Item Response Model with Lower Order Parameters for Peer Assessment

Masaki Uto[1(✉)] and Maomi Ueno[2]

[1] Nagaoka University of Technology, Niigata, Japan
uto@oberon.nagaokaut.ac.jp
[2] University of Electro-Communications, Tokyo, Japan
ueno@ai.is.uec.ac.jp

Abstract. 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. However, in previous models, the parameter estimation accuracy decreases as the number of raters increases because the number of rater parameters increases drastically. To solve that problem, 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.

Keywords: 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, has become popular in recent years [4]. Peer assessment presents many important benefits [3,4]. Therefore, peer assessment has been adopted into various learning processes.

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 [3]. Peer assessment enables realization of reliable assessment without burdening a teacher when the number of raters is sufficiently large [4]. However, it is difficult to increase the number of raters for each learner because one rater can only assess a few performances [3]. 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*. The reliability reveals a higher value if the ability of learners are obtainable with few errors when the performance tasks or raters are changed.

The reliability of peer assessment is known to depend on rater characteristics [4]. Therefore, the reliability is expected to be increased if the ability of learners is estimated considering the following rater characteristics [5]. 1) *Severity*:

C. Conati et al. (Eds.): AIED 2015, LNAI 9112, pp. 800–803, 2015.
DOI: 10.1007/978-3-319-19773-9_119

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

For this reason, some item response models which incorporate the rater characteristic parameters have been proposed [1,2,4,5]. However, in previous models, the number of rater parameters increases extremely as the number of raters increases because the models include high dimensional rater parameters. The accuracy of parameter estimation is known to be reduced when the number of parameters increases because the data size per parameter decreases. If the accuracy of parameter estimation is reduced, the reliability is necessarily reduced.

To solve the problem, this article presents a proposal of a new item response model for peer assessment. The model incorporates rater's consistency and severity parameters to maintain as few rater parameters as possible. The model presents the following advantages. 1) The model has fewer rater parameters than previous models. Therefore, the model can provide higher parameter estimation accuracy when the number of raters increases. 2) The model can improve the reliability because it can estimate the learner's ability parameter with higher accuracy and can consider the rater's consistency and severity characteristics.

This article also proposes a parameter estimation method using a hierarchical Bayes model for the proposed model. In addition, this article demonstrates the effectiveness of the proposed model through actual data experiments.

2 Proposed Model

This article assumes that peer assessment data consist of categories $k \in \{1..., K\}$ given by each rater $r \in \{1, \cdots, R\}$ to each work of learner $j \in \{1, \cdots, J\}$ for each assignment $i \in \{1, \cdots, I\}$. This article proposes an item response model for the data by extending the graded response model. The model gives the probability that rater r responds in category k to learner j's assignment i as follows.

$$P_{ijrk} = P^*_{ijrk-1} - P^*_{ijrk}, \quad \begin{cases} P^*_{ijrk} = [1 + \exp(-\alpha_i \alpha_r (\theta_j - b_{ik} - \varepsilon_r))]^{-1}, \\ P^*_{ijr0} = 1, \quad P^*_{ijrK} = 0. \end{cases}$$

Therein, θ_j is the ability of learner j, α_i is a discriminant parameter of assignment i, α_r reflects the consistency of rater r, b_{ik} denotes the difficulty in obtaining the score k for assignment i ($b_{i1} < \cdots < b_{iK-1}$), and ε_r represents the severity of rater r. For model identification, $\alpha_{r=1} = 1$, $\varepsilon_1 = 0$ and $\Pi_r \alpha_r = 1$ are assumed.

The unique feature of the proposed model is that each rater has only one consistency and severity parameter. If higher dimensional rater parameters are used, such as that described by Ueno et al. [4] and Patz et al. [1], then the number of rater parameters increases rapidly concomitantly with the increasing number of raters. In the proposed model, the number of rater parameters increases slowly as the number of raters increases. Therefore, the proposed model is expected to improve the parameter estimation accuracy because the estimation accuracy generally increases when the number of parameters decreases.

Another feature of the proposed model is introducing the rater consistency parameter. The models of Patz et al. [1] and Ueno et al. [4] use no consistency

parameters. However, the reliability is known to be increased if the ability of learners is estimated considering the rater consistency [5]. The parameter α_r used in the proposed model can optimally represent the rater consistency [5].

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.

To estimate the parameters in item response models, the Bayes estimation has generally been used. However, the accuracy of the Bayes estimation depends on parameters, called *hyperparameters*, which are arbitrarily determined by an analyst. Therefore, this article employes a parameter estimation method using a hierarchical Bayes model (HBM) for the proposed model. This method can estimate the hyperparameters from given data.

3 Model Comparison Using Information Criteria

This section presents model comparisons using information criteria to confirm whether the proposed model is suitable for actual peer assessment 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 by one author. 2) The reports were evaluated by 20 other raters who had attended the same e-learning course. The raters rated them using 5 categories.

Using the actual data, the BIC and DIC were calculated for the proposed model, the models of Patz et al. [1], Usami [5], and Ueno et al. [4], and the hierarchical rater model (HRM) [2]. In those information criteria, the BIC asymptotically selects the true model; the DIC selects the model to minimize the prediction error on future data. The model which maximizes the criteria is regarded as the optimal model. Here, the criteria were calculated with fixed hyperparameters. Only for the proposed model, the criteria were also calculated using HBM.

Table 1 presents results. Comparing the results of each model with the fixed hyperparameters, both information criteria selected the proposed model as the optimal model. It means that the proposed model was estimated as the best approximation of the true model and the best predictor of future data. Additionally, the proposed model with HBM further improved the performances.

4 Reliability Evaluation

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

The procedure is as follows: 1) For the proposed model, the model of Patz, Usami, and Ueno, and HRM, the rater and assignment parameters were estimated using the actual data. Here, the hyperparameters were fixed. Only for the proposed model, the estimation using HBM was also conducted. 2) A subset of raters and assignments which consists of 10 raters and 3 assignments was created. Then, the data corresponding to the subset was created from the actual data. Given the data and the estimated rater/assignment parameters, the learners' abilities were estimated. 3) For 100 different subsets of raters and assignments,

Table 1. Scores of Information Criteria

	BIC	DIC
Proposed (HBM)	-1503.37	-2525.93
Proposed	-1508.27	-2531.50
Patz et al.	-1694.58	-2573.47
Usami	-1593.63	-2572.51
Ueno et al.	-1614.76	-2537.65
HRM	-2397.28	-3409.92

Table 2. The reliability evaluation result

Proposed (HBM)	.845 (.065)
Proposed	.831 (.066)
Patz et al.	.789 (.076)
Usami	.816 (.070)
Ueno et al.	.805 (.075)
HRM	.663 (.117)

* Shaded(Underlined) texts represent the maximum(second largest) values.

the procedure 2 was conducted. Then, the Pearson's correlations among all the pairs of the estimated ability vectors were calculated. Finally, Tukey's multiple comparison test was conducted to compare the mean of the correlations.

From the definition of the reliability, a model which reveals higher correlation values can be regarded as reliable. Table 2 presents the mean and standard deviation of the correlations. The results can be summarized as follows: 1) Given the fixed hyperparameters, the proposed model revealed significantly higher reliability than the other models. 2) The proposed model with HBM demonstrated the highest reliability in all models.

5 Conclusion

This article proposed a new item response model which incorporates the rater's consistency and severity parameters to maintain as few rater parameters as possible. The actual data experiments demonstrated that: 1) the proposed model was the most suitable for the actual data; 2) the model improved the reliability; and 3) the parameter estimation using HBM further improved those performances.

References

1. Patz, R.J., Junker, B.W.: Applications and extensions of mcmc in irt: Multiple item types, missing data, and rated responses. Journal of Educational and Behavioral Statistics **24**, 342–366 (1999)
2. Patz, R.J., Junker, B.W., Johnson, M.S.: 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 (1999)
3. Piech, C., Huang, J., Chen, Z., Do, C., Ng, A., Koller, D.: Tuned models of peer assessment in MOOCs. In: Proceedings of Sixth International Conference of MIT's Learning International Networks Consortium (2013)
4. Ueno, M., Okamoto, T.: Item response theory for peer assessment. In: Eighth IEEE International Conference on Advanced Learning Technologies. ICALT 2008, pp. 554–558 (2008)
5. Usami, S.: 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 (2010)