

# Item Response Theory for Peer Assessment

Masaki Uto, and Maomi Ueno, *Member, IEEE*

**Abstract**—As an assessment method based on a constructivist approach, peer assessment has become popular in recent years. However, in peer assessment, a problem remains that reliability depends on the rater characteristics. For this reason, some item response models that incorporate rater parameters have been proposed. Those models are expected to improve the reliability if the model parameters can be estimated accurately. However, when applying them to actual peer assessment, the parameter estimation accuracy would be reduced for the following reasons. 1) The number of rater parameters increases with two or more times the number of raters because the models include higher-dimensional rater parameters. 2) The accuracy of parameter estimation from sparse peer assessment data depends strongly on hand-tuning parameters, called hyperparameters. To solve these problems, this article presents a proposal of a new item response model for peer assessment that incorporates rater parameters to maintain as few rater parameters as possible. Furthermore, this article presents a proposal of a parameter estimation method using a hierarchical Bayes model for the proposed model that can learn the hyperparameters from data. Finally, this article describes the effectiveness of the proposed method using results obtained from a simulation and actual data experiments.

**Index Terms**—Peer assessment, rater characteristics, reliability, item response theory, hierarchical Bayes model

## 1 Introduction

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

- 1) Learners take responsibility for their learning and become autonomous [2], [3], [4].
- 2) Treating assessment as a part of learning, mistakes can come to represent opportunities rather than failures [3].
- 3) Giving rater roles to learners raises their motivation [3], [4].
- 4) Transferable skills such as evaluation skills and discussion skills are practiced [3], [5].
- 5) By evaluating others, raters can learn from others' work, which induces self-reflection [2], [3], [5].
- 6) Learners can receive useful feedback even when they have no instructor [5]. Feedback from other learners who have similar backgrounds is readily understood [2].
- 7) When the learners are mature adults, evaluation by multiple raters is more reliable than that by a single instructor [2].
- 8) Even when the number of learners increases extremely as in massive open online courses, peer assessment can offer feedback for each learner [6], [7].

Therefore, peer assessment has been adopted into various learning processes. In addition, many peer as-

essment support systems have been developed [2], [8], [9], [10], [11], [12], [13], [14].

This article specifically examines the benefit of peer assessment to improve the reliability of assessment for learners' performance, such as essay writing. Although the assessment of learners' performance has become important, it is difficult for a single teacher to assess them when the number of learners increases. Peer assessment enables realization of reliable assessment without burdening a teacher when the number of raters is sufficiently large [2]. However, it is difficult to increase the number of raters for each learner because one rater can only assess a few performances [6], [15]. Therefore, the main issue of this article is to improve the reliability of peer assessment for sparse data. In this article, the reliability is defined as *stability of learners' ability estimation* [16]. The reliability reveals a higher value if the ability of learners is obtainable with few errors when the performance tasks or raters are changed.

The reliability of peer assessment is known to depend on rater characteristics [2], [6], [7], [17], [18]. Therefore, the reliability is expected to be increased if the ability of learners is estimated considering the following rater characteristics [6], [19], [20].

- 1) *Severity*: Because each rater has a different rating severity.
- 2) *Consistency*: Because a rater might not always be consistent in applying the same assessment criteria.

A similar problem has been described in essay testing situations where multiple raters evaluate several essays [21], [22]. To resolve the problem, some item response models have been proposed that incorporate the rater characteristic parameters [19], [23], [24]. For example, Patz et al. [23] proposed a generalized partial credit model (GPCM) [25] that incorporates a rater's severity parameter. Furthermore, Usami [19] has pointed out that

- M. Uto is with the Nagaoka University of Technology, Nagaoka-shi, Niigata, Japan.  
E-mail: uto@oberon.nagaokaut.ac.jp
- M. Ueno is with the University of Electro-Communications, Chofu-shi, Tokyo, Japan.  
E-mail: ueno@ai.is.uec.ac.jp

raters might not always assess performance consistently. Therefore, Usami [19] proposed a GPCM that incorporates rater consistency and severity parameters. The models described above can be regarded as extensions of the multi-facet Rasch model proposed by Linacre [24].

Ueno et al. [2] proposed a graded response model [26] that incorporates a rater's severity parameter for peer assessment. The study also proposed a rating consistency index that is calculable using the severity parameter. Furthermore, an approximate parameter estimation method for the model was proposed. However, higher estimation accuracy would not be obtained using the estimation method.

As another approach, a hierarchical rater model (HRM) has been proposed [21], [27], [28]. The HRM assumes that each learner's work has an ideal rating. The ideal ratings follow a polytomous item response model. Furthermore, the raters' actual ratings are assumed to follow the function of the ideal ratings and rater characteristic parameters.

In previously developed models, the ability of learners can be estimated considering rater characteristics. Therefore, the peer assessment reliability is expected to be improved if the model parameters can be estimated accurately. However, when applying them to actual peer assessment, the parameter estimation accuracy would be reduced for the following reasons.

- 1) In previous models, the number of rater parameters increases with two or more times the number of raters because the models include higher-dimensional rater parameters. The parameter estimation accuracy is known to be reduced when the number of parameters increases because the data size per parameter decreases [29].
- 2) As the parameter estimation method for previous models, Bayes estimation has been generally used. However, the accuracy of Bayes estimation is known to depend strongly on hand-tuning parameters, called hyperparameters, especially when the data are sparse [30]. Peer assessment data usually become sparse because each rater can assess only a few works [6], [15]. Therefore, the accuracy of parameter estimation would be reduced if the hyperparameters were determined arbitrarily, as in previous studies.

To resolve the problems, this article presents a proposal of a new item response model for peer assessment that incorporates rater consistency and severity parameters to maintain as few rater parameters as possible. Furthermore, this article presents a proposal of a parameter estimation method using a hierarchical Bayes model (HBM) for the proposed model that can learn the hyperparameters from data. The proposed method presents the following advantages.

- 1) The proposed model has fewer rater parameters than previous models have. Therefore, the proposed model can provide higher estimation ac-

curacy of the parameters and ability when the number of raters increases.

- 2) The proposed parameter estimation method estimates the hyperparameters from data. Therefore, the accuracy of parameter estimation from sparse peer assessment data is expected to be increased.
- 3) The reliability of peer assessment can be improved because the ability of learners is estimated with higher accuracy and considering the rater's consistency and severity characteristics.

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

## 2 e-Learning Environment

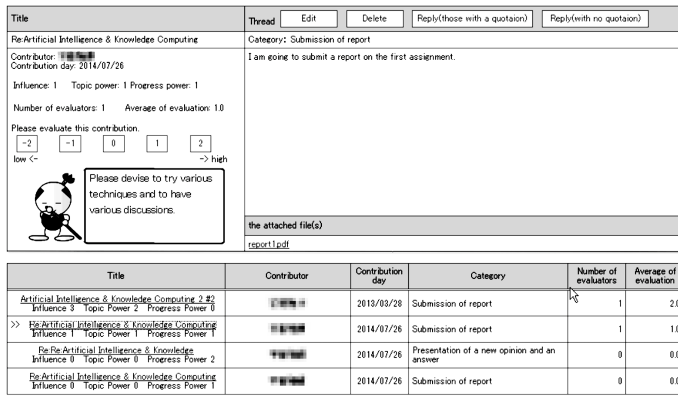
One author has developed a learning management system (LMS) called *Samurai* [31] that is used with huge numbers of e-learning courses. Here we describe the LMS Samurai structure briefly. LMS Samurai presents content sessions tailored for 90-min classes (the units are called *topics*). Learners choose from the array of topics and watch the topic lesson. Fifteen of these content sessions for 90-min classes are produced, constituting a two-unit course. Each content session provides instructional text screens, instructional images, instructional videos, and a practice test. How learners respond to the sessions and how long it takes them to complete the lesson are stored automatically in LMS's learning history database. Those data are analyzed using various data mining techniques. Learning is facilitated by an agent. In addition, LMS Samurai has a discussion board system that enables learners to submit reports, and enables them to assess and discuss one another.

One author offered an e-learning course on *statistics* from 2009 to 2011 using the LMS. This course was taken by 91 learners (32 in 2011, 34 in 2010, and 25 in 2009). In this course, five report assignments were provided and the learners should mutually peer assess their works. The total number of submissions about those assignments in the discussion board was 1554 (412 in 2011, 732 in 2010, and 410 in 2009). The learners actively assessed and provided formative comments for one another.

## 3 Peer Assessment

A main use of peer assessment in learning situations is giving formative comments among learners [1]. Another use of peer assessment is to improve the reliability of assessment for learners' performance, such as essay writing and programming. The assessment of learners' performance has become important because social constructivism, active learning, problem-based learning, and project-based learning have become popular in actual school education [32], [33].

Nevertheless, when the number of learners increases, it is difficult for a single teacher to assess them. Peer assessment enables realization of reliable assessment without burdening the teacher when the number of raters is sufficiently large [2]. However, it is difficult to



| Title  | Contributor | Contribution day | Category                                    | Number of evaluators | Average of evaluation |
|--|-------------|------------------|---|----------------------|-----------------------|
| Artificial Intelligence & Knowledge Computing 2 #2<br>Influence 3 Topic Power 2 Progress Power 0 |             | 2018/02/28       | Submission of report                        | 1                    | 2.0                   |
| Artificial Intelligence & Knowledge Computing<br>Influence 1 Topic Power 1 Progress Power 1      |             | 2014/07/26       | Submission of report                        | 1                    | 1.0                   |
| Artificial Intelligence & Knowledge Computing<br>Influence 0 Topic Power 0 Progress Power 2      |             | 2014/07/26       | Presentation of a new opinion and an answer | 0                    | 0.0                   |
| Artificial Intelligence & Knowledge Computing<br>Influence 0 Topic Power 0 Progress Power 1      |             | 2014/07/26       | Submission of report                        | 0                    | 0.0                   |

Fig. 1. Peer Assessment System.

increase the number of raters for each learner because one rater can only assess a few performances [6], [15]. Therefore, this article specifically examines improvement of the reliability of peer assessment for sparse data.

In addition, peer assessment is justified as an appropriate assessment method because the ability of learners would be defined naturally in the learning community as a social agreement [34].

The ability of learners obtained from peer assessment is generally used for feedback to the learners as grades or numerical scores. Furthermore, they have been used recently for recommending learners' works that obtained high scores [35], predicting rater reliability [15], selecting peer raters for each learner [36], and assigning weights to formative comments [6]. Consequently, the accuracy of peer assessment has become important.

### 3.1 Peer Assessment System

In LMS *Samurai* [31], peer assessment can be conducted using a discussion board system. The system enables learners to post their works and helps other learners to post ratings and comments for the posted works. Fig. 1 portrays a system interface by which a learner submitted a report. The lower half of Fig. 1 presents hyperlinks for other learners' comments. By clicking the hyperlink, detailed comments are displayed in the upper right of Fig. 1. The five buttons shown at the upper left are used for peer assessment. The buttons include  $-2$  (Bad),  $-1$  (Poor),  $0$  (Fair),  $1$  (Good), and  $2$  (Excellent). The learner who submitted the report can take the ratings and comments into consideration and rework it. The averaged rating score of the report is calculated from the peer assessment and stored in the system. This score is used to recommend excellent reports to the other learners in this system. This article attempts to improve the reliability of this rating score.

The rating data  $U$  obtained from the peer assessment system consist of categories  $k$  ( $k = 1, \dots, K$ ) given by each rater  $r$  ( $r = 1, \dots, R$ ) to each work of learner  $j$  ( $j = 1, \dots, J$ ) for each assignment  $i$  ( $i = 1, \dots, I$ ). In this article, the categories of the rating buttons  $[-2, 1, 0, 1, 2]$  are transformed into  $[1, 2, 3, 4, 5]$ . Here, let  $x_{ijr}$  be a response of rater  $r$  to learner  $j$ 's work for assignment

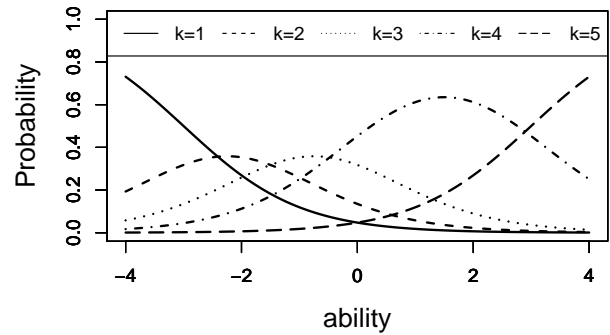


Fig. 2. Item characteristic curves of the graded response model for five categories.

*i*. The data  $U$  are described as

$$U = \{x_{ijr} | x_{ijr} \in \{1, \dots, K\}\} \\ (j = 1, \dots, J, i = 1, \dots, I, r = 1, \dots, R). \quad (1)$$

The data  $U$  consist of three-way data, which are learners  $\times$  raters  $\times$  assignments. This article assumes application of an item response model to the three-way data.

## 4 Item Response Theory

The item response theory (IRT) [37], which is a test theory based on mathematical models, has been used widely with the widespread use of computer testing. Reports of the literature describe that IRT offers the following benefits [2].

- 1) It is possible to assess ability while minimizing the effects of heterogeneous or aberrant items that have low estimation accuracy.
- 2) The learner's responses to different items can be assessed on the same scale.
- 3) Missing data can be estimated easily.

Traditionally, IRT 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 test and report assessment, has been attempted [20], [27], [38].

The following subsections describe the two representative polytomous item response models: the Graded Response Model (GRM) [26] and Generalized Partial Credit Model (GPCM) [25].

### 4.1 Graded Response Model

The GRM gives the probability that learner  $j$  responds in category  $k$  for item  $i$  as follows.

$$P_{ijk} = P_{ijk-1}^* - P_{ijk}^*, \quad (2)$$

$$\begin{cases} P_{ijk}^* = \frac{1}{1 + \exp(-\alpha_i(\theta_j - b_{ik}))} & k = 1, \dots, K-1, \\ P_{ij0}^* = 1, \\ P_{ijK}^* = 0. \end{cases} \quad (3)$$

In those equations,  $K$  represents the number of response categories,  $\alpha_i$  is a discrimination parameter of item  $i$ ,  $b_{ik}$  is a difficulty parameter that denotes the upper grade

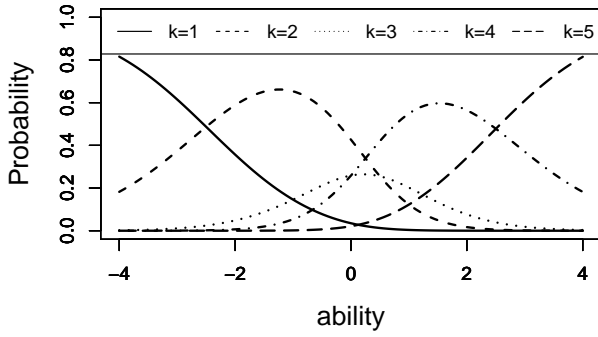


Fig. 3. Item characteristic curves of the generalized partial credit model for five categories.

threshold parameter for category  $k$  of item  $i$ , and  $\theta_j$  is the latent ability of learner  $j$ . Here, the order of the difficulty parameters is restricted by  $b_{i1} < b_{i2} < \dots < b_{iK-1}$ .

Fig. 2 portrays item characteristic curves of the GRM for five categories with  $\alpha_i = 1.0$ ,  $b_{i1} = -3.0$ ,  $b_{i2} = -1.5$ ,  $b_{i3} = 0.0$ ,  $b_{i4} = 3.0$  and  $K = 5$ . Its horizontal axis shows the learner's ability  $\theta_j$ ; the vertical axis shows the probability of the learner's response in each category. It is apparent from Fig. 2 that a learner who has lower ability tends to respond in a lower category. A learner who has a higher ability tends to respond in a higher category.

## 4.2 Generalized Partial Credit Model

The GPCM gives the probability of a response in category  $k$  of item  $i$  as follows.

$$P_{ijk} = \frac{\exp \sum_{m=1}^k [\alpha_i(\theta_j - \beta_{im})]}{\sum_{l=1}^K \exp \sum_{m=1}^l [\alpha_i(\theta_j - \beta_{im})]}, \quad (4)$$

where,  $\beta_{ik}$  is a step difficulty parameter that denotes a difficulty of transition between category  $k-1$  and category  $k$  of item  $i$ . Here,  $\beta_{i1} = 0$  for each  $i$  is given for model identification.

By decomposing the step difficulty parameter  $\beta_{ik}$  to  $\beta_i - d_{ik}$ , the response function of the GPCM is often described as follows.

$$P_{ijk} = \frac{\exp \sum_{m=1}^k [\alpha_i(\theta_j - \beta_i - d_{im})]}{\sum_{l=1}^K \exp \sum_{m=1}^l [\alpha_i(\theta_j - \beta_i - d_{im})]}, \quad (5)$$

where,  $\beta_i$  is called a positional parameter;  $d_{ik}$  is a threshold parameter. Here,  $d_{i1} = 0$  and  $\sum_{k=2}^K d_{ik} = 0$  for each  $i$  are given for model identification.

The partial credit model (PCM) [39] is a special case of the GPCM when  $\alpha_i = 1.0$  for all items. Moreover, the rating scale model [40] is a special case of PCM when  $d_{ik}$  has the same value over all items.

Fig. 3 depicts item characteristic curves of the GPCM for five categories with  $\alpha_i = 1.0$ ,  $\beta_{i2} = -2.5$ ,  $\beta_{i3} = 0.5$ ,  $\beta_{i4} = 0.0$ ,  $\beta_{i5} = 2.5$ , and  $K = 5$ . Its horizontal axis shows the learner's ability  $\theta_j$ ; the vertical axis shows the probability of the learner's response in category  $k$ . A feature of the GPCM is that the step difficulty parameters are not restricted in ascending order, in contrast to the difficulty parameter of the GRM. When the step difficulty parameters in the GPCM are not ordered in

ascending order, some response curves sink under the other curves, such as the category 3 in Fig. 3.

## 4.3 Comparison between GRM and GPCM

Both the GPCM and GRM are applicable to polytomous response data and have item parameters of similar kinds. Several studies that have compared the GPCM with the GRM have reported that the GRM is more useful than the GPCM. Baker et al. [41] applied the GRM and the GPCM to a psychological questionnaire. They have reported that the GRM demonstrated higher goodness of fit to the data and higher reliability than the GPCM. Moreover, Shojima [42] reported that the cases in which the GRM fit the data generated from the GPCM could be observed more frequently than the opposite case. Furthermore, Samejima [43] has proposed four criteria in evaluating polytomous item response models, and claimed that the GRM holds the following two desirable characteristics: 1) additivity and 2) generalizability to a continuous response model.

As the other salient feature, the GRM has less computational complexity of parameter estimation than the GPCM. In the parameter estimation method for the polytomous item response models (e.g., the generally used EM algorithm or the Markov Chain Monte Carlo [44], [45]), the likelihood must be calculated iteratively. The computational complexity of the likelihood in the GRM for one response datum is  $O(1)$ . It is much less than the complexity in the GPCM, which is  $O(k + \sum_{l=1}^K l)$ . The parameter estimation of the GRM is much faster than in the GPCM.

Based on the considerations presented above, employing the GRM is expected to be more desirable than using the GPCM.

## 5 Item Response Models that Incorporate Rater Parameters

As described in Section 3.1, the peer assessment data  $U$  consist of three-way data, which are learners  $\times$  raters  $\times$  assignments. The basic item response models, such as the GRM and the GPCM, are not applicable for the three-way data. To resolve the problem, some item response models that incorporate the rater parameters have been proposed.

### 5.1 GPCM Incorporating Rater Parameters

Patz et al. [23] proposed a rater parameter  $\rho_{ir}$ , which denotes rater  $r$ 's severity for assignment  $i$ . A GPCM that incorporates  $\rho_{ir}$  provides the probability that rater  $r$  responds in category  $k$  to learner  $j$ 's work for assignment  $i$  as follows.

$$P_{ijrk} = \frac{\exp \sum_{m=1}^k [\alpha_i(\theta_j - \beta_{im} - \rho_{ir})]}{\sum_{l=1}^K \exp \sum_{m=1}^l [\alpha_i(\theta_j - \beta_{im} - \rho_{ir})]}, \quad (6)$$

where,  $a_i$  is a discriminant parameter of assignment  $i$ ;  $\beta_{ik}$  is a step difficulty parameter that denotes the difficulty of transition between category  $k-1$  and  $k$  in assignment  $i$ . Here,  $\beta_{i1} = 0$  and  $\rho_{i1} = 0$  for each  $i$  are given for model identification.

Usami [19] has proposed a GPCM that incorporates a rater's consistency and severity parameters to resolve the difficulty that raters might not be consistent. The response probability of the model is defined as described below.

$$P_{ijrk} = \frac{\exp \sum_{m=1}^k [\alpha_i \alpha_r (\theta_j - (\beta_i + \beta_r) - d_{im} d_r)]}{\sum_{l=1}^K \exp \sum_{m=1}^l [\alpha_i \alpha_r (\theta_j - (\beta_i + \beta_r) - d_{im} d_r)]}, \quad (7)$$

where,  $\alpha_r$  reflects the consistency of rater  $r$ ,  $\beta_i$  is a positional parameter of assignment  $i$ ,  $\beta_r$  is a positional parameter of rater  $r$ ,  $d_{ik}$  is a threshold parameter of assignment  $i$  for category  $k$ , and  $d_r$  is a threshold parameter of rater  $r$ . For model identification,  $\prod_r \alpha_r = 1$ ,  $\sum_r \beta_r = 0$ ,  $\prod_r d_r = 1$ ,  $d_{i1} = 0$  and  $\sum_{k=2}^K d_{ik} = 0$  for each  $i$  are given.

The models presented above are regarded as extensions of the multi-facet Rasch model [24]. The multi-facet Rasch model defines the log odds ratio  $\ln(P_{ijrk}/P_{ijrk-1})$  as a linear combination of each facet like  $\theta_j - b_i - \beta_r$ .

## 5.2 GRM that Incorporates Rater Parameters

Ueno et al. [2] proposed a GRM that incorporates the rater's severity parameter for peer assessment. The model gives the response probability as presented below.

$$P_{ijrk} = P_{ijrk-1}^* - P_{ijrk}^*, \quad (8)$$

$$\begin{cases} P_{ijrk}^* = \frac{1}{1 + \exp(-\alpha_i(\theta_j - b_i - \varepsilon_{rk}))} & k = 1, \dots, K-1, \\ P_{ijr0}^* = 1, \\ P_{ijrK}^* = 0. \end{cases}$$

In those expressions,  $b_i$  represents the difficulty of assignment  $i$ ; and  $\varepsilon_{rk}$  denotes the severity of rater  $r$  for category  $k$ . Here,  $\varepsilon_{r1} < \varepsilon_{r2} < \dots < \varepsilon_{rK-1}$ . Additionally,  $\varepsilon_{i1} = -2.0$  is given for model identification.

The study also proposed the following rating consistency index.

$$R_r = \frac{1}{K} \exp\left(-\sum_{k=1}^K P(\hat{\varepsilon}_{rk}) \log P(\hat{\varepsilon}_{rk})\right). \quad (9)$$

Therein,  $P(\hat{\varepsilon}_{rk}) = \frac{1}{1 + \exp(\hat{\varepsilon}_{rk-1})} - \frac{1}{1 + \exp(\hat{\varepsilon}_{rk})}$ . The consistency index reveals a higher value when the severity parameters are distributed in a wide range and at even intervals.

## 5.3 Hierarchical Rater Model

The models described above have been proposed as item response models that directly incorporate the rater parameters. A different modeling called the hierarchical rater model (HRM) has been proposed [21], [27], [28].

The main ideas of HRM are the use of an ideal rating  $\xi_{ij}$  of each work and hierarchical structure data modeling. Concretely, the HRM assumes that learner  $j$ 's work for assignment  $i$  has the ideal rating  $\xi_{ij}$ . Rater  $r$ 's rating  $x_{ijr}$  follows the function of the ideal rating  $\xi_{ij}$  and the rater characteristic parameters. Patz et al. [21] proposed the following HRM.

1) The ideal rating  $\xi_{ij}$  to learner  $j$ 's work for assign-

ment  $i$  is given by the PCM below.

$$p(\xi_{ij} = k | \theta_j, \beta_i, \mathbf{d}_i) = \frac{\exp \sum_{m=1}^k [\theta_j - \beta_i - d_{im}]}{\sum_{l=1}^K \exp \sum_{m=1}^l [\theta_j - \beta_i - d_{im}]}. \quad (10)$$

Here,  $d_{i1} = 0$  and  $\sum_{k=2}^K d_{ik} = 0$  for each  $i$  are given for model identification.

2) Given the ideal rating  $\xi_{ij}$ , the rater  $r$ 's response  $x_{ijr}$  to learner  $j$ 's work for assignment  $i$  is assumed by the following signal detection model [46].

$$p(x_{ijr} = k | \xi_{ij}) \propto \exp\left\{\frac{-k + \xi_{ij} + \sigma_r}{2\psi_r^2}\right\}. \quad (11)$$

Therein,  $\sigma_r$  denotes a rater's severity. The reciprocal of  $\psi_r^2$  denotes a rater's consistency.

DeCarlo et al. [27] proposed another HRM. The model used the following latent class signal detection model [47] instead of the signal detection model.

$$p(x_{ijr} = k | \xi_{ij}) = p(x_{ijr} \geq k-1 | \xi_{ij}) - p(x_{ijr} \geq k | \xi_{ij}), \quad (12)$$

$$\begin{cases} p(x_{ijr} \geq k | \xi_{ij}) = \frac{1}{1 + \exp(d_{rk} - c_r \xi_{ij})} & k = 1, \dots, K-1, \\ p(x_{ijr} \geq 0 | \xi_{ij}) = 1, \\ p(x_{ijr} \geq K | \xi_{ij}) = 0. \end{cases}$$

In the equations presented above,  $c_r$  stands for a rater  $r$ 's consistency. In addition,  $d_{rk}$  signifies rater  $r$ 's severity for category  $k$ . Here,  $d_{r1} < d_{r2} < \dots < d_{rK-1}$ . The latent class signal detection model is regarded as the GRM with a discrete latent variable.

DeCarlo et al. [27] also used the following GPCM instead of the PCM.

$$p(\xi_{ij} = k | \theta_j, \beta_i) = \frac{\exp \sum_{m=1}^k [\alpha_i(\theta_j - \beta_{im})]}{\sum_{l=1}^K \exp \sum_{m=1}^l [\alpha_i(\theta_j - \beta_{im})]}. \quad (13)$$

Here,  $\beta_{i1} = 0$  for each  $i$  is given for model identification.

## 5.4 Other Statistical Models for Peer Assessment

Several statistical models have been used for peer assessment without the item response model [15], [48]. In these models, the generation process of rating data  $x_{ijr}$  is formulated as a normal distribution, which depends on the ideal rating  $\xi_{ij}$  and rater characteristics. However, the models cannot estimate the learner ability because they do not incorporate the learner's ability parameter.

In addition, the generalizability theory [49] has been used widely for analyzing the reliability of an assessment with multiple raters. The generalizability theory enables estimation of the reliability of a performance assessment, including expert and peer assessment, and enables analysis of the influence of the raters and assignments on the reliability. Moreover, Longford [50] proposed an extended framework of the generalizability theory for analyzing each rater's characteristics. However, these methods do not estimate the ability of learners directly considering the characteristics of raters and assignments.

Therefore, we are not concerned with these models and methods in this article.

## 5.5 Problems of the Previous Models

In the previous models, the ability of learners can be estimated considering rater characteristics. Therefore, the

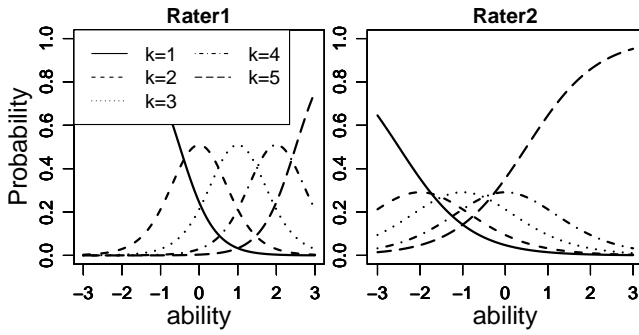


Fig. 4. Item characteristic curves of two raters.

reliability of peer assessment is expected to be improved if the model parameters can be estimated accurately. However, in the previous models, the number of rater parameters increases with two or more times the number of raters because the models include higher-dimensional rater parameters. The parameter estimation accuracy is known to be reduced when the number of parameters increases because the data size per parameter decreases [29]. In peer assessment, the number of raters increases concomitantly with an increasing number of learners. Therefore, the parameter estimation accuracy for the previous models would be reduced when applying them to actual peer assessment.

To solve the problems, this article presents a proposal of a new item response model for peer assessment. The proposed model incorporates a rater's consistency and severity parameters to maintain as few rater parameters as possible.

## 6 Proposed Model

This article presents a proposal of an item response model for peer assessment by extending the GRM as follows.

$$P_{ijrk} = P_{ijrk-1}^* - P_{ijrk}^*, \quad (14)$$

$$\begin{cases} P_{ijrk}^* = \frac{1}{1 + \exp(-\alpha_i \alpha_r (\theta_j - b_{ik} - \varepsilon_r))} & k = 1, \dots, K-1, \\ P_{ijr0}^* = 1, \\ P_{ijrK}^* = 0. \end{cases}$$

In those equations,  $b_{ik}$  denotes the difficulty in obtaining the score  $k$  for assignment  $i$  (here  $b_{i1} < b_{i2} < \dots < b_{iK-1}$ ), and  $\varepsilon_r$  represents the severity of rater  $r$ . Here,  $\alpha_{r=1} = 1$  and  $\varepsilon_1 = 0$  are assumed for model identification.

For explanation of the proposed rater parameters, Fig. 4 shows item characteristic curves of two raters with the assignment parameters  $\alpha_i = 1.5$ ,  $b_{i1} = -1.5$ ,  $b_{i2} = -0.5$ ,  $b_{i3} = 0.5$ , and  $b_{i4} = 1.5$ . The left panel shows the item characteristic curves of *Rater 1* who has  $\alpha_r = 1.5$  and  $\varepsilon_r = 1.0$ . The right panel shows the item characteristic curves of *Rater 2*, who has  $\alpha_r = 0.8$  and  $\varepsilon_r = -1.0$ . Fig. 4 presents a graph with the horizontal axis showing a learner's ability  $\theta_j$ . The vertical axis shows the rating probability in each category.

Fig. 4 shows that *Rater 1*, who has a higher consistency, can distinguish a learner's ability more accurately. Additionally, it is apparent that the item characteristic curves

TABLE 1  
Number of parameters in each model.

|             | Number of parameters    |
|-------------|-------------------------|
| Proposed    | $IK + 2(R - 1) + J$     |
| Patz1999    | $I(K + R - 1) + J$      |
| Usami2010   | $IK + 3(R - 1) + J$     |
| Ueno2008    | $2I + R(K - 1) - 1 + J$ |
| HRM-Patz    | $I(K - 1 + J) + 2R + J$ |
| HRM-DeCarlo | $I(K + J) + RK + J$     |

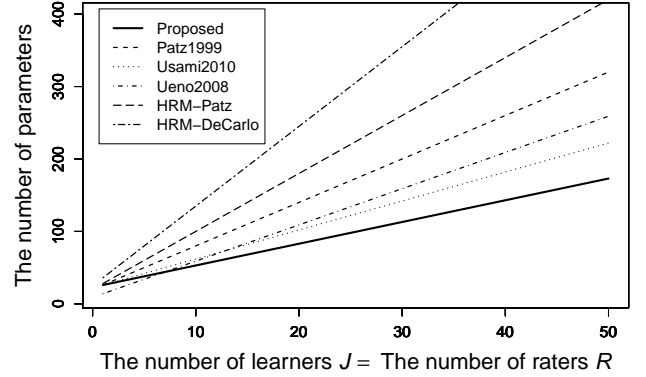


Fig. 5. Relations between the parameter number and the number of raters = learners for each model.

of *Rater 1* shifted to the right compared to those of *Rater 2*. Therefore, a higher ability is necessary to obtain a score from *Rater 1* than to obtain the same score from *Rater 2*.

### 6.1 Reducing the Number of Parameters

The unique feature of the proposed model is that each rater has only one consistency and severity parameter. Consequently, when the number of raters increases, the number of rater parameters in the proposed model increases more slowly than those in the models with higher dimensional rater parameters, such as Ueno et al. [2] and Patz et al. [23].

Table 1 presents the number of parameters in the proposed model and in the previous models. In Table 1, *Patz1999* denotes equation (6), *Usami2010* denotes (7), *Ueno2008* denotes (8), *HRM-Patz* denotes the combination of (10) and (11), and *HRM-DeCarlo* denotes the combination of (12) and (13).

According to Table 1, it is apparent that the proposed model has the minimum number of parameters when  $2R + 1 > 3I$ ,  $I > 2$ , and  $K = 5$ . The conditions are generally fulfilled because the number of raters is fundamentally greater than the number of assignments in peer assessment.

Here, Fig. 5 shows relations between the number of parameters and the number of raters  $R =$  learners  $J$  given  $K = 5$  and  $I = 5$ . The Fig. 5 horizontal axis shows  $R = J$ ; the vertical axis shows the number of parameters. Although the assumption of  $R = J$  is a strict restriction, this article assumes the most difficult condition to estimate the rater parameters in peer assessment.

According to Fig. 5, the proposed model has the minimum number of parameters when the number of raters = learners is large. In contrast, Ueno2008 has

the minimum number of parameters when the number of raters = learners is small. The horizontal value of the intersection point between the proposed model and Ueno2008 in Fig. 5 approaches zero when the number of assignments  $I$  or categories  $K$  decreases.

The parameters in the proposed model are fewer than in previous models as the raters and learners become more numerous. The accuracy of parameter estimation for a model, which has fewer parameters, is known to be generally higher because the model has a greater number of data per parameter [29]. Consequently, the proposed model can realize higher estimation accuracy than previous models if the suitability of the proposed model for peer assessment data is the same as those of the earlier models.

The following GPCM extension model, which has the same number of parameters as the proposed model, is possible.

$$P_{ijrk} = \frac{\exp \sum_{m=1}^k [\alpha_i \alpha_r (\theta_j - \beta_{im} - \rho_r)]}{\sum_{l=1}^K \exp \sum_{m=1}^l [\alpha_i \alpha_r (\theta_j - \beta_{im} - \rho_r)]} \quad (15)$$

Here,  $\rho_r$  denotes the severity of rater  $r$ . For model identification,  $\alpha_{r=1} = 1$ ,  $\rho_1 = 0$  and  $\beta_{i1} = 0$  for each  $i$  are given. However, this article proposed the GRM extension model because GRM is known to provide higher performance than GPCM, as described in Section 4.3.

## 6.2 Improving the Reliability

The other feature of the proposed model is introducing the rater's consistency parameter. Patz1999 and Ueno2008 use no consistency parameters. However, the reliability of peer assessment is known to be reduced if the learner's ability is estimated ignoring the rater's consistency and severity [19] [51]. Therefore, to obtain higher reliability, consideration of the rater's consistency is necessary. Usami [19] demonstrated that parameter  $\alpha_r$  used in Usami2010 can optimally represent the rater's consistency. Therefore, the parameter is used for this study.

In summary, the proposed model is expected to improve the reliability of peer assessment because the ability of learners can be estimated with higher accuracy and can be considered with the rater's consistency and severity characteristics.

However, if an extremely large rating data for each learner is obtainable, then models with higher dimensional parameter (e.g., a model incorporating the interaction among assignment, rater and learner) might realize higher reliability than the proposed model. As described in Section 3, collecting large rating data for each learner is generally difficult in actual situations [6], [15].

Additionally, it is noteworthy that the proposed model does not consider the learner's ability change in the process of peer assessment. The proposed model is assumed to be applied to peer assessment data collected during a short period, in which major ability change does not occur.

## 7 Parameter Estimation

To estimate the parameters in item response models, several previous studies used Bayes estimation. In Bayes estimation, parameters are regarded as random variables. Prior distributions are assumed for each parameter. The prior distributions reflect the uncertainty of the parameters before observing the data. The parameters in the prior distributions, called *hyperparameters*, are determined arbitrarily as reflecting an analyst's subjectivity.

Letting the set of parameters be  $\theta = \{\theta_1, \dots, \theta_J\}$ ,  $\alpha_i = \{\log \alpha_{i=1}, \dots, \log \alpha_{i=I}\}$ ,  $\mathbf{b} = \{b_{11}, \dots, b_{IK-1}\}$ ,  $\alpha_r = \{\log \alpha_{r=1}, \dots, \log \alpha_{r=R}\}$  and  $\varepsilon = \{\varepsilon_1, \dots, \varepsilon_R\}$ . Furthermore,  $g(\theta_j|\tau_\theta)$ ,  $g(\alpha_i|\tau_{\alpha_i})$ ,  $g(b_{ik}|\tau_b)$ ,  $g(\alpha_r|\tau_{\alpha_r})$  and  $g(\varepsilon_r|\tau_\varepsilon)$  denote the prior distributions. Here,  $\tau_\theta$ ,  $\tau_{\alpha_i}$ ,  $\tau_b$ ,  $\tau_{\alpha_r}$ , and  $\tau_\varepsilon$  are the hyperparameters. Then, the posterior distribution of the proposed model is described as follows.

$$g(\theta, \alpha_i, \mathbf{b}, \alpha_r, \varepsilon, |U) \propto L(U|\theta, \alpha_i, \mathbf{b}, \alpha_r, \varepsilon) g(\theta|\tau_\theta)g(\alpha_i|\tau_{\alpha_i})g(\mathbf{b}|\tau_b)g(\alpha_r|\tau_{\alpha_r})g(\varepsilon|\tau_\varepsilon). \quad (16)$$

Therein,

$$L(U|\theta, \alpha_i, \mathbf{b}, \alpha_r, \varepsilon) = \prod_{j=1}^J \prod_{i=1}^I \prod_{r=1}^R \prod_{k=1}^K (P_{ijrk})^{z_{ijrk}}, \quad (17)$$

$$z_{ijrk} = \begin{cases} 1 : x_{ijr} = k, \\ 0 : \text{otherwise.} \end{cases} \quad (18)$$

As the priors on  $\log \alpha_i$ ,  $\log \alpha_r$ ,  $\varepsilon_r$  and  $\theta_j$ , normal distributions are generally assumed. For example, the prior on  $\log \alpha_i$  is described as

$$\log \alpha_i \sim N(\mu_{\alpha_i}, \sigma_{\alpha_i}), \quad (19)$$

where  $N(\mu_{\alpha_i}, \sigma_{\alpha_i})$  denotes the normal distribution with mean  $\mu_{\alpha_i}$  and standard deviation  $\sigma_{\alpha_i}$ .

Here, the scale of  $\theta_j$  must be fixed for model identification. In this article, the standard normal distribution  $N(0, 1)$  is assumed as the scale of  $\theta_j$ . Therefore, the hyperparameters  $\tau_\theta = \{\mu_\theta, \sigma_\theta\}$  in the prior  $g(\theta_j|\tau_\theta)$  are fixed as  $\{0, 1\}$ . In the following sections, the notation  $g(\theta_j)$  is used instead of  $g(\theta_j|\tau_\theta)$  to represent that the hyperparameters  $\tau_\theta$  are fixed.

For the prior on  $\mathbf{b}$ , the multivariate normal distribution  $MN(\mu_b, \Sigma_b)$  is assumed. Here,  $\mu_b$  is a  $K$  dimensional mean vector;  $\Sigma_b$  is a covariance matrix.

In Bayes estimation, the point estimation of each parameter is generally provided as the expected value of the marginal posterior distribution [29], [30]. It is called *the expected a posteriori (EAP) estimate*. For example, the EAP estimates of  $\theta_0$  can be provided as the expectation of the marginal posterior distribution  $g(\theta_0|U)$ , where  $g(\theta_0|U)$  is calculated by marginalizing all parameters except  $\theta_0$  from the posterior distribution (16).

### 7.1 Hierarchical Bayes Model

The EAP estimation generally provides more robust estimation than the maximum likelihood estimation or maximum a posteriori (MAP) estimation [29], [44]. However, the accuracy of the Bayes estimation is known to depend on hyperparameters, especially when the data are sparse [30]. In peer assessment, gathering the large

amount of data is generally difficult because a rater can only evaluate a few works [6], [15]. Therefore, the estimation accuracy would be reduced if the hyperparameters were determined arbitrarily, as in previous studies. To solve the problem, this article presents a proposal of a parameter estimation method using a hierarchical Bayes model (HBM) for the proposed model. In this method, the hyperparameters can be learned from data in the parameter estimation process.

In the HBM, the hyperparameters are also regarded as random variables. Prior distributions are assumed for each hyperparameter. Therefore, the posterior distribution of the proposed model is described as follows.

$$g(\boldsymbol{\theta}, \boldsymbol{\alpha}_i, \tau_{\alpha_i}, \mathbf{b}, \tau_b, \boldsymbol{\alpha}_r, \tau_{\alpha_r}, \boldsymbol{\varepsilon}, \tau_\varepsilon | \mathbf{U}) \\ \propto L(\mathbf{U} | \boldsymbol{\theta}, \boldsymbol{\alpha}_i, \mathbf{b}, \boldsymbol{\alpha}_r, \boldsymbol{\varepsilon}) g(\boldsymbol{\theta}) g(\boldsymbol{\alpha}_i | \tau_{\alpha_i}) g(\tau_{\alpha_i}) \\ g(\mathbf{b} | \tau_b) g(\tau_b) g(\boldsymbol{\alpha}_r | \tau_{\alpha_r}) g(\tau_{\alpha_r}) g(\boldsymbol{\varepsilon} | \tau_\varepsilon) g(\tau_\varepsilon). \quad (20)$$

Here,  $g(\tau_{\alpha_i})$ ,  $g(\tau_b)$ ,  $g(\tau_{\alpha_r})$  and  $g(\tau_\varepsilon)$  denote prior distributions on the hyperparameters. We designate the priors as *hyperpriors*.

The conjugate priors are used as the hyperpriors. In this article, normal distributions are used as the priors on  $\log \alpha_i$ ,  $\log \alpha_r$  and  $\varepsilon_r$ . The conjugate prior on the mean of a normal distribution is a normal distribution  $N(\mu_0, \sigma_0)$ . The conjugate prior on the variance is an inverse gamma distribution  $IG(g_1, g_2)$ . Here,  $g_1$  is called a sharpness parameter;  $g_2$  is a scale parameter.

The conjugate prior on the mean vector  $\boldsymbol{\mu}_b$  is a multivariate normal distribution  $MN(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$ , and the prior on covariance  $\boldsymbol{\Sigma}_b$  is an inverse-Wishart distribution  $IW(\nu, \boldsymbol{\Sigma})$  which has the scale matrix  $\boldsymbol{\Sigma}$  and degrees of freedom  $\nu_0 \geq K$ .

## 7.2 MCMC

In the EAP estimation, the marginal posterior distributions must be calculated. However, when the models are complicated, such as the proposed model, it is generally impossible to derive the marginal posterior distribution analytically or to calculate it using a numerical analytical method such as the Gaussian quadrature integral because of a high-dimensional multiple integral. To resolve the problem, the Markov Chain Monte Carlo method (MCMC), which is a random-sampling-based estimation method, has been proposed. The MCMC effectiveness has been demonstrated in various fields [29] [52]. In item response theory, the MCMC has been used especially with complicated models such as the hierarchical Bayes IRT, multidimensional IRT and multilevel IRT [30].

One shortcoming of MCMC is the computational load. Although the EAP estimation using MCMC generally provides robust estimation as described in Subsection 7.1, the MCMC algorithm might not be feasible if the data are extremely large. When the data become large, other estimation methods which have asymptotic consistency, such as the MAP estimation and maximum marginal likelihood (MML) estimation, would also provide accurate parameter estimation. The MAP and MML estimation using Newton–Raphson method can

be solved with lower computational cost than that of MCMC. Therefore, the MAP or MML estimation using Newton–Raphson method might be preferred for extremely large data. However, the extremely large data are not assumed in this article because increasing the number of raters for each learner is difficult in actual peer assessment. Such sparse data justify the use of MCMC estimation.

The fundamental idea of MCMC is to define a Markov chain  $M_0, M_1, M_2, \dots$  with states  $M_t = (\boldsymbol{\theta}^t, \boldsymbol{\alpha}_i^t, \mathbf{b}^t, \boldsymbol{\alpha}_r^t, \boldsymbol{\varepsilon}^t)$ ; then to simulate observations from the Markov chain.

As a MCMC algorithm for the item response theory, Patz et al. [23] proposed the Metropolis Hastings within Gibbs sampling method (Gibbs/MH).

Based on the Gibbs/MH method, the procedures of the parameter estimation using HBM for the proposed model can be formulated as presented below.

1) Sample  $\boldsymbol{\theta}^t$  as follows.

- a) Draw each  $\theta_j^t \sim h(\theta_j^t | \theta_j^{t-1})$  independently for each  $j=1, 2, \dots, J$ . As the proposal distribution  $h(\theta_j^t | \theta_j^{t-1})$ , the following normal distribution  $N(\theta_j^{t-1}, \sigma_p)$  is used.

$$h(\theta_j^t | \theta_j^{t-1}) = \frac{1}{\sigma_p \sqrt{2\pi}} \exp \left[ -\frac{(\theta_j^t - \theta_j^{t-1})^2}{2\sigma_p^2} \right]. \quad (21)$$

The standard deviation  $\sigma_p$  of the proposal distribution is a smaller value than that of the prior distribution, such as 0.01.

- b) Calculate the following acceptance probability.

$$a(\theta_j^t | \theta_j^{t-1}) \\ = \min \left( \frac{L(\mathbf{U}_j | \theta_j^t, \boldsymbol{\theta}_{-j}^{t-1}, \boldsymbol{\xi}^{t-1}) g(\theta_j^t)}{L(\mathbf{U}_j | \boldsymbol{\theta}^{t-1}, \boldsymbol{\xi}^{t-1}) g(\theta_j^{t-1})}, 1 \right), \quad (22)$$

where  $\boldsymbol{\xi}^t = \{\boldsymbol{\alpha}_i^t, \mathbf{b}^t, \boldsymbol{\alpha}_r^t, \boldsymbol{\varepsilon}^t\}$ ,  $\boldsymbol{\theta}_{-j}^t = \{\boldsymbol{\theta}^t \setminus \theta_j^t\}$  and

$$L(\mathbf{U}_j | \theta_j^t, \boldsymbol{\theta}_{-j}^{t-1}, \boldsymbol{\xi}^{t-1}) \\ = \prod_{i=1}^I \prod_{r=1}^R \prod_{k=1}^K p(x_{ijrk} = k | \theta_j^t, \boldsymbol{\theta}_{-j}^{t-1}, \boldsymbol{\xi}^{t-1})^{z_{ijrk}}.$$

- c) Accept  $\theta_j^t$  with probability  $a(\theta_j^t | \theta_j^{t-1})$ , otherwise let  $\theta_j^t = \theta_j^{t-1}$ .

- 2) Sample each  $\boldsymbol{\alpha}_i^t$ ,  $\mathbf{b}^t$ ,  $\boldsymbol{\alpha}_r^t$ ,  $\boldsymbol{\varepsilon}^t$ , using the same procedure of 1). Here, to restrict the order of the difficulty parameter  $b_{ik}$ , the acceptance probability must be 0 if a drawn sample  $\mathbf{b}^t$  does not satisfy the order restriction.

- 3) The hyperparameters for  $\log \alpha_i$ , namely  $\mu_{\alpha_i}$ ,  $\sigma_{\alpha_i}$ , are drawn from the conditional probability distribution  $p(\mu_{\alpha_i}, \sigma_{\alpha_i} | \boldsymbol{\alpha}_i^t)$ . Concretely,

$$\mu_{\alpha_i} | \sigma_{\alpha_i}, \boldsymbol{\alpha}_i^t \sim N \left( \frac{I_0 \mu_0 + I \bar{\alpha}_i}{I + I_0}, \frac{\sigma_{\alpha_i}^2}{I + I_0} \right), \quad (23)$$

$$\sigma_{\alpha_i}^2 | \boldsymbol{\alpha}_i^t \sim IG(g_1 + \frac{I}{2}, \sigma_n^2), \quad (24)$$

where  $\bar{\alpha}_i = \sum_i \log \alpha_i^t / I$ ,  $\sigma_n = g_2 + \frac{\sum_i (\log \alpha_i - \bar{\alpha}_i)^2}{2} + \frac{I I_0 (\bar{\alpha}_i - \mu_0)}{I + I_0}$  and  $I_0$  is a small positive value. For further details related to the derivation of (23) and (24), see [30], [53]. To obtain random samples



from an inverse gamma distribution, the sampling algorithm proposed by [54] is useful.

- 4) The hyperparameters for  $\mathbf{b}^t$ ,  $\alpha_r^t$  and  $\epsilon^t$  are updated similarly. Samples of hyperparameters  $\mu_b$  and  $\Sigma_b$  are drawn from the following distributions.

$$\mu_b | \Sigma_b, \mathbf{b}^t \sim MN\left(\frac{I_0 \mu_0 + I \bar{\mathbf{b}}}{I + I_0}, \frac{\Sigma_b}{I + I_0}\right), \quad (25)$$

$$\Sigma_b | \mathbf{b}^t \sim IW(I + \nu, \Sigma^{*-1}). \quad (26)$$

Here,  $\Sigma^* = \Sigma_0 + \sum_i (\mathbf{b}_i - \bar{\mathbf{b}})(\mathbf{b}_i - \bar{\mathbf{b}})^t + \frac{I I_0}{I + I_0} (\bar{\mathbf{b}} - \mu_0)(\bar{\mathbf{b}} - \mu_0)^t$ , and  $\bar{\mathbf{b}} = \sum_i \mathbf{b}_i / I$ . For further details related to deriving the posterior, see [30], [53].

- 5) Repeat the procedures described above.

The EAP estimation is given by calculating the mean of the samples generated from the chain. The samples before a burn-in are discarded because the first samples tend to depend on the initial values. The pseudo-code of the MCMC algorithm for the proposed model is summarized in Algorithm 1.

---

**Algorithm 1** MCMC algorithm for the proposed model.

---

**Given** maximum chain length  $T$ , burn-in period  $B$ , interval  $E$ .

**Initialize** array for MCMC sample  $\mathbf{A} = \{\}$ .

**Initialize** all parameters  $\theta^0, \alpha_i^0, \mathbf{b}^0, \alpha_r^0, \epsilon^0$ , and hyperparameters for the priors on  $\alpha_i, \mathbf{b}, \alpha_r, \epsilon$ .

**for**  $t = 1$  to  $T$  **do**

**for each**  $\omega \in \{\{\theta, \alpha_i, \mathbf{b}, \alpha_r, \epsilon\} \setminus \{\alpha_{r=1}, \epsilon_1\}\}$  **do**

    Sample  $\omega^t \sim N(\omega^{t-1}, \sigma_p)$ .

    Accept  $\omega^t$  with probability  $\alpha(\omega^t, \omega^{t-1})$ .

**end for**

**for each hyperparameter**  $h$  **do**

    Set  $h \leftarrow h^{new}$  drawn from (23)(24)(25)(26).

**end for**

**if**  $t \geq B$  and  $t \% E = 0$  **then**

    Add  $\{\theta^t, \alpha_i^t, \mathbf{b}^t, \alpha_r^t, \epsilon^t\}$  to  $\mathbf{A}$ .

**end if**

**end for**

**return** Averaged value of  $\mathbf{A}$

---

## 8 Simulation Experiment

To evaluate the parameter estimation accuracy of the proposed model and the previous models, the following simulation experiment was conducted.

- 1) In the proposed model, Patz1999, Usami2010, Ueno2008, HRM-Patz, HRM-DeCarlo and *Expanded GPCM* that denotes (15), the true parameter values were generated randomly from the distributions in Table 2. In Table 2,  $LN(\mu, \sigma)$  denotes a lognormal distribution with mean  $\mu$  and standard deviation  $\sigma$ .
- 2) Data were sampled randomly given  $I = 5$ ,  $K = 5$ ,  $R = J = 5, 10, 20, 50$  and the generated parameters in procedure 1).
- 3) Using the data, the parameters were estimated using MCMC. Here, the parameter estimation of the proposed model was conducted in the following settings.

TABLE 2

True priors used for the simulation experiment.

|   |
|---|
| $\log \alpha_i \sim N(0.1, 0.4)$<br>$\log \alpha_r, \log c_r, \beta_i, \beta_r, b_i \sim N(0.0, 0.5)$<br>$\epsilon_r, \rho_{ir}, \beta_{ik}, d_{ik}, d_r, \sigma_r, \theta \sim N(0.0, 1.0)$<br>$\psi_r \sim LN(0.4, 0.2)$<br>$\mathbf{b}_{ik}, \epsilon_{rk}, \mathbf{d}_{rk} \sim MN(\mu_b, \Sigma_b)$<br>$\mu = \{-2.00, -0.75, 0.75, 2.00\}$<br>$\Sigma = \begin{pmatrix} 0.16 & 0.10 & 0.04 & 0.04 \\ 0.10 & 0.16 & 0.04 & 0.04 \\ 0.04 & 0.04 & 0.16 & 0.10 \\ 0.04 & 0.04 & 0.10 & 0.16 \end{pmatrix}$ |
|---|

- a) The true hyperparameters in Table 2 were used.
  - b) The hyperparameters were learned using HBM.
  - c) For the prior on  $\log \alpha_i$ ,  $\log \alpha_r$  and  $\epsilon_r$ , the true variance  $\times 2$  and true mean were used. Moreover, for the prior on  $b_{ik}$ , the true covariance  $\times 2$  and true mean vector were used.
  - d) The hyperparameters were generated randomly from the following procedure. Let  $\mu_\tau$  and  $\sigma_\tau$  be the true mean and variance in the prior on  $\log \alpha_i$ ,  $\log \alpha_r$ ,  $\epsilon_r$ . Let  $\mu_b^*$  and  $\Sigma_b^*$  be the true mean vector and covariance matrix in the prior on  $\mathbf{b}$ . Here, the means and variances for the prior on  $\log \alpha_i$ ,  $\log \alpha_r$ ,  $\epsilon_r$  were selected randomly from  $N(\mu_\tau, 0.5)$  and  $LN(\sigma_\tau, 0.5)$ . The mean vector for the prior on  $b_{ik}$  was selected from  $MN(\mu_b^*, \Sigma_b^*)$ . The covariance matrix was selected from  $z \Sigma_b^*$ , where  $z \sim uniform(0, 2)$ .
- However, in the models aside from the proposed model, the true hyperparameters were given. Here, the standard deviation of the proposal distribution used for MCMC was 0.01. The burn-in period was 30,000. The EAP estimates were calculated as the mean of the samples obtained from 30,000 period to 50,000 period at intervals of 1000.
- 4) The root mean square deviations (RMSEs) between the estimated parameters and the true parameters were calculated.
  - 5) After repeating the procedure described above 20 times, the average and standard deviation of the RMSE values were calculated.

In this experiment, all models can estimate the parameters and abilities with high accuracy if sufficient rating data exist for each learner. However, as described in Section 3, it is generally difficult to increase the number of raters for each learner. In practice, each rater can assess, at most, several dozen works. The main purpose of these experiments is to evaluate the estimation accuracy of the parameters and ability when several dozen  $J = R$  are given.

### 8.1 Accuracy of Parameter Estimation

Table 3 shows the RMSE of the rater and assignment parameter estimation.

According to results obtained using the proposed model with the true and wrong hyperparameters, the estimation accuracy depended on the hyperparameters.

TABLE 3  
Parameter and ability estimation accuracy of each model.

|                                    | rater and assignment parameters |              |              |              | ability     |              |              |              |
|------------------------------------|---------------------------------|--------------|--------------|--------------|-------------|--------------|--------------|--------------|
|                                    | $J = R = 5$                     | $J = R = 10$ | $J = R = 20$ | $J = R = 50$ | $J = R = 5$ | $J = R = 10$ | $J = R = 20$ | $J = R = 50$ |
| Proposed model                     |                                 |              |              |              |             |              |              |              |
| with true hyperparameters          | .233 (.030)                     | .156 (.023)  | .126 (.012)  | .087 (.014)  | .285 (.061) | .182 (.064)  | .144 (.037)  | .104 (.028)  |
| with learned hyperparameters       | .248 (.030)                     | .172 (.027)  | .148 (.019)  | .109 (.023)  | .297 (.073) | .189 (.050)  | .152 (.034)  | .112 (.037)  |
| with $2\sigma_r^2$ and $2\Sigma_b$ | .255 (.036)                     | .199 (.033)  | .162 (.023)  | .134 (.021)  | .317 (.063) | .220 (.061)  | .193 (.046)  | .165 (.041)  |
| with random hyperparameters        | .421 (.097)                     | .330 (.077)  | .297 (.076)  | .269 (.084)  | .355 (.154) | .337 (.125)  | .287 (.090)  | .251 (.092)  |
| Previous models                    |                                 |              |              |              |             |              |              |              |
| Patz1999                           | .449 (.044)                     | .341 (.041)  | .240 (.025)  | .157 (.019)  | .337 (.090) | .238 (.078)  | .200 (.046)  | .163 (.050)  |
| Usami2010                          | .405 (.064)                     | .275 (.055)  | .192 (.025)  | .166 (.016)  | .322 (.094) | .237 (.075)  | .191 (.063)  | .159 (.032)  |
| Ueno2008                           | .229 (.034)                     | .186 (.022)  | .162 (.017)  | .128 (.013)  | .273 (.051) | .226 (.050)  | .200 (.058)  | .138 (.050)  |
| HRM-Patz                           | .620 (.055)                     | .476 (.029)  | .294 (.022)  | .195 (.024)  | .461 (.095) | .447 (.109)  | .372 (.057)  | .361 (.030)  |
| HRM-DeCarlo                        | .595 (.264)                     | .592 (.271)  | .565 (.239)  | .533 (.220)  | .926 (.189) | .804 (.199)  | .772 (.112)  | .715 (.058)  |
| Expanded GPCM                      | .324 (.034)                     | .225 (.037)  | .155 (.036)  | .123 (.031)  | .308 (.085) | .239 (.076)  | .179 (.064)  | .132 (.044)  |

\* Shaded cells in the table represent minimum values.

Furthermore, the proposed model using HBM revealed the closest accuracy using the true hyperparameters.

According to the results of each model with the true hyperparameters, the proposed model revealed the minimum RMSE in all cases except for Ueno2008 with  $J = R = 5$ . Ueno2008 had the minimum RMSE because it has the minimum number of parameters when  $J = R = 5$ .

From these results, it is apparent that the proposed model realizes higher accuracy of parameter estimation than the other models when the number of raters increases. Furthermore, parameter estimation using HBM is expected to provide higher performance in practice because the true hyperparameters are practically unknown.

According to the result, the RMSEs of the proposed model were lower than those of the expanded GPCM in all cases, although these two models have the same number of parameters. In this experiment, the true and prior distributions for the category parameters  $b_{ik}$  in the proposed model have smaller variance than  $\beta_{ik}$  in the expanded GPCM. This article selected the distributions on  $b_{ik}$  to represent the ascending order restriction of the parameters. As a result, the RMSE of the category parameters in the proposed model tends to be lower because both the true and estimated values of the parameters are distributed within a smaller range than those of the parameters in the expanded GPCM. The estimates of all the parameters and ability are mutually dependent. Therefore, the RMSE of the proposed model tends to be lower than that of the expanded GPCM.

## 8.2 Accuracy of Ability Estimation

Table 3 shows the RMSE of the learner's ability estimation.

Comparing the ability estimation accuracy and the parameter estimation accuracy in Table 3, a similar tendency of the parameter estimations can be confirmed. Concretely,

- 1) the proposed model provided higher accuracy of ability estimation than the other models when the number of raters increases,
- 2) the ability estimation accuracy depends on the hyperparameters,
- 3) the ability estimation using HBM provides the closest accuracy using true hyperparameters.

The results showed that if the proposed model is

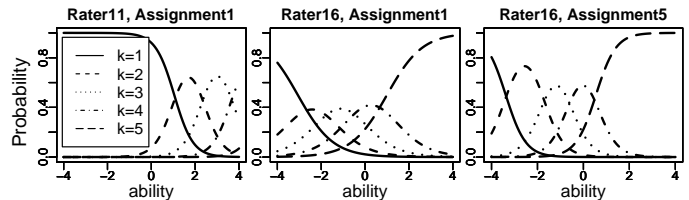


Fig. 6. Item characteristic curves of the proposed model estimated using actual data.

suitable for given data, then the proposed model can estimate the ability with the highest accuracy even if the raters and assignments are changed. Therefore, the proposed model is expected to realize the highest reliability of peer assessment if the model is suitable for peer assessment data.

## 9 Actual Data Experiment

Actual data experiments were conducted to evaluate the suitability and reliability of the proposed model for an actual peer assessment.

### 9.1 Actual Data

The actual data were gathered using the following procedures.

- 1) 20 learners' reports for 5 assignments were collected from an e-learning course offered from 2009 to 2011 on *statistics* as described in Section 2. The 20 learners were selected randomly from the learners who submitted all 5 report assignments. The details were 8 learners from 2009, 8 learners from 2010, and 4 learners from 2011.
- 2) The 20 learners' reports for 5 assignments were evaluated by 20 other raters who had attended the same e-learning course. The raters rated the reports using the 5 categories based on a rubric that the author offered.

### 9.2 Example of Parameter Estimation

This subsection presents a parameter estimation example in the proposed model using actual data.

Parameter estimation using HBM was conducted by the MCMC using the same procedure as the simulation experiment. Table 4 presents the estimated parameters and hyperparameters. Furthermore, Fig. 6 depicts item characteristic curves of the *Rater 11* for the *Assignment 1* and *16* for the *Assignment 1* and *5*.

TABLE 4  
Estimated parameters and hyperparameters.

|              | $\hat{\alpha}_i$ | $\hat{b}_{i1}$ | $\hat{b}_{i2}$ | $\hat{b}_{i3}$ | $\hat{b}_{i4}$ |
|--------------|------------------|----------------|----------------|----------------|----------------|
| Assignment 1 | 1.031            | -1.840         | -0.537         | 0.790          | 2.226          |
| Assignment 2 | 1.331            | -1.984         | -0.427         | 0.887          | 1.891          |
| Assignment 3 | 1.128            | -2.404         | -0.866         | 0.861          | 2.354          |
| Assignment 4 | 1.681            | -1.659         | -0.369         | 0.890          | 2.203          |
| Assignment 5 | 1.904            | -2.151         | -0.519         | 0.626          | 1.774          |

|          | $\hat{\alpha}_r$ | $\hat{\varepsilon}_r$ |          | $\hat{\alpha}_r$ | $\hat{\varepsilon}_r$ |
|----------|------------------|-----------------------|----------|------------------|-----------------------|
| Rater 1  | 1.000            | 0.000                 | Rater 11 | 2.268            | 2.985                 |
| Rater 2  | 1.130            | 0.009                 | Rater 12 | 1.063            | 0.119                 |
| Rater 3  | 1.359            | -0.333                | Rater 13 | 1.336            | -0.578                |
| Rater 4  | 1.326            | -0.617                | Rater 14 | 1.599            | -0.434                |
| Rater 5  | 1.108            | -0.493                | Rater 15 | 1.776            | -1.085                |
| Rater 6  | 1.800            | -0.252                | Rater 16 | 1.202            | -1.230                |
| Rater 7  | 0.989            | -0.512                | Rater 17 | 1.045            | -1.180                |
| Rater 8  | 0.975            | -0.700                | Rater 18 | 1.500            | -1.076                |
| Rater 9  | 1.357            | 0.093                 | Rater 19 | 1.068            | -1.387                |
| Rater 10 | 1.270            | -0.337                | Rater 20 | 1.009            | 0.803                 |

|            | $\hat{\theta}$ |            | $\hat{\theta}$ |
|------------|----------------|------------|----------------|
| Learner 1  | 0.302          | Learner 11 | -0.204         |
| Learner 2  | -0.256         | Learner 12 | -0.369         |
| Learner 3  | 0.852          | Learner 13 | -0.610         |
| Learner 4  | -0.271         | Learner 14 | -0.593         |
| Learner 5  | 0.033          | Learner 15 | -0.194         |
| Learner 6  | 0.298          | Learner 16 | -0.645         |
| Learner 7  | -0.679         | Learner 17 | 0.019          |
| Learner 8  | 0.402          | Learner 18 | -0.628         |
| Learner 9  | -0.254         | Learner 19 | -0.515         |
| Learner 10 | -0.169         | Learner 20 | -0.565         |

$\log \alpha_i \sim N(0.270, 0.493^2)$ ,  $\mathbf{b}_i \sim MN(\boldsymbol{\mu}_b, \boldsymbol{\Sigma}_b)$   
 $\boldsymbol{\mu}_b = \{-2.147, -0.699, 0.817, 2.208\}$   
 $\boldsymbol{\Sigma}_b = \begin{pmatrix} 0.136 & 0.027 & 0.076 & 0.105 \\ 0.027 & 0.135 & 0.051 & 0.050 \\ 0.076 & 0.051 & 0.141 & 0.049 \\ 0.105 & 0.050 & 0.049 & 0.131 \end{pmatrix}$   
 $\log \alpha_r \sim N(0.227, 0.382^2)$ ,  $\varepsilon_r \sim N(-0.274, 1.059^2)$

According to Table 4 and Fig. 6, the rater characteristics can be regarded as explained below.

- *Rater 11* evaluated with high consistency and with severe criteria. *Rater 11* tended to give the lowest score to a learner who has ability below the average.
- *Rater 16* has the average-valued consistency and low value of severity.

Moreover, it is apparent that *Assignment 5* has a higher value of the discriminant parameter and can distinguish learners' ability more accurately than *Assignment 1*.

The proposed model can estimate the learner's ability considering these rater and assignment characteristics.

### 9.3 Model Comparison using Information Criteria

This subsection presents model comparisons using information criteria to ascertain whether the proposed model is suitable for the actual data. The procedures of this experiment are described below.

- 1) Using the actual data, the parameters of the proposed model, Patz1999, Usami2010, Ueno2008, HRM-Patz, HRM-DeCarlo, and Expanded GPCM were estimated. Here, the hyperparameters in Table 2 are given. In the proposed model, parameter estimation using HBM was also conducted.
- 2) Several information criteria were calculated for each model. Concretely, BIC [55], Marginal Likelihood (ML), DIC [56], [57] and WAIC [58] were calculated. Here, ML was estimated using Monte

TABLE 5  
Scores of Each Information Criterion.

| $R = J = 5$   | BIC     | ML             | DIC            | WAIC           |
|---------------|---------|----------------|----------------|----------------|
| Proposed(HBM) | -188.82 | -108.38        | -238.35        | -116.53        |
| Proposed      | -191.14 | <u>-109.24</u> | <u>-239.15</u> | <u>-117.57</u> |
| Patz1999      | -226.44 | -118.51        | -262.58        | -126.85        |
| Usami2010     | -208.35 | -120.15        | -266.68        | -129.86        |
| Ueno2008      | -187.65 | -113.78        | -243.97        | -120.35        |
| HRM-Patz      | -285.98 | -143.44        | -297.14        | -273.74        |
| HRM-DeCarlo   | -327.17 | -143.92        | -295.60        | -563.98        |
| Expanded GPCM | -199.89 | -118.68        | -258.41        | -127.15        |

| $R = J = 10$  | BIC            | ML             | DIC            | WAIC           |
|---------------|----------------|----------------|----------------|----------------|
| Proposed(HBM) | -367.41        | -220.96        | -476.95        | -237.57        |
| Proposed      | <u>-368.38</u> | <u>-223.44</u> | <u>-486.37</u> | <u>-241.54</u> |
| Patz1999      | -452.70        | -231.85        | -519.18        | -254.06        |
| Usami2010     | -391.86        | -229.95        | -495.37        | -244.52        |
| Ueno2008      | -396.91        | -228.46        | -486.71        | -242.40        |
| HRM-Patz      | -736.82        | -431.22        | -881.37        | -1230.85       |
| HRM-DeCarlo   | -844.12        | -443.24        | -886.29        | -1793.42       |
| Expanded GPCM | -371.54        | -226.65        | -492.88        | -243.08        |

| $R = J = 20$  | BIC      | ML       | DIC      | WAIC     |
|---------------|----------|----------|----------|----------|
| Proposed(HBM) | -1498.09 | -1218.14 | -2506.43 | -1250.82 |
| Proposed      | -1500.46 | -1220.45 | -2511.77 | -1253.29 |
| Patz1999      | -1694.58 | -1244.63 | -2573.47 | -1280.17 |
| Usami2010     | -1555.66 | -1229.95 | -2523.76 | -1259.95 |
| Ueno2008      | -1614.76 | -1234.67 | -2537.65 | -1266.92 |
| HRM-Patz      | -2383.86 | -1700.11 | -3401.43 | -4263.99 |
| HRM-DeCarlo   | -2838.99 | -1997.94 | -3975.69 | -7184.62 |
| Expanded GPCM | -1501.59 | -1223.00 | -2519.72 | -1257.36 |

\* Shaded cells represent maximum scores.

\*\* Underlined texts represent second largest scores.

Carlo integration because the exact calculation is intractable as a result of the high-dimensional integral. In those criteria, the ML and BIC, an asymptotic approximation to ML, are more important because these criteria are known to realize the consistent model selection [55]. The consistent model selection means that the probability of selecting the true model goes to 1 as the data size approaches infinity. The DIC and WAIC select the model to minimize the generalization error, which is regarded as the prediction error on future data. In those criteria, the model which maximizes the score is regarded as the optimal model.

- 3) The procedure 1) ~ 2) was conducted using data that reduced the number of learners  $J =$  raters  $R$  to 5 and 10. Data of  $J = R = 5$  are defined as  $x_{i,0,0} \sim x_{i,5,5}$ . Data of  $J = R = 10$  are  $x_{i,0,0} \sim x_{i,10,10}$ .

Table 5 presents results. Comparing the results of each model with the fixed hyperparameters, the proposed model was estimated as the optimal model in almost all cases. When  $J = R = 5$ , Ueno2008 had higher BIC than the proposed model because Ueno2008 incorporates the minimum number of parameters. However, as described in Section 6.1, the proposed model has practically the minimum number of parameters in peer assessment.

Furthermore, according to Table 5, the proposed model with HBM provided higher performances than that with the fixed hyperparameters in all cases.

In conclusion, the proposed model is expected to be the most suitable for the actual data because the model was estimated as the best approximation of the true model and the best predictor of future data.

## 9.4 Reliability Evaluation

This section evaluates the reliability of peer assessment using the actual data.

This article defined the reliability as *stability of the learner's ability estimation* [16]. From the definition, a model that can estimate the ability with little error when using the different assignments' and raters' ratings is regarded as a reliable model. Consequently, in this experiment, the reliability was evaluated using the following procedure.

- 1) Using the actual data, the rater and assignment parameters in the proposed model, Patz1999, Usami2010, Ueno2008, HRM-Patz, HRM-DeCarlo and Expanded GPCM were estimated. Here, the hyperparameters are given as shown in Table 2. In the proposed model, the parameter estimation using HBM was also conducted. Furthermore, to evaluate the effectiveness of the rater's consistency parameter, the proposed model without consistency parameter  $\alpha_r$  was assumed.
- 2) First, we created an *assignment group*, which consists of arbitrarily selected 3 assignments from all five assignments. Here, we designate all patterns of the assignment groups ( ${}_5C_3 = 10$  patterns) as *the set of assignment groups*. Similarly, we created a *rater group*, which consists of 10 arbitrarily selected raters from all 20 raters. We chose 10 rater groups from all patterns of the rater groups ( ${}_{20}C_{10} = 184756$  patterns). The 10 rater groups are designated as *the set of rater groups*.
- 3) By choosing one rater group from *the set of rater groups* and one assignment group from *the set of assignment groups*, all the pairs of a rater  $\times$  assignment group were created ( $10 \times 10 = 100$  pairs). Then, the data corresponding to each rater  $\times$  assignment group were created from the actual data.
- 4) Using the data for each rater  $\times$  assignment group, the learners' abilities  $\theta$  were estimated. In this estimation, the rater and assignment parameters estimated in procedure 1) were given. From this procedure, we obtained 100 patterns of estimated ability vector  $\hat{\theta}$  corresponding to 100 different combinations of raters and assignments.
- 5) We calculated the Pearson's correlation among all the pairs of the estimated ability vector  $\hat{\theta}$  ( ${}_{100}C_2 = 4950$  pairs). Then, the mean of the correlation values was calculated.
- 6) Tukey's multiple comparison test was conducted to compare the mean of the correlations among the models.

Here, the same experiment was conducted using a method by which the ability is given as the averaged value of the raw ratings. We designate this method as the *Averaged Score*.

In the experiment, the correlation is expected to reveal a higher value if the model is suitable for the real data

and the parameters are estimated with high accuracy.

Table 6 presents the result. In Table 6,  $\mu$  and  $\sigma$  respectively stand for the mean and standard deviation of the Pearson's correlation values. In addition,  $t$  denotes the test statistic.

According to Table 6, the proposed model, Patz1999, Usami2010, Ueno2008, HRM-Patz and Expanded GPCM had higher correlation values than the Averaged Score. Results show that the item response models were effective to improve the reliability of peer assessment. HRM-DeCarlo revealed lower correlation than the Averaged Score because HRM-DeCarlo had too many parameters and because the parameter estimation accuracy was extremely low.

Furthermore, when the fixed hyperparameters were given, it is apparent that the proposed model revealed significantly higher correlation than the other models. Here, the proposed model without the consistency parameter  $\alpha_r$  revealed significantly lower reliability than the proposed model. The use of the rater consistency parameter  $\alpha_r$  is fundamentally effective for improving the reliability of the proposed model. In addition, Table 6 presents the proposed model with HBM, which demonstrated the highest correlation in all models.

These results demonstrate that the proposed model can realize higher reliability than the other models. Parameter estimation using HBM can also improve the reliability.

## 10 Conclusion

This article proposed the new item response model for peer assessment that can realize higher reliability of peer assessment. The proposed model incorporates the rater's consistency and severity parameters to maintain as few rater parameters as possible. Consequently, when the number of raters increases, the number of rater parameters in the proposed model increases more slowly than those in the previous models. In addition, this article proposed a parameter estimation method for the proposed model using the hierarchical Bayes model. Although the accuracy of the Bayes estimation using sparse data depends strongly on the hyperparameters, the proposed estimation method can improve the accuracy because the hyperparameters are learned from the data. Therefore, the proposed method is expected to improve the reliability of peer assessment because it can estimate the ability of learners with higher accuracy and considering the rater's consistency and severity characteristics.

Furthermore, this article demonstrated the effectiveness of the proposed method through several experiments. In the simulation experiment, we demonstrated that the proposed model can provide the highest estimation accuracy of the parameters and ability when the number of raters increased. Additionally, we demonstrated that the accuracy of the Bayes estimation depended on the hyperparameters and that the estimation accuracy using the hierarchical Bayes model was close to

TABLE 6  
Result of the reliability evaluation.

|                        | Proposed (HBM)<br>$\mu = .834$<br>$\sigma = .068$ | Proposed<br>$\mu = .829$<br>$\sigma = .069$ | Proposed without $a_r$<br>$\mu = .802$<br>$\sigma = .072$ | Patz1999<br>$\mu = .789$<br>$\sigma = .076$ | Usami2010<br>$\mu = .818$<br>$\sigma = .069$ | Ueno2008<br>$\mu = .805$<br>$\sigma = .075$ | HRM-Patz<br>$\mu = .653$<br>$\sigma = .104$ | HRM-DeCarlo<br>$\mu = .576$<br>$\sigma = .135$ | Expanded GPCM<br>$\mu = .821$<br>$\sigma = .065$ | Averaged Score<br>$\mu = .621$<br>$\sigma = .146$ |
|------------------------|---|---|---|---|--|---|---|--|--|---|
| Proposed               | $t = 3.224$<br>( $p < .05$ )                      |   |   |   |  |   |   |  |  |   |
| Proposed without $a_r$ | $t = 17.468$<br>( $p < .01$ )                     | $t = 14.244$<br>( $p < .01$ )               |   |   |  |   |   |  |  |   |
| Patz1999               | $t = 24.685$<br>( $p < .01$ )                     | $t = 21.461$<br>( $p < .01$ )               | $t = 7.217$<br>( $p < .01$ )                              |   |  |   |   |  |  |   |
| Usami2010              | $t = 8.764$<br>( $p < .01$ )                      | $t = 5.540$<br>( $p < .01$ )                | $t = 8.704$<br>( $p < .01$ )                              | $t = 15.921$<br>( $p < .01$ )               |  |   |   |  |  |   |
| Ueno2008               | $t = 15.655$<br>( $p < .01$ )                     | $t = 12.431$<br>( $p < .01$ )               | $t = 1.813$<br>(-)  | $t = 9.030$<br>( $p < .01$ )                | $t = 6.892$<br>( $p < .01$ )                 |   |   |  |  |   |
| HRM-Patz               | $t = 97.895$<br>( $p < .01$ )                     | $t = 94.671$<br>( $p < .01$ )               | $t = 80.427$<br>( $p < .01$ )                             | $t = 73.210$<br>( $p < .01$ )               | $t = 89.131$<br>( $p < .01$ )                | $t = 82.240$<br>( $p < .01$ )               |   |  |  |   |
| HRM-DeCarlo            | $t = 139.111$<br>( $p < .01$ )                    | $t = 135.887$<br>( $p < .01$ )              | $t = 121.643$<br>( $p < .01$ )                            | $t = 114.426$<br>( $p < .01$ )              | $t = 130.348$<br>( $p < .01$ )               | $t = 123.456$<br>( $p < .01$ )              | $t = 41.216$<br>( $p < .01$ )               |  |  |   |
| Expanded GPCM          | $t = 7.287$<br>( $p < .01$ )                      | $t = 4.063$<br>( $p < .01$ )                | $t = 10.181$<br>( $p < .01$ )                             | $t = 17.398$<br>( $p < .01$ )               | $t = 1.477$<br>(-)                           | $t = 8.368$<br>( $p < .01$ )                | $t = 90.608$<br>( $p < .01$ )               | $t = 131.824$<br>( $p < .01$ )                 |  |   |
| Averaged Score         | $t = 114.898$<br>( $p < .01$ )                    | $t = 111.674$<br>( $p < .01$ )              | $t = 97.430$<br>( $p < .01$ )                             | $t = 90.213$<br>( $p < .01$ )               | $t = 106.135$<br>( $p < .01$ )               | $t = 99.243$<br>( $p < .01$ )               | $t = 17.003$<br>( $p < .01$ )               | $t = 24.213$<br>( $p < .01$ )                  | $t = 107.611$<br>( $p < .01$ )                   |   |

the accuracy achieved using the true hyperparameters.

In the actual data experiments, to confirm the validity of the proposed model for actual peer assessment data, the model comparisons using information criteria were conducted. Results show that the proposed model was expected to be the most suitable for the data because the model was estimated as the best approximation of the true model and the best predictor of future data. In addition, this article demonstrated that the proposed model realized the highest reliability of peer assessment. In the actual data experiments, the proposed model with the parameter estimation using the hierarchical Bayes model revealed higher performance than the proposed model with fixed hyperparameters.

The analyses described in this article used the Gibbs/MH method as the MCMC algorithm for parameter estimation because the algorithm is simple and easy to implement. Recently, several newer MCMC algorithms (e.g., the Hamiltonian Monte Carlo [59] and the no-U-turn sampler [60]) have been proposed. They are known to be more efficient than the Gibbs/MH. Developing an efficient MCMC algorithm for the proposed model remains as a future task.

In addition, as discussed in Section 6.2, the proposed model ignores that a learner's ability changes in the process of peer assessment. It is another future task to formulate an item response model that incorporates such ability change as that in the dynamic item response model [61].

## Appendix

The MCMC program for the parameter estimation of the proposed model can be downloaded from <https://bitbucket.org/uto/peerassessmentirt.git>. The source code was written in Java.

## References

- [1] K. J. Topping, E. F. Smith, I. Swanson, and A. Elliot, "Formative peer assessment of academic writing between postgraduate students," *Assessment & Evaluation in Higher Education*, vol. 25, no. 2, pp. 149–169, 2000.
- [2] M. Ueno and T. Okamoto, "Item response theory for peer assessment," in *Proc. IEEE International Conference on Advanced Learning Technologies*, 2008, pp. 554–558.
- [3] S. Bostock, "Student peer assessment in higher education: A meta-analysis comparing peer and teacher marks," *The Review of Educational Research*, vol. 70, no. 3, pp. 287–322, 2000.
- [4] R. L. Weaver and H. W. Cotrell, "Peer evaluation: A case study," *Innovative Higher Education*, vol. 11, no. 1, pp. 25–39, 1986.
- [5] J. Hamer, K. T. Ma, and H. H. Kwong, "A method of automatic grade calibration in peer assessment," in *Proc. 7th Australasian Computing Education Conference*, vol. 42, 2005, pp. 67–72.
- [6] H. Suen, "Peer assessment for massive open online courses (MOOCs)," *The International Review of Research in Open and Distributed Learning*, vol. 15, no. 3, pp. 313–327, 2014.
- [7] N. B. Shah, J. Bradley, S. Balakrishnan, A. Parekh, K. Ramchandran, and M. J. Wainwright, "Some scaling laws for MOOC assessments," in *Proc. ACM KDD Workshop on Data Mining for Educational Assessment and Feedback*, 2014.
- [8] P. Davies, "Review in computerized peer-assessment. will it affect student marking consistency?" in *Proc. International Computer Assisted Assessment Conference*, 2007, pp. 143–151.
- [9] S. S. J. Lin, E. Z. F. Liu, and S. M. Yuan, "Web-based peer assessment: feedback for students with various thinking-styles," *Journal of Computer Assisted Learning*, vol. 17, no. 4, pp. 420–432, 2001.
- [10] A. Bhalerao and A. Ward, "Towards electronically assisted peer assessment: A case study," *Association for Learning Technology Journal*, vol. 9, pp. 26–37, 2001.
- [11] S. Trahasch, "From peer assessment towards collaborative learning," in *Proc. Frontiers in Education Conference*, vol. 2, 2004, pp. 16–20.
- [12] 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," *Computers & Education*, vol. 45, no. 2, pp. 187–202, 2005.
- [13] J. Sitthiworachart and M. Joy, "Effective peer assessment for learning computer programming," in *Proc. 9th Annual SIGCSE Conference on Innovation and Technology in Computer Science Education*, 2004, pp. 122–126.
- [14] K. Cho and C. D. Schunn, "Scaffolded writing and rewriting in the discipline: A web-based reciprocal peer review system," *Computers & Education*, vol. 48, no. 3, pp. 409–426, 2007.
- [15] C. Piech, J. Huang, Z. Chen, C. Do, A. Ng, and D. Koller, "Tuned models of peer assessment in MOOCs," in *Proc. 6th International Conference of MIT's Learning International Networks Consortium*, 2013.
- [16] S. Kim, "A note on the reliability coefficients for item response model-based ability estimates," *Psychometrika*, vol. 77, no. 1, pp. 153–162, 2012.
- [17] Z. Wang and L. Yao, "The effects of rater severity and rater distribution on examinees' ability estimation for constructed-response items," Educational Testing Service Research Report, Tech. Rep., 2007.

- [18] S. J. Lurie, A. C. Nofziger, S. Meldrum, C. Mooney, and R. M. Epstein, "Effects of rater selection on peer assessment among medical students," *The International Journal of Medical Education*, vol. 40, no. 11, pp. 1088–1097, 2006.
- [19] 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 Japanese Journal of Educational Psychology*, vol. 58, no. 2, pp. 163–175, 2010.
- [20] E. Muraki, C. Hombo, and Y. Lee, "Equating and linking of performance assessments," *Applied Psychological Measurement*, vol. 24, pp. 325–337, 2000.
- [21] R. J. Patz, B. W. Junker, and M. S. Johnson, "The hierarchical rater model for rated test items and its application to large-scale educational assessment data," *Journal of Educational and Behavioral Statistics*, vol. 27, no. 4, pp. 341–366, 1999.
- [22] N. Dato and D. Gruijter, "Two simple models for rater effects," *Applied Psychological Measurement*, vol. 8, no. 2, pp. 213–218, 1984.
- [23] R. J. Patz and B. W. Junker, "Applications and extensions of MCMC in IRT: Multiple item types, missing data, and rated responses," *Journal of Educational and Behavioral Statistics*, vol. 24, pp. 342–366, 1999.
- [24] J. M. Linacre, *Many-faceted Rasch Measurement*. MESA Press, 1989.
- [25] E. Muraki, "A generalized partial credit model," in *Handbook of Modern Item Response Theory*. Springer Verlag, 1997, ch. 9, pp. 153–164.
- [26] F. Samejima, "Estimation of latent ability using a response pattern of graded scores," *Psychometrika*, no. 17, pp. 1–100, 1969.
- [27] L. T. DeCarlo, Y. K. Kim, and M. S. Johnson, "A hierarchical rater model for constructed responses, with a signal detection rater model," *Journal of Educational Measurement*, vol. 48, no. 3, pp. 333–356, 2011.
- [28] Y. Lu and X. Wang, "A hierarchical bayesian framework for item response theory models with applications in ideal point estimation," *The society for political methodology*, Saint Louis, Missouri, USA, Tech. Rep., 2006.
- [29] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer Verlag, 2006.
- [30] J. P. Fox, *Bayesian item response modeling: Theory and applications*. Springer Verlag, 2010.
- [31] M. Ueno, "Data mining and text mining technologies for collaborative learning in an ILMS "samurai"," in *Proc. IEEE International Conference on Advanced Learning Technologies*, 2004, pp. 1052–1053.
- [32] J. Cheaney and T. Ingebritsen, "Problem-based learning in an online course: A case study," *The International Review of Research in Open and Distributed Learning*, vol. 6, no. 3, pp. 1–18, 2006.
- [33] H.-J. Lee and C. Lim, "Peer evaluation in blended team project-based learning: What do students find important?" *Educational Technology & Society*, vol. 15, no. 4, pp. 214–224, 2012.
- [34] J. Lave and E. Wenger, *Situated Learning: Legitimate Peripheral Participation*. Cambridge University Press, 1991.
- [35] M. Ueno and M. Uto, "Learning community using social network service," in *Proc. International Conference Web Based Communities*, 2011, pp. 109–119.
- [36] R. Crespo, A. Pardo, and C. Kloos, "An adaptive strategy for peer review," in *Proc. Frontiers in Education Conference*, vol. 2, 2004, pp. 7–13.
- [37] F. Lord, *Applications of item response theory to practical testing problems*. Lawrence Erlbaum Associates, 1980.
- [38] M. Matteucci and L. Stracqualursi, "Student assessment via graded response model," *Statistica*, vol. 4, pp. 435–447, 2006.
- [39] G. Masters, "A rasch model for partial credit scoring," *Psychometrika*, vol. 47, no. 2, pp. 149–174, 1982.
- [40] D. Andrich, "A rating formulation for ordered response categories," *Psychometrika*, vol. 43, no. 4, pp. 561–573, 1978.
- [41] J. G. Baker, "A comparison of graded response and rasch partial credit models with subjective well-being," *Journal of Educational and Behavioral Statistics*, vol. 25, pp. 253–270, 2000.
- [42] K. Shojima, "Selection of item response model by genetic algorithm," *Behaviormetrika*, vol. 34, no. 1, pp. 1–26, 2007.
- [43] F. Samejima, "Evaluation of mathematical models for ordered polychotomous responses," *Behaviormetrika*, vol. 23, no. 1, pp. 17–35, 1996.
- [44] F. B. Baker and S. H. Kim, *Item Response Theory: Parameter Estimation Techniques*. CRC Press, 2004.
- [45] W. V. D. Linden and R. Hambleton, *Handbook of modern item response theory*. Springer Verlag, 1996.
- [46] W. W. Peterson, T. G. Birdsall, and W. C. Fox, "The theory of signal detectability," *Transactions IRE Profession Group on Information Theory*, vol. 4, pp. 171–212, 1954.
- [47] L. T. DeCarlo, "A model of rater behavior in essay grading based on signal detection theory," *Journal of Educational Measurement*, vol. 42, no. 1, pp. 53–76, 2005.
- [48] I. M. Goldin, "Accounting for peer reviewer bias with bayesian models," in *Proc. International Conference on Intelligent Tutoring Systems*, 2012.
- [49] L. Cronbach, R. Nageswari, and G. Gleser, "Theory of generalizability: A liberation of reliability theory," *The British Journal of Statistical Psychology*, vol. 16, pp. 137–163, 1963.
- [50] N. Longford, "Reliability of essay rating," in *Models for Uncertainty in Educational Testing*. Springer Verlag, 1995, ch. 2, pp. 17–46.
- [51] K. Cho, C. D. Schunn, and R. Wilson, "Validity and reliability of scaffolded peer assessment of writing from instructor and student perspectives," *Journal of Educational Psychology*, vol. 98, no. 4, pp. 891–901, 2006.
- [52] S. Brooks, A. Gelman, G. Jones, and X. Meng, *Handbook of Markov Chain Monte Carlo*. CRC Press, 2011.
- [53] K. P. Murphy, "Conjugate bayesian analysis of the gaussian distribution," University of British Columbia, Vancouver, British Columbia, Canada, Tech. Rep., 2007.
- [54] H. Tanizaki, "A simple gamma random number generator for arbitrary shape parameters," *Economics Bulletin*, vol. 3, pp. 1–10, 2008.
- [55] G. Schwarz, "Estimating the dimensions of a model," *The Annals of Statistics*, vol. 6, pp. 461–464, 1978.
- [56] D. J. Spiegelhalter, N. G. Best, B. P. Carlin, and A. Van Der Linde, "Bayesian measures of model complexity and fit," *Journal of the Royal Statistical Society*, vol. 64, no. 4, pp. 583–639, 2002.
- [57] A. Gelman, J. Hwang, and A. Vehtari, "Understanding predictive information criteria for bayesian models," *Statistics and Computing*, vol. 24, no. 6, pp. 997–1016, 2014.
- [58] S. Watanabe, "Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory," *Journal of Machine Learning Research*, pp. 3571–3594, 2010.
- [59] R. M. Neal, "MCMC using Hamiltonian dynamics," in *Handbook of Markov Chain Monte Carlo*. CRC Press, 2011, ch. 5, pp. 113–162.
- [60] M. D. Hoffman and A. Gelman, "The No-U-Turn sampler: Adaptively setting path lengths in hamiltonian monte carlo," *Journal of Machine Learning Research*, vol. 15, pp. 1593–1623, 2014.
- [61] X. Wang, J. O. Berger, and D. S. Burdick, "Bayesian analysis of dynamic item response models in educational testing," *The Annals of Applied Statistics*, vol. 7, no. 1, pp. 126–153, 2013.



**Masaki Uto** received a Ph.D. degree from the University of Electro Communications in 2013. He has been an Assistant Professor of the Nagasaki University of Technology since 2014. His research interests include e-learning, e-testing, machine learning, and data mining.



**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 Graduate School of Information Systems at the University of Electro-Communications since 2013. He received Best Paper awards from ICTAI2008, ED-MEDIA 2008, 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.