of different spot-checking peer assessment models, which were designed based on the scientific motivation questionnaire developed by Glynn, Brickman, Armstrong and Taasoobshirazi (2011). The survey items in Questionnaire 1 covered four topics: intrinsic motivation, self-determination, self-efficacy, and career motivation of students. Students were required to choose a score from 1 to 5 as their feedback for every survey item in the questionnaire, with a score of 5 indicating the highest satisfaction and a score of 1 representing the most negative attitude of students.

In Questionnaire 1, we used the analysis of variance (ANOVA) method to determine the impact of the proposed SC-PA model on students. The Cronbach's alpha value of the experimental group questionnaire, which is the most commonly used reliability analysis method in social science research (Cho, 2016), is 0.97 and 0.90 for the control groups, which shows that the questionnaire is reliable.

Survey items 1 to 4 in Questionnaire 1 are about intrinsic motivation, which relates to intrinsic satisfaction with participating in peer assessment. As shown in Table 10, there was no significant difference between the two groups (F=0.4739, $P > 0.05$). In other words, the SC-PA model does not enhance student satisfaction with participating in peer assessment.

Survey items 5 to 8 in Questionnaire 1 are about self-determination, which refers to the control students believing they have over their learning of assessment.In Table 11, the results (F=3.5227, $P < 0.05$) show that there is a significant difference between improvements in the two groups. Our publicized punishment rule changes the spot-checking probability status into punished probability, making students more willing to participate in peer assessment and enhance their self-determination.

Table 9
Survey Items in Exp.2

	Index	Survey Item
Questionnaire 1	1	Peer assessment is interesting.
	2	I am curious about other students' solutions to the same problem.
	3	I believe that I will do well in peer assessment.
	4	I enjoy reviewing other students' submissions.
	5	I have put enough efforts into peer assessment.
	6	I have carefully considered each rubric set by the teacher to give a reasonable score to the submission assigned to me.
	7	I have spent a lot of time for peer assessment.
	8	I can master the skills of peer assessment.
	9	I believe that the score I gave to a submission is consistent with the score given by the teacher.
	10	I would like to perform better than other students in peer assessment.
	11	I agree that reviewing other students' submissions is a good way for me to consolidate the knowledge.
	12	I believe that the score I gave to a submission will not be questioned by other people.
	13	I believe that learning the skills of peer assessment will be also helpful in my future work.
	14	Getting a good score for my submission in a peer assessment activity is important to me.
	15	My career will involve the assessing of others' works.
	16	I will utilize the skills of peer assessment in my career.
Response scale:		□ Never □ Rarely □ Sometimes □ Usually □ Always
Questionnaire 2	1	I think displaying the probability state that I will be spot-checked plays a vital role in motivating me to give good scores to other students' peer assessment submissions.
	2	The higher probability of spot-checking motivates me to complete the peer assessment task more seriously.
	3	I think showing the probability statuses that students will be spot-checked offers great help in ensuring the fairness of peer assessment and guaranteeing students can receive more helpful feedback from their peers.
	4	I think the peer assessment activities hold for this course are relatively frequent.
	5	Frequent review activities can be annoying to me.
	6	I agree that with the increase in the number of peer assessment activities, the accuracy of reviewing my results given by me will accordingly improve.
Response scale:		□ Yes □ No

Table 10
ANOVA Analysis of Students' Intrinsic Motivation

Group type	Mean	Standard Deviation	F	p
Experimental Group	2.08	1.39	0.4739	0.6227
Control Group	2.17	1.07		

Survey items 9 to 12 in questionnaire 1 are about self-efficacy, which refers to students' beliefs that they can achieve well in the assessment. The self-efficacy questionnaire measures the strength of an individual's belief in completing a peer assessment task. As for results (F=3.5227, $P < 0.05$) shown in Table 12, there is a significant difference between the two groups. The SC-PA model makes students more focus more on peer assessment and makes them believe that their review grades' scores are closer to their true scores.

Survey items 13 to 16 in Questionnaire 1 are about career motivation, which relates to participation in peer assessment as a means of achieving tangible goals, such as a career. As for the results (F=1.394, $P > 0.05$) shown in Table 13, there is no significant difference between the two groups. Students who participated in peer assessment through our spot-checking model did not significantly enhance their career motivation.

Table 11
ANOVA Analysis of Students' Self-determination

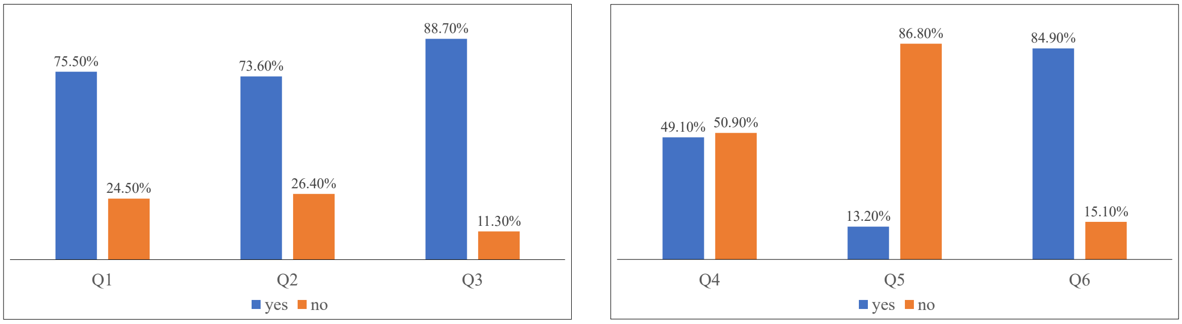
Group type	Mean	Standard Deviation	F	p
Experimental Group	2.94	0.92	3.242*	*0.0398
Control Group	2.75	0.84		

Table 12
ANOVA Analysis of Students' Self-efficacy

Group type	Mean	Standard Deviation	F	p
Experimental Group	2.41	1.01	3.522*	*0.030
Control Group	2.16	0.85		

Table 13
ANOVA Analysis of Students' Career Motivation

Group type	Mean	Standard Deviation	F	p
Experimental Group	2.45	1.02	1.394	0.2400
Control Group	2.36	0.86		



(a) Questionnaire 2 Q1-Q3

(b) Questionnaire 2 Q4-Q6

Figure 4: Statistics of Students' Feedbacks in Questionnaire 2

Overall, the evaluation results of Questionnaire 1 support the hypothesis that the SC-PA model can better help students improve their self-determination and self-efficacy, keep them focused on peer assessment, and make them more confident in their peer assessment results to improve the accuracy of peer assessment.

Since students in Exp. 2 participated in two different types of peer assessments, we asked the students about their perceptions of the two peer assessment methods in the Exp. 2 Questionnaire. In addition, they participated in twice as many peer assessments as those in Exp. 1. Therefore, we also investigated the effect of the number of frequent peer assessment activities on students using the questionnaire. For each question in Exp. 2, the students could choose "yes" or "no" as the answer.

Survey items 1 to 3 in Questionnaire 2 provide feedback on comparing the two spot-checking peer assessment models. Figure 4 shows that most students thought that the SC-PA model had better results than the classical peer assessment method in terms of motivational effect and fairness. 73.6% of the students believed that observing the spot-checking probability status made them feel in the state of being checked, and thus, was more effective in motivation. In addition, 88.7% of the students think that feedback from the SC-PA model is more valuable and fairer than the classical feedback.

Survey items 4 to 6 in Questionnaire 2 provided feedback on students' perceptions of the number of peer assessments. In Exp.2, we conducted eight peer assessment activities. We speculate that multiple peer assessments might drain the students' energy and thus lead to decreased motivation. However, based on the survey results, this contradicts our assumptions. As shown in Figure 4, more than half of the students thought that holding eight peer assessment activities in one session was acceptable. Only 13.2% of students felt that the frequent peer assessments annoyed them, and 84.9% of the students felt that their grading accuracy would gradually increased as the number of peer assessments increased, which indicates that peer assessments positively affected students' learning.

It is worth mentioning that we also surveyed four teachers who used the system for peer assessment activities. All teachers reported that their grading quality had improved with the help of our SC-PA model. In addition, all teachers expressed their willingness to continue their future teaching practice in our system and recommended it to more teachers for peer assessment activities.

5. Discussion and summary

This paper proposes a novel spot-checking peer assessment model, named SC-PA, which is leveraged by Stackelberg game theory to improve the accuracy and incentive of peer assessment. Empirical studies were conducted on the students from GuangxiUiversity to evaluate the performance of the proposed model. In terms of the accuracy of peer assessment, we found that students with respect to the SC-PA model performed more carefully in peer assessment than their counterparts with respect to the classical spot-checking model. In addition, the SC-PA model is able to gradually decrease the proportion of students adopting the "all correct" strategy. The study results respond to the Research Question 1 in Section 2. We also analyzed the incentive effect of displaying the spot-checking probability statuses for the students in the SC-PA model. We found that students with high probabilities of being spot-checked on average provided more precise review results than those with low spot-checking probabilities. In addition, the scoring accuracy of students with high review reliability increased constantly with the iteration of peer assessment activities. All these results answer the Research Question 2, which is raised in Section 2. By conducting a questionnaire 1 survey, we found that students who participated in the peer assessment activities associated with the SC-PA model showed better self-determination and self-efficacy than those who participated in the classical spot-checking peer assessment activities. Meanwhile, the feedback of students for Questionnaire 2 demonstrated that most students deemed the SC-PA model to perform better than the classical spot-checking peer assessment model in terms of motivational effectiveness and fairness. These results address the Research Question 3 in Section 2. The experimental results also refute our hypothesis that multiple peer assessment activities may drain students' energy and thus lead to a decline in their motivation in peer review. Feedback from most students shows that their motivation to review their peers' submissions increases gradually as the number of peer assessment activities increases, which proves the effectiveness of peer assessment in improving students' initiative in learning. In addition, the teachers involved in our peer assessment activities provided positive comments on the proposed SC-PA model.

This study has some limitations. We have only introduced the punishment incentive rule in the SC-PA model. The consideration of adding rewarding rules, such as rewarding students who review seriously, is one of our future studies. Moreover, more reviewing behaviors of students, such as reviewing time of a submission, can also be considered in our future studies to further improve the effectiveness of spot-checking based peer assessment.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (Nos. 62067001, U1811261), Higher Education Undergraduate Teaching Reform Project in Guangxi (No. 2020JGA116), and the Special funds for Guangxi BaGui Scholars. This work is partially supported by the Guangxi Natural Science Foundation (No. 2019JJA170045), and the Fundamental Research Funds for the Central Universities (No. N2117001).

References

Alfaro, L., Shavlovsky, M & Polychronopoulos, V., 2016. Incentives for truthful peer grading. CORR arxiv 1604.

Archibald, R.C., 1938. A semicentennial history of the American Mathematical Society, 1888-1938: With biographies and bibliographies of the past presidents. volume 1. American Mathematical Society.

Burguillo, J.C., 2010. Using game theory and competition-based learning to stimulate student motivation and performance. Computers & Education 55, 566–575. doi:10.1016/j.compedu.2010.02.018.

Capuano, N., Caballé, S., Percannella, Gennaro & Ritrovato, P., 2020. Fopa-mc: Fuzzy multi-criteria group decision making for peer assessment. Soft Computing 24, 17679–17692. doi:10.1007/s00500-020-05155-5.

Carbonara, A.U., Datta, A., Sinha, A., Zick, Y., 2015. Incentivizing peer grading in moocs: An audit game approach, in: In T.-F. International Joint (Ed.) Conference on Artificial Intelligence.

Chang, S.C., Hsu, Ting-Chia & Jong, M.S.Y., 2020. Integration of the peer assessment approach with a virtual reality design system for learning earth science. Computers & Education 146, 103758. doi:10.1016/j.compedu.2019.103758.

Chiong, Raymond & Jovanovic, J., 2012. Collaborative learning in online study groups: An evolutionary game theory perspective. Journal of Information Technology Education: Research 11, 81–101. doi:10.28945/1574.

Cho, E., 2016. Making reliability reliable: A systematic approach to reliability coefficients. Organizational Research Methods 19, 651–682. doi:10.1177/1094428116656239.

Elbeck, M., DeLong, Debbie & Zank, G., 2016. A conceptual framework of cognitive game theory to motivate student learning. Journal of Higher Education Theory & Practice 16.

Fang, H., Wang, Y., Jin, Qun & Ma, J., 2017. Rankwithta: A robust and accurate peer grading mechanism for moocs, in: 2017 IEEE 6th International Conference on Teaching, Assessment, and Learning for Engineering (TALE), IEEE. pp. 497–502.

Formanek, M., Wenger, M., Buxner, S., Impey, Chris & Sonam, T., 2017. Insights about large-scale online peer assessment from an analysis of an astronomy mooc. Computers & Education 113, 243–262. doi:10.1016/j.compedu.2017.05.019.

Fudenberg, D., Tirole, J., 1991. Game theory. MIT press.

Gamage, D., Whiting, M.E., Rajapakshe, T., Thilakarathne, H., Perera, I., Fernando, S., 2017. Improving assessment on moocs through peer identification and aligned incentives, in: Proceedings of the Fourth (2017) ACM Conference on Learning@ Scale, pp. 315–318.

Glynn, S.M., Brickman, P., Armstrong, N., Taasoobshirazi, G., 2011. Science motivation questionnaire ii: Validation with science majors and nonscience majors. Journal of research in science teaching 48, 1159–1176. doi:10.1002/tea.20442.

Henderson, N., Kumaran, V., Min, W., Mott, B., Wu, Z., Boulden, D., Lord, T., Reichsman, F., Dorsey, C., Wiebe, E., et al., 2020. Enhancing student competency models for game-based learning with a hybrid stealth assessment framework. International Educational Data Mining Society .

Hovardas, T., Tsivitanidou, O.E., Zacharia, Z.C., 2014. Peer versus expert feedback: An investigation of the quality of peer feedback among secondary school students. Computers & Education. 71, 133–152. doi:10.1016/j.compedu.2013.09.019.

Hsia, L., Hwang, I.H..G., 2016. Effects of different online peer-feedback approaches on students' performance skills, motivation and self-efficacy in a dance course. Computers & Education. 96, 55–71. doi:10.1016/j.compedu.2016.02.004.

Hsu, T.C., 2016. Effects of a peer assessment system based on a grid-based knowledge classification approach on computer skills training. Journal of Educational Technology and Society 19, 100–111.

Hwang, G.J., Hung, Chun-Ming & Chen, N.S., 2014. Improving learning achievements, motivations and problem-solving skills through a peer assessment-based game development approach. Educational Technology Research and Development 62, 129–145. doi:10.1007/s11423-013-9320-7.

Lan, A.S., Vats, D., Waters, A.E., Baraniuk, R.G., 2015. Mathematical language processing: Automatic grading and feedback for open response mathematical questions, in: Proceedings of the Second (2015) ACM conference on learning@ scale, pp. 167–176.

Leitmann, G., 1978. On generalized stackelberg strategies. Journal of Optimization Theory and Applications 26, 637–643. doi:10.1007/BF00933155.

Li, L., Liu, X., Steckelberg, A.L., 2010. Assessor or assessee: How student learning improves by giving and receiving peer feedback. British Journal of Educational Technology 41, 525–536. doi:10.1111/j.1467-8535.2009.00968.x.

Lin, G.Y., 2018. Anonymous versus identified peer assessment via a facebook-based learning application: Effects on quality of peer feedback, perceived learning, perceived fairness, and attitude toward the system. Computers & Education 116, 81–92. doi:10.1016/j.compedu.2017.08.010.

Luaces, O., Díez, J., Alonso-Betanzos, A., Troncoso, A., Bahamonde, A., 2015. A factorization approach to evaluate open-response assignments in moocs using preference learning on peer assessments. Knowledge-Based Systems 85, 322–328. doi:10.1016/j.knosys.2015.05.019.

Moccozet, L., Tardy, C., Opprecht, W., Léonard, M., 2013. Gamification-based assessment of group work, in: 2013 International Conference on Interactive Collaborative Learning (ICL), IEEE. pp. 171–179.

Noorani, S.F., Manshaei, M.H., Montazeri, M.A., Zhu, Q., 2018. Game-theoretic approach to group learning enhancement through peer-to-peer explanation and competition. IEEE Access 6, 53684–53697. doi:10.1109/ACCESS.2018.2871155.

Noorbehbahani, F., Kardan, A., 2011. The automatic assessment of free text answers using a modified bleu algorithm. Computers & Education 56, 337–345. URL: https://www.sciencedirect.com/science/article/pii/S0360131510002058, doi:https://doi.org/10.1016/j.compedu.2010.07.013.

Piech, C., Huang, J., Chen, Z., Do, C., Ng, A., Koller, D., 2013. Tuned models of peer assessment in moocs. arXiv preprint arXiv:1307.2579 .

Rico-Juan, J.R., Gallego, A.J., Calvo-Zaragoza, J., 2019. Automatic detection of inconsistencies between numerical scores and textual feedback in peer-assessment processes with machine learning. Computers & Education 140, 103609. URL: https://www.sciencedirect.com/science/article/pii/S0360131519301629, doi:https://doi.org/10.1016/j.compedu.2019.103609.

Tenorio, T., Bittencourt, I.I., Isotani, S., Pedro, A., Ospina, P., 2016. A gamified peer assessment model for on-line learning environments in a competitive context. Computers in Human Behavior 64, 247–263. doi:10.1016/j.chb.2016.06.049.

Topping, K., 1998. Peer assessment between students in colleges and universities. Review of educational Research 68, 249–276. doi:10.3102/00346543068003249.

Vallam, R.D., Bhatt, P., Mandal, D., Narahari, Y., 2021. Improving teacher-student interactions in online educational forums using a markov chain based stackelberg game model. arXiv preprint arXiv:2112.01239 .

Von Stackelberg, H., 2010. Market structure and equilibrium. Springer Science & Business Media.

Vu, T.T., Dall'Alba, G., 2007. Students' experience of peer assessment in a professional course. Assessment & Evaluation in Higher Education 32, 541–556. doi:10.1080/02602930601116896.

Walsh, T., 2014. The peerrank method for peer assessment. arXiv preprint arXiv:1405.7192 .

Wang, T., Jing, X., Li, Q., Gao, J., Tang, J., 2019. Improving peer assessment accuracy by incorporating relative peer grades. International Educational Data Mining Society .

Wang, W., An, B., Jiang, Y., 2018. Optimal spot-checking for improving evaluation accuracy of peer grading systems, in: Proceedings of the AAAI Conference on Artificial Intelligence. doi:10.1609/aaai.v32i1.11336.

Wang, W., An, B., Jiang, Y., 2020. Optimal spot-checking for improving the evaluation quality of crowdsourcing: Application to peer grading systems. IEEE Transactions on Computational Social Systems 7, 940–955. doi:10.1109/TCSS.2020.2998732.

Wang, Y., Fang, H., Jin, Q., Ma, J., 2022. Sspa: An effective semi-supervised peer assessment method for large scale moocs. Interactive Learning Environments 30, 158–176. doi:10.1080/10494820.2019.1648299.

Wright, J.R., Thornton, C., Leyton-Brown, K., 2015. Mechanical ta: Partially automated high-stakes peer grading, in: Proceedings of the 46th ACM Technical Symposium on Computer Science Education, pp. 96–101.

Wu, W., Daskalakis, C., Kaashoek, N., Tzamos, C., Weinberg, M., 2015. Game theory based peer grading mechanisms for moocs, in: Proceedings of the Second (2015) ACM Conference on Learning@ Scale, pp. 281–286.

Xu, J., Li, Q., Liu, J., Lv, P., Yu, G., 2021. Leveraging cognitive diagnosis to improve peer assessment in moocs. IEEE Access 9, 50466–50484. doi:10.1109/ACCESS.2021.3069055.

Zarkoob, H., Fu, H., Leyton-Brown, K., 2019. Report-sensitive spot-checking in peer-grading systems. arXiv preprint arXiv:1906.05884 .