Item Response Theory Without Restriction of Equal Interval Scale for Rater's Score

Masaki Uto(✉) and Maomi Ueno

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

Abstract. With the spread of large-scale e-learning environments such as MOOCs, peer assessment has been used recently to measure learner ability. Nevertheless, peer assessment presents the important difficulty that the ability assessment accuracy depends strongly on rater characteristics. To resolve that difficulty, item response theory (IRT) models that incorporate rater characteristic parameters have been proposed. However, those models rely upon the assumption of an equal interval scale for raters' scores although the scales are known to vary across raters. To resolve the difficulty, this study proposes a new IRT model without the restriction of an equal interval scale for raters. The proposed model is expected to improve model fitting to peer assessment data. Furthermore, the proposed model can realize more robust ability assessment than conventional models can. This study demonstrates the effectiveness of the proposed model through experimentation with actual data.

Keywords: Educational measurement · E-learning
Item response theory · Peer assessment · Rating scale

1 Introduction

Peer assessment, which is mutual assessment among learners, has become popular with the widespread use of large-scale e-learning environments such as massive open online courses (MOOCs) [1–3]. Peer assessment has been adopted in various learning and assessment situations because it provides many benefits (e.g., [2,3]). One important use of peer assessment is for summative assessment, which provides a measure of learner ability [4,5]. Peer assessment is justified as an appropriate summative assessment method because the learner ability is definable naturally in the learning community as a social agreement [3,6]. Furthermore, even when learners are numerous, as in MOOCs, peer assessment can be conducted by assigning a few peer-raters to each learner, although assessment by instructors becomes difficult [2,4,7,8].

Peer assessment, however, presents the difficulty that the assessment accuracy of learner ability depends strongly on rater characteristics such as rating

C. Penstein Rosé et al. (Eds.): AIED 2018, LNAI 10948, pp. 363–368, 2018.
https://doi.org/10.1007/978-3-319-93846-2_68

severity and consistency [2,3,9,10]. Item response theory (IRT) models incorporating rater characteristic parameters have been proposed to resolve that difficulty, (e.g., [3,9,11,12]). A traditional model is the many facet Rasch model (MFRM) [12], which is defined as a partial credit model [13] incorporating a rater severity parameter. Additionally, an extension of this model using the generalized partial credit model [14] has been proposed [11]. Furthermore, to resolve the difficulty that raters are not always consistent, a graded response model [15] incorporating rater consistency and severity parameters has been proposed recently [3]. Those IRT models are known to provide more accurate ability assessment than average or total scores do because they can estimate the ability considering some rater characteristics [3].

However, when the diversity of raters' assessment skills increases as in peer assessment, the rating scales are known to vary across raters [10,16,17]. For example, some raters presumably overuse a few restricted categories, avoid some specific categories, and use all categories uniformly. However, earlier IRT models have been incapable of representing such rater characteristics because they assume an equal interval scale for raters' scores. Consequently, the models will not fit peer assessment data well. Low model fit generally reduces the ability assessment accuracy [3].

To resolve that difficulty, this study proposes a new IRT model without the restriction of the equal interval scale for raters. Specifically, the proposed model is defined as a generalized partial credit model that incorporates a rater severity parameter for each rating category. The proposed model is expected to improve the model fitting to peer assessment data because differences in the scale among raters can be represented. Furthermore, the proposed model can realize more robust ability assessment than conventional models because the introduction of the unequal interval scales for raters enables more precise representation of the characteristics of aberrant raters, who use extremely different rating scales from those used by others. This study demonstrates the effectiveness of the proposed model through the use of actual data experiments.

2 Proposed Model

The rating data $\boldsymbol{U}$ obtained from peer assessment consist of rating category $k \in \mathcal{K} = \{1, \cdots, K\}$ given by peer-rater $r \in \mathcal{J} = \{1, \cdots, J\}$ to the outcome of learner $j \in \mathcal{J}$ for task $t \in \mathcal{T} = \{1, \cdots, T\}$. Letting u_{tjr} be a response of rater r to learner j's outcome for task t, the data $\boldsymbol{U}$ are described as $\boldsymbol{U} = \{u_{tjr} \mid u_{tjr} \in \mathcal{K} \cup \{-1\}, t \in \mathcal{T}, j \in \mathcal{J}, r \in \mathcal{J}\}$, where $u_{tjr} = -1$ denotes missing data. This study was conducted to estimate the learner ability accurately from the peer assessment data $\boldsymbol{U}$ using item response theory (IRT) [18].

The proposed model is defined as a generalized partial credit model that incorporates the rater severity parameter for each rating category and the rater consistency parameter. The model provides the response probability P_{ijrk} as

$$P_{ijrk} = \frac{\exp \sum_{m=1}^{k} \left[\alpha_r \alpha_i (\theta_j - \beta_i - \beta_r - d_{rm})\right]}{\sum_{l=1}^{K} \exp \sum_{m=1}^{l} \left[\alpha_r \alpha_i (\theta_j - \beta_i - \beta_r - d_{rm})\right]}. \tag{1}$$

where θ_j represents the latent ability of learner j, α_i denotes the discrimination parameter for task i, β_i denotes the difficulty of task i, α_r signifies the consistency of rater r, β_r denotes the severity of rater r, and d_{rk} represents the severity of rater r to give category k. Here, $\alpha_{r=1} = 1$, $\beta_{r=1} = 0$, $d_{r1} = 0$, and $\sum_{k=2}^{K} d_{rk} = 0$ are assumed for model identification.

In the proposed model, d_{rk} controls the intervals between adjacent categories for each rater. Furthermore, the intervals determine the rater's response probability for each category. Specifically, as interval $d_{rk+1} - d_{rk}$ becomes larger, the response probability for category k increases. As interval $d_{rk+1} - d_{rk}$ becomes smaller, the probability of responding with category k decreases.

The proposed model can represent such differences in the rating scale among raters although earlier IRT models with rater parameters (e.g., [3,11,12]) incorporate the assumption of an equal interval scale for raters' scores. The scales generally vary among raters in peer assessment, as described in Sect. 1. Therefore, the proposed model is expected to provide higher model fitting to peer assessment data than the conventional models. Because better model fitting generally improves the ability assessment accuracy [3], the proposed model is expected to provide higher accuracy than the conventional models provide.

3 Actual Data Experiment

This section presents a description of evaluation of the effectiveness of the proposed model using actual peer assessment data. Actual data were gathered using the following procedures. (1) 30 university students were enrolled in this study as participants. (2) They were asked to complete four essay-writing tasks that were set in the national assessment of educational progress (NAEP) 2002 [19] and 2007 [20]. (3) After the participants completed all tasks, they were asked to evaluate the essays of all other participants for all four tasks. The assessments were conducted using a rubric that includes five rating categories.

Using the peer assessment data, we conducted the following experiment. (1) The parameters of the proposed model, MFRM [12], the model proposed by Patz and Junker [11] (designated as *Patz1999*), and that proposed by Uto and Ueno [3] (designated as *Uto2016*) were estimated using the MCMC algorithm. The widely applicable information criteria (WAIC) and log marginal likelihood (ML) were also calculated for each model. (2) Given the estimated task and rater parameters, the learner ability was re-estimated from each rater's data. Then, we calculated the RMSE between the ability values estimated from each rater's data and those estimated using complete data in Procedure 1. The average value of the RMSE over all raters was calculated for each model. In addition, this index was calculated for a method by which the ability is given as the averaged value of the raw ratings (designated as *Averaged*).

Table 1 presents results. As shown in Table 1, the proposed model was selected as the best model by both information criteria. Results show that the proposed model presented the lowest RMSE value. Here, we conducted multiple comparisons using the Dunnet method to ascertain whether the RMSE value of the

Table 1. Information criteria and ability assessment accuracies

	Information criteria		RMSE		
	WAIC	ML	Mean	SD	Test statistic
Proposed	**−4396.07**	**−4324.23**	**0.313**	0.053	–
MFRM	−4646.46	−4615.25	0.379	0.075	2.745 ($p = 0.024$)
Patz1999	−4646.08	−4575.41	0.464	0.067	6.348 ($p < 0.001$)
Uto2016	−4434.82	−4385.57	0.382	0.065	2.897 ($p = 0.016$)
Averaged	–	–	0.499	0.157	6.997 ($p < 0.001$)

Table 2. Rater parameters estimated from actual data

Rater	α_r	β_r	d_{r2}	d_{r3}	d_{r4}	d_{r5}	Rater	α_r	β_r	d_{r2}	d_{r3}	d_{r4}	d_{r5}
1	1.000	0.000	−1.169	−0.154	0.152	1.171	16	1.249	0.148	−0.111	−1.637	−0.295	2.043
2	0.638	0.132	−0.383	−0.460	−0.163	1.007	17	1.261	−0.413	−1.231	−0.846	0.567	1.509
3	1.267	0.393	−0.991	−0.308	0.477	0.822	18	1.670	0.206	−1.307	−0.299	0.393	1.213
4	1.115	0.025	−1.695	−0.416	0.051	2.059	19	1.770	0.455	−2.278	−0.459	1.829	0.908
5	0.963	−0.334	−1.740	−0.372	0.740	1.372	20	1.261	0.698	−1.506	−0.599	0.340	1.764
6	0.928	−0.078	−1.774	−0.145	0.386	1.532	21	0.745	0.004	−1.137	0.083	0.623	0.431
7	0.746	0.856	−0.357	−0.546	0.882	0.022	22	1.354	0.249	−2.051	−0.308	0.755	1.604
8	1.809	0.301	−1.511	−0.680	0.701	1.489	23	1.153	0.188	−1.493	−1.501	0.927	2.068
9	1.091	0.793	−1.857	−0.034	0.414	1.477	24	0.568	0.231	−1.376	−0.458	0.792	1.042
10	0.797	−0.111	−0.445	−0.089	0.133	0.401	25	0.829	−0.126	−0.536	0.030	0.236	0.270
11	1.137	−0.262	−1.645	−0.584	0.626	1.602	26	0.571	0.773	−1.027	0.106	0.268	0.653
12	1.029	−0.182	−1.780	−0.651	0.603	1.828	27	0.920	−0.079	−0.941	0.130	−0.374	1.185
13	0.858	0.648	−1.171	−0.129	0.694	0.606	28	0.855	−0.397	−0.589	−0.943	−0.441	1.973
14	0.881	0.235	−1.935	−0.017	0.595	1.358	29	1.338	0.118	−1.423	−0.253	0.494	1.182
15	1.374	−0.128	−1.480	−0.897	0.618	1.759	30	0.834	−0.285	−1.741	0.715	−0.067	1.092

proposed model is significantly lower than that of the other models, or not. The results, which are shown in *Test statistic* column of Table 1, demonstrate that the RMSE of the proposed model was significantly lower than those of the conventional models.

The proposed model outperformed the conventional model when assessing raters with various rating scales. To emphasize this point, Table 2 presents rater parameters estimated using the proposed model. From the table, we can confirm the large variety of rating scales among the raters. The proposed model can represent those rater characteristics appropriately, although the conventional models cannot represent them. Therefore, in this experiment, the proposed model presented the highest model fitting and ability assessment accuracy.

4 Conclusion

This study proposed a new IRT model without the restriction of the equal interval scale for raters' scores. Experiments conducted with actual data demonstrated that the proposed model can improve the model fitting and ability

assessment accuracy when raters have different rating scales. Although this study specifically addressed only peer assessment accuracy, the proposed model is useful for various purposes such as evaluating assessment skills, creating peer assessment groups, and selecting optimal peer-raters for each learner. Such applications are left as subjects for future work.

References

1. Moccozet, L., Tardy, C.: An assessment for learning framework with peer assessment of group works. In: Proceedings of International Conference on Information Technology Based Higher Education and Training, pp. 1–5 (2015)
2. Shah, N.B., Bradley, J., Balakrishnan, S., Parekh, A., Ramchandran, K., Wainwright, M.J.: Some scaling laws for MOOC assessments. In: ACM KDD Workshop on Data Mining for Educational Assessment and Feedback (2014)
3. Uto, M., Ueno, M.: Item response theory for peer assessment. IEEE Trans. Learn. Technol. **9**(2), 157–170 (2016)
4. Staubitz, T., Petrick, D., Bauer, M., Renz, J., Meinel, C.: Improving the peer assessment experience on MOOC platforms. In: Proceedings of Third ACM Conference on Learning at Scale, New York, NY, USA, pp. 389–398 (2016)
5. Terr, R., Hing, W., Orr, R., Milne, N.: Do coursework summative assessments predict clinical performance? a systematic review. BMC Med. Educ. **17**(1), 40 (2017)
6. Lave, J., Wenger, E.: Situated Learning - Legitimate Peripheral Participation. Cambridge University Press, New York (1991)
7. Uto, M., Thien, N.D., Ueno, M.: Group optimization to maximize peer assessment accuracy using item response theory. In: André, E., Baker, R., Hu, X., Rodrigo, M.M.T., du Boulay, B. (eds.) AIED 2017. LNCS (LNAI), vol. 10331, pp. 393–405. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-61425-0_33
8. Nguyen, T., Uto, M., Abe, Y., Ueno, M.: Reliable peer assessment for team project based learning using item response theory. In: Proceedings of International Conference on Computers in Education, pp. 144–153 (2015)
9. Eckes, T.: Introduction to Many-Facet Rasch Measurement: Analyzing and Evaluating Rater-Mediated Assessments. Peter Lang Publishing Inc., Frankfurt (2015)
10. Myford, C.M., Wolfe, E.W.: Detecting and measuring rater effects using many-facet Rasch measurement: Part I. J. Appl. Measur. **4**, 386–422 (2003)
11. Patz, R.J., Junker, B.: Applications and extensions of MCMC in IRT: multiple item types, missing data, and rated responses. J. Educ. Behav. Stat. **24**, 342–366 (1999)
12. Linacre, J.: Many-Faceted Rasch Measurement. MESA Press, Chicago (1989)
13. Masters, G.: A Rasch model for partial credit scoring. Psychometrika **47**(2), 149–174 (1982)
14. Muraki, E.: A generalized partial credit model: application of an EM algorithm. Appl. Psychol. Measur. **16**(2), 159–176 (1992)
15. Samejima, F.: Estimation of latent ability using a response pattern of graded scores. Psychometrika Monography **17**, 1–100 (1969)
16. Kassim, N.L.A.: Judging behaviour and rater errors: an application of the many-facet Rasch model. GEMA Online J. Lang. Stud. **11**(3), 179–197 (2011)
17. Rahman, A.A., Ahmad, J., Yasin, R.M., Hanafi, N.M.: Investigating central tendency in competency assessment of design electronic circuit: analysis using many facet Rasch measurement (MFRM). Int. J. Inf. Educ. Technol. **7**(7), 525–528 (2017)

18. Lord, F.: Applications of Item Response Theory to Practical Testing Problems. Erlbaum Associates, Hillsdale (1980)
19. Persky, H., Daane, M., Jin, Y.: The nation's report card: Writing 2002. Technical report, National Center for Education Statistics (2003)
20. Salahu-Din, D., Persky, H., Miller, J.: The nation's report card: Writing 2007. Technical report, National Center for Education Statistics (2008)