Group optimization to maximize peer assessment accuracy using item response theory and integer programming

Masaki Uto, Duc-Thien Nguyen, and Maomi Ueno, *Member, IEEE*

Abstract—With the wide spread of large-scale e-learning environments such as MOOCs, peer assessment has been popularly used to measure learner ability. When the number of learners increases, peer assessment is often conducted by dividing learners into multiple groups to reduce the learner's assessment workload. However, in such cases, the peer assessment accuracy depends on the method of forming groups. To resolve that difficulty, this study proposes a group formation method to maximize peer assessment accuracy using item response theory and integer programming. Experimental results, however, have demonstrated that the method does not present sufficiently higher accuracy than a random group formation method does. Therefore, this study further proposes an external rater assignment method that assigns a few outside-group raters to each learner after groups are formed using the proposed group formation method. Through results of simulation and actual data experiments, this study demonstrates that the proposed external rater assignment can substantially improve peer assessment accuracy.

Index Terms—Peer assessment, item response theory, group formation, e-learning, MOOCs, collaborative learning

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], [2], [3]. Peer assessment has been adopted in various learning and assessment situations (e.g., [3], [4], [5], [6], [7], [8], [9]) because it provides many important benefits [1], [2], [3], [10], [11], [12], [13], [14] such as 1) Learners take responsibility for their learning and become autonomous. 2) Assigning rater roles to learners raises their motivation. 3) Transferable skills such as evaluation skills and discussion skills are practiced. 4) By evaluating others, raters can learn from others' work, which induces self-reflection. 5) Learners can receive useful feedback even when they have no instructor.

One common use of peer assessment in higher education is for summative assessment [15], [16], [17]. Peer assessment is justified as an appropriate assessment method because the abilities of learners are definable naturally in the learning community as a social agreement [2], [18]. The importance of this usage has been increasing concomitantly with the wider use of large-scale e-learning environments such as MOOCs [13], [14], [15]. In such environments, evaluation by a single instructor becomes difficult because the number of learners is extremely large. Peer assessment can be conducted without burdening an instructor's or a learner's workload if learners are divided into small groups within which the members assess each other, or if only a few peer-raters are assigned to each learner [14], [16], [17].

Peer assessment, however, entails the difficulty that the assessment accuracy of learner ability depends on rater characteristics such as rating severity and consistency [1], [2], [13], [14], [19], [20], [21], [22], [23]. To resolve that difficulty, item response theory (IRT) [24] models incorporating rater parameters have been proposed (e.g., [1], [2], [23], [25], [26], [27], [28]). The IRT models are known to provide more accurate ability assessment than average or total scores do because they can estimate the ability along with consideration of rater characteristics [2].

In learning contexts, peer assessment has often been adopted for group learning situations such as collaborative learning, active learning, and project-based learning (e.g., [13], [15], [16], [19], [29], [30], [31]). Specifically, learners are divided into multiple groups in which they work together. Peer assessment is conducted within the groups. However, in such peer assessment, the ability assessment accuracy depends also on a way to form groups. For example, when a group consists of learners who can do accurate mutual assessment, their abilities can be estimated accurately from the obtained assessment data. By contrast, if a group consists of learners who tend to assess others randomly, then accurate ability assessment is expected to be difficult. Therefore, group optimization is important to improve the assessment accuracy when peer assessment is conducted within groups.

Only one report of the relevant literature describes a study [31] that proposed a group formation method particularly addressing peer assessment accuracy. However, the purpose of this method is to form groups while providing equivalent assessment accuracy to all learners to the greatest degree possible. Although the method can reduce differences in accuracy among learners, it does not maximize the accuracy.

- *M. Uto is with the University of Electro-Communications, Choufu-shi, Tokyo, Japan.*
  E-mail: uto@ai.lab.uec.ac.jp
- *D.T. Nguyen is with the University of Electro-Communications, Choufu-shi, Tokyo, Japan.*
- *M. Ueno is with the University of Electro-Communications, Choufu-shi, Tokyo, Japan.*

Manuscript received April 19, 2005; revised August 26, 2015.

To resolve that shortcoming, this study proposes and evaluates a new group formation method that maximizes peer assessment accuracy based on IRT. Specifically, the method is formulated as an integer programming problem, a class of mathematical optimization problems for which variables are restricted to integers, that maximizes the lower bound of the Fisher information measure: a widely used index of ability assessment accuracy in IRT. The method is expected to improve the ability assessment accuracy because groups are formed so that the learners in the same group can assess one another accurately. However, experimentally obtained results demonstrated that the method did not present sufficiently higher accuracy than that of a random group formation method. The result suggests that it is generally difficult to assign raters with high Fisher information to all learners when peer assessment is conducted only within groups.

To alleviate that shortcoming, this study further proposes an external rater assignment method that assigns a few optimal outside-group raters to each learner after forming groups using the method presented above. We formulate the method as an integer programming problem that maximizes the lower bound of the Fisher information for each learner given by assigned outside-group raters. Simulations and actual data experiments demonstrate that assigning a few optimal external raters using the proposed method can improve the peer assessment accuracy considerably.

It is noteworthy that many group formation methods have been proposed for improving the effectiveness of collaborative learning (e.g., [31], [32], [33], [34], [35], [36], [37], [38]). This study does not specifically examine learning effectiveness. However, groups that are formed to maximize the assessment accuracy are expected to be effective to improve learning because receiving accurate assessments generally promotes effective learning [19]. For that reason, group formation for improving peer assessment accuracy can be regarded as an important research effort in the field of educational technology.

2 Peer Assessment Data

This study uses a learning management system (LMS) called *SamurAI* [39] as a peer assessment platform.

The LMS SamurAI stores huge numbers of e-learning courses, where each course comprises 15 content sessions tailored for 90-min classes. Each class comprises instructional text screens, images, videos, practice tests, and report-writing tasks. To submit reports and conduct peer assessment, this system offers a discussion board system. Fig. 1 portrays a system interface by which a learner submits a report. The lower half of Fig. 1 presents a hyperlink 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 to assign ratings. The buttons represent -2 (Bad), -1 (Poor), 0 (Fair), 1 (Good), and 2 (Excellent). The system calculates the averaged rating score of each report and uses it to recommend excellent reports to other learners [40]. Other studies have used such scores for various purposes such as grading learners [41], [42], evaluating rater reliability [43], predicting learners' future performance [44], [45], and

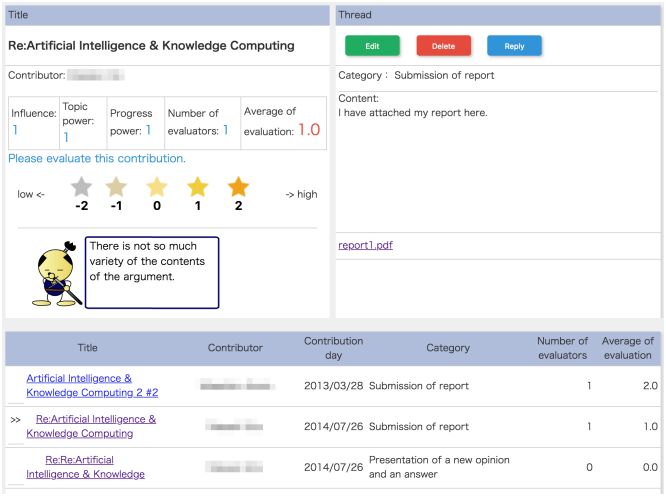


Fig. 1. Peer assessment system implemented in LMS *SamurAI*.

assigning weights to formative comments [13]. This article describes our attempts at improving the score accuracy.

The rating data $\boldsymbol{U}$ obtained from the peer assessment system consist of rating categories $k \in \mathcal{K} = \{1, \cdots, K\}$ given by each peer-rater $r \in \mathcal{J} = \{1, \cdots, J\}$ to each learning outcome of learner $j \in \mathcal{J}$ for each task $t \in \mathcal{T} = \{1, \cdots, T\}$. Letting u_{tjr} be a response of rater r to learner j's outcome for task t, the data $\boldsymbol{U}$ are described as

$$\boldsymbol{U} = \{u_{tjr} \mid u_{tjr} \in \mathcal{K} \cup \{\text{-1}\}, t \in \mathcal{T}, j \in \mathcal{J}, r \in \mathcal{J}\}, \quad (1)$$

where $u_{tjr} = -1$ denotes missing data. This study uses five categories $\{1, 2, 3, 4, 5\}$ transformed from the rating buttons $\{-2, -1, 0, 1, 2\}$ in the peer assessment platform above.

As described in Section 1, peer assessment is often conducted by dividing learners into multiple groups. This study assumes that peer assessment groups are created for each task $t \in \mathcal{T}$. Here, let x_{tgjr} be a dummy variable that takes the value of 1 if learner j and peer r are included in the same group $g \in \mathcal{G} = \{1, \cdots, G\}$ for assessment of task t, and which takes the value of 0 otherwise. Then, peer assessment groups for task t can be described as shown below.

$$\boldsymbol{X}_t = \{x_{tgjr} \mid x_{tgjr} \in \{0, 1\}, g \in \mathcal{G}, j \in \mathcal{J}, r \in \mathcal{J}\} \quad (2)$$

Consequently, when peer assessment is conducted among group members, the rating data u_{tjr} become missing data if learners j and r are not in the same group ($\sum_{g=1}^{G} x_{tgjr} = 0$).

This study is intended to assess the learner ability from the peer assessment data $\boldsymbol{U}$ accurately by optimizing the group formation $\boldsymbol{X} = \{\boldsymbol{X}_1, \cdots, \boldsymbol{X}_T\}$. For that purpose, we use item response theory.

3 Item Response Theory

Item response theory (IRT) [24], a test theory based on mathematical models, has been used widely for educational testing. Actually, IRT represents the probability that a learner responds to a test item as a function of the latent ability of the learner and item characteristics such as difficulty and discrimination. The use of IRT provides the following benefits. 1) A learner's responses to different test items

can be assessed on the same scale. 2) Missing data can be handled easily.

Many IRT models are applicable to ordered-categorical data such as peer assessment data. The representatives are the Rating Scale Model (RSM) [46], Partial Credit Model (PCM) [47], Generalized Partial Credit Model (GPCM) [48], and Graded Response Model (GRM) [49]. Although those traditional IRT models are applicable to two-way data consisting of learners $\times$ test items, they are inapplicable to the peer assessment data directly because they are three-way data comprising learners $\times$ raters $\times$ tasks, as defined in Section 2.

To resolve that difficulty, IRT models that incorporate rater parameters have been proposed (e.g., [1], [2], [23], [25], [26], [27], [28]). These models treat item parameters in traditional IRT models as task parameters. For example, an item difficulty parameter is regarded as a task difficulty parameter.

The following subsection introduces an IRT model for peer assessment [2], which is known to realize the highest ability assessment accuracy in the related models when the number of raters (= learners) increases.

3.1 Item response theory for peer assessment

The IRT model for peer assessment [2] has been formulated as a GRM that incorporates rater parameters. The model defines the probability that rater r responds in category k to learner j's outcome for task t as

$$P_{tjrk} = P^*_{tjrk-1} - P^*_{tjrk} \tag{3}$$
$$\begin{cases} P^*_{tjr0} = 1, \\ P^*_{tjrk} = \frac{1}{1+\exp(-\alpha_t\gamma_r(\theta_j-\beta_{tk}-\varepsilon_r))}, 1 < k < K-1 \\ P^*_{tjrK} = 0. \end{cases}$$

The following are used in those equations: γ_r reflects the consistency of rater r; ε_r represents the severity of rater r; α_t is a discrimination parameter of task t; and β_{tk} denotes the difficulty in obtaining category k for task t (with $\beta_{t1} < \cdots < \beta_{tK-1}$).

Fig. 2 presents examples of item response curves (IRCs) for three raters (designated as Rater 1, 2 and 3) having different characteristics. We can draw the IRCs for a rater r by plotting the probability P_{tjrk} with changing ability θ_j given parameter values of the rater and task t. In this example, the parameters for Rater 1 were $\gamma_r = 1.2$ and $\epsilon_r = 1.5$, those for Rater 2 were $\gamma_r = 1.2$ and $\epsilon_r = -1.5$, and those for Rater 3 were $\gamma_r = 0.8$ and $\epsilon_r = -1.5$, respectively. The task parameters were set as $\alpha_t = 1.0$, $\beta_{t1} = -1.5$, $\beta_{t2} = -0.5$, $\beta_{t3} = 0.5$, and $\beta_{t4} = 1.5$. The left panel of Fig. 2 portrays the IRCs of Rater 1. The central panel shows the IRCs of Rater 2. The right panel shows the IRCs of Rater 3. The horizontal axis shows the learner ability. The first vertical axis shows the response probability for each category.

This IRT model presents the severity of each rater as ϵ_r. As the parameter value increases, the IRCs shift to the right. For instance, Fig. 2 shows that the IRCs of Rater 1, who has high severity, shifted rightward compared to those of Rater 2. That tendency reflects that raters with higher severity tend to assign low scores consistently.

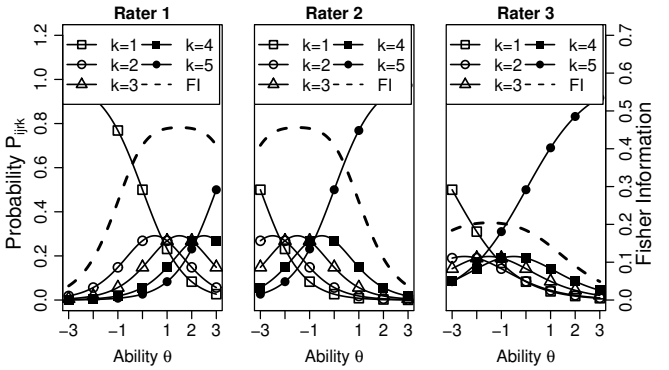


Fig. 2. Item response curves of the IRT model with rater parameters for three raters.

Furthermore, the model presents the consistency of each rater as γ_r. The lower the parameter value becomes, the smaller the differences in the response probabilities among the categories, as in the IRCs of Rater 3. Therefore, a rater with a lower consistency parameter has a stronger tendency to assign different scores to learners with the same ability level. Those raters generally engender low ability assessment accuracy because their scores do not necessarily reflect the true ability of a learner.

The interpretation of the task parameters is the same as that of the item parameters in GRM.

The IRT models with rater parameters are known to provide higher ability assessment accuracy than average or total rating scores do because they can estimate the ability considering the rater characteristics [50], [51], [52], [53]. Additionally, the IRT model introduced into this subsection is known to achieve the highest peer assessment accuracy in the related models when the number of raters increases [2]. This study assumes that group formation becomes increasingly necessary as the number of learners (=raters) increases. For that reason, this study uses this model.

The authors have examined the effectiveness of those IRT models by their application to actual peer assessment data collected using LMS SamurAI [1], [2]. However, the influence of the means of forming groups has been ignored. As described in Section 1, the assessment accuracy depends on a group formation when the peer assessment is conducted only among group members. This study improves the assessment accuracy by optimizing the group formation based on the IRT model.

3.2 Model identifiability

The IRT model above entails a non-identifiability problem, meaning that the parameter values cannot be determined uniquely because different sets of them provide the same response probability [54], [55]. Although the GRM parameters are identifiable by fixing the distribution of the ability [56], [57], this model still has indeterminacy of the scale for $\alpha_t\gamma_r$ and that of the location for $\beta_{tk} + \epsilon_r$, even if the ability distribution is fixed. Specifically, the response probability P_{tjrk} with α_t and γ_r engenders the same value of P_{tjrk} with $\alpha'_t = \alpha_t c$ and $\gamma'_r = \frac{\gamma_r}{c}$ for any constant c because $\alpha'_t\gamma'_r = (\alpha_t c)\frac{\gamma_r}{c} = \alpha_t\gamma_r$. Similarly, the response probability with β_{tk} and ϵ_r engenders the same value of P_{tjrk} with

$\beta'_{tk} = \beta_{tk} + c$ and $\epsilon'_r = \epsilon_r - c$ for any constant c because $\beta'_{tk} + \epsilon'_r = (\beta_{tk} + c) + (\epsilon_r - c) = \beta_{tk} + \epsilon_r$. The scale indeterminacy, as in the $\alpha_t\gamma_r$ case, is known to be removed by fixing one parameter or restricting the product of some parameters [56]. Furthermore, the location indeterminacy, as in the $\beta_{tk} + \epsilon_r$ case, is solvable by fixing one parameter or restricting the mean of some parameters [48], [55], [56]. This study uses the restrictions $\prod_{r=1}^{R} \gamma_r = 1$ and $\sum_{r=1}^{R} \varepsilon_r = 0$ for model identification.

It is noteworthy that, because no identification problem exists, restrictions on the rater parameters are not required when the task parameters are known and the distribution of the ability is fixed.

3.3 Model assumption

This model requires several assumptions. One important assumption is local independence, which is a common assumption in IRT (e.g., [55], [58], [59]). This assumption implies that ratings for a learner become locally independent among all raters and tasks given the ability of the learner. An earlier report described that local independence among raters is not satisfied when inter-rater agreement is high (e.g., [25], [60], [61]). When dependence among raters is assumed to be strong, IRT models that can consider their effects, such as the rater bundle model [61] and the hierarchical rater models [25], [62], might be appropriate.

Another assumption of this model is that no interaction occurs between raters and tasks. For example, if rater severity differs across tasks, then the assumption is not satisfied. In such a case, incorporating different rater severity parameters for tasks, such as introduced into [26], might be desirable.

Those assumptions are evaluated in the actual data experiment section.

3.4 Fisher information

In IRT, the standard error estimate of ability assessment is defined as the inverse square root of the Fisher information (FI). More information implies less error of the assessment. Therefore, FI can be regarded as an index of the ability assessment accuracy under the assumptions that the model is correct and that the ratings are a valid reflection of the targeted learning outcome.

In the IRT model for peer assessment [2], FI of rater r in task t for a learner with ability θ_j is calculable as

$$I_{tr}(\theta_j) = \alpha_t^2 \gamma_r^2 \sum_{k=1}^{K} \frac{\left(P^*_{tjrk-1}Q^*_{tjrk-1} - P^*_{tjrk}Q^*_{tjrk}\right)^2}{P^*_{tjrk-1} - P^*_{tjrk}}, \quad (4)$$

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

Fig. 2 depicts the FI function for the three example raters introduced into 3.1. The dotted lines and the right vertical axis show FI values. A comparison between Rater 1 and Rater 2, who have different severities with the same consistency, shows that the severe (or lenient) rater tends to give higher FI values for high (or low) ability levels. That tendency reflects the fact that severe (or lenient) raters do not distinguish low (or high) ability learners because their ratings for such learners are biased to the lowest (or highest) score. Fig. 2 also shows that FI of Rater 3, who has low consistency, is extremely low overall. That result reflects the fact that inconsistent raters engender low ability assessment accuracy because their ratings do not necessarily reflect the true ability, as described in 3.1.

The FI of multiple raters for learner j in task t is definable by the sum of the information of each rater under the local independence assumption. Therefore, when peer assessment is conducted within group members, the information for learner j in task t is calculable as shown below.

$$I_t(\theta_j) = \sum_{\substack{r=1 \\ r\neq j}}^{J} \sum_{g=1}^{G} I_{tr}(\theta_j) x_{tgjr} \quad (5)$$

A high value of FI $I_t(\theta_j)$ signifies that the group members can assess learner j accurately. Therefore, if we form groups to provide great amounts of FI for each learner, then the ability assessment accuracy can be maximized. Based on this idea, the next section presents a proposal of a group formation method to maximize the peer assessment accuracy.

4 Group formation using item response theory and integer programming

4.1 Group formation method

We formulate the group formation optimization method as an integer programming problem that maximizes the lower bound of FI for each learner. Hereinafter, this method is designated as *PropG*. Specifically, *PropG* for task t is formulated as the following integer programming problem.

$$\text{maximize} \quad y_t \quad (6)$$

$$\text{subject to} \quad \sum_{\substack{r=1 \\ r\neq j}}^{J} \sum_{g=1}^{G} I_{tr}(\theta_j) x_{tgjr} \geq y_t, \qquad \forall j, \quad (7)$$

$$\sum_{g=1}^{G} x_{tgjj} = 1, \qquad \forall j, \quad (8)$$

$$n_l \leq \sum_{j=1}^{J} x_{tgjj} \leq n_u, \qquad \forall g, \quad (9)$$

$$x_{tgjr} = x_{tgrj}, \qquad \forall g, j, r, \quad (10)$$

$$x_{tgjr} \in \{0, 1\}, \qquad \forall g, j, r. \quad (11)$$

The first constraint requires that FI for each learner j be larger than a lower bound y_t. The second constraint restricts each learner as belonging to one group. The third constraint controls 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 so that the numbers of learners in respective groups become as equal as possible. Here, $\lfloor\ \rfloor$ and $\lceil\ \rceil$ respectively denote floor and ceiling functions. If the remainder of J/G equals to zero, then the numbers of group members become equal for all groups; otherwise, they differ among groups. In the latter case, the difference in numbers between groups is equal to or less than one. This integer programming maximizes the lower bound of FI for learners. Therefore, by solving the problem, one can obtain groups that provide as much FI as possible to each learner.

TABLE 1
Prior distributions for IRT model parameters

$$\log \alpha_t, \log \gamma_r \sim N(0.0, 0.4)$$
$$\epsilon_r, \theta_j \sim N(0.0, 1.0)$$
$$\beta_{tk} \sim MN(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \begin{cases} \boldsymbol{\mu} = \{-2.0, -0.75, 0.75, 2.0\} \\ \boldsymbol{\Sigma} = \begin{vmatrix} 0.16 & 0.10 & 0.04 & 0.04 \\ 0.10 & 0.16 & 0.10 & 0.04 \\ 0.04 & 0.10 & 0.16 & 0.10 \\ 0.04 & 0.04 & 0.10 & 0.16 \end{vmatrix} \end{cases}$$

As another approach, it might be possible to make the peer assessment completely adaptive so that raters with the highest FI are sequentially assigned to each learner. However, just as the traditional adaptive testing with an insufficiently large or diverse item bank does (e.g., [63], [64], [65], [66]), this approach increases the assessment errors as the process proceeds because the number of learners assignable to each rater is limited. Consequently, this approach tends to pose biased assessment accuracies for learners. However, *PropG* resolves this difficulty because the assignment is optimized to maximize the lower bound of FI for learners.

PropG is inspired by automated uniform test assembly methods using integer programming and IRT, which have been studied extensively in educational testing fields (e.g., [67], [68], [69], [70], [71]).

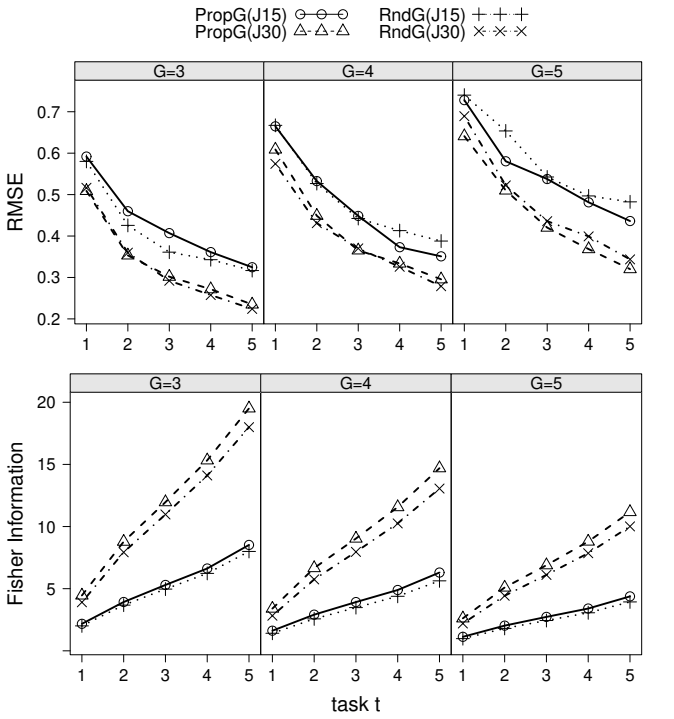


Fig. 3. RMSE and FI values of group formation methods in the simulation experiment.

4.2 Evaluation of group formation methods

The ability assessment accuracy is expected to be improved considerably if *PropG* can form groups to give sufficiently high FI to each learner. To evaluate this point, we conducted the following simulation experiment.

1) For $J \in \{15, 30\}$ and $T = 5$, the true IRT model parameters were generated randomly from the distributions presented in Table 1. The values of J and T were chosen to match the conditions of two actual e-learning courses offered by one author from 2007 to 2013 using LMS SamurAI. Specifically, $J = 15$ and 30 were used because the average numbers of learners in each course were 12.9 (standard deviation=4.2) and 32.9 (standard deviation=14.6), respectively. Also, $T = 5$ was used because the maximum number of tasks was 5. Furthermore, the parameter distributions in Table 1 assume correlation of β_{tk} among categories because an increase of β_{tk} tends to increase $\beta_{t,k+1}$ as a result of the order restriction $\beta_{t,k+1} > \beta_{tk}$.
2) For the first task $t = 1$, learners were divided into $G \in \{3, 4, 5\}$ groups using *PropG* and a random group formation method (designated as *RndG*). For *PropG*, the FI values were calculated using the true parameter values. The number of groups is usually determined so that each group comprises 3–14 members while maintaining the number as equal as possible for all groups [32], [72], [73], [74]. This experiment used $G = 3, 4$, and 5 because the number of group members falls within this range when $J \in \{15, 30\}$. Here, *PropG* was solved using *IBM ILOG CPLEX Optimization Studio* [75]. We used a feasible solution when the optimal solution was not obtained within 10 min.
3) Given the created groups and the true model parameters, peer assessment data were sampled randomly for the current task t based on the IRT model.
4) Given the true rater and task parameters, the learner ability was estimated from the data generated to date. Here, the expected a posteriori (EAP) estimation using Gaussian quadrature [76] was used for the estimation.
5) The root mean square error (RMSE), and average bias between the estimated ability and the true ability were calculated. We also calculated the FI given to each learner.
6) Procedures 2) – 5) were repeated for the remaining tasks.
7) After 10 repetitions of the procedures described above, the average values of RMSE, average bias, and FI obtained from Procedure 5) were calculated. In this experiment, $PropG$ provided the optimal solutions within 10 min for 98% of the group formations when $J = 15$, and for 78% of them when $J = 30$.

Fig. 3 presents RMSE and FI results. The horizontal axis shows the task index; the vertical axis shows the RMSE (upper panels) and FI (lower panels). The lines represent the results of *PropG* and *RndG* for each number of learners. Results demonstrate that FI increases and RMSE decreases with the decreasing number of groups G or with increasing numbers of tasks or learners because the number of data for each learner increases. Generally, the increase of data per learner is known to engender improvement of the

TABLE 2
Distributions of rater parameters for each ability level in group formation methods

	RndG							
	$\log\gamma_r$				ϵ_r			
θ	≤ -0.4	$(-0.4, 0]$	$(0, 0.4]$	> 0.4	≤ -1	$(-1, 0]$	$(0, 1]$	> 1
≤ -1	0.17	0.31	0.32	0.20	0.14	0.33	0.38	0.14
$(-1, 0]$	0.15	0.32	0.34	0.18	0.16	0.33	0.36	0.15
$(0, 1]$	0.16	0.33	0.33	0.18	0.16	0.30	0.38	0.16
> 1	0.14	0.32	0.34	0.19	0.15	0.35	0.36	0.14
	PropG							
	$\log\gamma_r$				ϵ_r			
θ	≤ -0.4	$(-0.4, 0]$	$(0, 0.4]$	> 0.4	≤ -1	$(-1, 0]$	$(0, 1]$	> 1
≤ -1	0.19	0.34	0.29	0.18	0.30	0.39	0.26	0.05
$(-1, 0]$	0.14	0.31	0.34	0.21	0.16	0.34	0.39	0.11
$(0, 1]$	0.13	0.32	0.37	0.17	0.11	0.33	0.37	0.20
> 1	0.19	0.33	0.30	0.17	0.04	0.23	0.48	0.25

ability assessment accuracy [2]. Furthermore, we confirmed that the average biases were extremely close to zero in all cases. Specifically, the minimum value was -0.08 and the maximum value was 0.02, which indicates that there was no overestimation or underestimation of the ability.

Comparing the group formation methods, *PropG* presents higher FI than *RndG* in all cases. To examine the reason, we analyzed the relation between learner ability and the assigned rater parameters. For this analysis, we divided the values of the ability and the rater parameters into four levels $\leq -\sigma$, $(-\sigma, 0]$, $(0, \sigma]$, and $> \sigma$, where $\sigma = 0.4$ for $\log\gamma_r$ and $\sigma = 1$ for θ_j and ϵ_r. Subsequently, we calculated the proportion that raters with each parameter level were assigned to learners with each ability level. Table 2 presents the results. Results show that the distributions of the rater severity parameter differ between the group formation methods, although those of the rater consistency parameter are mutually similar. Specifically, *PropG* tends to assign severe raters to high-ability learners and lenient raters to low-ability learners. As explained in 3.4, severe (or lenient) raters tend to provide higher FI to high (or lower) ability level. For these reasons, *PropG* presents higher FI than *RndG* does.

Fig. 3, however, shows that *PropG* does not decrease RMSE sufficiently because it does not improve FI much. To improve FI dynamically, the proportion of high consistent raters for each learner should be increased because those raters tend to give high FI overall. However, the experimentally obtained results indicate that it is difficult to form groups to increase the proportion.

As described in the experimental procedure 7), we repeated the simulation procedures 10 times for each setting. To examine effects of the number of repetitions, we conducted the same experiment for 5 and 20 repetitions given $G = 5$. Fig. 4 shows the RMSE for each repetition. According to Fig. 4, when the repetition count is 5, $RndG$ for $J = 30$ provides the higher RMSE than $RndG$ for $J = 15$ in $t = 1$ although the amount of rating data for $J = 30$ is larger than that for $J = 15$, which suggests that few repetitions, such as 5 times, might produce unstable results. In addition, 10 and 20 repetitions presented the same tendencies discussed in this subsection. Because the experiments conducted in this study require high computational cost and time, we set the number of repetitions to 10.

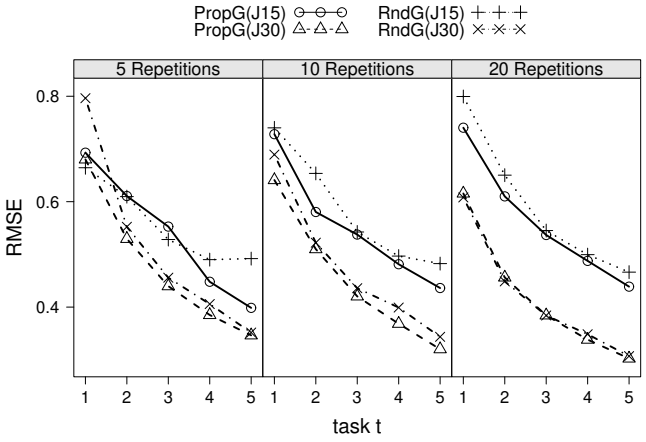


Fig. 4. RMSE values of group formation methods for each number of repetitions.

Although this experiment used the true IRT parameter values to calculate FI in *PropG*, these values are practically unknown. Use of *PropG* when the parameters are unknown is proposed in Section 6.

5 External Rater Assignment

The preceding section explained the difficulty of assigning raters with high FI to all learners when peer assessment is conducted only within groups. To overcome this shortcoming, this study further proposes the assignment of outside-group raters to each learner, given the groups created using *PropG*.

The proposed external rater assignment method is formulated as an integer programming problem that maximizes the lower bound of information for learners given by the assigned outside-group raters. Specifically, given a group formation $\boldsymbol{X}_t$, the proposed method for task t is defined as shown below.

$$\text{maximize :} \quad y'_t \tag{12}$$

$$\text{subject to :} \quad \sum_{r \in \boldsymbol{C}_{tj}} I_{tr}(\theta_j) z_{tjr} \geq y'_t, \qquad \forall j \tag{13}$$

$$\sum_{r \in \boldsymbol{C}_{tj}} z_{tjr} = n^e, \qquad \forall j \tag{14}$$

$$\sum_{j=1}^{J} z_{tjr} \leq n^J, \qquad \forall r \tag{15}$$

$$z_{tjj} = 0, \qquad \forall j \tag{16}$$

$$z_{tjr} \in \{0, 1\}, \qquad \forall j, r \tag{17}$$

Here, $\boldsymbol{C}_{tj} = \{r \mid \sum_{g=1}^{G} x_{tgjr} = 0\}$ is the set of outside-group raters for learner j in task t given a group formation $\boldsymbol{X}_t$. In addition, z_{tjr} is a variable that takes 1 if external rater r is assigned to learner j in task t; it takes 0 otherwise. Furthermore, n^e denotes the number of external raters assigned to each learner; n^J is the upper limit number of outside-group learners assignable to each rater. Here, n^e and n^J must satisfy $n^J \geq n^e$. The increase of n^J makes it easier to assign optimal raters to each learner, although differences in the assessment workload among the learners increases.

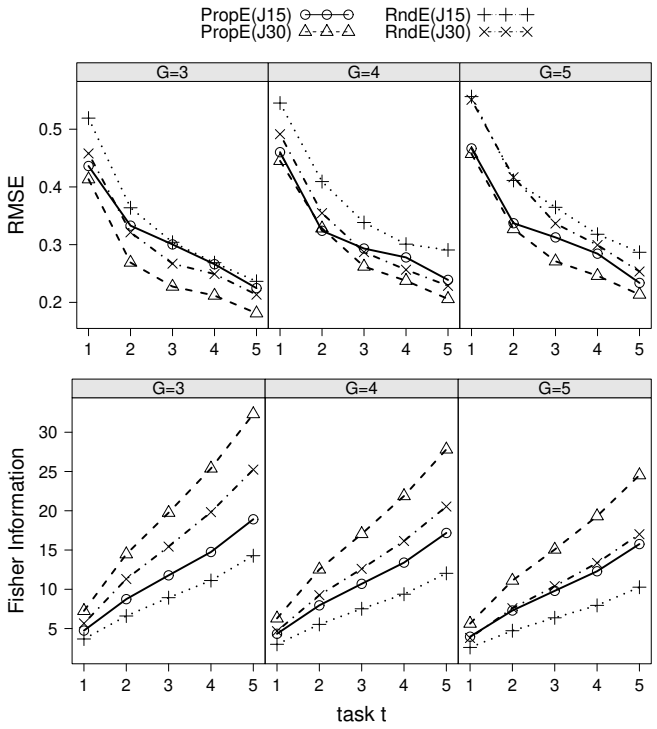


Fig. 5. RMSE and FI values of external rater assignment methods for each G and t in the simulation experiment.

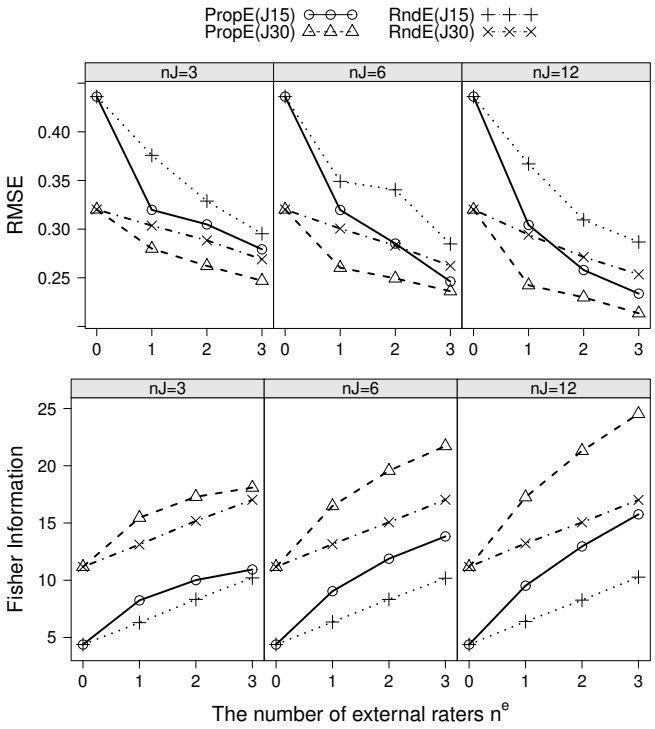


Fig. 6. RMSE and FI values of external rater assignment methods for each n^J and n^e in the simulation experiment.

In the integer programming problem, the first constraint restricts that the FI for each learner given by the assigned outside-group raters must exceed a lower bound y'_t. The second constraint requires that n^e number of outside-group raters must be assigned to each learner. The third constraint restricts that each learner can assess at most n^J number of outside-group learners. The objective function is defined as the maximization of the lower bound of the information for learners given by assigned external raters. Therefore, by solving the proposed method, an external rater assignment z_{tjr} is obtainable so that n^e outside-group raters with high FI are assigned to each learner.

5.1 Simulation experiment of external rater assignment method

Using the proposed method, each learner can be assessed not only by the group members but also by optimal outside-group raters. Therefore, ability assessment accuracy is expected to be improved considerably. To confirm that capability, we conducted the following simulation experiment, which is similar to that conducted in 4.2.

1) For $J \in \{15, 30\}$ and $T = 5$, the true model parameters were generated randomly from the distributions in Table 1.
2) For the first task $t = 1$, learners were divided into $G \in \{3, 4, 5\}$ groups using *PropG*. Then, given the created groups, $n^e \in \{1, 2, 3\}$ outside-group raters were assigned to each learner using the proposed external rater assignment method (designated as *PropE*) and a random assignment method (designated as *RndE*). Here, we changed the value of n^J for $\{3, 6, 12\}$ to evaluate its effects. In *PropG* and *PropE*, FI was calculated using the true parameter values. In *PropG*, we used a feasible solution when the optimal solution was not obtained within 10 min. *PropE* provided the optimal solutions within 10 min for all settings.
3) Peer assessment data were sampled randomly for current task t following the IRT model, given the true model parameters, the formed groups and the rater assignment.
4) The following procedures were identical to procedures 4) – 7) of the previous experiment.

We first examine the respective effects of the numbers of tasks, groups and learners on performance of the external rater assignment methods. Fig. 5 shows the RMSE and FI for each t, G and J when $n^J = 12$ and $n^e = 3$. Results show that the accuracy of the external rater assignment methods tends to increase concomitantly with decreasing number of groups and increasing number of tasks or learners because the number of rating data for each learner increases. This tendency is consistent with that of the group formation methods, as explained in 4.2.

Additionally, to analyze effects of n^e and n^J, Fig. 6 shows the RMSE and FI for each n^e and n^J when $G = 5$ and $t = 5$. The horizontal axis shows the values of n^e: the vertical axis and each line are the same as in Fig. 5. Here, the results for $n^e = 0$ indicate those of *PropG*. According to the results, both external rater assignment methods reveal higher FI and the lower RMSE than *PropG* in all cases, which suggests that the addition of the external raters is effective to improve the ability assessment accuracy. Furthermore, Fig.

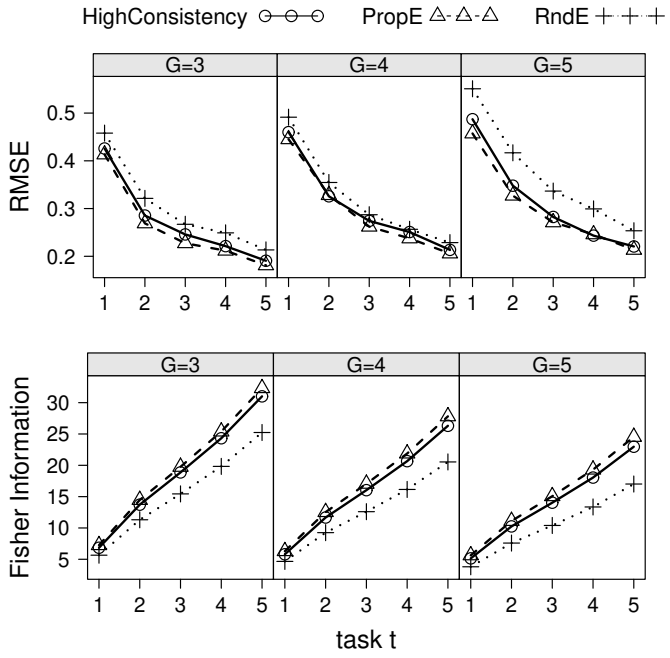


Fig. 7. RMSE and FI values of a rater-consistency-based external rater assignment method for each G and t in the simulation experiment with $J = 30$.

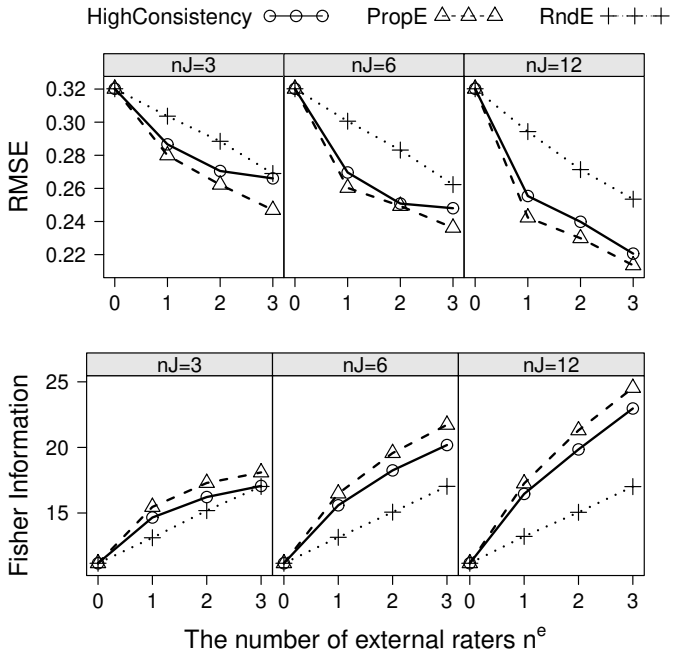


Fig. 8. RMSE and FI values of a rater-consistency-based external rater assignment method for each n^J and n^e in the simulation experiment with $J = 30$.

TABLE 3
Distributions of rater parameters for each ability level in external rater assignment methods.

	RndE							
	$\log \gamma_r$				ϵ_r			
θ	≤ -0.4	$(-0.4, 0]$	$(0, 0.4]$	> 0.4	≤ -1	$(-1, 0]$	$(0, 1]$	> 1
≤ -1	0.19	0.34	0.29	0.18	0.30	0.39	0.26	0.05
$(-1, 0]$	0.14	0.31	0.34	0.21	0.16	0.34	0.39	0.11
$(0, 1]$	0.13	0.32	0.37	0.17	0.11	0.33	0.37	0.20
> 1	0.19	0.33	0.30	0.17	0.04	0.23	0.48	0.25
	PropE							
	$\log \gamma_r$				ϵ_r			
θ	≤ -0.4	$(-0.4, 0]$	$(0, 0.4]$	> 0.4	≤ -1	$(-1, 0]$	$(0, 1]$	> 1
≤ -1	0.12	0.21	0.30	0.37	0.29	0.40	0.28	0.03
$(-1, 0]$	0.08	0.19	0.30	0.42	0.17	0.32	0.42	0.09
$(0, 1]$	0.08	0.20	0.33	0.39	0.11	0.33	0.39	0.17
> 1	0.12	0.21	0.31	0.36	0.03	0.21	0.50	0.26

6 shows that FI of the external rater assignment methods increase monotonically with increasing number of assigned external raters n^e. Also, RMSE tends to decrease as n^e increases.

The average biases were close to zero for all settings. Concretely, the minimum value was -0.07; the maximum value was 0.06, which means that there was no systematic overestimation or underestimation of ability.

Comparison of the external rater assignment methods reveals that the proposed method presented higher FI than the random assignment method in all cases. To examine the reason, we analyzed the relation between learner ability and the assigned rater parameters using the same procedures in 4.2. Table 3 presents results for $n^e = 3$ and $n^J = 12$. Results show that *PropE* reveals a higher proportion of consistent raters than *RndE* does. Because consistent raters generally give substantially high FI, *PropE* can improve FI dynamically. Consequently, the RMSEs of *PropE* are lower than those of $RndE$ in all cases. Furthermore, Fig. 6 shows that the performance of *PropE* tends to become better as increasing n^J. It reflects the fact that the increase of n^J facilitates better rater assignment.

The differences in FI between *PropE* and *RndE* are small when $n^J = 3$ and $n^e = 3$. As n^J decreases and/or n^e increases, assigning optimal raters becomes difficult even if the proposed method is used because the number of assignable raters for each learner decreases. Particularly, $n^J = n^e$ is the most difficult situation to assign optimal raters because all learners must be assigned to n^e number of outside-group learners even if some of them have extremely low FI. For that reason, the proposed method does not improve FI much when $n^J = 3$ and $n^e = 3$.

From those results, we infer that the proposed external rater assignment method can improve the peer assessment accuracy efficiently when a large value of n^J and a small value of n^e are given.

It is noteworthy that, from Table 3 and the discussion presented above, assigning external raters with high consistency might provide higher performance. The proposed method can be changed easily to assign the N most consistent raters for all learners by replacing FI function $I_{tr}(\theta_j)$ in Eq. (13) to the consistency parameter γ_r. To compare the performance of this method with the proposed method, we conducted the same experiment as that conducted in this subsection using the N most consistent raters assignment method for $J = 30$. Fig. 7 and 8 show the results. In the figures, the plots of *HighConsistency* portray the results of the N most consistent raters assignment method; the other plots are the same as those in Fig. 5 and 6. The N most

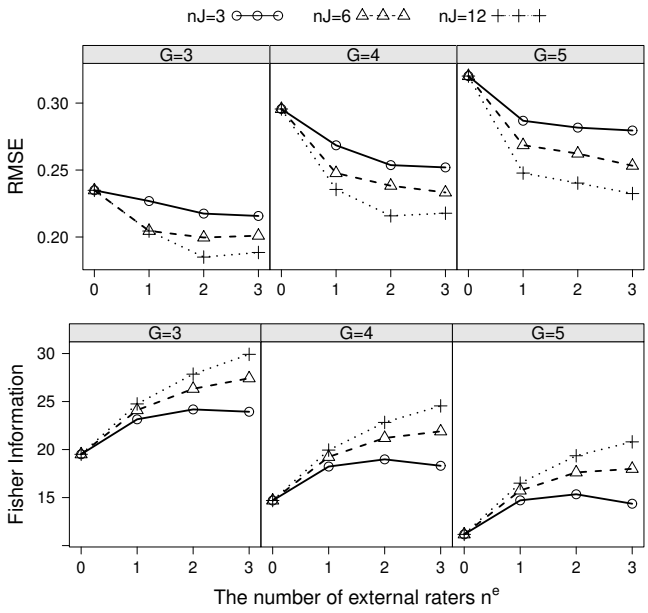


Fig. 9. RMSE and FI values of $PropExRm$.

consistent raters assignment method shows higher FI and smaller RMSE compared with *RndE*. However, comparing the N most consistent raters assignment method with *PropE*, it reveals lower FI and higher RMSE in all cases. The reason is that *PropE* directly maximizes FI for learners; it then achieves higher accuracy of learner ability estimations than the N most consistent raters assignment method does.

5.2 Effectiveness of external rater introduction

In the experiment described above, we demonstrated that the proposed external rater assignment method provided higher ability assessment accuracy than *PropG* did. The major reasons of the improvement are the increase of assigned raters and the introduction of optimal external raters. Although the experiment described earlier demonstrated the effectiveness of increasing raters, the effects of introducing optimal external rater were not examined directly. Therefore, this subsection explains evaluation of those effects using a simulation experiment.

For this evaluation, we introduce another external rater assignment method that assigns optimal outside-group raters without increasing the total number of raters for each learner. Specifically, the method first assigns n^e external raters by the proposed external rater assignment method. Then n^e internal-group members with the lowest FI were removed. Hereinafter, we designate the method as *PropExRm*. If *PropExRm* outperforms *PropG*, then the effectiveness of the optimal external rater introduction can be confirmed.

To compare the accuracy, we conducted the same simulation experiment as in 5.1 using *PropExRm* as the external rater assignment method for $J = 30$. Fig. 9 presents results for $t = 5$. The horizontal axis shows the number of n^e; the vertical axis shows the RMSE and FI values. Each line represents the result for each n^J. Results show that *PropExRm* reveals higher FI and the lower RMSE than *PropG* ($n^e = 0$) in all cases, although the number of raters for each learner is not increased. The results demonstrate that the introduction of optimal external raters is effective to improve the peer assessment accuracy.

FI does not increase monotonically with increasing n^e when $n^J = 3$, unlike in earlier experiments. *PropExRm* can remove internal-group raters who have higher FI than the added external raters have. Therefore, the possibility of removing internal raters with high FI increases as n^e increases. Additionally, assigning external raters with high FI becomes difficult as n^e increases and/or n^J decreases because assignable raters are reduced, as discussed before. Therefore, FI of $n^e = 3$ is less than that of $n^e = 2$ when $n^J = 3$.

6 Proposed Method with Parameter Estimation and Evaluation

6.1 Method

PropG and *PropE* require estimated IRT model parameter values to calculate FI. Although the experiments described above used the true parameter values for the calculation, they are practically unknown. Therefore, this section presents a description of how to use *PropG* and *PropE* when the IRT parameters are unknown in actual e-learning situations.

We consider the following two assumptions for using *PropG* and *PropE* in an e-learning course.

1) More than one task is offered in the course.
2) All tasks were used in past e-learning courses at least once. Past learners' peer assessment data corresponding to the tasks were collected.

Although the second assumption might not necessarily be satisfied in practice, it is necessary to estimate the task parameters. LMS SamurAI stores peer assessment data corresponding to all the tasks offered in past courses [2]. In such cases, the task parameters can be estimated from the data.

Given task parameter estimates, we can use *PropG* and *PropE* through the following procedures under the first assumption.

1) For the first task, peer assessment is conducted using randomly formed groups.
2) The rater parameters and learner ability are estimated from the obtained peer assessment data.
3) For the next task, group formation and external rater assignment are conducted using *PropG* and *PropE* given the parameter estimates.
4) Repeat procedures 2) and 3) for remaining tasks.

As described in 3.2, when the ability distribution is fixed, the restrictions on the rater parameters for model identification are not required in the parameter estimation of Procedure 2) because the task parameters are given.

6.2 Simulation experiments

To evaluate *PropG* and *PropE* with parameter estimation, the following simulation experiment was conducted.

1) For $J \in \{15, 30\}$ and $T = 5$, true model parameters were generated randomly following the distributions in Table 1.

2) For the first task $t = 1$, $G \in \{3, 4, 5\}$ groups were created randomly.
3) Given the formed groups and true parameters, rating data for task $t = 1$ were sampled randomly.
4) From the generated data, the rater parameters and learner abilities were estimated using the Markov chain Monte Carlo (MCMC) algorithm [2]. In the estimation, the true task parameters were given.
5) The RMSE between the estimated ability and the true ability were calculated. We also calculated FI for each learner.
6) For the next task, $G \in \{3, 4, 5\}$ groups were formed by *PropG* and *RndG*. Furthermore, given the groups formed by *PropG*, $n^e \in \{1, 2, 3\}$ external raters were assigned to learners by *PropE* and *RndE* under $n^J \in \{3, 6, 12\}$. Here, *PropG* and *PropE* used the true task parameters obtained in Procedure 1) and the current estimates of ability and rater parameters to calculate FI.
7) Given the formed groups and rater assignment, peer assessment data for the current task were sampled randomly. Rating data were sampled from the IRT model given the true parameter values obtained in procedure 1).
8) Given the true task parameters, the learner ability and rater parameters were estimated from the data up to the current task.
9) The RMSE and FI were calculated using the same procedure as that used for 5).
10) For the remaining tasks, Procedures 6) – 9) were repeated.
11) After repeating the procedures described above 10 times, the average values of the RMSE and FI were calculated.

Fig. 10 presents results obtained using the respective group formation methods. Figs. 11 and 12 present results obtained using the external rater assignment methods. Here, Fig. 11 presents results for each $t \geq 2$ and G when $n^J = 12$ and $n^e = 3$. Also, Fig. 12 shows those for each n^e and n^J when $G = 5$ and $t = 5$. According to the results, we can confirm a similar tendency with the results of the previous simulation experiments in all cases. Specifically, the following tendency can be confirmed.

1) *PropG* does necessarily not outperform *RndG*.
2) Both the external rater assignment methods present higher accuracy than that provided by *PropG*.
3) *PropE* can improve the assessment accuracy more efficiently than *RndE* when a large value of n^J and a small value of n^e are given.

Results show that *PropG* and *PropE* with parameter estimation work appropriately.

7 Actual Data Experiment

This section evaluates the effectiveness of *PropG* and *PropE* using actual peer assessment data.

7.1 Actual data

Actual data were gathered using the following procedures.

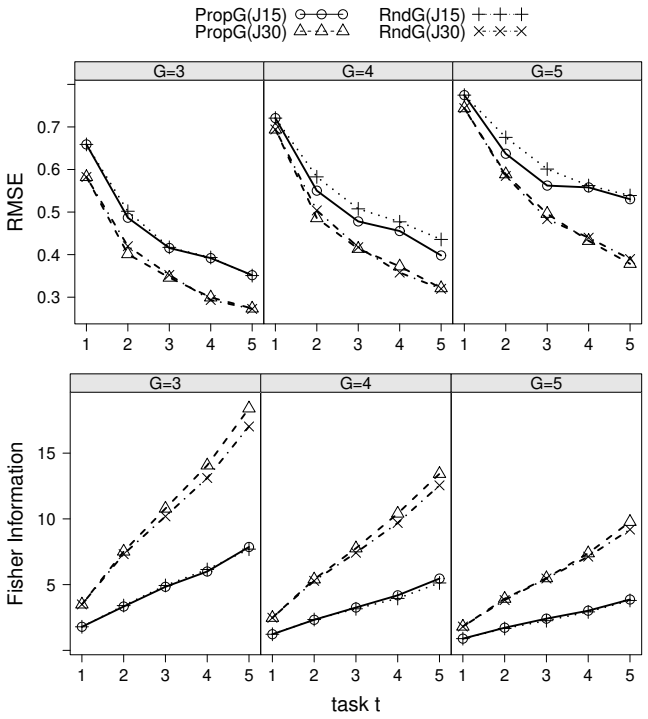


Fig. 10. RMSE and FI values of group formation methods in simulation experiment with parameter estimation.

1) As subjects for this study, 34 university students were recruited. All were majoring in various science fields such as statistics, materials, chemistry, mechanics, robotics, and information science. They included 19 undergraduate, 13 master course, and 2 doctor course students.
2) They were asked to complete four essay writing tasks offered in the National Assessment of Educational Progress (NAEP) [77] and 2007 [78]. No specific or preliminary knowledge was needed to complete the tasks.
3) After the participants completed all tasks, they were asked to evaluate the essays of all other participants for all four tasks. Assessments were conducted using a rubric that we created based on the assessment criteria for grade 12 NAEP writing [78]. The rubric consists of five rating categories with corresponding scoring criteria.

Furthermore, we collected additional rating data for task parameter estimation. The data consist of ratings assigned by 5 graduate school students to the essays gathered in the experiment above. Hereinafter, the data are designated as *five raters' data*.

Ability estimation using the peer assessment data might be biased because the given task parameters estimated from the five raters' data would not fit well if characteristics of the peer assessment data and the five raters' data were to differ extremely. Therefore, it is desirable that characteristics of the two datasets be similar. To evaluate the similarity, we compare descriptive statistics for the two datasets. Table 4 shows the average and standard deviation of ratings and the appearance rate of each rating category in each dataset.

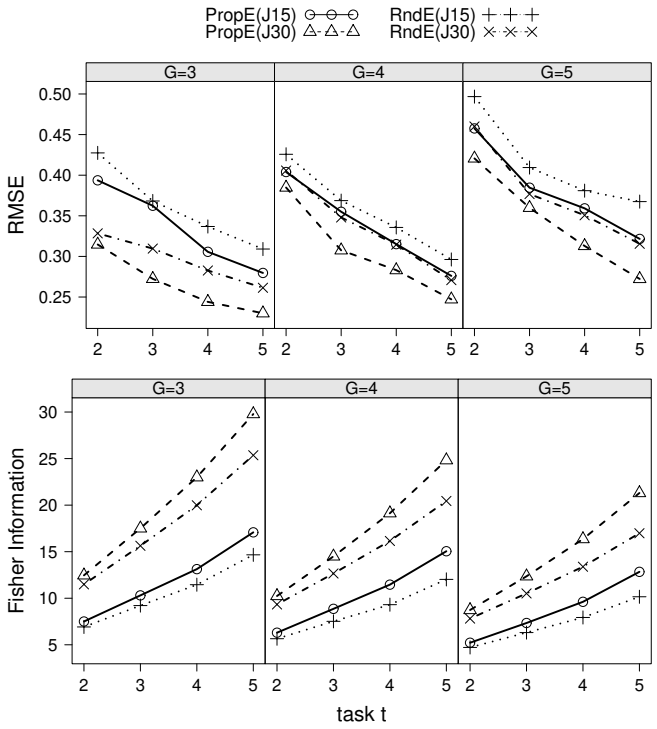


Fig. 11. RMSE and FI values of external rater assignment methods for each G and t in simulation experiment with parameter estimation.

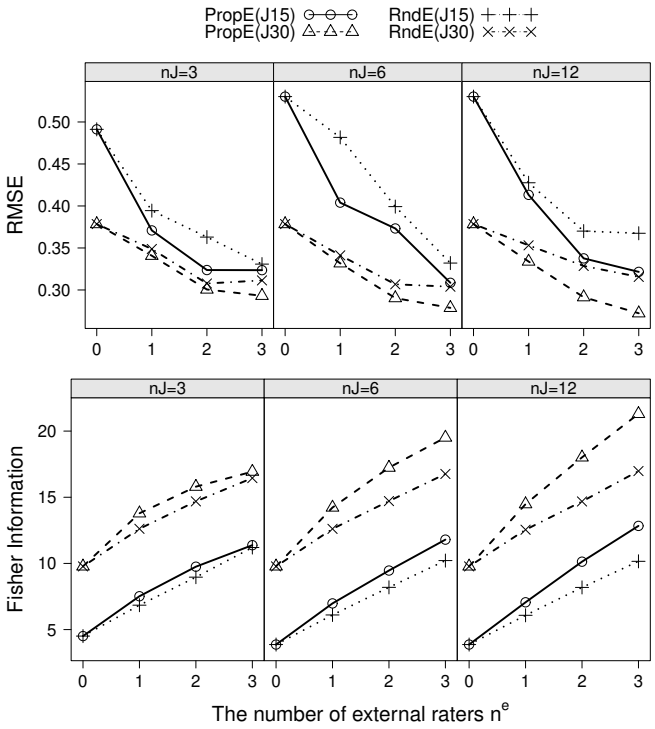


Fig. 12. RMSE and FI values of external rater assignment methods for each n^J and n^e in a simulation experiment with parameter estimation.

TABLE 4
Descriptive statistics for each actual dataset

Data	Avg.	SD	Appearance rate of each category				
			1	2	3	4	5
Peer assessment	2.21	1.01	4.50	19.49	36.83	29.35	9.84
Five raters'	2.04	1.01	5.29	25.59	36.62	24.85	7.65

Furthermore, we calculated the correlation of the average scores for each learner using the peer assessment data and the five raters' data. Results show that the correlation value was 0.69; it was significantly correlated at the 0.001 level. The results suggest that the characteristics of the two datasets are similar.

7.2 Evaluation of model fitting

As discussed in 3.3, the IRT model in Eq. (3) includes the assumption of local independence. Therefore, we examined this assumption using the $Q3$ statistics [79], which is a well known method for empirically examining local dependence. Here, let E_{tjr} be the residual between the observed rating u_{tjr} and the expected rating $\sum_{k=1}^{K} k \cdot P_{tjrk}$. Then, the $Q3$ statistics for two task-rater pairs, (t, r) and (t', r'), are defined as the Pearson correlation coefficient between the residuals, $\boldsymbol{E}_{tr}$ and $\boldsymbol{E}_{t'r'}$ (where $\boldsymbol{E}_{tr} = \{E_{t1r} \cdots, E_{tJr}\}$). A high correlation value signifies that the task-rater pairs are locally dependent. Therefore, we calculated this index for all task-rater pairs and tested the significance using Student's t-test with significance inferred at the 0.05 level.

Results demonstrate that 93% of the pairs had no significant correlation in both datasets. The results suggest that the local independence assumption is satisfied in almost all cases. Furthermore, to examine the rater dependencies, we analyzed the results among raters in the same task. Consequently, 97% of the rater pairs revealed no significant correlation in both datasets. That amount indicates that the rater dependencies are negligibly small in the datasets.

Additionally, we examined another model assumption: that no interaction exists between tasks and raters. We used generalizability theory [80] to test the assumption. Generalizability theory can estimate the effects of the error sources (such as learners, raters, and tasks) and their mutual interactions on ratings using analysis of variance. It gives high variance estimates to the sources and interactions when observed ratings depend strongly on them. In the peer assessment data, the variance estimate of the task-rater interaction accounted only for 2% of the total variance. Furthermore, it was 3% in the five raters' data. The results suggest that the effect of the interaction is negligible.

From the analysis described above, we confirmed that the assumptions of the IRT model were approximately satisfied. This fact validates the use of the model in this experiment.

7.3 Experimental procedures and results

Using the actual data, we conducted the following experiments, which are similar to those in 6.2.

1) The task parameters in the IRT model were estimated using the five raters' data.
2) Given the task parameter estimates, the rater parameters and learner ability were estimated using the full peer assessment data.

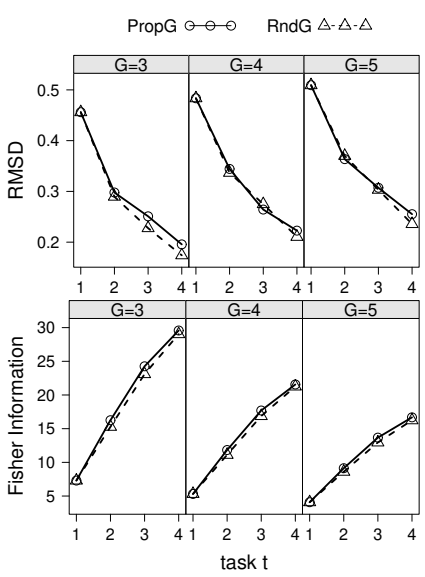


Fig. 13. RMSE and FI values of group formation methods in the actual data experiment.

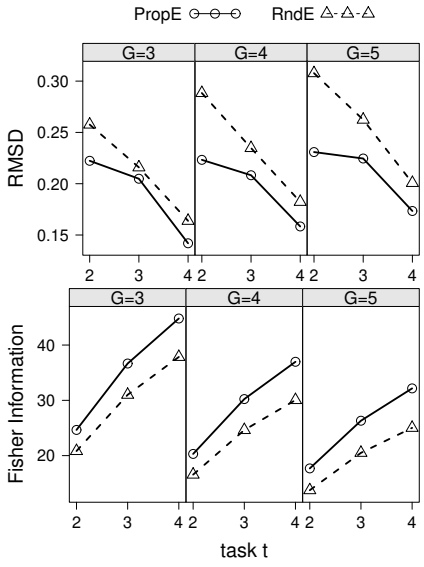


Fig. 14. RMSE and FI values of external rater assignment methods for each G and t in the actual data experiment.

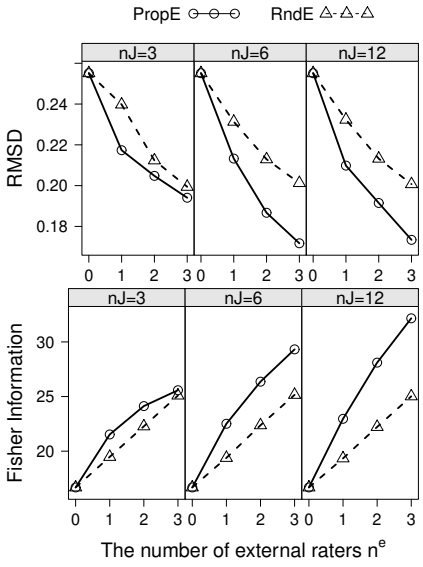


Fig. 15. RMSE and FI values of external rater assignment methods for each n^J and n^e in the actual data experiment.

3) For the first task, $G \in \{3, 4, 5\}$ groups were created randomly.
4) The peer assessment data u_{1jr} were changed to missing data if learner r and learner j were not in the same group.
5) From the peer assessment data for the first task, the rater parameters and learner ability were estimated given the task parameters estimated in Procedure 1).
6) The Root Mean Square Deviation (RMSD) between the ability estimates and that estimated from the complete data in Procedure 2) was calculated. We also calculated FI for each learner.
7) For the next task, $G \in \{3, 4, 5\}$ groups were formed by *PropG* and *RndG*. Then, given the groups formed by *PropG*, $n^e \in \{1, 2, 3\}$ external raters were assigned to learners by *PropE* and *RndE* under $n^J \in \{3, 6, 12\}$. Here, *PropG* and *PropE* used the task parameters obtained in Procedure 1) and the current estimates of ability and rater parameters to calculate FI.
8) Given the group formations and external rater assignments, the peer assessment data u_{tjr} were changed to missing data if learner j and r are not in the same group and if learner r is not the external rater of learner j.
9) Given the task parameter estimates, the learner ability and rater parameters were estimated from the peer assessment data up to the current task.
10) The RMSD and FI were calculated using the same procedure as 6).
11) For the remaining tasks, procedures 7) – 10) were repeated.
12) After repeating the procedures described above 10 times, the average values of the RMSD and FI were calculated.

Fig. 13 presents results of each group formation method. Figs. 14 and 15 show those of the external rater assignment methods. Fig. 14 presents results for each $t \geq 2$ and $G \in \{3, 4, 5\}$ when $n^J = 12$ and $n^e = 3$. Fig. 15 shows those for each n^e and n^J when $G = 5$ and $t = 4$. Results show similar tendencies to those obtained from the simulation experiments. Specifically, comparing the group formation methods, *PropG* does not improve the accuracy much because the improvement of FI is not significant. The assessment accuracy is improved drastically by introducing external raters. Furthermore, the proposed external rater assignment method realizes the higher accuracy than the random assignment method when n^J is large and n^e is small.

In this experiment, *PropE* improved the RMSD from about 0.02 to 0.05 from *RndE*, and from about 0.05 to 0.10 from *PropG* and *RndG*. To examine the effects of these improvements, we evaluate the accuracy of learner rankings based on the ability estimates. Providing accurate learner rankings is important because they are often used to determine the final grades of learners (e.g., [81], [82], [83]).

We evaluated the ranking accuracy given the ability estimates as follows.

1) We calculated the learner rankings based on learner abilities estimated from the full peer assessment data.
2) Similarly, we calculated the learner rankings based on the ability estimates using each method (namely, *RndG*, *PropG*, *RndE*, and *PropE*) for $G \in \{3, 4, 5\}$ and $t = 4$. Here, $n^e = 3$ and $n^J = 12$ were given for *PropE* and *RndE*.
3) We calculated the percent correct and the mean absolute error (MAE) between the ranking of 1) and that of 2).
4) We calculated the average percent correct and the MAE of 10 repetitions.

TABLE 5
Accuracy of learner ranking for each method.

Index	G	PropE	RndE	PropG	RndG
Percent correct	3	15.9%	14.7%	14.3%	14.1%
	4	15.1%	12.4%	9.7%	11.5%
	5	14.4%	10.3%	8.2%	9.8%
MAE	3	3.04	3.34	3.65	3.58
	4	3.31	3.65	4.15	4.25
	5	3.66	3.93	4.61	4.67

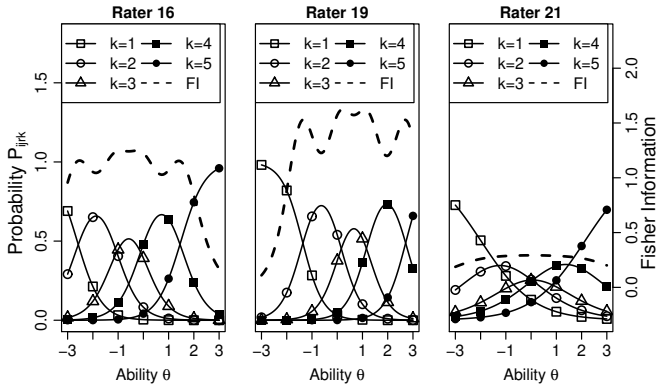


Fig. 16. Item response curves of three raters in actual data experiments.

Table 5 presents the results. The results demonstrate that *PropE* achieves the highest percent correct and the lowest MAE among all methods. Especially, when $G = 5$, *PropE* improves the percent correct by about 4 to 6% compared to the other methods. These results suggest that improvement of RMSD by *PropE* has a non-negligible effect on increasing the accuracy of learner rankings (grading).

7.4 Example of ability estimation and rater assignment

This subsection presents an example of rater assignment by the proposed method and the estimated IRT model parameters. Table 6 shows group members and external raters for each learner in task 4, along with estimated parameter values obtained through experimentation given $G = 5$, $n^J = 6$ and $n^e = 3$. Furthermore, the *assigned count* row shows how often each learner was assigned to the others in task 4.

The table shows that the learners have different rater characteristics. As examples, Fig. 16 depicts the IRCs of Rater 16, 19, and 21 for task 4. The horizontal axis shows a learner's ability θ_j: the first vertical axis shows the response probability of the rater for each category; the second vertical axis shows FI. According to Table 6 and Fig. 16, the characteristics of each rater can be interpreted as 1) Rater 16 is a lenient rater with high-valued consistency. The rater tends to provide higher FI for low ability levels. 2) Rater 19 is more severe than Rater 16 with high-valued consistency. The rater tends to assign higher FI for high ability levels. 3) Rater 21 is an extremely inconsistent rater. Therefore, FI is low overall.

Table 6 also shows that *PropG* and *PropE* assign raters in considering their characteristics and the learner ability. For example, *PropE* tends to assign lenient raters (such as Rater 16 and 17) to the low ability learners (such as Learners 6, 16 and 23) because those raters have higher FI for low ability levels. Conversely, it tends to assign severe raters (such as Rater 8 and 19) to high ability learners (such as Learners 2, 4, 12 and 13) because those raters provide higher FI for high ability levels. Moreover, it does not assign inconsistent raters (such as Rater 7 and 24) to anybody because their FI values are low overall.

TABLE 6
Parameter estimates and assigned raters for each learner given $t = 4$, $G = 5, n^J = 6$, and $n^e = 3$

Learner	$\hat{\gamma}_r$	$\hat{\epsilon}_r$	$\hat{\theta}_j$	Group members	External raters	Assigned count
1	0.85	-0.61	0.50	{6,11,20,25,26,31}	{4,12,15}	6
2	0.75	-1.01	0.74	{7,9,12,13,21}	{8,20,22}	5
3	0.80	0.39	0.27	{8,18,19,22,30,32}	{5,28,33}	6
4	1.19	-0.42	0.75	{14,17,23,24,33,34}	{8,9,20}	12
5	1.07	-0.53	-0.12	{10,15,16,27,28,29}	{11,17,32}	12
6	1.13	0.34	-0.23	{1,11,20,25,26,31}	{16,17,32}	6
7	0.49	1.42	0.76	{2,9,12,13,21}	{19,20,32}	5
8	1.86	0.37	0.50	{3,18,19,22,30,32}	{9,28,33}	12
9	1.06	1.27	0.20	{2,7,12,13,21}	{18,23,33}	11
10	0.56	0.17	0.07	{5,15,16,27,28,29}	{18,19,23}	6
11	1.31	-0.17	0.64	{1,6,20,25,26,31}	{8,9,22}	12
12	0.87	-0.28	0.91	{2,7,9,13,21}	{8,20,22}	11
13	0.80	0.95	0.52	{2,7,9,12,21}	{18,19,23}	5
14	1.00	0.41	0.25	{4,17,23,24,33,34}	{12,15,19}	6
15	1.60	-0.61	0.13	{5,10,16,27,28,29}	{4,11,19}	12
16	1.62	-0.64	-0.90	{5,10,15,27,28,29}	{11,17,32}	12
17	1.55	-0.77	0.67	{4,14,23,24,33,34}	{5,9,22}	12
18	1.23	0.24	0.30	{3,8,19,22,30,32}	{4,5,23}	12
19	1.88	0.60	0.26	{3,8,18,22,30,32}	{15,16,17}	12
20	0.99	0.47	0.09	{1,6,11,25,26,31}	{18,28,33}	12
21	0.74	0.00	0.50	{2,7,9,12,13}	{8,20,22}	5
22	1.36	0.41	0.22	{3,8,18,19,30,32}	{5,23,28}	12
23	1.35	0.27	-1.01	{4,14,17,24,33,34}	{11,16,32}	12
24	1.09	0.20	-0.03	{4,14,17,23,33,34}	{11,16,32}	6
25	0.75	-0.37	0.24	{1,6,11,20,26,31}	{4,12,15}	6
26	0.86	-0.15	0.39	{1,6,11,20,25,31}	{5,28,33}	6
27	0.88	-0.91	-0.08	{5,10,15,16,28,29}	{4,12,17}	6
28	1.20	-0.13	0.03	{5,10,15,16,27,29}	{18,19,23}	12
29	1.18	-0.70	0.06	{5,10,15,16,27,28}	{4,11,12}	6
30	0.78	0.80	0.04	{3,8,18,19,22,32}	{15,16,17}	6
31	0.91	-0.69	0.71	{1,6,11,20,25,26}	{8,9,22}	6
32	1.18	-1.17	0.73	{3,8,18,19,22,30}	{9,20,33}	12
33	1.01	-0.14	-0.01	{4,14,17,23,24,34}	{5,18,28}	12
34	0.92	-0.23	0.22	{4,14,17,23,24,33}	{12,15,16}	6

Task	$\hat{\alpha}_t$	$\hat{\beta}_{t1}$	$\hat{\beta}_{t2}$	$\hat{\beta}_{t3}$	$\hat{\beta}_{t4}$
1	1.53	-1.75	-0.57	0.88	2.03
2	1.47	-2.63	-0.83	0.71	2.27
3	1.49	-2.45	-0.91	0.68	2.03
4	1.14	-1.98	-0.48	0.60	2.13

Furthermore, Table 6 shows that the proposed external rater assignment method can engender unbalanced assessment workload among learners. Specifically, consistent raters tend to have a higher workload than inconsistent raters do because they generally give high FI values. We can reduce this imbalance by decreasing n^J, although the ability assessment accuracy tends to decline, as demonstrated in the earlier experiments. This result suggests that n^J should be set as large as possible within the acceptable range of the unbalanced assessment workload.

8 Conclusion

This study proposed methods to improve peer assessment accuracy when the assessment is conducted by dividing

learners into multiple groups using IRT and integer programming. Specifically, we first proposed the group formation method, which maximizes the lower bound of FI for each learner. The experimentally obtained results, however, showed that the method did not improve the accuracy sufficiently compared to a random group formation method.

To resolve that difficulty, we further proposed the external rater assignment method, which assigns a few optimal 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.

The proposed method requires estimated IRT parameter values to calculate the Fisher information, even if they are practically unknown. This study examined the usage of the proposed method with parameter estimation assuming an application to an actual e-learning situation. Through the simulation and actual data experiments, we demonstrated that the usage worked appropriately.

In this study, the simulation and actual data experiments were conducted assuming small numbers of learners to match the scale of the authors' past e-learning courses. Our future studies will evaluate the effectiveness of the proposed method when applied to large-scale peer assessment data. To use an extremely large dataset, some improvement of computational efficiency of the proposed method might be necessary. This represents another issue for future study.

Furthermore, as discussed in Section 1, the proposed method is expected to be effective for learning improvement, although this study examined only the peer assessment accuracy. Evaluation of that assumption is left as a task for future study.

APPENDIX

Programs for the parameter estimation of the IRT model, the proposed group optimization method, and the proposed external rater assignment method can be downloaded from the Bitbucket repository https://bitbucket.org/uto/group_optimization_irt.git. The programs were written in Java. They require *IBM ILOG CPLEX Optimization Studio* [75]. Additionally, the numerical data associated with the experiments in this study have been deposited to the same repository.

ACKNOWLEDGMENTS

This work was supported by JSPS KAKENHI Grant Number 17H04726.

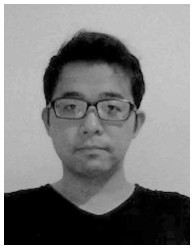
REFERENCES

[1] M. Ueno and T. Okamoto, "Item Response Theory for Peer Assessment," in *Proc. 8th IEEE Int. Conf. Adv. Learn. Technol.*, pp. 554–558, July 2008.
[2] M. Uto and M. Ueno, "Item response theory for peer assessment," *IEEE Trans. Learn. Technol.*, vol. 9, pp. 157–170, April 2016.
[3] P. Davies, "Review in computerized peer-assessment. will it have an effect on student marking consistency?," in *Proc. 11th CAA Int. Comput. Assisted Conf.*, pp. 143–151, 2007.
[4] S. S. Lin, E. Z.-F. Liu, and S.-M. Yuan, "Web-based peer assessment: feedback for students with various thinking-styles," *J. Comput. Assist. Learn.*, vol. 17, no. 4, pp. 420–432, 2001.
[5] A. Bhalerao and A. Ward, "Towards electronically assisted peer assessment: a case study," *ALT-J: research learning technology*, vol. 9, no. 1, pp. 26–37, 2001.
[6] S. Trahasch, "From peer assessment towards collaborative learning," in *Frontiers in Education, 2004. FIE 2004. 34th Annual*, pp. F3F–16, IEEE, 2004.
[7] Y.-T. Sung, K.-E. Chang, S.-K. Chiou, and H.-T. Hou, "The design and application of a web-based self- and peer-assessment system," *Comput. & Educ.*, vol. 45, no. 2, pp. 187–202, 2005.
[8] J. Sitthiworachart and M. Joy, "Effective peer assessment for learning computer programming," in *ACM SIGCSE Bulletin*, vol. 36, pp. 122–126, ACM, 2004.
[9] K. Cho and C. D. Schunn, "Scaffolded writing and rewriting in the discipline: A web-based reciprocal peer review system," *Comput. & Educ.*, vol. 48, no. 3, pp. 409–426, 2007.
[10] S. Bostock, "Student peer assessment," *Learn. Technol.*, 2000.
[11] R. L. Weaver and H. W. Cotrell, "Peer evaluation: A case study," *Innov. High. Educ.*, vol. 11, no. 1, pp. 25–39, 1986.
[12] J. Hamer, K. T. Ma, and H. H. Kwong, "A method of automatic grade calibration in peer assessment," in *Proc. 7th Australas. Conf. Comput. Educ.*, pp. 67–72, Australian Computer Society, Inc., 2005.
[13] H. K. Suen, "Peer assessment for massive open online courses (MOOCs)," *The Int. Rev. Res. Open Distributed Learn.*, vol. 15, no. 3, pp. 312–327, 2014.
[14] N. B. Shah, J. Bradley, S. Balakrishnan, A. Parekh, K. Ramchandran, and M. J. Wainwright, "Some Scaling Laws for MOOC Assessments," in *KDD Workshop on Data Mining for Educational Assessment and Feedback (ASSESS 2014)*, 2014.
[15] L. Moccozet and C. Tardy, "An assessment for learning framework with peer assessment of group works," in *Proc. 14th Int. Conf. Inf. Technol. High. Educ. Train.*, pp. 1–5, 2015.
[16] T. Staubitz, D. Petrick, M. Bauer, J. Renz, and C. Meinel, "Improving the peer assessment experience on MOOC platforms," in *Proceedings of the Third (2016) ACM Conference on Learning@ Scale*, pp. 389–398, ACM, 2016.
[17] A. ArchMiller, J. Fieberg, J. Walker, and N. Holm, "Group peer assessment for summative evaluation in a graduate-level statistics course for ecologists," *Assess. & Eval. High. Educ.*, pp. 1–13, 2016.
[18] J. Lave and E. Wenger, *Situated Learning. Legitimate Peripheral Participation*. Cambridge University Press, 1991.
[19] C. H. Lan, S. Graf, K. R. Lai, and K. Kinshuk, "Enrichment of peer assessment with agent negotiation," *IEEE Trans. Learn. Technol.*, vol. 4, no. 1, pp. 35–46, 2011.
[20] S. Usami, "A polytomous item response model that simultaneously considers bias factors of raters and examinees: Estimation through a Markov chain Monte Carlo algorithm," *The Jpn. J. Educ. Psychol.*, vol. 58, no. 2, pp. 163–175, 2010.
[21] Z. Wang and L. Yao, "The Effects of Rater Severity and Rater Distribution on Examinees' Ability Estimation for Constructed Response Items," *ETS Res. Rep. Ser.*, vol. 2013, no. 2, pp. 1–22, 2013.
[22] S. J. Lurie, A. C. Nofziger, S. Meldrum, C. Mooney, and R. M. Epstein, "Effects of rater selection on peer assessment among medical students," *Med. Educ.*, vol. 40, no. 11, pp. 1088–1097, 2006.
[23] T. Eckes, *Introduction to Many-Facet Rasch Measurement: Analyzing and Evaluating Rater-Mediated Assessments*. Peter Lang, 2011.
[24] F. M. Lord, *Applications of item response theory to practical testing problems*. Lawrence Erlbaum Associates, Inc., 1980.
[25] R. J. Patz, B. W. Junker, M. S. Johnson, and L. T. Mariano, "The hierarchical rater model for rated test items and its application to large-scale educational assessment data," *J. Educ. Behav. Stat.*, vol. 27, no. 4, pp. 341–384, 2002.
[26] R. J. Patz and B. W. Junker, "Applications and Extensions of MCMC in IRT: Multiple Item Types, Missing Data, and Rated Responses," *J. Educ. Behav. Stat.*, vol. 24, no. 4, pp. 342–366, 1999.
[27] J. M. Linacre, *Many-faceted Rasch measurement*. Chicago: MESA Press, 1989.
[28] M. Uto and M. Ueno, "Empirical comparison of item response theory models with rater's parameters," *Heliyon, Elsevier*, vol. 4, no. 5, pp. 1–32, 2018.
[29] Y. J. Hung, "Group peer assessment of oral English performance in a taiwanese elementary school," *Stud. Educ. Eval.*, vol. 59, pp. 19 – 28, 2018.

[30] J. W. Strijbos, "Assessment of (computer-supported) collaborative learning," *IEEE Transactions on Learn. Technol.*, vol. 4, pp. 59–73, Jan 2011.
[31] T. Nguyen, M. Uto, Y. Abe, and M. Ueno, "Reliable Peer Assessment for Team-project-based Learning using Item Response Theory," in *Proc. 23rd Int. Conf. Comp. Educ.*, pp. 144–153, 2015.
[32] Y. S. Lin, Y. C. Chang, and C. P. Chu, "Novel approach to facilitating tradeoff multi-objective grouping optimization," *IEEE Trans. Learn. Technol.*, vol. 9, pp. 107–119, April 2016.
[33] I. Srba and M. Bielikova, "Dynamic Group Formation as an Approach to Collaborative Learning Support," *IEEE Trans. Learn. Technol.*, vol. 8, pp. 173–186, April 2015.
[34] Y. Pang, R. Mugno, X. Xue, and H. Wang, "Constructing collaborative learning groups with maximum diversity requirements," in *Proc. 15th IEEE Int. Conf. Adv. Learn. Technol.*, pp. 34–38, July 2015.
[35] M. I. Dascalu, C. N. Bodea, M. Lytras, P. O. De Pablos, and A. Burlacu, "Improving e-learning communities through optimal composition of multidisciplinary learning groups," *Comput. Hum. Behav.*, vol. 30, pp. 362–371, 2014.
[36] J. Moreno, D. A. Ovalle, and R. M. Vicari, "A genetic algorithm approach for group formation in collaborative learning considering multiple student characteristics," *Comput. & Educ.*, vol. 58, no. 1, pp. 560–569, 2012.
[37] R. Hübscher, "Assigning students to groups using general and context-specific criteria," *IEEE Trans. Learn. Technol.*, vol. 3, no. 3, pp. 178–189, 2010.
[38] Y. T. Lin, Y. M. Huang, and S. C. Cheng, "An automatic group composition system for composing collaborative learning groups using enhanced particle swarm optimization," *Comput. & Educ.*, vol. 55, no. 4, pp. 1483–1493, 2010.
[39] M. Ueno, "Data mining and text mining technologies for collaborative learning in an ILMS "SamurAI"," in *Proc. 4th IEEE Int. Conf. Adv. Learn. Technol.*, pp. 1052–1053, Aug 2004.
[40] M. Ueno and M. Uto, "Learning community using social network service," in *Proc. International Conference Web Based Communities*, pp. 109–119, 2011.
[41] F. Dochy, M. Segers, and D. Sluijsmans, "The use of self-, peer and co-assessment in higher education: A review," *Stud. High. Educ.*, vol. 24, no. 3, pp. 331–350, 1999.
[42] P. M. Sadler and E. Good, "The impact of self-and peer-grading on student learning," *Educ. Assess.*, vol. 11, no. 1, pp. 1–31, 2006.
[43] C. Piech, J. Huang, Z. Chen, C. Do, A. Ng, and D. Koller, "Tuned models of peer assessment in MOOCs," in *Proc. 6th Int. Conf. Educ. Dat. Min.*, pp. 153–160, 2013.
[44] M. Ueno, "On-line contents analysis system for e-learning," in *IEEE International Conference on Advanced Learning Technologies*, pp. 762–764, 2004.
[45] M. Ueno, "Intelligent LMS with an agent that learns from log data," in *Proceedings of E-Learn: World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education 2005*, pp. 3169–3176, 2005.
[46] D. Andrich, "A rating formulation for ordered response categories," *Psychom.*, vol. 43, no. 4, pp. 561–573, 1978.
[47] G. N. Masters, "A Rasch model for partial credit scoring," *Psychom.*, vol. 47, no. 2, pp. 149–174, 1982.
[48] E. Muraki, "A generalized partial credit model: Application of an EM algorithm," *Appl. Psychol. Meas.*, vol. 16, pp. 159–176, June 1992.
[49] F. Samejima, "Estimation of Latent Ability Using a Response Pattern of Graded Scores," *Psychom. Monogr.*, no. 17, pp. 1–100, 1969.
[50] G. Engelhard, "The measurement of writing ability with a many-faceted Rasch model," *Appl. Meas. Educ.*, vol. 5, no. 3, pp. 171–191, 1992.
[51] T. Eckes, "Examining rater effects in TestDaF writing and speaking performance assessments: A many-facet Rasch analysis," *Lang. Assess. Q.*, vol. 2, no. 3, pp. 197–221, 2005.
[52] L. Tesio, M. Ponzio, G. Dati, P. Zaratin, M. A Battaglia, A. Simone, and M. Grzeda, "Funding medical research projects: Taking into account referees' severity and consistency through many-faceted Rasch modeling of projects' scores," *J. Appl. Meas.*, vol. 16, pp. 129–152, 2015.
[53] J. M. Casabianca and E. W. Wolfe, "The impact of design decisions on measurement accuracy demonstrated using the hierarchical rater model," *Psychol. Test Assess. Model.*, vol. 59, no. 4, pp. 471–492, 2017.
[54] E. San Martín, J. González, and F. Tuerlinckx, "On the unidentifiability of the fixed-effects 3pl model," *Psychom.*, vol. 80, no. 2, pp. 450–467, 2015.
[55] W. J. van der Linden, *Handbook of Item Response Theory, Volume One: Models*. CRC Press, 2016.
[56] J.-P. Fox, *Bayesian item response modeling: Theory and applications*. Springer, 2010.
[57] W. J. van der Linden, *Handbook of Item Response Theory, Volume Two: Statistical Tools*. CRC Press, 2016.
[58] M. L. Nering and R. Ostini, *Handbook of Polytomous Item Response Theory Models*. Routledge, Taylor & Francis Group, 2010.
[59] S. P. Reise and D. A. Revicki, *Handbook of Item Response Theory Modeling: Applications to Typical Performance Assessment*. Routledge, 2014.
[60] L. T. Mariano and B. W. Junker, "Covariates of the rating process in hierarchical models for multiple ratings of test items," *J. Educ. Behav. Stat.*, vol. 32, no. 3, pp. 287–314, 2007.
[61] M. Wilson and M. Hoskens, "The rater bundle model," *J. Educ. Behav. Stat.*, vol. 26, no. 3, pp. 283–306, 2001.
[62] L. T. DeCarlo, Y. K. Kim, and M. S. Johnson, "A hierarchical rater model for constructed responses, with a signal detection rater model," *J. Educ. Meas.*, vol. 48, no. 3, pp. 333–356, 2011.
[63] D. J. Weiss, "Improving measurement quality and efficiency with adaptive testing," *Appl. Psychol. Meas.*, vol. 6, no. 4, pp. 473–492, 1982.
[64] B. Babcock and D. J. Weiss, "Termination criteria in computerized adaptive tests: Variable-length CATs are not biased," in *GMAC Conference on Computerized Adaptive Testing*, pp. 1–21, 2009.
[65] X. Zhou, *Designing P-Optimal Item Pools in Computerized Adaptive Tests with Polytomous Items*. PhD thesis, Michigan State University, 2012.
[66] L. Zhang, C. A. Lau, and S. Wang, "Influence of item pool characteristics on repeated measures for student growth in computerized adaptive testing," in *The annual meeting of the National Council on Measurement in Education*, pp. 1–41, 2013.
[67] W. J. van der Linden, *Linear models for optimal test design*. Springer-Verlag New York, 2005.
[68] E. Boekkooi-Timminga, *The Construction of Parallel Tests from IRT-Based Item Banks*, vol. 15. J. Educational Statistics, 1990.
[69] T. Ishii, P. Songmuang, and M. Ueno, "Maximum clique algorithm and its approximation for uniform test form assembly," *IEEE Transactions on Learn. Technol.*, vol. 7, pp. 83–95, Jan 2014.
[70] P. Songmuang and M. Ueno, "Bees algorithm for construction of multiple test forms in e-testing," *IEEE Transactions on Learn. Technol.*, vol. 4, no. 3, pp. 209–221, 2011.
[71] T. Ishii and M. Ueno, "Algorithm for uniform test assembly using a maximum clique problem and integer programming," in *Artificial Intelligence in Education*, pp. 102–112, Springer International Publishing, 2017.
[72] D. M. Sluijsmans, G. Moerkerke, J. J. van Merriënboer, and F. J. Dochy, "Peer assessment in problem based learning," *Stud. Educ. Eval.*, vol. 27, no. 2, pp. 153–173, 2001.
[73] T. Papinczak, L. Young, and M. Groves, "Peer assessment in problem-based learning: A qualitative study," *Adv. Heal. Sci. Educ.*, vol. 12, no. 2, pp. 169–186, 2007.
[74] Y. Cho, S. Je, Y. S. Yoon, H. R. Roh, C. Chang, H. Kang, and T. Lim, "The effect of peer-group size on the delivery of feedback in basic life support refresher training: a cluster randomized controlled trial," *BMC Med. Educ.*, vol. 16, no. 1, p. 167, 2016.
[75] IBM Corp., *IBM ILOG CPLEX Optimization Studio: CPLEX User's Manual*. IBM Corp., 12.6 ed., 2015.
[76] F. B. Baker and S.-H. Kim, *Item response theory: Parameter estimation techniques*. Marcel Dekker, Inc, 2004.
[77] H. Persky, M. Daane, and Y. Jin, "The nation's report card: Writing 2002," tech. rep., National Center for Education Statistics, 2003.
[78] D. Salahu-Din, H. Persky, and J. Miller, "The nation's report card: Writing 2007," tech. rep., National Center for Education Statistics, 2008.
[79] W. M. Yen, "Effects of local item dependence on the fit and equating performance of the three-parameter logistic model," *Appl. Psychol. Meas.*, vol. 8, no. 2, pp. 125–145, 1984.
[80] R. L. Brennan, *Generalizability Theory*. Springer Verlag, 2001.
[81] University of Pennsylvania, "Course syllabus: Operations strategy." https://syllabi-media.s3.amazonaws.com/prod/2017A-OIDD615006-84fb5a70.pdf. Accessed: 2018-12-12.
[82] University of Southern California, "Course syllabus: Introduction to the legal environment of business."

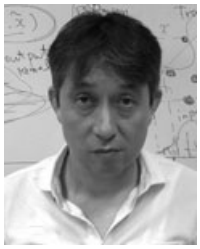
https://msbfile03.usc.edu/digitalmeasures/kfields/schteach/Syllabus%20FBE%20403%20(Fall%202015)-1.pdf. Accessed: 2018-12-12.

[83] University of Alberta, "Course syllabus: Corporate finance." https://catalogue.ualberta.ca/Syllabus/Download?Subject=FIN&Catalog=501&filename=2014-WINTER-FIN501-LEC-X50.pdf. Accessed: 2018-12-12.



Masaki Uto received a Ph.D. degree from the University of Electro-Communications in 2013. He has been an Assistant Professor of the University of Electro-Communications since 2015. His research interests include e-learning, e-testing, machine learning, and data mining.

Duc-Thien Nguyen received a B.E. degree from Hanoi University of Technology in 2004 and an M.E. degree from the University of Electro-Communications in 2014, where he is currently pursuing a Ph.D. degree. His research interests include e-testing, Bayesian statistics, and machine learning.



Maomi Ueno received a Ph.D. degree in Computer Science from the Tokyo Institute of Technology in 1994. He has been a Professor of the University of Electro-Communications since 2013. He received Best Paper awards from ICTAI2008, ED-MEDIA2008, e-Learn2004, e-Learn2005, and e-Learn2007. His interests are e-learning, e-testing, e-portfolio, machine learning, data mining, Bayesian statistics, and Bayesian networks. He is a member of the IEEE.