Group Optimization to Maximize Peer Assessment Accuracy Using Item Response Theory

Masaki Uto(✉), Nguyen Duc Thien, and Maomi Ueno

University of Electro-Communications, Tokyo, Japan
{uto,thien,ueno}@ai.is.uec.ac.jp

Abstract. As an assessment method based on a social constructivist approach, peer assessment has become popular in recent years. When the number of learners increases as in MOOCs, peer assessment is often conducted by dividing learners into multiple groups to reduce the learner's assessment workload. However, in this case, a difficulty remains that the assessment accuracies of learners in each group depends on the assigned rater. To solve that problem, this study proposes a group optimization method to maximize peer assessment accuracy based on item response theory using integer programming. Experimental results, however, showed that the proposed method does not necessarily present higher accuracy than a random group formation. Therefore, we further propose an external rater selection method that assigns a few outside-group raters to each learner. Simulation and actual data experiments demonstrate that introduction of external raters using the proposed method improves the peer assessment accuracy considerably.

Keywords: Peer assessment · Item response theory · Group formation · Rater selection · Ability measurement

1 Introduction

As an assessment method based on a social constructivist approach, peer assessment, which is mutual assessment among learners, has become popular in recent years [1–3]. Peer assessment can provide the following important benefits [1,2,4].

1. Treating assessment as a part of learning, mistakes can come to represent opportunities rather than failures.
2. Assigning rater roles to learners raises their motivation. Moreover, evaluating others enhances learning from others' work, which induces self-reflection.
3. Transferable skills such as evaluation skills and discussion skills are practiced.
4. Feedback from others who have similar backgrounds is readily understood.
5. Even when the number of learners increases extremely as in massive open online courses (MOOCs), feedbacks and assessment can be offered for each learner.

E. André et al. (Eds.): AIED 2017, LNAI 10331, pp. 393–405, 2017.
DOI: 10.1007/978-3-319-61425-0_33

6. When the learners are mature adults, evaluation by multiple raters is more reliable than that by a single instructor.

Therefore, peer assessment has been adopted into various learning and assessment situations (e.g., [1,3,5]).

One important use of peer assessment is providing formative comments to learners to enhance learning [6,7]. For that purpose, peer assessment has usually been adopted into group learning situations such as collaborative learning, active learning, and project-based learning (e.g., [4,7,8]). Another use of peer assessment is summative assessment [7–9]. The importance of this usage has been increasing with the widespread of large-scale e-learning environments such as MOOCs [7,10,11]. In such environments, evaluation by a single instructor becomes difficult because the number of learners increases extremely. On the other hand, peer assessment can be conducted without burdening the learner's assessment workload if learners are divided into small groups, in which the members assess each other, or only a few peer-raters are assigned to each learner [8,9,11]. Furthermore, peer assessment is justified as an appropriate assessment method because the ability of learners would be defined naturally in the learning community as a social agreement [2,12]. From the above points, this study specifically examines the utilization of peer assessment for summative assessment.

Peer assessment, however, has a problem that the assessment accuracy of a learner's ability depends on rater characteristics such as rating severity and consistency [1,2,4,10,11,13]. To solve the problem, item response theory (IRT)[14] models that incorporate rater characteristic parameters have been proposed (e.g., [1,2,13,15]). The IRT models are known to provide higher assessment accuracy than using the average ratings because they can estimate the ability of learners considering rater characteristics [2].

On the other hand, as described before, peer assessment is often conducted by dividing learners into multiple groups to reduce the learner's assessment workload when the number of learners increases. In such cases, a difficulty persists that assessment accuracies of learners in each group depend on the rater characteristics of the group members. For example, when a group consists of inconsistent peer-raters, the assessment accuracy of the learners in the group will be decreased. To resolve that shortcoming, this study develops a group optimization method to maximize the peer assessment accuracy.

Only one report of the relevant literature describes a study [16] that proposed a group formation method particularly addressing peer assessment accuracy. The purpose of the present study is to provide all learners with assessments that are as equivalently accurate as possible. For that purpose, the study proposed a method that forms groups such that each learner is assessed by peer-raters who are as diverse as possible. The method is expected to reduce differences in assessment accuracy among learners. However, the method does not necessarily maximize the accuracy.

To resolve that shortcoming, this study proposes a new group formation method to maximize peer assessment accuracy based on IRT. Specifically, the method is formulated as an integer programming problem that maximizes the lower bound of the Fisher information measure, a widely used index of ability assessment accuracy in IRT, for each learner. The proposed method is expected to form groups so that the learners in the same group can assess each other accurately. However, experimentally obtained results showed that the proposed method does not necessarily provide higher accuracy than a random group formation method. The result suggests that it is generally impossible to assign raters with high Fisher information to all learners when peer assessment is conducted only within a group.

To resolve the problem, the study proposes an external rater selection method that assigns a few outside-group raters to each learner. The proposed method is formulated as an integer programming problem that maximizes the lower bound of the Fisher information for each learner given by assigned outside-group raters. The proposed method is expected to improve the ability assessment accuracy dynamically because learners can be assessed by outside-group raters who can accurately assess them. Through simulation and experiments using actual data, we demonstrate the effectiveness of the proposed method. Although external evaluation is known to be important for organizations, our results justified it from data.

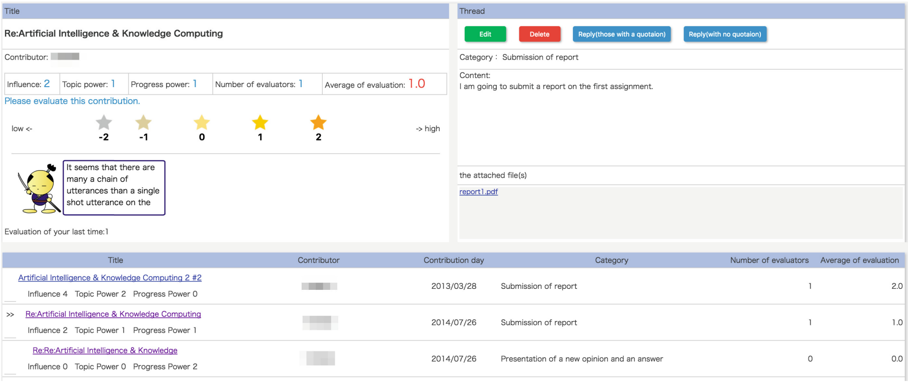
It is noteworthy that many group formation methods have been proposed for improving the effectiveness of collaborative learning (e.g., [17,18]). This study does not specifically examine learning effectiveness. However, the use of groups created using the proposed method are expected to be effective to improve learning because receiving accurate assessment is known to promote effective learning [4]. Therefore, group optimization for improving peer assessment accuracy can be regarded as an important research effort in the field of educational technology.

2 Peer Assessment

One author has developed a learning management system (LMS) called *Samurai* [19]. This study uses the system as a peer assessment platform. Hereinafter, we describe the system structure briefly.

LMS Samurai stores huge numbers of e-learning courses. Each course consists of 15 content sessions tailored for 90-min classes (the units are designated as *topics*). Each topic comprises instructional text screens, images, videos, and practice tests. In some courses, report writing assignments are given to learners. LMS Samurai has a discussion board system that enables learners to submit reports and to conduct peer assessment.

Figure 1 portrays a system interface by which a learner submits a report. The lower half of Fig. 1 presents hyperlinks for other learners' comments. By clicking a hyperlink, detailed comments are displayed in the upper right of Fig. 1. The five star buttons shown at the upper left are used for assigning ratings. The buttons include −2 (Bad), −1 (Poor), 0 (Fair), 1 (Good), and 2 (Excellent). The learner who submitted the report can take the ratings and comments into consideration and rework the report accordingly.



Title

Re:Artificial Intelligence & Knowledge Computing

Contributor: [redacted]

Influence: 2	Topic power: 1	Progress power: 1	Number of evaluators: 1	Average of evaluation: 1.0

Please evaluate this contribution.

It seems that there are many a chain of utterances than a single shot utterance on the

Evaluation of your last time:1

Thread

Category : Submission of report

Content:
I am going to submit a report on the first assignment.

the attached file(s)

report1.pdf

Title	Contributor	Contribution day	Category	Number of evaluators	Average of evaluation
Artificial Intelligence & Knowledge Computing 2 #2 Influence 4 Topic Power 2 Progress Power 0	[redacted]	2013/03/28	Submission of report	1	2.0
>> Re:Artificial Intelligence & Knowledge Computing Influence 2 Topic Power 1 Progress Power 1	[redacted]	2014/07/26	Submission of report	1	1.0
Re:Re:Artificial Intelligence & Knowledge Influence 0 Topic Power 0 Progress Power 2	[redacted]	2014/07/26	Presentation of a new opinion and an answer	0	0.0

Fig. 1. Peer assessment system.

As described in Sect. 1, peer assessment is often conducted by dividing learners into multiple groups. Peer assessment groups can be described as

$$\boldsymbol{X} = \{x_{iggr} | x_{iggr} \in \{0, 1\}\}. \tag{1}$$

Here, x_{igjr} is a variable that takes the value of 1 if learner $j \in \{1, \cdots, J\}$ and peer-rater $r \in \{1, \cdots, J\}$ are included in the same group $g \in \{1, \cdots, G\}$ for assessment of assignment $i \in \{1, \cdots, I\}$. It takes the value of 0 otherwise.

The rating data $\boldsymbol{U}$ obtained from the peer assessment consist of rating categories $k \in \{1 \ldots, K\}$ given by each peer-rater r to each learning outcome of learner j for each assignment i. Letting u_{ijr} be a response of rater r to learner j's outcome for assignment i, the data $\boldsymbol{U}$ are described as

$$\boldsymbol{U} = \{u_{ijr} | u_{ijr} \in \{-1, 1, \cdots, K\}\}. \tag{2}$$

Here, $u_{ijr} = -1$ denotes missing data. When peer assessment is conducted only among group members, the data u_{ijr} for j and r corresponding to $\sum_{g=1}^{G} x_{igjr} = 0$ are missing data. This study uses five categories {1, 2, 3, 4, 5} transformed from the rating buttons {−2, −1, 0, 1, 2} in the system.

This study applies item response theory to the peer assessment data for improving the accuracy of learner ability assessment.

3 Item Response Theory

The item response theory (IRT) [14], a test theory based on mathematical models, has been used widely in areas of educational testing. Actually, IRT is known to realize an accurate assessment of learners' ability by facilitating consideration of test item characteristics (e.g., difficulty and discrimination). Traditionally, IRT has been applied to tests of which the responses can be scored automatically as correct or wrong. In recent years, however, application of polytomous

IRT models to performance assessments such as essay tests and report assessment has been attempted.

Peer assessment data $\boldsymbol{U}$ are three-way data, as defined in Sect. 2. However, traditional IRT models are not directly applicable to such multi-way data [1,2]. To resolve that difficulty, IRT models that incorporate rater characteristic parameters have been proposed (e.g.,[1,2,13,15]). The following subsections introduce an IRT model proposed for peer assessment [2].

3.1 Item Response Theory for Peer Assessment

The IRT model for peer assessment [2] has been proposed as a graded response model (GRM). It is a representative polytomous IRT model that incorporates rater characteristic parameters. The model defines the probability that rater r responds in category k to learner j's work for assignment i as

$$P_{ijrk} = P^*_{ijrk-1} - P^*_{ijrk}, \tag{3}$$

$$P^*_{ijrk} = \left[1 + \exp(-\alpha_i \alpha_r (\theta_j - b_{ik} - \varepsilon_r))\right]^{-1}, \tag{4}$$

where $P^*_{ijr0} = 1$ and $P^*_{ijrK} = 0$. In those equations, θ_j denotes the ability of learner j; α_r reflects the consistency of rater r; ε_r represents the severity of rater r; α_i is a discrimination parameter of assignment i; and b_{ik} denotes the difficulty in obtaining the score k for assignment i. Here, the order of b_{ik} is restricted by $b_{i1} < \cdots < b_{iK-1}$. Furthermore, $\alpha_{r=1} = 1$ and $\varepsilon_1 = 0$ are given for model identification.

For explanation of the rater parameters, Fig. 2 shows item characteristic curves of two raters with assignment parameters $\alpha_i = 1.5$, $b_{i1} = -1.5$, $b_{i2} = -0.5$, $b_{i3} = 0.5$, and $b_{i4} = 1.5$. The left panel presents item characteristic curves of *Rater 1*, who has $\alpha_r = 1.5$ and $\varepsilon_r = 1.0$. The right panel shows item characteristic curves of *Rater 2*, who has $\alpha_r = 0.8$ and $\varepsilon_r = -1.0$. Figure 2 presents a graph with the horizontal axis showing a learner's ability θ_j. The vertical axis shows the rating probability for each category.

According to Fig. 2, the higher the rater consistency parameter is, the greater the differences in the response probability among the rating categories are. That fact reflects that a rater who has a higher consistency can distinguish differences of performance more accurately and consistently. Additionally, Fig. 2 shows that the item response function of *Rater 1*, who has higher severity, shifted to the right compared to those of *Rater 2*, which means that a higher performance is necessary to obtain a score from *Rater 1* than to obtain the same score from *Rater 2*.

This IRT model is known to realize higher accuracy of ability assessment than the other models when the number of raters increases [2]. This study assumes that a group formation is necessary because of an increasing number of learners (=raters). Therefore, we employ the IRT model.

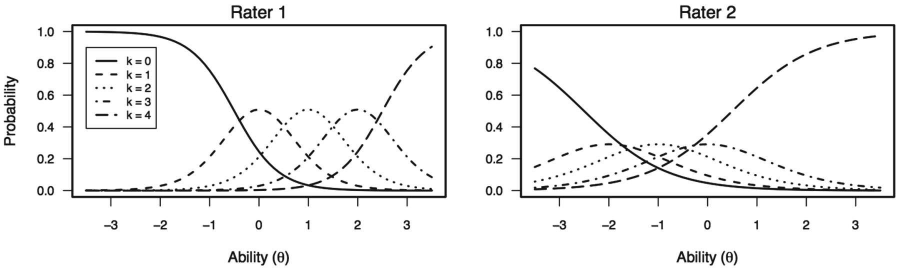


Fig. 2. Item characteristic curves of two raters.

3.2 Fisher Information

In IRT, the standard error of ability estimation is defined as the inverse square root of the Fisher information. Because more information implies less error of measurement, the Fisher information has been widely used as an index of the ability assessment accuracy.

The Uto and Ueno [2] model provides the Fisher information of rater r in assignment i for a learner with ability θ_j as

$$I_{ir}(\theta_j) = \alpha_i^2 \alpha_r^2 \sum_k \frac{\left(P^*_{ijrk-1} Q^*_{ijrk-1} - P^*_{ijrk} Q^*_{ijrk}\right)^2}{P_{ijrk}}, \tag{5}$$

where $Q^*_{ijrk} = 1 - P^*_{ijrk}$.

When peer assessment is conducted among group members, the information for learner j in assignment i is definable by the sum of the information of each rater in the same group as follows.

$$I_i(\theta_j) = \sum_{\substack{r=1 \\ r \neq j}}^{J} \sum_{g=1}^{G} I_{ir}(\theta_j) x_{igjr} \tag{6}$$

A high Fisher information $I_i(\theta_j)$ represents that the assigned raters will accurately assess the ability of learner j. Therefore, if we form groups to provide as much information $I_i(\theta_j)$ to each learner as possible, then the ability assessment accuracy is expected to be improved.

4 Group Optimization Based on Item Response Theory

This study proposes a group formation method to maximize the peer assessment accuracy based on IRT. Specifically, the proposed method is formulated as an integer programming method that maximizes the lower bound of the Fisher information for each learner.

4.1 Group Optimization Method

The group optimization method for assignment i is formulated as shown below.

$$\begin{aligned}
\text{maximize :} \quad & y_i & \\
\text{subject to :} \quad & \sum_{\substack{r=1 \\ r\neq j}}^{J} \sum_{g=1}^{G} I_{ir}(\theta_j) x_{igjr} \geq y_i, & \forall j \\
& \sum_{g=1}^{G} x_{igjj} = 1, & \forall j \\
& \sum_{g=1}^{G} (1 - x_{igjj}) \sum_{r=1}^{J} x_{igjr} = 0, & \forall j \\
& n_l \leq \sum_{j=1}^{J} x_{igjj} \leq n_u, & \forall j \\
& n_l \leq \sum_{g=1}^{G} x_{igjj} \sum_{r=1}^{J} x_{igjr} \leq n_u, & \forall j \\
& x_{igjr} = x_{igrj}, & \forall g, j, r \\
& x_{igjr} \in \{0, 1\}, & \forall g, j, r
\end{aligned}$$

The first constraint requires that the Fisher information for each learner j must be larger than a lower bound y_i. The second and third constraints restrict each learner as belonging to one group. The fourth and fifth constraints control the number of learners in a group. Here, n_l and n_u represent the lower and upper bounds of the number of learners in group g. In this study, $n_l = \lfloor J/G \rfloor$ and $n_u = \lceil J/G \rceil$ are used to equalize the number of learners across groups. Here, $\lfloor \ \rfloor$ and $\lceil \ \rceil$ respectively indicate floor and ceiling functions. This integer programming maximizes the lower bound of the Fisher information for each learner. By solving the problem, we will obtain groups that provide as much Fisher information as possible to each learner.

It is noteworthy that the proposed method requires the estimated parameters of the IRT model. This study assumes that provisional values of the parameters can be given. Examples of their estimation are explained in Sect. 7.

4.2 Evaluation of Group Optimization Method

To confirm the effectiveness of the proposed method, the following simulation experiment was conducted.

1. For $J = 30$ and $I \in \{3, 5\}$, the true parameters were generated randomly.
2. For each assignment i, learners were divided into $G = \{4, 5\}$ groups using the proposed method (designated as *MxFiG*) and a random group formation

Table 1. The average and standard deviation (in parentheses) of the RMSE values in the simulation experiments.

J	I	G			$n^R = 1$		$n^R = 2$		$n^R = 3$	
			RndG	MxFiG	ExRnd	ExFi	ExRnd	ExFi	ExRnd	ExFi
30	3	4	0.368 (0.043)	0.360 (0.068)	0.343 (0.060)	0.297 (0.055)	0.325 (0.058)	0.287 (0.048)	0.318 (0.059)	0.262 (0.038)
		5	0.438 (0.052)	0.408 (0.078)	0.374 (0.079)	0.321 (0.050)	0.333 (0.059)	0.304 (0.054)	0.306 (0.055)	0.291 (0.048)
	5	4	0.252 (0.025)	0.264 (0.065)	0.253 (0.072)	0.235 (0.044)	0.230 (0.057)	0.216 (0.034)	0.216 (0.057)	0.197 (0.037)
		5	0.298 (0.043)	0.307 (0.045)	0.299 (0.045)	0.253 (0.051)	0.259 (0.048)	0.241 (0.039)	0.244 (0.038)	0.225 (0.037)

method (designated as *RndG*). The proposed method was solved using *IBM ILOG CPLEX Optimization Studio*. A feasible solution is used if the optimal solution could not be obtained within five minutes.

3. Given the created groups and the true model parameters, rating data were sampled randomly.
4. The ability of learners was estimated from the generated data given the true parameters of raters and assignments. The expected a posteriori (EAP) estimation was used for the estimation.
5. The root mean square deviation (RMSE) between the estimated ability and the true ability were calculated.
6. After repeating the procedures described above 10 times, the average and standard deviations of the RMSE values were calculated.

Table 1 presents the results. Table 1 shows that the proposed method did not necessarily outperform the random method. The results suggest the general impossibility of assigning raters with high Fisher information to all learners when peer assessment is conducted only among group members.

To confirm that point, Fig. 3 shows the Fisher information for each learner in groups created using the proposed method, given that $J = 30$ and $G = 5$. In the figure, the horizontal axis shows the ability of learner θ. The vertical axis shows the Fisher information $I_i(\theta_j)$. Each datapoint represents an individual learner; the symbols of the data points represent groups to which each learner belongs. According to Fig. 3, we can confirm that high Fisher information is not necessarily provided to all learners.

5 External Rater Selection

The previous section presented a demonstration that the ability assessment accuracy cannot necessarily be improved if peer assessment is conducted only within a group. To overcome that shortcoming, this study further proposes an external rater selection method that assigns a few outside-group raters to each learner.

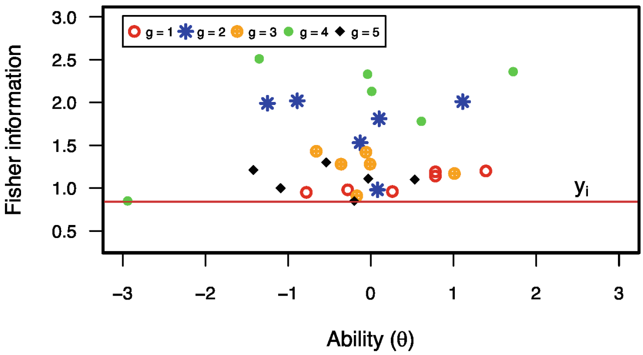


Fig. 3. Fisher information for each learner in groups.

5.1 External Rater Selection Method

The external rater selection method assigns outside-group raters to each learner while providing as much Fisher information as possible. Concretely, the method is formulated as an integer programming problem that maximizes the lower bound of information for learners. Given a group formation $\boldsymbol{X}$, the proposed method for assignment i is defined as follows.

$$\begin{aligned} &\text{maximize :} && y_i && \\ &\text{subject to :} && \sum_{r \in \boldsymbol{C}_{ij}} I_{ir}(\theta_j) z_{ijr} \geq y_i, && \forall j \\ & && \sum_{r \in \boldsymbol{C}_{ij}} z_{ijr} = n^R, && \forall j \\ & && \sum_{j=1}^{J} z_{ijr} \leq n^J, && \forall r \\ & && z_{ijj} = 0, && \forall j \\ & && z_{ijr} \in \{0, 1\}, && \forall j, r \end{aligned}$$

Here, $\boldsymbol{C}_{ij} = \{r \mid r \in \{1, \cdots, J\}. \sum_{g=1}^{G} x_{igjr} = 0\}$ is the set of outside-group raters for learner j in assignment i given a group formation $\boldsymbol{X}$. Also, z_{ijr} is a variable that takes 1 if rater r is assigned to learner j in assignment i; it takes 0 otherwise. The upper limit number of external raters for each learners is n^R. n^J is the upper limit number of outside-group learners assigned to each rater.

The first constraint indicates that the Fisher information for each learner must exceed a lower bound y_i. The second constraint requires that each learner be evaluated by n^R number of external raters. The objective function is defined as the maximization of the lower bound of the information for learners given by assigned external raters.

The proposed method will assign external raters with high Fisher information to each learner. Therefore, the ability assessment accuracy is expected to be improved dynamically, merely by introducing a few external raters.

5.2 Evaluation of External Rater Selection Method

To evaluate the effectiveness of the proposed method, we conducted similar simulation experiments to those explained in Subsect. 4.2. In this experiment, after forming groups by the proposed group optimization method in Procedure 2, $n^R \in \{1, 2, 3\}$ number of external raters were assigned to all learners. The proposed method (designated as *ExFi*), and a random selection method (designated as *ExRnd*) were used as external rater selection methods. For the proposed method, $n^J = 12$ was given.

The results are presented in the columns of *ExFi* and *ExRnd* in Table 1. Table 1 shows that both external rater selection methods improved the ability assessment accuracy as the number of external raters increased.

A comparison of the results of *ExFi* and *ExRnd* revealed that the proposed method provides higher accuracy in all cases. The results confirmed that introducing the external raters with high Fisher information by the proposed method efficiently improves the accuracy of ability assessment.

6 Actual Data Experiments

Actual data experiments were conducted to evaluate the proposed method.

For the experiments, actual peer assessment data were gathered as follows. (1) 34 university students were collected as participants. (2) They were asked to complete four essay writing assignments that were set in the national assessment of educational progress (NAEP) 2002 [20] and 2007 [21]. (3) After the participants completed all assignments, they were asked to evaluate the essays of all other participants for all four assignments. The assessments were conducted using a rubric that we created based on the assessment criteria for grade 12 NAEP writing [21]. The rubric consists of five rating categories with corresponding scoring criteria.

Using the data, we conducted the following experiments.

1. The parameters in the IRT model were estimated using the Markov chain Monte Carlo algorithm [2].
2. For the number of groups $G \in \{3, 4, 5\}$, groups were formed by *MxFiG* and *RndG*. Then, given the groups formed by *MxFiG*, external raters were assigned by *ExRnd* and *ExFi*.
3. The rating data u_{ijr} were changed to missing data when rater r did not assess learner j's work for assignment i in the formed groups and external rater allocations.
4. Given the parameters of raters and assignments that were estimated in Procedure 1, the abilities of learners were estimated from the missing data.

Table 2. The average and standard deviation (in parentheses) of the RMSE values in the actual data experiment.

J	I	G			$n^R = 1$		$n^R = 2$		$n^R = 3$	
			RndG	MxFiG	ExRnd	ExFi	ExRnd	ExFi	ExRnd	ExFi
34	4	3	0.199 (0.027)	0.214 (-)	0.203 (0.009)	0.180 (-)	0.191 (0.014)	0.170 (-)	0.191 (0.012)	0.130 (-)
		4	0.241 (0.036)	0.259 (-)	0.236 (0.013)	0.210 (-)	0.226 (0.014)	0.197 (-)	0.208 (0.023)	0.175 (-)
		5	0.287 (0.035)	0.323 (-)	0.295 (0.024)	0.255 (-)	0.272 (0.018)	0.206 (-)	0.251 (0.021)	0.192 (-)

5. We calculated the RMSEs between the ability estimated from the complete data and those estimated from missing data.
6. For random methods (*RndG* and *ExRnd*), we repeated the procedure described above 10 times. Then the average and standard deviation of the RMSE were calculated. For proposed methods (*MxFiG* and *ExFi*), we did not repeat the procedure because the optimal solution can be determined uniquely.

Table 2 presents the results. Comparing the group formation methods, the proposed group formation method did not necessarily present higher accuracy than the random method, as was true of the simulation results.

According to the results of the external rater selection methods, the accuracies of both methods increased as the number of external raters increased. Comparison of the selection methods shows that the proposed method revealed higher accuracy than the random method in all cases. Specifically, the proposed method with one external rater revealed almost equivalent accuracy to that of the random method with three external raters. Results show that the proposed method is effective for improving the accuracy of ability assessment.

7 Conclusion

This study proposed methods to improve peer assessment accuracy when the assessment is conducted by dividing learners into multiple groups. Specifically, we first proposed the IRT-based group optimization method, which maximizes the lower bound of the Fisher information for each learner. The experimentally obtained results, however, showed that the proposed method does not necessarily provide higher accuracy than a random group formation method.

To resolve the problem, we further proposed the external rater selection method, which assigns a few outside-group raters to each learner. Concretely, the method was formulated as an integer programming problem that maximizes the lower bound of information provided for learners by assigned outside-group raters. The simulation and actual data experiments demonstrate that introducing a few optimal external raters improved the ability assessment accuracy dynamically. Although external evaluation is generally important for organizations, the results justified it from data.

As described in Subsect. 4.1, the proposed methods require the estimated parameters of IRT models. An approach to estimate the assignment parameters is to use peer assessment data collected from past learners of the same course. To estimate the rater parameters and ability, peer assessment for the first assignment might be conducted using other grouping methods. Given the parameters estimated by the data, the proposed methods are useful from the second assignment. Moreover, re-estimating the parameters after every peer assessment using all previous data will be more appropriate.

In this study, we specifically examined only the peer assessment accuracy. However, as discussed in Sect. 1, the proposed methods would also be effective for learning improvement. Evaluation of that assumption is left as a task for future study.

References

1. Ueno, M., Okamoto, T.: Item response theory for peer assessment. In: Proceedings of IEEE International Conference on Advanced Learning Technologies, pp. 554–558(2008)
2. Uto, M., Ueno, M.: Item response theory for peer assessment. IEEE Trans. Learn. Technol. **9**(2), 157–170 (2016)
3. Davies, P.: Review in computerized peer-assessment. Will it affect student marking consistency? In: Proceedings of 11th CAA International Computer Assisted Conference, pp. 143–151(2007)
4. Lan, C.H., Graf, S., Lai, K.R., Kinshuk, K.: Enrichment of peer assessment with agent negotiation. IEEE Trans. Learn. Technol. **4**(1), 35–46 (2011)
5. Cho, K., Schunn, C.D.: Scaffolded writing and rewriting in the discipline: a web-based reciprocal peer review system. Comput. Educ. **48**(3), 409–426 (2007)
6. Topping, K.J., Smith, E.F., Swanson, I., Elliot, A.: Formative peer assessment of academic writing between postgraduate students. Assess. Eval. High. Educ. **25**(2), 149–169 (2000)
7. Moccozet, L., Tardy, C.: An assessment for learning framework with peer assessment of group works. In: Proceedings of International Conference on Information Technology Based Higher Education and Training, pp. 1–5 (2015)
8. Staubitz, T., Petrick, D., Bauer, M., Renz, J., Meinel, C.: Improving the peer assessment experience on mooc platforms. In: Proceeings of Third ACM Conference on Learning at Scale, New York, NY, USA 389–398 (2016)
9. ArchMiller, A., Fieberg, J., Walker, J., Holm, N.: Group peer assessment for summative evaluation in a graduate-level statistics course for ecologists. Assess. Eval. High. Educ. 1–13 (2016)
10. Suen, H.: Peer assessment for massive open online courses (MOOCs). Int. Rev. Res. Open Distrib. Learn. **15**(3), 313–327 (2014)
11. Shah, N.B., Bradley, J., Balakrishnan, S., Parekh, A., Ramchandran, K., Wainwright, M.J.: Some scaling laws for MOOC assessments. In: ACM KDD Workshop on Data Mining for Educational Assessment and Feedback (2014)

12. Lave, J., Wenger, E.: Situated Learning. Legitimate Peripheral Participation. Cambridge University Press, New York, Port Chester, Melbourne, Sydney (1991)
13. Eckes, T.: Introduction to Many-Facet Rasch Measurement: Analyzing and Evaluating Rater-Mediated Assessments. Peter Lang Pub Inc., Bern (2015)
14. Lord, F.: Applications of Item Response Theory to Practical Testing Problems. Erlbaum Associates, Mahwah (1980)
15. Patz, R.J., Junker, B.W., Johnson, M.S., Mariano, L.T.: The hierarchical rater model for rated test items and its application to large-scale educational assessment data. J. Educ. Behav. Stat. **27**(4), 341–366 (1999)
16. Nguyen, T., Uto, M., Abe, Y., Ueno, M.: Reliable peer assessment for team project based learning using item response theory. In: Proceedings of International Conference on Computers in Education, pp. 144–153 (2015)
17. Pang, Y., Mugno, R., Xue, X., Wang, H.: Constructing collaborative learning groups with maximum diversity requirements. In: 15th IEEE International Conference on Advanced Learning Technologies, pp. 34–38, July 2015
18. Lin, Y.S., Chang, Y.C., Chu, C.P.: Novel approach to facilitating tradeoff multi-objective grouping optimization. IEEE Trans. Learn. Technol. **9**(2), 107–119 (2016)
19. Ueno, M.: Data mining and text mining technologies for collaborative learning in an ILMS "samurai". In: Proceedings of IEEE International Conference on Advanced Learning Technologies, pp. 1052–1053 (2004)
20. Persky, H., Daane, M., Jin, Y.: The nation's report card: writing 2002. Technical report. National Center for Education Statistics (2003)
21. Salahu-Din, D., Persky, H., Miller, J.: The nation's report card: writing 2007. Technical report. National Center for Education Statistics (2008)