Computers, Materials & Continua CMC, vol.65, no.3, pp.2309-2319, 2020

A Sentinel-Based Peer Assessment Mechanism for Collaborative Learning

Cong Wang[1], Mingming Zhao[2], Qinyue Wang[2, 3] and Min Li[2, *]

Abstract: This paper introduces a novel mechanism to improve the performance of peer assessment for collaborative learning. Firstly, a small set of assignments which have being pre-scored by the teacher impartially, are introduced as "sentinels". The reliability of a reviewer can be estimated by the deviation between the sentinels' scores judged by the reviewers and the impartial scores. Through filtering the inferior reviewers by the reliability, each score can then be subjected into mean value correction and standard deviation correction processes sequentially. Then the optimized mutual score which mitigated the influence of the subjective differences of the reviewers are obtained. We perform our experiments on 200 learners. They are asked to submit their assignments and review each other. In the experiments, the sentinel-based mechanism is compared with several other baseline algorithms. It proves that the proposed mechanism can effectively improve the accuracy of peer assessment, and promote the development of collaborative learning.

Keywords: Smart education, peer assessment, collaborative learning.

1 An overview of peer assessment

Smart education, which is represented by Massive Open Online Courses, (MOOC), Small Private Online Courses (SPOC) and flip class, is changing the landscape of education profoundly [Reich (2015)]. Unlike traditional learning forms, in a smart education environment, the interaction among learners may contribute to achieve the educational goal by influencing educational motivation and aspirations through peer relationships, this process is called "collaborative learning". As a key mechanism of collaborative learning, peer assessment means that learners are also play the role of reviewers, to evaluate others' assignments and provide feedbacks. One's score is estimated according to the grades received from other learners. In this process one can improve one's understanding of the course material by evaluating the assignments of others. But learners in a smart education environment are quite different in learning style, profile and prior knowledge [Yang, Zhou

[1] Sichuan Police College, Luzhou, 646000, China.

[2] Sichuan Normal University, Chengdu, 610068, China.

[3] University of Tasmania, Sandy Bay, TAS, 7001, Australia.

[*] Corresponding Author: Min Li. Email: limin@sicnu.edu.cn.

Received: 31 January 2020; Accepted: 07 July 2020

and Yang (2019)], thus the quality of peer assessment relies a lot on the grading experience, ability and the subjective factors of the reviewers [Mustafaraj and Jessica (2015)]. A similar scene of peer assessment in collaborative learning is the peer review process for academic journals or conferences. As widely known, peer reviewing is a key component for the academic community. Tough the reviewers are believed to be trained rigorously for scientific purpose, the effectiveness and the impartiality of peer reviewing are still not recognized fully [Kobren, Saha and McCallum (2019)]. Several potential benefits of peer assessment have been proven [Stover (1976)]:

1. It brings fast and detail feedback on learners' assignments;
2. It helps learners acquire a better understanding on the course through reviewing others' assignments;
3. It may enlighten learners about the advantages and disadvantages of their own assignments.

Most objective questions, such as single/multiple choice, true/false and blank-filling questions, can be graded automatically by computers. But for open-ended questions without standard answers, for instance, programming assignments and essays, reviewers need to learn how to grade the answers accurately [Chinmay, Koh and Huy (2013)]. Peer assessment is considered as a valuable approach for this scene. In this situation learners can also play the role of reviewers, participate in the grading process by reviewing a certain number of assignments, and score them on the basis of specific rubrics or benchmarks established by the instructors [Godlee and Jefferson (1999)]. The final score of an assignment is usually aggregated from the scores received from the reviewers. But several aspects, including the statistical approaches, the personal knowledge of the reviewers, the pre-processing methods and even the degree of transparency, influence the final results of peer assessments to some degree [Price and Peter (2017)]. Among them the subjective factors are commonly believed to have high influence on peer assessment. The subjective factors of reviewers which polarize the reviewing results can be categorized into 2 aspects:

1. The individual judgment rules of the reviewers can rarely fit exactly with that of the instructors. Some lenient reviewers tend to score higher for encouraging learners, while some other reviewers may grade the assignments rigorously to spur learners to study harder.
2. Owing to the lack of experience, interest, or just reluctant to spend times on reviewing, even for malicious purpose [Zhao, Wang and Li (2020)], scores grading by the reviewers cannot be certainty considered impartial and honest [Chinmay, Koh and Huy (2013)].

For example, reviewer A, B and C are asked to review a programming assignment according to the established rules including correctness, readability, complexity and robustness, and grade the assignment as follows:

Table 1: A reviewing example

reviewer	correctness	readability	complexity	robustness	score
A	30	10	15	10	65
B	35	10	20	20	85
C	30	15	15	15	75

correctness (0~40); readability (0~20); complex (0~20); robustness (0~20).

From Tab. 1, it can be seen that each reviewer has its own understanding on the rules, which results in obvious different feedbacks for a same assignment, under the same grading guidelines. In light of this, in recent years, more and more studies have concentrated on technological solutions to improve the accuracy of peer assessment for eliminating biases caused by reviewers' subjective influences.

In this paper, we propose a novel approach for peer assessment. Firstly, we select a small set of pre-scored assignments, which are called "sentinels", and mixed them secretly with other normal assignments before sending to reviewers. Then we split the grading task into two phases, the reviewer grading phase and the assignment grading phase. In the reviewer grading phase, each reviewer's reliability is estimated according to the score they given to the hidden sentinels. The subjective scales of reviewers are also calibrated by weighting the reviewers in this phase. Then, in the assignment grading phase, a weighted average score is obtained as the final results. The experiments show that the proposed approach achieves higher accuracy than traditional methods.

2 Related works

A study on an online course named Human Computer Interaction (HCI) indicates that, though learners' grades exhibit agreement with teacher-given grades [Chinmay, Koh and Huy (2013)], obvious room still remains for the peer assessment to improve. It is estimated that 43% of assignments' grades given from learners fell over 10% from the corresponding teacher's grades. For some specific assignments, the learners' grades are deviated from the teachers' grades about 70pp. Thus, a key challenge still lies in how to estimate the reliabilities and correct the biases of the reviewers.

Several statistical models have been proposed for peer assessment [Nicola, Vincenzo and Francesco (2017); Piech, Huang and Chen (2013); Uto and Maomi (2016)]. The statistical methods commonly assume that the deviation between the grading scores and the ground truth obeys a specific distribution. By estimating the parameters of the distribution, the characters of the grading deviation can be captured for achieving higher assessment accuracy. However, those studies indicate only the personal-thinking of the reviewers can influence the assessment accuracy. The influences caused by the unfaithful reviewers are always ignored since they can hardly be modeled by statistical methods.

Several studies [Walsh (2013); Lu, Warren, Jermaine et al. (2015); Sunahase, Yukino and Hisashi (2017)] assume that the higher score a learner achieves, the more reliable this learner is [Mi and Yeung (2015)]. A common potential limitation of these studies is that the reviewer and the learner set must be placed in one-to-one correspondence, i.e., all the reviewers must submit their assignments as learners for grading, and all the learners are required to play the role of reviewers to grading others' assignments. But in reality, this condition cannot always be satisfied. For example, there exists a number of volunteers in learning community who dedicate to help other learners [Almatrafi and Aditya (2019)]. As a part of core members of learning community, the volunteers are also called as super user, and often serve as reliable reviewers for peer assessment. SSPA is another semi-supervised reliability estimating method which evaluate the similarity between reviewers through considering a chain of reviewers [Wang, Hui and Qun (2019)]. SSPA may get unreliable

grading results while the chain grows too long. And more, SSPA probably fail while the cooperative graph is disconnected since the reviewer chain will be broken in this case.

Raman et al. [Raman and Joachims (2014)] suggest an effective method OPG for peer assessment through ranking the assignments instead of scoring them. OPG algorithm does not require an equivalence between the reviewer and the learner set. However, their experiments require much more reviewers than learners, thus for OPG, how to enroll enough reviewers for peer assessment is still a real challenge [Xiong and Hoi (2018)]. Moreover, OPG cannot identify precise preferences such as "much better" or "a little better". Thus, FOPA algorithm is proposed to improve OPG by fuzzy group decision making [Nicola, Vincenzo and Francesco (2017)] to model the precise preferences. Experiments show that FOPA can achieve better results than other algorithms. But an accuracy falling is also shown when the learners are more than 100.

3 Methodology

3.1 The basic idea of sentinel-based peer assessment mechanism

To improve the accuracy of peer assessment in collaborative learning, this study selects a small set of assignments as sentinels for representing different levels of the assignments. For example, selects a good, a medium and a bad assignment respectively. Notice that the sentinels are selected carefully rather than randomly since a random selection may leads to poor discriminations among sentinels. Generally, the assignments are roughly graded into different quality levels, and in each level a specific number of representative assignments are selected as sentinels. Then, the sentinels are graded by teachers or other verified reliable reviewers, and the scores of the sentinels are recorded as the impartial scores, or the ground truths.

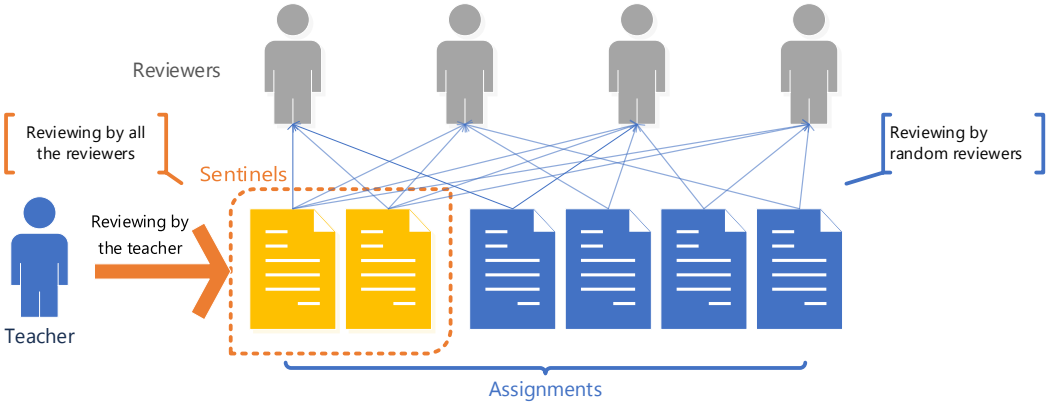

Figure 1: The sketch map of sentinel-based peer assessment mechanism

As shown in Fig. 1, in the peer assessment process, each reviewer will receive a specific number of assignments for grading. The assignments are selected from two parts: the majority are selected from the normal assignments and others are selected from the sentinels. This process is double-blind, only the instructor knows which assignments are the sentinels. From the perspective of the reviewers, all the assignments they received look similar.

Through analyzing the deviations between the reviewer-graded sentinel scores and the corresponding impartial scores, each reviewer' s reliability can be evaluated for filtering inferiors. Furthermore, the deviations are also used to calibrate the subjective offsets to avoid inflating/deflating the scores. The whole process of the mechanism is shown as Fig. 2:

Figure 2: The workflow of sentinel-based peer assessment mechanism

Obviously, two key issues should be illustrated in the mechanism. One is how to evaluate the reliability of the reviewers and filter the inferiors; the other is how to calibrate the raw scores received from the reviewers. The two issues will be discussed further down.

3.2 Reviewers filtering

As discussed in previous section, too many inferiors may decrease the grading accuracy. To avoid this situation, we adopt a white list mechanism: Only the gradings from reliable reviewers are considered in the following procedures. To filter inferiors, the Borda count method is adopted. Borda count is classified as an ordinal voting method since each rank on the ballot is worth a certain number of points. For example, three sentinels A_1, A_2, and A_3 has been graded by the teacher and sorted in descending order $\{A_2 > A_1 > A_3\}$, It can be determined that the first rank gains 3 points, the second rank gains 2 points while the third rank gains 1 point, thus for the teacher, the Borda counts of the sentinels $\{A_1, A_2, A_3\}$ are $\{2, 1, 3\}$ respectively. Then the sentinels are sent to reviewer R for grading. R believes that A_1 is better than A_3, while A_3 is better than A_2, the sequence of the sentinels graded by R is $\{A_1 > A_3 > A_2\}$. Thus, for reviewer R, the Borda counts of the sentinels set $\{A_1, A_2, A_3\}$ are $\{3, 1, 2\}$ respectively. The reason why Borda count is used here for filtering is:

Intuitively, answering the question "How many points should be scored to assignment A?" is sometimes difficult for the lack of a standard answer. Why an assignment should get 85 points rather than 84 may puzzle a part of reviewers. The more possible choices exist, the more puzzled the reviewer becomes. Judging "Is A better than B?" is usually easier than the above question since only two possible answer, YES or NO can be chosen. Thus, for a responsible reviewer, it can generally be ensured that the score of A is higher than the score of B if A is really better than B.

Firstly, the Manhattan distance D_R of reviewer R is defined as:

$$D_R = \sum_{i \in sentinels} |\bar{B}_i - B_{iR}| \quad (1)$$

where $\bar{B}_i$ and B_{iR} mean the Borda counts graded by the teacher and reviewer R on sentinel i. For a given filtering threshold t, reviewers with higher Manhattan distance than t will be deemed as inferiors, and therefore their reviewing results should also be discarded. The threshold t may be a specific number or a percentage quantile.

3.3 Weighting reviewers

An appropriate weight to evaluate each reviewer's reliability is also necessary for improving the performance of peer assessment. In this study the deviations between the

reviewer gradings on the sentinels and the corresponding impartial scores are used for weighting. A simple inverse strategy is used here. For reviewer *R*, it grades sentinel *i* with score x_{iR} while the corresponding impartial score is x_i, the weight w_R of reviewer *R* can be represented as:

$$w_R = \frac{1}{\sum_{i \in sentinels} |x_i - x_{iR}|} \tag{2}$$

3.4 The calibrations on reviewing results

The personal criteria of the reviewers also need to be considered since the subjective factors may inflate/deflate the scores of the assignments. In this section a calibration method with 2 steps is proposed to adjust the subjective bias.

Let's assume that there are *a* reviewers and *b* assignments, note it DOES NOT require the reviewers to submit their assignments. The score of assignment *i* graded by reviewer *j* is recorded as x_{ij}. Then the score matrix *Y* can be represented as:

$$Y = \begin{bmatrix} x_{11} & \cdots & x_{1a} \\ \cdots & \cdots & \cdots \\ x_{b1} & \cdots & x_{ba} \end{bmatrix}_{b \times a} \tag{3}$$

Each column of score matrix *Y* represents the scores graded by a specific reviewer while each row represents the scores of a specific assignment received from the reviewers. It is noticeable that the matrix is sparse, i.e., only a few elements of *Y* have definite values since each reviewer only grade a few assignments. We define x_{ij}=0 temporarily while reviewer *j* doesn't grade assignment *i.* The other aspects worth noticing is that the sentinels also exist in the matrix since they are graded by all the reviewers.

The first step of the calibration is estimating the baselines of the scores. Let's assume that there are *s* sentinels, now we can get the mean score $\bar{x}_j$ of reviewer *j*:

$$\bar{x}_j = \frac{\sum_{i \in sentinels} x_{ij}}{s} \tag{4}$$

The subjective offset u_j of each reviewer *j* can be estimated as:

$$u_j = x - \bar{x}_j \tag{3}$$

where x means the mean value of the impartial scores of the sentinels. An intuitive meaning of u_j is the subjective inflating/deflating degree of reviewer *j*.

To avoid possible overflow, the oversized scores are limited to the max value (e.g. 100 for centesimal system) while the negative value is set to be 1 (not set to be zero since it may be mixed up with the unreviewed elements).

In the second step, the excessive randomness will be calibrated through processing the deviations. The standard deviation σ_j of reviewer *j* can be obtained by:

$$\sigma_j = \sqrt[2]{\frac{\sum_{i \in sentinels} (x_{ij} - \bar{x}_j)^2}{s}} \tag{6}$$

And the mean deviation of all the reviewers is:

$$\bar{\sigma} = \frac{\sum_{j=1}^{a} \sigma_j}{a} \tag{7}$$

Now the calibrated value x'_{ij} of each score x_{ij} can be represented as:

$$x'_{ij} = x_{ij} + u_j \times t_j \tag{8}$$

For the purpose of calibrating too random score, a deflation factor t_j for reviewer j is introduced as:

$$t_j = \frac{\bar{\sigma}}{\sigma_j} \tag{9}$$

The gradings from the reviewers with large standard deviations should be deflated as:

$$x''_{ij} = x'_{ij} + \left(\hat{x}_j - x'_{ij}\right) \times t_j, s.t. \sigma_j > \bar{\sigma} \tag{10}$$

where

$$\hat{x}_j = \frac{\sum_{i=1}^{b} x'_{ij}}{\left|\{x'_{ij} | x'_{ij} \neq 0\}\right|} \tag{11}$$

After the two-step calibration process, the score matrix turns to be:

$$Y' = \begin{bmatrix} x''_{11} & \dots & x''_{1a} \\ \dots & \dots & \dots \\ x''_{b1} & \dots & x''_{ba} \end{bmatrix}_{b \times a} \tag{12}$$

At last, the fixed score g_i of each assignment i can be computed as the weighted mean value:

$$g_i = \begin{cases} x_i, i\ has\ been\ graded\ by\ teacher \\ \frac{\sum_{x''_{ij} \neq 0} w_j \times x''_{ij}}{\sum_{x''_{ij} \neq 0} w_j}, otherwise \end{cases} \tag{13}$$

4 Experiments

4.1 Dataset

We perform the experiments according to the course "Data Structure" of Sichuan Normal University. 200 learners from 5 different classes participate in the experiments as both learners and reviewers. In our experiments, each learner is asked for doing a same project: coding for building a linked list and implementing 4 basic functions: insert, delete, find and get_length. Learners need to finish the project in class and submit their assignments within the allotted time. In the next class, a scoring criterion is distributed to learners to help them reviewing. Each learner receives about 7 assignments which include 3 sentinels, and each assignment are sent to no less than 4 reviewers. No more assignments are distributed to a specific reviewer for the purpose of lightening the reviewers' workloads [Xiong and Hoi (2018)]. Similarly, learners are required to grade these assignments in time from 4 aspects: correctness/bugs, readability/coding style, time/space complexity and robustness. The full scores of the 4 aspects are 50, 20, 20 and 10 respectively. The whole reviewing process is double-blind, i.e. the reviewers don't know who are the submitters of the allocated assignments, and the learners also don't know their assignments are graded by who. We also carefully avoid that a learner and his/her reviewers come from a same class. Each assignment is graded by a same teacher impartially for evaluating the

performance of the mechanism. Upon deadline, 191 learners submitted their feedbacks at least in part, and among them 178 learners finished all the sentinel's grading missions. The deviations between the raw scores and the impartial scores are shown in Fig. 3:

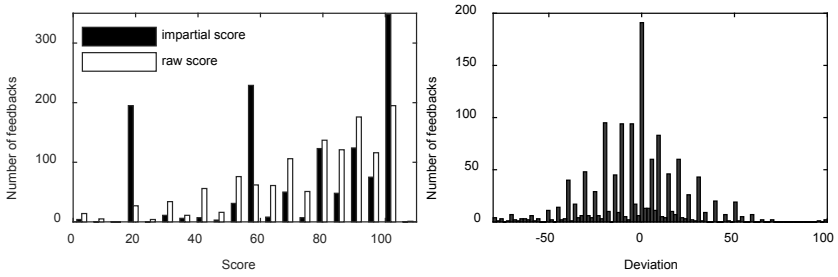

Figure 3: The overview of the raw score and the impartial score

It can be seen in Fig. 3(a) that, there is an obvious difference between the distribution of the raw scores graded by the reviewers and the impartial scores. The 3 significant peaks of the impartial scores include the 3 sentinels respectively. Note that the sentinels are counted several times since from the reviewers' perspective they are independent and different assignments. Clearly for the reviewers, only a few low-quality assignments are graded correctly. For more details, the deviations between the raw scores and the impartial scores are shown in Fig. 3(b). An obvious normal distribution of the deviations between the raw scores and the impartial scores is emerged in this figure. There are about 14% raw scores graded by the reviewers are as same as the impartial scores. And more, more than 30% raw scores are deviated with more than 20 points against the impartial scores. Which indicates that a proper calibration is necessary for improving the reviewing quality of peer assessment.

4.2 Evaluation index

In the experiments, the mean absolute error of the final results (MAE) and the standard deviations are considered as the key performances for the evaluations. Each assignment i in the test set is graded by the teacher with an impartial score s_i, then the formal definition of MAE can be written as:

$$MAE = \frac{\sum_{i=1}^{a}|s_i - g_i|}{a} \tag{14}$$

MAE are used to measure the accuracy of the algorithms, while the standard deviation of the final results (SD) are used to measure the robustness:

$$SD = \sqrt[2]{\frac{\sum_{i=1}^{a}(abs(s_i - g_i) - MAE)^2}{a}} \tag{15}$$

4.3 Determining the filtering threshold t

The only parameter needs to be considered in this work is the threshold t for inferior filtering. To determine the value of t, we set t to be 0%, 25%, 50% (roughly corresponding to filter 0, 1 and 2 inferiors). We do not set t up to be 75% for the lack of feedbacks. In our dataset, a small number of reviewers do not submit enough feedbacks in time. Under this

situation, a too large value of *t* may lead to a dilemma that for some assignments all the feedbacks on them are filtered. The MAE for different *t* is shown in Tab. 2:

Table 2: The relationship between the threshold and the MAE

	t=0%	*t*=25%	*t*=50%
MAE	13.26	11.97	11.67

It is shown that, along with the increasing of threshold *t*, the MAE decreases gradually. If no inferior is filtered, the MAE is similar with the arithmetic mean, and when 50% inferiors is filtered, the MAE drops down to 11.67. Thus, in this study, the threshold is set to 50%. It can ensure that once the number of feedbacks of an assignment is no less than 2, a meaningful score of this assignment can be obtained definitely.

4.4 The comparisons between the proposed model and the contrast algorithms

In order to evaluate the performance of this study, the sentinel-based mechanism is compared with several benchmark algorithms: PeerRank, SSPA and simple arithmetic mean score.

Table 3: The performance improvement of sentinel-based model

	Score		Improvement	
	MAE	SD	MAE	SD
Sentinel-based	11.67	15.12	12.45%	14.43%
PeerRank	14.21	18.51	-6.6%	-4.75%
SSPA	13.14	16.97	1.4%	3.96%
Arithmetic mean	13.33	17.67	-	-

Tab. 3 lists the MAE and the SD of each algorithm. In our experiments, the PeerRank algorithm uses its default settings. For SSPA, owing to the existence of the sentinels, the similarities of all reviewers can be estimated by Eq. (2) of that study directly. In our study, the threshold is set to 50%, which means only the top half part of the reviewers is taken into consideration. We do not consider motivating reviewers for more reliable feedbacks through extra bonus points for two reasons: Firstly, the score represents the quality of the assignment rather than the quality of the feedback; secondly, it is not suitable for the situation that the reviewers set is not equal to the learners set. In this table, the arithmetic mean of the score and the corresponding standard deviation is listed as baselines. It is shown in MAE value that, PeerRank is at the bottom and even worse than the baseline algorithm, while SSPA performs little better than the arithmetic mean. Our sentinel-based model shows about 12.45% improvement on MAE value, which shows the best performance among the algorithms. From the perspective of the robustness, PeerRank also worse than the simple arithmetic mean; SSPA achieves a better level than the arithmetic mean in some degree, while the sentinel-based model also performs better than SSPA. The experiments validate the effectiveness of the model for both accuracy and robustness. Which proves that the proposed model can grade the assignments closer to the teacher's judgment. For collaborative learning it is a more suitable method for peer assessment.

5 Conclusion and future works

To improve the efficiency of collaborative learning and lighten the workload of teachers while facing excessive number of learners, a novel semi-supervised peer assessment mechanism is proposed in this paper. In this study, only a small set of assignments need to be graded by the teacher. This small set is defined as sentinels and thereafter mixed secretly with other assignments before distributing to the reviewers for grading. During the reviewing process, reviewers generally reflect their true level if they are not aware of the existence of the sentinels. Through filtering inferiors and calibrating the raw scores according to the reviewing results of the sentinels, the weighted mean scores of the assignments are obtained to evaluate the qualities of the assignments. Experiments prove that the proposed method has better performances than traditional methods.

Several problems also attract us to study in future works:

First of all, the scoring rule is a key factor which influences the reviewed result. But for subjective items, how to make an appropriate and detailed rule is still a challenging problem. Moreover, currently the weight of each key performance index is established empirically. We plan to introduce scale analysis to develop specific questionnaires for different kind of assignments, and weight each item through factor analysis or PCA.

Secondly, we notice that owing to the randomness of the assignment distribution process, the reviewers allocated to a number of assignments are almost all inferiors, which is a potential factor for pulling down the reviewing quality. Thus, another unsolved problem is how to optimize the distribution process by utilizing the prior information.

At last, the incentive mechanism, i.e., how to stimulate reviewers for more reliable feedbacks is also an opening issue and worth to study.

Funding Statement: This study is sponsored by the National Natural Science Foundation of China (61602331) and the Opening Foundation for the Key Laboratory of Sichuan Province (NDSMS201606).

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

Almatrafi, O.; Aditya, J. (2019): Systematic review of discussion forums in massive open online courses (MOOCs). *IEEE Transactions on Learning Technologies*, vol. 12, no. 3, pp. 413-428.

Chinmay, K.; Koh, P.; Huy, L. (2013): Peer and self assessment in massive online classes. *ACM Transactions on Computer-Human Interaction*, vol. 20, no. 6, pp. 33.

Godlee, F.; Jefferson, T. (1999): *Peer Review in Health Sciences.*

Kobren, A.; McCallum, A. (2019): Paper matching with local fairness constraints. *Proceedings of the 25th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 1247-1257.

Lu, Y.; Warren, J.; Jermaine, C.; Chaudhuri, S.; Rixner, S. (2015): Grading the graders: motivating peer graders in a MOOC. *Proceedings of the 24th International Conference on World Wide Web*, pp. 680-690.

Mi, F.; Yeung, D. (2015): Probabilistic graphical models for boosting cardinal and ordinal peer grading in MOOCs. *Proceedings of the 29th AAAI Conference on Artificial Intelligence*, pp. 454-460.

Mustafaraj, E.; Jessica, B. (2015): The visible and invisible in a MOOC discussion forum. *Proceedings of the 2th ACM Conference on Learning @ Scale*, pp. 351-354.

Nicola, C.; Vincenzo, L.; Francesco, O. (2017): A fuzzy group decision making model for ordinal per assessment. *IEEE Transactions on Learning Technologies*, vol. 10, no. 2, pp. 247-259.

Piech, C.; Huang, J.; Chen, Z. (2013): Tuned models of peer assessment in MOOCs. *Proceedings of the 6th International Conference on Educational Data Mining*, pp. 153-160.

Price, S.; Peter, A. (2017): Computational support for academic peer review: a perspective from artificial intelligence. *Communications of the ACM*, vol. 60, no. 3, pp. 70-79.

Raman, K.; Joachims, T. (2014): Methods for ordinal peer grading. *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1037-1046.

Reich, J. (2015): Rebooting MOOC research. *Science*, vol. 347, no. 6217, pp. 34-35.

Stover, V. (1976): The impact of self-grading on performance and evaluation in a constitutional law course. *Teaching Political Science*, vol. 3, no. 3, pp. 303-310.

Sunahase, T.; Yukino, B.; Hisashi, K. (2017): Pairwise HITS: quality estimation from pairwise comparisons in creator-evaluator crowd sourcing process. *Proceedings of The Thirty-First AAAI Conference On Artificial Intelligence*, pp. 977-984.

Uto, M.; Maomi, U. (2016): Item response theory for peer assessment. *IEEE Transactions on Learning Technologies*, vol. 9, no. 2, pp. 157-170.

Walsh, T. (2013): The peerrank method for peer assessment. *Proceedings of the Twenty-First European Conference on Artificial Intelligence*, pp. 909-914.

Wang, Y.; Hui, F.; Qun, J. (2019): SSPA: an effective semi-supervised peer assessment method for large scale MOOCs. *Interactive Learning Environments*, vol. 44, no. 2, pp. 1-19.

Xiong, Y.; Hoi, K. (2018): Assessment approaches in massive open online courses: possibilities, challenges and future directions. *International Review of Education*, vol. 64, no. 2, pp. 241-263.

Xiong, Y.; Hoi, K. (2018): Assessment approaches in massive open online courses: possibilities, challenges and future directions. *International Review of Education*, vol. 64, no. 2, pp. 241-263.

Yang, Y.; Zhou, D.; Yang, X. (2019): A multi-feature weighting based k-means algorithm for MOOC learner classification. *Computers, Materials & Continua*, vol. 59, no. 2, pp. 625-633.

Zhao, M.; Wang, C.; Li, M. (2020): Peer grading algorithm against malicious evaluation for collaborative learning. *Application Research of Computers*, vol. 37, no. 8 (Preprinted).