Ogata, H. et al. (Eds.) (2015). Proceedings of the 23rd International Conference on Computers in Education. China: Asia-Pacific Society for Computers in Education

Reliable Peer Assessment for Team-project-based Learning using Item Response Theory

Thien D. NGUYEN [a*], Masaki UTO [b], Yu ABE [c] & Maomi UENO [a]
[a] *University of Electro-Communications, Japan*
[b] *Nagaoka University of Technology, Japan*
[c] *Recruit Career Co. Ltd., Japan*
* thien@ai.is.uec.ac.jp

Abstract: This study addresses the reliability of peer assessment in team-project-based learning using item response models that incorporate rater characteristic parameters. The following difficulties can arise when applying item response models to peer assessment in team-project-based learning. (1) Earlier item response models incorporate the assumption that peer rating data consist of three-way data, which are learners × performance tasks × raters. However, the data used for this article are four-way data, which are learners × project tasks × raters × dimensions of abilities because multiple dimensions of learner abilities are generally assessed in project-based learning. Previous models are not directly applicable to four-way data. (2) In team-project-based learning, the learners are partitioned into several teams and are peer assessed within each team. In this case, the peer assessment reliability depends on the team assembly because the ability estimation accuracy for each learner depends on the characteristics of peer raters within the same team. If the teams are assembled randomly or fixed for all project tasks, then differences in the accuracy of ability estimation among learners would increase. To resolve this problem, this article presents a proposal of reliable peer assessment in team-project-based learning. Concretely, we extend the previous item response model that incorporates rater characteristic parameters to apply to four-way data. Furthermore, we formulate a team assembly problem for team-project-based learning as an integer programming problem. The team assembly method maximizes the difference between teams that are assembled for a project task and those assembled for previous project tasks.

Keywords: Team-project-based learning, peer assessment, item response theory, team assembly, reliability

1. Introduction

In recent years, assessment has been facing a shift from traditional testing to authentic assessment (Dochy et al., 2006). Authentic assessment is designed to assess higher-order skills and thinking processes instead of factual knowledge and lower-order cognitive skills (Jonsson, 2007). In the context of authentic assessment, learner performance and learning activities are assessed to capture such abilities by letting learners solve realistic or authentic problems (Jonsson, 2007).

Assessment in project (or problem)-based learning has been attempted to assess performances of learners in authentic problems (Lee and Lim, 2012, Whatley, 2012). Project-based learning brings together learning through experimentation and learning by doing (Whatley, 2012). In recent years, team-project-based learning that emphasizes interactions among learners through solving authentic problems has attracted much attention for cultivating and assessing learners' social abilities (Lee and Lim, 2012). During team-project-based learning, teamwork abilities including communication, leadership, collaboration, and interpersonal relations are performed, in addition to personal abilities such as critical reasoning, creative thinking, and responsibility (Lee and Lim, 2012). Therefore, assessing performances of learners in team-project-based learning can enable measurement of those authentic abilities.

The assessment of learning processes is more important than that of learning outcomes to measure learners' authentic abilities performed through team-project-based learning. In those assessments, multiple dimensions of learner's abilities (e.g., ability of communication, leadership, and

responsibility) are generally assessed using an evaluation criterion or scoring rubric. However, it is difficult for a few instructors to observe and assess all the processes occurring within team projects when numerous teams are being evaluated, although the instructors might be able to assess the learning outcomes (Admiraal et al., 2014; Lee and Lim, 2012; Shah et al., 2014).

Peer assessment, which is mutual assessment among learners (Topping et al., 2000), is an effective method to monitor and assess processes and outcomes of team projects without burdening instructors (Lee and Lim, 2012; Suen, 2014; Ueno and Okamoto, 2008; Wang and Yao, 2007). Peer assessment presents many important benefits (Piech et al., 2013; Ueno and Okamoto, 2008). Furthermore, peer assessment can be justified as an appropriate assessment method because the learner ability would be defined naturally in the learning community as a social agreement (Lave and Wenger, 1991). Therefore, although peer assessment has been adopted into various learning processes, it has been pointed out that reliability of peer assessment is generally lower than that of instructor assessment unless a sufficient number of peer raters are available for each learner (Piech et al., 2013; Suen, 2014). In this study, the reliability is defined as the stab*ility of learners' ability estimation* (Kim, 2012). Reliability takes a higher value if the learner abilities are obtainable with few errors when the project tasks or raters are changed.

One factor affecting the decrease in reliability is that rater's ratings in peer assessment depend on rater characteristics such as the rating consistency and severity (Lurie et al., 2006; Shah et al., 2014; Suen, 2014; Ueno and Okamoto, 2008; Wang and Yao, 2007). Therefore, the reliability of peer assessment would be improved if the ability of learners were estimated considering the rater characteristics (Usami, 2010; Suen, 2014; Muraki et al., 2000). To realize such ability estimation, several item response models that incorporate rater characteristic parameters have been proposed (Patz et al., 1999; Ueno and Okamoto, 2008; Usami, 2010; Uto and Ueno, 2015). Item response theory (Lord, 1980), a test theory based on mathematical models, has generally been used in areas of educational testing and assessment such as entrance exams or certification tests. Traditional item response models enable the estimation of examinees' abilities considering the characteristics of test items (e.g., item difficulty and discrimination). Furthermore, item response models that incorporate rater parameters can estimate the ability of learners considering not only the characteristics of items (or project tasks) but also those of raters. Earlier studies (Ueno and Okamoto, 2008; Uto and Ueno, 2015) have demonstrated that the ability of learners as estimated by those item response models was more reliable than a score obtained using traditional scoring methods such as an averaged or summed raw score in peer assessment.

This study was conducted to improve the reliability of peer assessment in team-project-based learning using the item response models. However, the following problems can occur when applying item response models to peer assessment in team-project-based learning.

1. Previous item response models incorporate the assumption that the peer rating data consist of each rater's ratings for each learner's outcome in each performance task. Therefore, the data consist of three-way data. However, the data assumed for this study are four-way data, which are learners × project tasks × raters × dimensions of abilities because multiple dimensions of learner abilities are generally assessed in project-based learning. Previous models are not directly applicable to four-way data.
2. In team-project-based learning, the learners are partitioned into several teams and are peer assessed within each team. In this case, the reliability of peer assessment depends on the team assembly because the ability estimation accuracy for each learner depends on the characteristics of peer raters within the same team. If the teams are randomly assembled or fixed for all project tasks, then differences in the accuracy of ability estimation among learners would increase.

To resolve those difficulties, we extend the item response model proposed by Uto and Ueno (2015) to apply to four-way data. Furthermore, we propose a team assembly method for team-project-based learning. This article assumes that several project tasks exist and that the teams are changed after each project tasks. The team assembly method maximizes the difference between teams that are assembled for a current project task and those for previous project tasks. The assembly method is formulated as an integer programming problem. The features of the proposed method are the following.

1. The proposed item response model is expected to improve the reliability of peer assessment that measures multiple abilities of learners.

2. The proposed team assembly method is expected to realize more equivalent accuracy of ability estimation for learners because each learner is assessed by as varied a group of peer-raters as possible through multiple tasks.

In addition, this article demonstrates the effectiveness of the proposed method through simulation and actual data experiments.

2. Peer Assessment in Team-project-based Learning

This article assumes that several project tasks $\{t \mid t = 1, \cdots, T\}$ exist for team-project-based learning. For each project task $t \in \{1, \cdots, T\}$, learners $\{j \mid j = 1, \cdots, J\}$ are divided into some teams $\{g \mid g = 1, \cdots, G\}$ that consist of a few learners. The teams are shuffled after each project task. Peer assessment is conducted within the team. In the peer assessment, the peer raters assess multiple dimensions of peer-learner abilities $\{d \mid d = 1, \cdots, D\}$, which the assessment aims to measure (e.g., ability of communication, leadership, and responsibility), using K categories $\{k \mid k = 1, \cdots, K\}$ based on an evaluation criterion.

From the above, the peer rating data $\boldsymbol{U}$ consists of categories $k \in \{1, \cdots, K\}$ given by each peer rater $r \in \{1, \cdots, R\}$ to each learner $j \in \{1, \cdots, J\}$ on each ability $d \in \{1, \cdots, D\}$ for each task $t \in \{1, \cdots, T\}$. Therefore, let x_{tjdr} be a response of rater r to learner j's ability d for task t, the data $\boldsymbol{U}$ are described as shown below.

$$\boldsymbol{U} = \{ x_{tjdr} \in \{-1,1,2, \cdots, K\} \mid j = 1, \cdots, J; t = 1, \cdots, T;\ r = 1, \cdots, R; d = 1, \cdots, D\} \quad (1)$$

Here, $x_{tjdr} = -1$ denotes missing data.

As described above, the learners are divided into G teams for each task t. The number of team G is assigned by an analyst. Here, we assume that the numbers of learners of respective teams are equivalent. Therefore, the number of learners in each team g for each task t n_{tg} is constrained with $n_l \leq n_{tg} \leq n_u : \forall t, \forall g$. Therein, n_l is an integral in the range of $J/G - 1 < n_l \leq J/G$, and n_u is an integral in the range of $J/G \leq n_u < J/G + 1$.

This article presents a proposal of an item response model for the peer assessment data $\boldsymbol{U}$ and a team assembly method to realize reliable and fair peer assessment in team project learning.

3. Item Response Theory

Item response theory (Lord, 1980), a test theory based on mathematical models, has been used widely with the widespread use of computer testing. Traditionally, item response theory has been applied to test items of which the responses can be scored automatically as correct or wrong, such as multiple-choice items. In recent years, however, applying polytomous item response models to performance assessments such as essay tests and report assessment has been attempted (Matteucci and Stracqualursi, 2006; Muraki et al., 2000).

However those basic item response models are not applicable for the peer assessment data because the data generally consists of three (or more)-way data which are learners × raters × tasks. To resolve the problem, some item response models that incorporate the rater parameters have been proposed (Patz et al., 1999; Ueno and Okamoto, 2008; Usami, 2010; Uto and Ueno, 2015). For the analyses described herein, we use the item response model for peer assessment proposed by Uto and Ueno (2015).

3.1 Item Response Theory for Peer Assessment

The reliability of peer assessment is known to be improved if the learner ability is estimated considering the rater characteristics, especially the rater severity and consistency (Muraki et al., 2000; Suen, 2014; Usami, 2010). Therefore, Uto and Ueno (2015) proposed an item response model that incorporates the rater consistency and severity parameters. This model provides the probability P_{tjrk} that rater r responds in category k to learner j's work for task t as follows.

$$P_{tjrk} = P^*_{tjrk-1} - P^*_{tjrk} \quad (2)$$

$$\begin{cases} P^*_{tjrk} = \dfrac{1}{1+\exp\left(-\alpha_t\,\alpha_r\left(\theta_j - b_{tk} - \varepsilon_r\right)\right)};\ k = 1,\cdots,K-1 \\ P^*_{tjr0} = 1 \\ P^*_{tjrK} = 0 \end{cases} \tag{3}$$

In those equations, α_t is a discrimination parameter of task t, b_{tk} denotes the difficulty in obtaining the score k for task t (here $b_{t1} < \cdots, < b_{tK-1}$), α_r is the consistency of rater r, ε_r represents the severity of rater r, and θ_j is the latent ability of learner j. Here, $\alpha_{r=1} = 1$, $\varepsilon_1 = 0$ and $\Pi_r\alpha_r = 1$ are assumed for model identification.

A unique feature of this model is that the model parameters are fewer than in other previous models as the raters and learners become more numerous. The parameter estimation accuracy is generally higher for a model that incorporates fewer parameters (Bishop, 2006).

3.2 Item Response Model for Peer Assessment that Measures Multi-Dimensional Abilities

As described before, the peer assessment data assumed in this article consist of four-way data, which are learners × project tasks × raters × dimensions of abilities. The item response model presented above is not directly applicable to the four-way data. To resolve this difficulty, this article presents a proposal of an item response model for the four-way data by extending the above item response model.

This article assumes that the dimensions of abilities are mutually independent. Then, the proposed model provides the probability P_{dtjrk} that rater r responds in category k to learner j's ability d in project task t as shown below.

$$P_{dtjrk} = P^*_{dtjrk-1} - P^*_{dtjrk} \tag{4}$$

$$\begin{cases} P^*_{dtjrk} = \dfrac{1}{1+\exp\left(-\alpha_{dt}\,\alpha_{dr}\left(\theta_{dj} - b_{dtk} - \varepsilon_{dr}\right)\right)};\ k = 1,\cdots,K-1 \\ P^*_{dtjr0} = 1 \\ P^*_{dtjrK} = 0 \end{cases} \tag{5}$$

In those equations, α_{dt} and b_{dtk} are the discrimination and difficulty parameters of task t for ability dimension d, α_{dr} and ε_{dr} are the consistency and severity of rater r for ability dimension d, and θ_{dj} is learner j's ability d. Here,$\alpha_{d,r=1} = 1$, $\varepsilon_{d1} = 0$ and $\Pi_r\alpha_{dr} = 1$ are assumed for model identification.

The proposed model is based on the idea of multi-unidimensional item response model, which is known as the special case of the multidimensional item response model (Sheng and Wikle, 2007).

Bayes estimation is useful to estimate the parameters of the proposed model as used in previous studies. As an estimation algorithm for Bayes estimation, the Markov Chain Monte Carlo method (MCMC), which is a random-sampling-based estimation method, is useful (Brooks et al., 2011).

For the analyses in this article, we assume that the ability and parameters in the proposed model are estimated using peer rating data collected from peer assessment in team-project-based learning as described in Section 2. Using the proposed model, learners' authentic abilities performed through team-project-based learning can be estimated accurately considering the characteristics of performance tasks and raters. The estimated abilities might be useful for providing feedback to learners or for helping instructors to evaluate the learner abilities.

4. Team Assembly Method

The proposed item response model is expected to improve the reliability of peer assessment that measures multiple dimensions of learner's abilities. However, in team-project-based learning, the reliability also depends on the team assembly because the ability estimation accuracy for each learner depends on the characteristics of peer raters within the same team. For example, if learners in a team have low rater consistency characteristics, then the learners within the team would not be given higher accuracy of ability estimation than the opposite case.

A naive method to solve the problem is assembling teams for a project task so that the difference between the teams and teams assembled for previous project tasks can be maximized. The accuracy of ability estimation for learners would be close to equivalent as the number of tasks increases because

each learner is assessed by varied a selection of peer-raters as possible through multiple tasks. However, it is difficult for instructors to assemble teams following the team assembly strategy when the number of learners or tasks increases. This article presents a proposal of a team assembly method using integer programming.

Here, let a_{tgj} be a dummy variable that takes the value 1 if learner j belongs to a team g in task t, and 0 if it is not. The team assembly aims to assemble teams for a current task t so that the difference between the teams and teams assembled for the previous tasks $t' \in \{1, \cdots, t-1\}$ can be maximized. The difference between two teams is defined as the Hamming distance $||\boldsymbol{a}_{tg} - \boldsymbol{a}_{t'g'}||$, which indicates the difference between two bit strings. Here, $\boldsymbol{a}_{tg} = \{a_{tg1}, \cdots, a_{tgJ}\}$.

Using the notations described above, we formulate the problem of assembling the teams $\boldsymbol{a}_t = \{\boldsymbol{a}_{t1}, \cdots, \boldsymbol{a}_{tG}\}$ for a current project task t given the teams assembled for all previous tasks $t' \in \{1, \cdots, t-1\}$ as the following integer programming problem.

Maximize:

$$\sum_{\mathrm{t}'} \sum_{\mathrm{g}} z_{t'g} \tag{6}$$

Subject to:

$$\begin{gathered} ||\boldsymbol{a}_{tg} - \boldsymbol{a}_{t'g'}|| \geq z_{t'g};\ \forall\, \mathrm{g}, \forall\, \mathrm{t}', \forall\, \mathrm{g}' \\ \sum_{t'} \sum_{g} c\left(a_{t'gj} : a_{t'gj'}\right) + \sum_{g} c\left(a_{tgj} : a_{tgj'}\right) \leq n_o;\ \forall j, \forall j', j \neq j' \\ n_l \leq \sum_{j} a_{tgj} \leq n_u;\ \forall g \\ \sum_{g} a_{tgj} = 1;\ \forall j \end{gathered} \tag{7}$$

The first constraint requires the Hamming distances between $\boldsymbol{a}_{tg}$ which is team g for the current task t and $\boldsymbol{a}_{t'g'}: \forall g'$, which are the teams assembled for the previous task t' having the minimum value $z_{t'g}$. The proposed team assembly method maximizes the summation of the minimum Hamming distance for all previous tasks $t' \in \{1, \cdots, t-1\}$ and for all the teams g in the current task t.

The second constraint requires that the frequency with which each learner pair appears in the same team through all tasks be no more than n_o. The constraint is necessary because their appearance frequency cannot be controlled solely with the above Hamming distance constraint. The frequency with which the same learner pairs appear in the same team should be reduced to increase the diversity of learner–rater combinations. In the second constraint, $c\ (a_{tgj} : a_{tgj'})$ denotes a function which returns 1 if $a_{tgj} = 1 \wedge\ a_{tgj'} = 1$ and takes 0 otherwise. Here, n_o is the maximum frequency with which each learner pair appears in same team over all tasks; it is given by an analyst.

The third constraint requires the range of the number of learners in each team. By the fourth constraint, each learner belongs to only one team for each task.

By solving the integer programming, we can obtain the teams for task t that maximize the difference between the teams and those for previous tasks.

5. Simulation Experiments

5.1 Difference in Reliability among Teams

The following simulation experiment was conducted to confirm whether the reliabilities of teams mutually differ.

1. Given the number of learners $J = 200$, categories $K = 3$, tasks $T = 1$ and ability dimensions $D = 1$, all parameters in the proposed item response model were generated randomly from the distributions presented in Table 1. In Table 1, $N\ (\mu, \sigma)$ denotes a normal distribution with mean μ and standard deviation σ; $MN\ (\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes a multidimensional normal distribution with mean vector $\boldsymbol{\mu}$ and co-variance matrix $\boldsymbol{\Sigma}$.

2. The 200 generated learners were divided randomly into 50 teams. Here, the number of learners on each team was four.
3. Using the generated model parameters, the reliability of each team was calculated. A reliability coefficient for item response models can be estimated as follows (Kim, 2012; Samejima, 1994).

$$\rho^2_{\hat{\theta}\theta} = \frac{\sigma^2_{\hat{\theta}} - \sigma^2_e}{\sigma^2_{\hat{\theta}}} = \frac{\sigma^2_{\theta}}{\sigma^2_{\theta} + \sigma^2_e}$$

Therein, σ^2_e is defined as $\int g(\theta)/I(\theta)\, d\theta$ and calculable using numerical integration. Here, $g(\theta)$ is the probabilistic distribution function for θ and σ^2_{θ} is the variance of the distribution. In addition, $I(\theta)$ is the Fisher information function at an ability level θ. The proposed item response model gives the Fisher information of rater r in task t at an ability level θ_{dj} as follows.

$$I_{tr}(\theta_{dj}) = \alpha^2_{\mathrm{dt}}\alpha^2_{\mathrm{dr}} \sum_{\mathrm{k}} \frac{(\mathrm{p}^*_{\mathrm{dtjrk}-1}\mathrm{q}^*_{\mathrm{dtjrk}-1} - \mathrm{p}^*_{\mathrm{dtjrk}}\mathrm{q}^*_{\mathrm{dtjrk}})^2}{\mathrm{p}_{\mathrm{dtjrk}}}$$

Therein, $\mathrm{q}^*_{\mathrm{dtjrk}} = 1 - \mathrm{p}^*_{\mathrm{dtjrk}}$. Furthermore, the information for learner j at ability θ_{dj} over all tasks is calculated by summing the information of all the raters that are assigned to learner j. Therefore, the information is defined as $I(\theta_{dj}) = \sum_t \sum_r I_{tr}(\theta_{dj}) \cdot z_{tjr}$, where z_{tjr} is a dummy variable which takes the value 1 if rater r and learner j belong to same team in project task t, and 0 if it is not.

Figure 1 presents results. Its horizontal axis shows the teams. The vertical axis shows the reliability of each team. According to Figure 1, it is apparent that the reliability depends on the teams. The next subsection demonstrates whether the proposed team assembly method enables the reliability for each learner to be equivalent.

Table 1: Prior distributions used for the simulation experiment and Bayes estimation

$log\, \alpha_{di} \sim \mathrm{N}(0.1, 0.4)$
$log\, \alpha_{dr} \sim \mathrm{N}(0.0, 0.5)$
$\varepsilon_{dr}, \theta_{dj} \sim \mathrm{N}(0.0, 1.0)$
$b_{dik} \sim MN(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
$\begin{cases} \boldsymbol{\mu} = \{-1.5, 1.5\} \\ \boldsymbol{\Sigma} = \begin{Bmatrix} 0.64 & 0.10 \\ 0.10 & 0.64 \end{Bmatrix} \end{cases}$

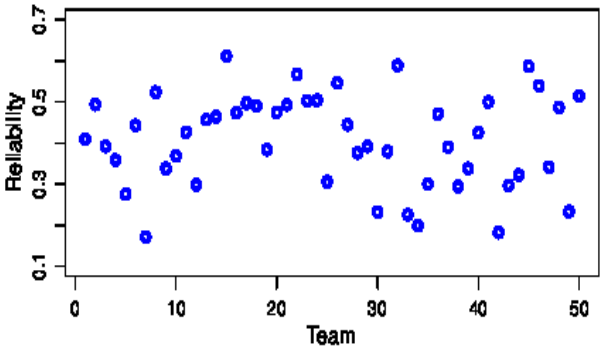


Figure 1. Difference in reliability among teams.

5.2 Evaluation of Team Assembly Method

This subsection demonstrates the features of the proposed team assembly method through the following simulation experiment:

1. For the number of learners $J \in \{15, 30\}$ and tasks $T \in \{3, 5\}$, all the parameters in the proposed item response model were generated randomly using the same method as that used in subsection 5.1. Here, the number of categories $K = 3$ and ability dimensions $D = 1$ were fixed.
2. The teams for the first task were assembled randomly. The number of teams was three for $J = 15$ and six for $J = 30$. Consequently, the number of learners in each team was five.
3. For each task $t \in \{2, \cdots, T\}$, teams were assembled using 1) the proposed method, 2) random assembly method, and 3) fixed method that provides the same teams assembled for the first task. Here, $n_0 = 2$ for $T = 3$ and $n_0 = 3$ for $T = 5$ were given for the proposed method. *IBM ILOG CPLEX Optimization Studio* was used to solve the proposed method.
4. From the obtained teams, the averaged value of the Hamming distances among all the team pairs, the averaged and maximum values of the frequency with which each learner pair appears in the same team. The average and standard deviation of the Fisher information for each learner were calculated.

Table 2 presents results: The proposed method realized a larger Hamming distance and smaller appearance frequency of each learner pair than the other methods, which means that the proposed team assembly method can increase the diversity of rater–learner combinations.

Furthermore, the proposed method tended to give smaller variances of Fisher information for learners than the other methods. The Fisher information $I(\theta)$ can be regarded as the stability of ability estimation at a specific ability level θ. Therefore, using the proposed team assembly method, the ability estimation accuracy for learners might be more equivalent than that of other methods.

In addition, the same experiment was repeated 10 times. Consequently, the percentage of instances in which the proposed method revealed a larger Hamming distance and a smaller appearance frequency of each learner pair than the other methods was 100%. Furthermore, the percentage of instances in which the proposed method provided a smaller variance of Fisher Information for learners than the other methods was 83%.

These results show that the proposed team assembly method can increase the diversity of rater–learner combinations and realize fairer peer assessment.

Table 2: Performances of the proposed team assembly method

		$J = R = 15$				$J = R = 30$			
		Hamming Distance	Appearance Frequency of each Learner Pair		Fisher Information	Hamming Distance	Appearance Frequency of each Learner Pair		Fisher Information
		Mean (SD)	Mean (SD)	Max	Mean (SD)	Mean (SD)	Mean (SD)	Max	Mean (SD)
Proposed	$T = 3$	6.000 (0.000)	1.250 (0.433)	2	5.731 (2.480)	8.000 (0.000)	1.071 (0.258)	2	5.191 (1.198)
	$T = 5$	6.000 (0.000)	1.515 (0.657)	3	7.616 (1.348)	8.000 (0.000)	1.214 (0.418)	3	10.296 (2.023)
Random	$T = 3$	4.667 (0.943)	1.324 (0.498)	3	5.299 (3.083)	6.000 (0.000)	1.216 (0.449)	3	4.537 (1.242)
	$T = 5$	4.600 (1.281)	1.705 (0.786)	4	7.327 (1.228)	5.600 (0.800)	1.420 (0.652)	4	10.303 (2.131)
Fixed	$T = 3$	0.000 (0.000)	3.000 (0.000)	3	6.312 (3.201)	0.000 (0.000)	3.000 (0.000)	3	4.956 (2.012)
	$T = 5$	0.000 (0.000)	5.000 (0.000)	5	7.009 (2.468)	0.000 (0.000)	5.000 (0.000)	5	10.351 (3.250)

6. Application to Actual Data Experiment

This section presents application of the proposed method to actual peer assessment in team-project-based learning.

6.1 Actual Data

The experiment was conducted using the following procedures. 1) First, 24 university students were recruited as study subjects. 2) They were divided into four teams for the first team project. 3) After the team project, the subjects were asked to assess the peer learners within the same team using three categories based on the evaluation criterion prepared by one of the authors. The evaluation criteria consist of three perspectives corresponding to the abilities that we aim to measure. The points of views are presented in Table 5. 4) For the next task, the subjects were divided into different teams that were assembled using the proposed team assembly method. 5) Repeat 3) and 4) until all the three project tasks were finished.

In these experiments, we assigned one expert assessor for each team to monitor and assess the entire process of the project work. The experts assessed the subjects using the same evaluation criteria that the subjects had used for peer assessment.

Table 3 presents the assembled teams for each task. The integers in each cell are the identification numbers of learners. According to Table 3, each learner pair appeared in the same team at most twice in all three tasks.

Table 3: Teams assembled by the proposed team assembly method for actual data experiment

	Team 1	Team 2	Team 3
Task 1	{1,8,9,16,17,24}	{2,7,10,15,18,23}	{3,6,11,14,19,22}
Task 2	{1,2,3,4,5,6}	{7,8,9,10,11,12}	{13,14,15,16,17,18}
Task 3	{1,5,9,10,13,22}	{4,6,7,15,17,19}	{3,8,12,18,20,24}

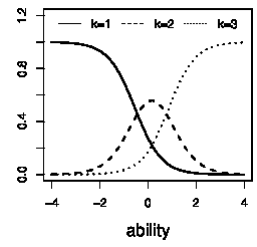
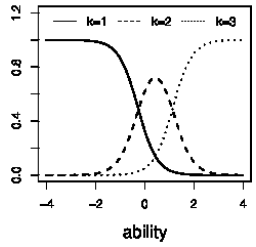
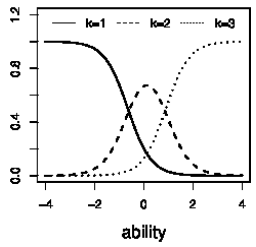
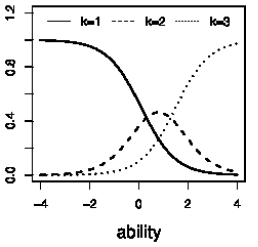
6.2 Examples of Estimated Parameters of the Item Response Model

This subsection presents an example of interpretations for the parameters in the proposed item response model. Table 4 presents item characteristic curves of two peer raters for two project tasks on two dimensions of abilities. According to Table 4, the characteristics of raters, tasks and abilities can be regarded as explained below.

1. *Rater 1* assessed with slightly higher consistency than *Rater 2*.
2. *Rater 2* assessed with slightly severe criteria and tended to give the lowest score to learners who have ability below the average.
3. *Task 2* had somewhat higher discriminant characteristics than *Task 1*.
4. *Task 1* can distinguish *Ability 2* more accurately than *Ability 1*.

The proposed model can estimate the learner's abilities considering these characteristics.

Table 4: Item characteristic curves of the proposed item response model

Rater 1			Rater 2
Task 1		Task 2	Task 1
Ability 1	Ability 2	Ability 1	Ability 1
			

6.3 Evaluation of Reliability

To evaluate the effectiveness of the proposed item response model, the following experiment was conducted.

1. Using the actual peer assessment data, the ability of learners was estimated using the proposed item response model (designated as $\widehat{\boldsymbol{\theta}}_{peer}$). Furthermore, the averaged raw score for each learner was also calculated (designated as $\boldsymbol{\mu}_{peer}$).
2. For each task $t \in \{1,2,3\}$, the ability of learners was estimated using the item response model (designated as $\widehat{\boldsymbol{\theta}}_{peer,task(t)}$). Here, the rater and task parameters, as estimated by the complete peer assessment data, were given. Furthermore, the average score for each task was also calculated (designated as $\boldsymbol{\mu}_{peer,task(t)}$).
3. Using the expert assessment data, the ability of learners was estimated using the proposed item response model (designated as $\widehat{\boldsymbol{\theta}}_{expert}$). In addition, the averaged raw score for each learner was calculated (designated as $\boldsymbol{\mu}_{expert}$).
4. The Pearson's correlations were calculated between $\widehat{\boldsymbol{\theta}}_{peer}$ and $\widehat{\boldsymbol{\theta}}_{peer,task(t)}$ $\forall t$ and $\widehat{\boldsymbol{\theta}}_{expert}$. In addition, the correlations between $\boldsymbol{\mu}_{peer}$ and $\boldsymbol{\mu}_{peer,task(t)}$ $\forall t$ and $\boldsymbol{\mu}_{expert}$ were calculated.

Table 5 presents the results, where r ($\boldsymbol{a}$: $\boldsymbol{b}$) denotes the correlation between two vectors $\boldsymbol{a}$ and $\boldsymbol{b}$. According to Table 5, the correlation between $\widehat{\boldsymbol{\theta}}_{peer}$ and $\widehat{\boldsymbol{\theta}}_{peer,task(t)}$ was higher than $\boldsymbol{\mu}_{peer}$ and $\boldsymbol{\mu}_{peer,task(t)}$ in all cases, demonstrating that the proposed item response model can estimate the learner's true ability more accurately than the average score method.

Furthermore, the correlation between $\widehat{\boldsymbol{\theta}}_{peer}$ and $\widehat{\boldsymbol{\theta}}_{expert}$ was higher than the correlation between $\boldsymbol{\mu}_{peer}$ and $\boldsymbol{\mu}_{expert}$, which indicates that the proposed item response model can improve the correlation between the results of peer assessment and expert assessment.

The results presented above demonstrate that the proposed item response model can provide learner abilities with smaller errors even when the project tasks and raters are changed. The proposed item response model can improve the reliability of peer assessment in team-project-based learning.

Table 5: Evaluation results of reliability using actual data

Ability 1: Idea (Whether the respondent gave new ideas or opinions)

$r(\widehat{\boldsymbol{\theta}}_{peer}: \widehat{\boldsymbol{\theta}}_{peer,task1})$	$r(\widehat{\boldsymbol{\theta}}_{peer}: \widehat{\boldsymbol{\theta}}_{peer,task2})$	$r(\widehat{\boldsymbol{\theta}}_{peer}: \widehat{\boldsymbol{\theta}}_{peer,task3})$	$r(\widehat{\boldsymbol{\theta}}_{peer}: \widehat{\boldsymbol{\theta}}_{expert})$
0.980	0.965	0.975	0.846
$r(\boldsymbol{\mu}_{peer}: \boldsymbol{\mu}_{peer,task1})$	$r(\boldsymbol{\mu}_{peer}: \boldsymbol{\mu}_{peer,task2})$	$r(\boldsymbol{\mu}_{peer}: \boldsymbol{\mu}_{peer,task3})$	$r(\boldsymbol{\mu}_{peer}: \boldsymbol{\mu}_{expert})$
0.949	0.925	0.926	0.809

Ability 2: Attitude of listening (Whether the respondent listened to other learners' opinions carefully)

$r(\widehat{\boldsymbol{\theta}}_{peer}: \widehat{\boldsymbol{\theta}}_{peer,task1})$	$r(\widehat{\boldsymbol{\theta}}_{peer}: \widehat{\boldsymbol{\theta}}_{peer,task2})$	$r(\widehat{\boldsymbol{\theta}}_{peer}: \widehat{\boldsymbol{\theta}}_{peer,task3})$	$r(\widehat{\boldsymbol{\theta}}_{peer}: \widehat{\boldsymbol{\theta}}_{expert})$
0.982	0.976	0.989	0.900
$r(\boldsymbol{\mu}_{peer}: \boldsymbol{\mu}_{peer,task1})$	$r(\boldsymbol{\mu}_{peer}: \boldsymbol{\mu}_{peer,task2})$	$r(\boldsymbol{\mu}_{peer}: \boldsymbol{\mu}_{peer,task3})$	$r(\boldsymbol{\mu}_{peer}: \boldsymbol{\mu}_{expert})$
0.968	0.946	0.970	0.811

Ability 3: Facilitation (Whether the respondent proposed how to proceed the discussion or fix digressions)

$r(\widehat{\boldsymbol{\theta}}_{peer}: \widehat{\boldsymbol{\theta}}_{peer,task1})$	$r(\widehat{\boldsymbol{\theta}}_{peer}: \widehat{\boldsymbol{\theta}}_{peer,task2})$	$r(\widehat{\boldsymbol{\theta}}_{peer}: \widehat{\boldsymbol{\theta}}_{peer,task3})$	$r(\widehat{\boldsymbol{\theta}}_{peer}: \widehat{\boldsymbol{\theta}}_{expert})$
0.981	0.987	0.973	0.900
$r(\boldsymbol{\mu}_{peer}: \boldsymbol{\mu}_{peer,task1})$	$r(\boldsymbol{\mu}_{peer}: \boldsymbol{\mu}_{peer,task2})$	$r(\boldsymbol{\mu}_{peer}: \boldsymbol{\mu}_{peer,task3})$	$r(\boldsymbol{\mu}_{peer}: \boldsymbol{\mu}_{expert})$
0.943	0.948	0.932	0.834

7. Conclusion

This article proposed a method to realize reliable and fair peer assessment for team-project-based learning. Concretely, we extended the previous item response model, which incorporates rater characteristic parameters for application to four-way data, which are learners × tasks × raters × dimensions of abilities. Furthermore, we proposed a team assembly method for team-project-based learning that maximizes the difference between teams assembling for a current project task and those assembled for previous project tasks. The assembly method was formulated as an integer programming problem.

In addition, this article demonstrated the following features of the proposed methods through simulation and actual data experiments.

1. The proposed item response model can realize more reliable ability estimation than the average score method for measuring multiple dimensions of learner's abilities.
2. The proposed team assembly method can increase the diversity of rater–learner combinations and realize more equivalent accuracy of ability estimation for learners.

For the proposed item response model, we assumed that the dimensions of learner's abilities are mutually independent. However, they might be mutually dependent in some actual situations. Construction of an item response model that can accommodate that dependency is a remaining task for future research.

Moreover, the proposed team assembly method was a naive approach. More intelligent team assembly methods are expected to be possible using the item response theory. We would like to construct such a team assembly method in future studies.

Acknowledgements

This work was supported by JSPS KAKENHI Grant Number 15K16256.

References

Admiraal, W., Huisman, B., & de Ven, M. (2014). Self and peer assessment in massive open online courses. *International Journal of Higher Education*, *3*(3), 119-128.

Bishop, C. M. (2006). Pattern Recognition and Machine Learning (Information Science and Statistics). Springer–Verlag New York, Inc.

Brooks, S., Gelman, A., Jones, G., & Meng, X. (2011). Handbook of Markov Chain Monte Carlo. Chapman & Hall/CRC Handbooks of Modern Statistical Methods. CRC Press.

Dochy, F., Gijbels, D., & Segers, M. (2006). Learning and the emerging new assessment culture. In Filip Dochy, Lieven Verschaffel, M. B. and Vosniadou, S., editors, Instructional psychology: past, present, and future trends: sixteen essays in honour of Eric de Corte, Advances in Learning and Instruction Series, pages 191–208. Elsevier, Amsterdam; Boston, Array Edition.

Jonsson, A. S. G. (2007). Research review. *Educational Research Review*, *2*(2), 130-144.

Kim, S. (2012). A note on the reliability coefficients for item response model-based ability estimates. *Psychometrika*, *77*(1), 153-162.

Lave, J. & Wenger, E. (1991). Situated Learning. Legitimate Peripheral Participation. Cambridge University Press, New York, Port Chester, Melbourne, Sydney.

Lee, H.-J. & Lim, C. (2012). Peer evaluation in blended team project-based learning: What do students find important? *Educational Technology & Society*, *15*(4), 214-224.

Lord, F. (1980). Applications of item response theory to practical testing problems. Erlbaum Associates.

Lurie, S. J., Nofziger, A. C., Meldrum, S., Mooney, C., & Epstein, R. M. (2006). Effects of rater selection on peer assessment among medical students. *Medical Education*, *40*(11), 1088-1097.

Matteucci, M. & Stracqualursi, L. (2006). Student assessment via graded response model. *Statistica*, *4*, 435-447.

Muraki, E., Hombo, C., & Lee, Y. (2000). Equating and linking of performance assessments. *Applied Psychological Measurement*, *24*(4), 325-337.

Patz, R. J., Junker, B. W., & Johnson, M. S. (1999). The hierarchical rater model for rated test items and its application to large-scale educational assessment data. *Journal of Educational and Behavioral Statistics*, *27*(4), 341-366.

Piech, C., Huang, J., Chen, Z., Do, C., Ng, A., & Koller, D. (2013). Tuned models of peer assessment in MOOCs. In Proceedings of Sixth International Conference of MIT's Learning International Networks Consortium.

Samejima, F. (1994). Estimation of reliability coefficients using the test information function and its modifications. *Applied Psychological Measurement*, (18), 229-244.

Shah, N. B., Bradley, J., Balakrishnan, S., Parekh, A., Ramchandran, K., & Wainwright, M. J. (2014). Some scaling laws for MOOC assessments. In ACM KDD Workshop on Data Mining for Educational Assessment and Feedback (ASSESS).

Sheng, Y. & Wikle, C. K. (2007). Comparing multi-dimensional and uni-dimensional item response theory models. *Educational and Psychological Measurement*, *67*(6), 899-919.

Suen, H. (2014). Peer assessment for massive open online courses (MOOCs).*The International Review of Research in Open and Distributed Learning*, *15*(3), 313-327.

Topping, K. J., Smith, E. F., Swanson, I., & Elliot, A. (2000). Formative Peer Assessment of Academic Writing Between Postgraduate Students. *Assessment & Evaluation in Higher Education*, 25(2), 149-169.

Ueno, M. & Okamoto, T. (2008). Item response theory for peer assessment. In Advanced Learning Technologies, 2008. ICALT '08. Eighth IEEE International Conference on, pages 554-558.

Usami, S. (2010). A polytomous item response model that simultaneously considers bias factors of raters and examinees: Estimation through a Markov chain Monte Carlo algorithm. *The Japanese Journal of Educational Psychology*, *58*(2), 163-175.

Uto, M. & Ueno, M. (2015). Item response model with lower order parameters for peer assessment. In Proceedings of 17th International Conference on Artificial Intelligence in Education, 9112: 800-803.

Wang, Z. & Yao, L. (2007). The effects of rater severity and rater distribution on examinees' ability estimation for constructed-response items. Technical report, ETS Research Report.

Whatley, J. (2012). Evaluation of a team-project-based learning module for developing employability skills. Issues in Informing Science and Information Technology (IISIT), 9, 75-92.