Shih, JL. et al. (Eds.) (2023). Proceedings of the 31st International Conference on Computers in Education. Asia-Pacific Society for Computers in Education

Tackling Unserious Raters in Peer Evaluation: Behavior Analysis and Early Detection with Learner Model

Changhao LIANG[a*], Izumi HORIKOSHI[b], Rwitajit MAJUMDAR[b] & Hiroaki OGATA[b]
[a]*Graduate School of Informatics, Kyoto University, Japan*
[b] *Academic Center for Computing and Media Studies, Kyoto University, Japan*
* liang.changhao.84c@st.kyoto-u.ac.jp

Abstract: Peer evaluation of individual or group work is often adopted in team-based learning design. However, some raters may not take the evaluation process seriously and exhibit behaviors such as using the same score, rushing through evaluations, or not evaluating during the presentation. This study investigates the issue of unserious peer evaluation in group presentations, focusing on their behavior patterns. Using evaluation behavior analysis indicators, we identified unserious raters who exhibited low reliability in the peer evaluation process. Further, we conducted a preliminary analysis to detect unserious raters based on learner model data available before the peer evaluation process. This information can assist teachers in providing personalized prompts and interventions prior to the peer evaluation process, thus enhancing the evaluation quality of these students with timely prompts to them.

Keywords: Peer evaluation, Evaluation behavior analysis (EBA), Team-based learning (TBA), Data-driven study, Learner model

1. Introduction

Evaluation is an essential aspect in collaborative learning, but teachers may struggle to properly evaluate each student (Amarasinghe et al., 2021). Peer evaluation offers formative feedback that encourages reflection and overcomes the limitations of traditional evaluation (Ohland et al., 2012). It has become widely adopted in online settings where student-centered learning is prevalent and can enhance both learning and interpersonal skills (Kasch et al., 2021). However, some raters may not take the evaluation process seriously, as Horikoshi and Tamura (2021) discovered. Such evaluations may involve using the same score, rushing through evaluations, or not evaluating during the presentation. These low-quality ratings can make peer evaluation results less reliable and lower the learning outcome.

Nevertheless, the reliability of unserious peer raters can be improved by proper interventions (Van Zundert et al., 2010). Current studies have made attempts to calibrate scores based on student engagement and previous peformance (Piech et al., 2013), or train evaluation skills during the peer evaluation process (Gorham et al., 2023). These approaches can be too late to nudge timely interventions to the ongoing evaluation activity. Fewer work addresses predicting problem raters early before the assessment to facilitate possible interventions to improve their evaluation behaviors.

This study investigates the issue of unserious peer evaluation in group presentations, focusing on their behavior patterns. Using behavioral indicators, we identified unserious raters who exhibited low reliability in the peer evaluation process. Further, we conducted a preliminary analysis to examine how the learner model data from their learning logs and their prior peer evaluation behaviors can be used for early detection. This information can assist teachers in providing personalized prompts and interventions prior to the peer evaluation process, thus enhancing the evaluation quality of these students in a timely manner.

2. Research Background

2.1 Peer evaluation in Team-Based Learning (TBL)

In peer evaluation, students provide ratings and feedback on each other's work, which is formative and can promote their performance in subsequent tasks (Ohland et al., 2012; Gorham et al., 2023). Research has shown that peer evaluation encourages students to think deeply and critically about their own work and contributes to the development of "internal feedback" skills, where learners reflect on and regulate their own learning processes (Nicol et al., 2014; To & Panadero, 2019).

Team-Based Learning (TBL) is an educational strategy that involves multiple rounds of group work with peer evaluation. It was first introduced in medical education (Michaelsen et al., 2002). During each round of TBL, students start by exploring the learning topic individually before working in teams to complete tasks (Parmelee et al., 2012). Group presentations and peer evaluation conclude each round, where students assess the products or outcomes of their peers' learning experiences and reflect themselves as a formative process (Topping, 1998). Moreover, under the data-driven environment, previous rounds' learning log data enables targeted interventions by teachers (Johnson, 2017).

2.2 Evaluation Behavior Analysis (EBA)

The process of peer evaluation generates behavior indicators that record key information, such as the identity of the evaluator, the timing of the evaluation, the items assessed, and the corresponding scores (Horikoshi & Tamura, 2021). The behavior indicators stem from "paradata" in the web survey research field, which refers to the log data generated during the evaluation process and is related to the quality of survey responses (Couper & Kreuter, 2013). For instance, shorter response times are associated with a "lack of motivation to answer accurately caused by continuous survey" (Yan & Tourangeau, 2008), and individuals who answer quickly as "speeders" can lead to poor responses (Zhang & Conrad, 2014).

The web survey research and peer evaluation research share the goal of measuring inappropriate behaviors in digital evaluation platforms. Therefore, to effectively analyze and visualize the quality of peer evaluation based on behaviors, the Evaluation Behavior Analysis (EBA) method has been developed. It involves extracting data from peer evaluations and utilizing it to gain insights into students' evaluation behaviors. Using the EBA method, instructors can identify patterns and trends in the evaluation behavior of students. Horikoshi et al. (2022) have defined feature variables that capture the key aspects of evaluation behavior, which are presented in Table 1.

Table 1. *Definition of feature variables of evaluation behaviors (Horikoshi et al., 2022)*

Behavior Indicator	Definition	Proposed constructs
Evaluation Time (ET)	Time span from clicking the first evaluation item to the last item.	Speed: how much time the rater spent on the evaluation
Mean of the Timestamp (tM)	Average elapsed time since the start of the presentation.	Timeliness: whether the rater evaluated immediately after the presentation
SD of the Timestamp (tSD)	Standard deviation of the timestamps for all evaluations.	Coherence: whether the rater evaluated evenly throughout or within a short time
Click Count (CC)	Total number of times the evaluation items were clicked.	Certainty: how many changes the rater made
Mean of the Score (sM)	Average score for all the evaluation items scored by the reviewer	Leniency: rater tendency to assign higher or lower scores

SD of the Score (sSD)	Standard deviation for the scores of all evaluation items from the rater.	Straightlining: whether the rater used only similar scores (Kim et al., 2019)

Compared to the conventional perspective of peer evaluation quality, which primarily emphasizes scores and compliance with others (Cho & Schunn, 2007; Fukazawa, 2010), the EBA indicators focus on the process of the peer evaluation. These evaluation behavior indicators go beyond mere consistency and provide insights into various aspects of peer evaluation performance (see Proposed construct in Table 1). Such insights can inform instructional design and enable targeted feedback. By identifying strengths and weaknesses in specific behavior indicators, EBA allows for the recognition of areas that need intervention, thus promoting the development of peer evaluation skills in line with the goal of formative assessment and learning enhancement.

2.3 Data-driven peer evaluation with learner model

Peer evaluation systems offer learners a scaffold to evaluate their group members and receive real-time feedback with reduced bias, enhanced individualism and privacy protection (Ismail et al., 2016; Cleynen et al., 2020). In online evaluation systems, both the evaluation outputs and the evaluation processes of raters can be traced, providing valuable data for learning analytics applications as part of the learner model attribute. The concept of the "learner model" encompasses domain-specific and domain-independent information, quantified as learning evidence that varies according to the learning context (Boticki et al., 2019). These indicators can derive from learning behaviors recorded on learning management platforms (LMS) such as e-book reading logs, academic scores, previous experiences in group work, and other relevant data. In the context of TBL, the learner model can be dynamic and continuously updated with the accumulation of data from each round. To support this process, Group Learning Orchestration Based on Evidence (GLOBE) (Liang et al., 2021) was proposed as an infrastructure that provides data-driven support for group work based on learner model data. Peer evaluation plays a significant role in GLOBE, serving as a module for collecting peer ratings and feedback (Liang et al., 2022a), while also contributing to the modeling of effective group work and task experiences (Janssen & Kirschner, 2020). By synchronizing the evaluation data with other collaboration attributes from the prior phase, the learner model can be utilized for subsequent rounds of TBL.

The data-driven perspective has been adopted to assess the quality of peer evaluation in individual tasks. For instance, Piech et al. (2013) developed tuned models of rating reliability based on students' previous performance in individual design assignments. Besides, there are studies focusing on written reviews for writing artifacts. Cho & Schunn (2007) considered consistency with others to model reviewers' capabilities, while Patchan et al. (2016) extracted features from review texts, such as sentimental tendencies and comment types, using semantic analysis to build a regression model. Regarding peer evaluation in group work, Liang et al. (2022a) demonstrated that the accumulated learner model, incorporating data on group work and task experiences, can estimate the consistency of peer evaluation using GLOBE. However, for iterative TBL with multiple rounds of group work, the detection of evaluation behaviors on rating scores has yet to be extensively investigated.

3. Method

In this study, we conduct two analyses to answer the research questions. First, we examined the behavior patterns of unserious raters based on evaluation behavior indicators using clustering analysis. Then, to investigate the potential of using the learner model data from learning logs for early detection, a preliminary analysis of classification was conducted. The research questions are as follows.

- RQ1: What are the peer evaluation behavior patterns?

- RQ2: How can the learner model data be used to early detect unserious raters?

3.1 Participant and context

The data of this study comes from a course of a Japanese university with a 4-week experiment. The course is for students beyond sophomore in computer science, with 35 students enrolled this year. The experiment employed an adapted TBL and jigsaw design, which is shown in Figure 1.

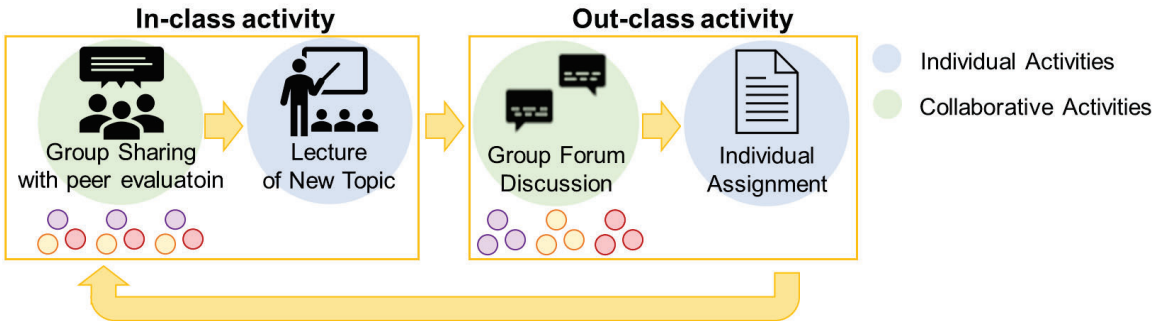

Figure 1. Workflow of the class.

In the first week of the experiment, a lecture on a new topic was delivered and BookRoll, an e-book reading tool that allows instructors to upload learning materials before each class and enables students to engage in various activities during their reading (Ogata et al., 2015), was introduced. Out-of-class activities included reviewing lectures on BookRoll, participating in forum discussions, and completing assignments to summarize them. Starting from the second week, in-class activities began with group sharing of the previous week's assignments. Each student presented the outcome from their forum discussion group in a jigsaw group. Both the forum discussion groups and jigsaw groups were created by the group formation system of GLOBE (Liang et al., 2021) based on each student's reading engagement in BookRoll. In each jigsaw group, the audience provided peer ratings on the individual presentation through the peer evaluation system. The jigsaw group then became the forum discussion group for the following week. Following this, a lecture on the topic of the second week was delivered. This workflow was repeated twice in the first three weeks, and as an assignment in the third week, students created a presentation to the whole class, summarizing what they had learned so far and presented it in the final week's class. The behavior pattern analysis in this study is based on the peer evaluation of this final presentation.

3.2 Data collection and preprocessing

In order to evaluate the final group presentation, students were instructed to assign a score on a 5-star scale to each group in the peer evaluation system (Liang et al., 2022a). The rubric was displayed at the top of the rating section in the peer evaluation system for reference (see Figure 2). The system also recorded a log of the timestamp and rating score each time a rating button was clicked by a student. To ensure privacy, the identity of each student was anonymized from the log. Using the clicking logs, six evaluation behavior indicators introduced in Table 1 were calculated. These indicators are used for visualization and clustering.

To detect unserious behavior prior to the peer evaluation, data from the learner model was collected. In this study, the following learner model data was available before the final peer evaluation of presentations:

- **Reading engagement (RE)**, which includes reading time, operation times, completion rate, and the number of red markers, yellow markers, and memos on the e-book platform BookRoll (Ogata et al., 2015). A comprehensive coefficient was calculated by averaging the percentage rank of the aforementioned indicators to represent reading engagement.

- **Forum engagement (FE)**, which consolidates the number of forum posts and characters in the out-class forum discussion. The percentage rank of the former indicators was consolidated to represent the forum engagement.
- **Prior evaluation behavior indicators**, which refer to indicators collected during the peer evaluation of individual presentations in jigsaw groups in the second (round 1) and third week (round 2). The six indicators introduced in Table 1 were collected for the first two rounds as the input indicators for classification.

Figure 2. Peer evaluation system with rubrics.

As some of the prior evaluation behavior indicators were found to be highly correlated and estimating the same facet, as also mentioned in Horikoshi et al. (2022), we performed dimension reduction through factor analysis. Based on the factor analysis, we combined ET and tSD as "**time feature (TF)**" (explaining 99.01% of variance for round 1 and 98.63% for round 2), and sM and sSD as "**scoring feature (SF)**" (explaining 84.91% of variance for round 1 and 91.97% for round 2). Additionally, since the extent of polarization in tM was deemed important in the pattern, we derived a new indicator (**tDEV**) from tM, which represents the z-score of tM and describes the deviation of rating time from the mean. **CC** is treated as an independent indicator due to its low correlation with other features. We used eight prior behavior indicators (four for each round) for the classification modeling.

3.3 Data analysis

To answer RQ1, we performed a clustering analysis to differentiate unserious raters from the participants. This analysis entailed clustering the students according to their evaluation behavior indicators, which were obtained from the final round of peer evaluation (for group presentation). We utilized the K-means method to obtain two distinct clusters, with the highest silhouette score. Subsequently, we examined the behavior patterns of the students by analyzing the distributions of each evaluation behavior indicator within each cluster.

For RQ2, we approached it as a binary classification problem to determine if the rater is unserious in evaluating the final group presentation. To accomplish this, we tested five commonly used machine learning classification models for numerical data and evaluated their

performance using the Area Under Curve (AUC) (Fawcett, 2006), with values ranging from 0 to 1. Furthermore, we conducted a feature ablation analysis (Gabrilovich & Markovitch, 2004) based on the information gain (IG) of ten input indicators as discussed in the previous section, to figure out the predictive indicators for the classification.

4. Result

4.1 Behavior patterns clustering

Figure 3 illustrates the distribution of EBA indicators for each cluster. It is evident that students in cluster C1 possess longer ET, more CC, and give a wider range of scores with lower sM and higher sSD. Although tM does not show a significant difference between the two clusters, the dispersion differs. Raters in C1 participated in peer evaluations during the presentation, and their distribution of timestamps appears to be more normalized. On the other hand, students in cluster C2 have shorter ET, fewer CC, polarized tM, and smaller tSD. Regarding scores, they tend to provide full marks, indicated by high sM and minimal sSD.

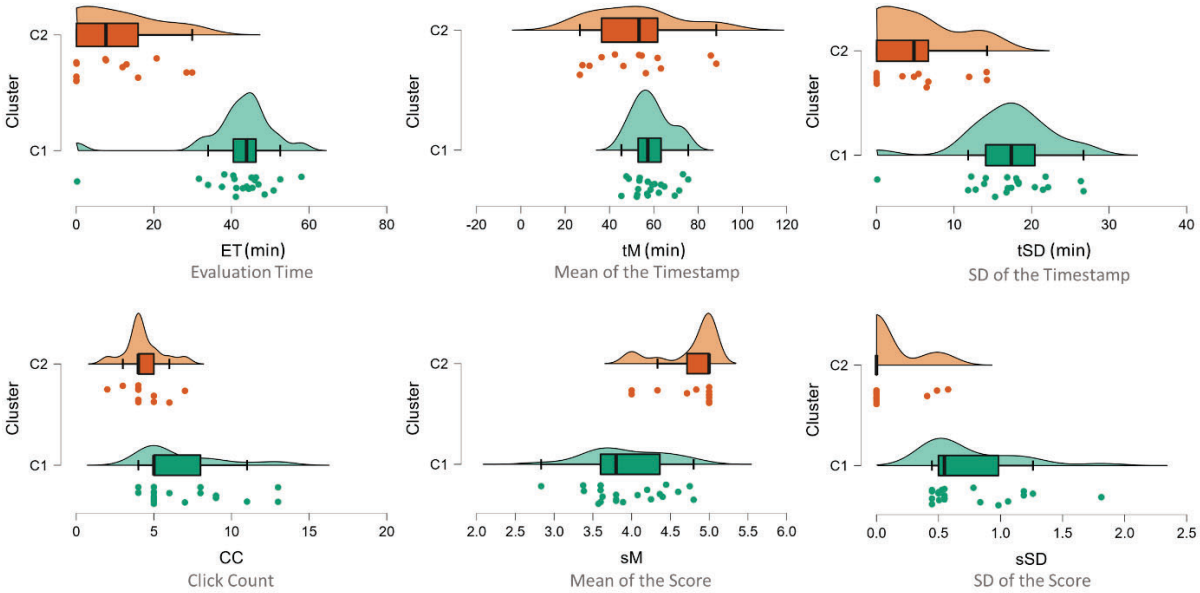

Figure 3. Distribution of EBA indicators of the two clusters.

Table 2. *Statistical analysis of EBA indicators for clustering*

	Cluster	N	Mean	SD	p
ET	C1	21	41.716	11.259	< .001***
	C2	14	9.658	10.748	
tM	C1	21	58.870	8.416	0.077
	C2	14	50.775	19.494	
tSD	C1	21	17.137	5.670	< .001***
	C2	13	5.173	5.401	
CC	C1	21	6.762	2.791	< .001***
	C2	14	4.071	1.492	
sM	C1	21	3.941	0.508	< .001***
	C2	14	4.777	0.378	
sSD	C1	21	0.746	0.370	< .001***
	C2	13	0.113	0.218	

****p* < .001.

4.2 Early detection of unserious raters

Figure 4 presents a performance comparison of various classification methods when using the top N input indicators ranked by IG, and Table 2 listed these indicators in the order of their IG in the classification modeling. Our analysis suggests that neural network and logistic regression models outperform other methods when utilizing the top five to six input indicators with high information gains. The AUC scores were 0.738 for the 5-feature condition (Neural Network) and 0.731 for the 6-feature condition (Logistic Regression).

As for predictive indicators, we observed that the deviation rating timestamp for round 2, indicating a straightlining pattern, had the highest IG. Additionally, SF for both rounds exhibited high information gains. Interestingly, all four prior behavior indicators for round 2 ranked in the top six indicators of the classification model. We also observed a significant difference between the two groups in SF for round 1 and TF for round 2. Meanwhile, the reading behaviors of the two groups that occurred before the assessment started. The RE feature also provided valuable information for distinguishing between different classes in a classification, underscoring the importance of integrating learning model data in predictive modeling. Conversely, the tDEV, CC, and TF of round 1 had low IG, which could be attributed to the unfamiliarity with the system in the first round as students needed time to get accustomed to it.

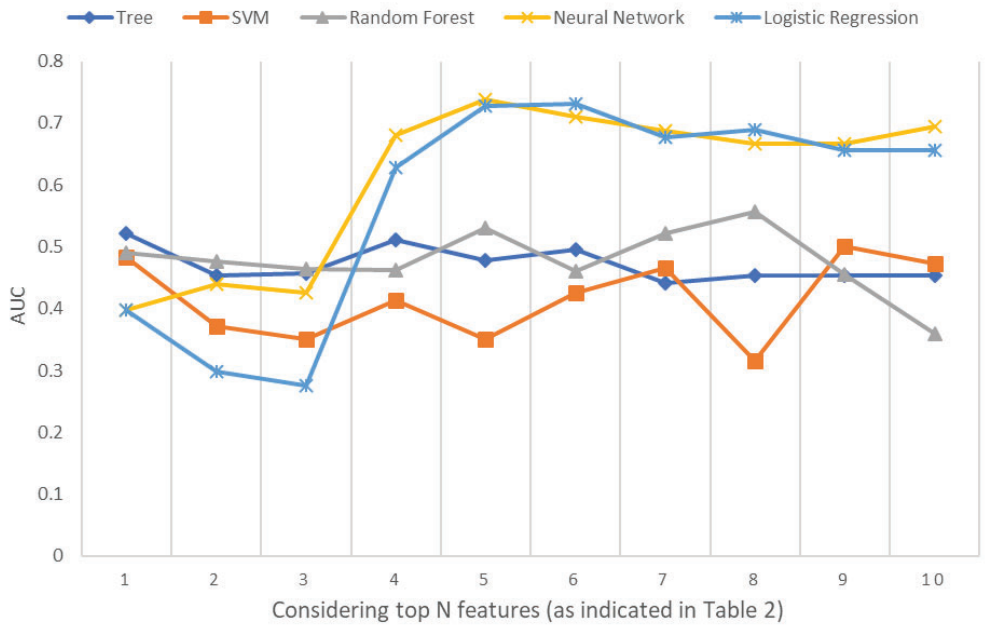

Figure 4. Prediction accuracy of classification based on learner model

Table 2. *Input indicators for the classification modelling ranked by information gain*

Rank	Indicator	Information Gain (IG)	t
1	tDEV-2	0.226	0.974
2	SF-2	0.211	0.865
3	RE	0.205	0.971
4	SF-1	0.178	3.251**
5	CC-2	0.154	1.976
6	TF-2	0.077	2.498*
7	tDEV-1	0.071	0.974
8	FE	0.055	1.264
9	CC-1	0.049	0.773
10	TF-1	0.031	0.397

*p < .05, **p < .01.

5. Discussion

The findings of this study emphasize the significance of integrating learning model data in peer evaluation of TBL. By using EBA indicators, we can analyze the time and scoring features of peer evaluation as the presentation progresses. These indicators can reveal behavior patterns suggested by Horikoshi & Tamura (2021) such as modifying the evaluation, spending time on the evaluation, or evaluating all evaluation items earlier or many evaluation items later. The clustering corroborates these patterns and identifies characteristics of unserious raters.

To provide a clearer understanding of the behavior pattern, Figure 5 displays plots of the evaluation behavior of typical raters in the two clusters, indicating the timestamps, scores, and rating intervals. The x-axis represents the elapsed time from the start of the first group presentation, and the y-axis denotes the candidate number of peer ratings. It can be observed that typical students in C1 tend to rate each candidate group across the group presentation sessions with even intervals. Moreover, they use different scores with noticeable variations. In contrast, typical students in C2 exhibit a straightlining and speedy pattern (Zhang & Conrad, 2014; Kim et al., 2019), completing their rating very quickly either at the beginning or the end of the session. In summary, C1 raters spend more time evaluating their peers, give a diverse range of scores with less agreement among themselves, and exhibit a more even distribution of timestamps when giving their evaluations. C2 raters, on the other hand, spend less time evaluating their peers, give higher scores with less variance, and show a polarized distribution of timestamps for their evaluations. These differences suggest that C1 raters demonstrate more thoughtful and critical evaluations, while C2 raters appear to be more lenient and less engaged in the evaluation process.

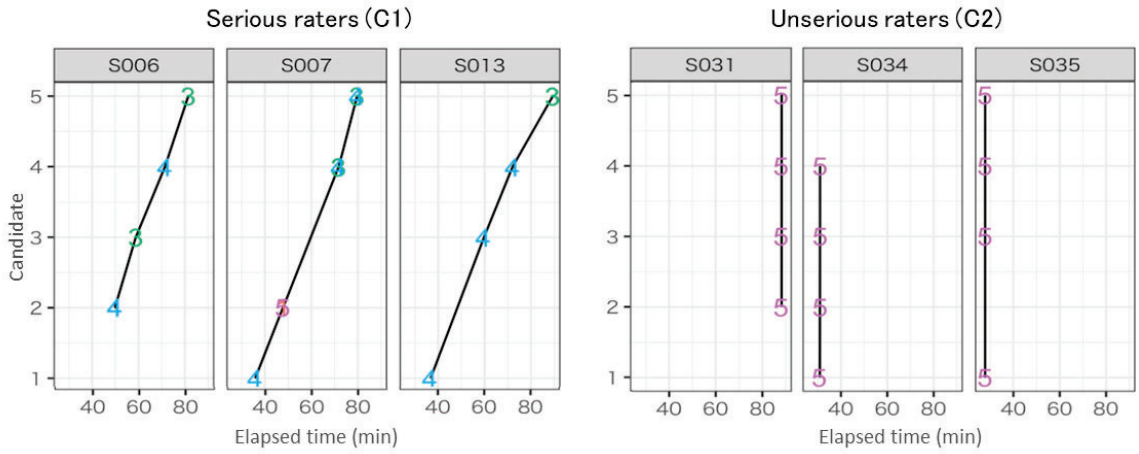

Figure 5. Visualization evaluation behavior of typical raters in two clusters.

Moreover, this study presents the potential of using learner model data collected from all phases of TBL in previous rounds to early predict unserious raters. Our analysis shows that scoring features in each round of TBL play a significant role in the detection model. Time features, which describe the time distribution and frequency of the ratings, can also be predictive when TBL is conducted over multiple rounds and raters become familiar with the system. Furthermore, the engagement of students in individual reading activities can serve as a predictor of unserious raters, while their forum engagement appears to be less relevant. This discrepancy may be attributed to forum posts being compulsory and formatted as part of the course grade, resulting in minimal variation among learners. In summary, the prediction model is expected to empower instructors to provide remedial instructions or give automatic nudges to these at-risk students, improving the reliability of peer evaluation as a formative assessment in TBL. These prompts can be delivered through group awareness tools (Strauß & Rummel, 2021) and email interventions (Damgaard & Nielsen, 2018).

The study also contributes to TBL design by introducing the potential of data-driven scaffolding with multiple rounds of group learning, where learning log data from previous rounds can be utilized for various learning analytics purposes. One example of data-driven

support is provided for the detection of unserious raters in peer evaluation, which is intended to improve the quality of peer assessment. Furthermore, continuous data support has broader applications in group learning. The learning logs of previous activities can provide data for creating groups (Liang et al., 2022b) and calibrating peer rating scores (Piech et al., 2013). The accumulated data can also be useful for data visualization platforms for reflecting on teaching interventions (Kuromiya et al., 2020).

However, there are several limitations to this study. The sample size of learners observed was relatively small, which might limit the generalizability of the findings. Moreover, it should be noticed that the current predictive model's AUC did not achieve a high level, and the model needs to be validated using a different student population. Besides behavior indicators, we plan to incorporate the consistency of the ratings, including the agreement with instructor-assigned grades and average student-assigned grades (Fukazawa, 2010), into the prior evaluation behavior indicators. Further, considering more predictors in the model, such as learning outcomes, collaborative skills, and personality variables (Piech et al., 2013; Sánchez et al., 2021), could also enhance its effectiveness. Qualitative observations and self-reports can offer valuable insights into the reasons behind unserious patterns, and exploring how the presented EBA estimated from logs connects to the observations is another promising topic. Lastly, since this research only involved one trial of a group presentation, conducting additional studies with more rounds of TBL and peer-evaluated group presentations is anticipated to address remaining issues and enhance the robustness of the findings.

6. Conclusion

In conclusion, this study discusses the issue of unserious raters in peer evaluation of group learning. We propose a method to describe unserious peer raters by detecting trends based on the clustering of EBA indicators. The results reveal typical behavior patterns of unserious raters: straightlining, speeding, and giving all full marks. Next, a preliminary evaluation is conducted for classifiers that can identify groups of unserious raters. The results revealed typical time and scoring features associated with these raters, as well as predictive indicators for early detection. Overall, these findings have implications for improving the effectiveness and reliability of peer evaluation in group learning contexts. Further investigation is required to explore the actual quality of ratings and validate the classification model.

Acknowledgements

This research was supported by the following grants: JSPS KAKENHI 20K20131, 20H01722, 22H03902, NEDO JPNP20006, JPNP18013, and JST JPMJSP2110.

References

Amarasinghe, I., Hernández-Leo, D., & Ulrich Hoppe, H. (2021). Deconstructing orchestration load: comparing teacher support through mirroring and guiding. *International Journal of Computer-Supported Collaborative Learning*, *16*(3), 307-338.

Cho, K., & Schunn, C. D. (2007). Scaffolded writing and rewriting in the discipline: A web-based reciprocal peer review system. *Computers & Education*, *48*(3), 409-426.

Damgaard, M. T., & Nielsen, H. S. (2018). Nudging in education. *Economics of Education Review*, *64*, 313-342.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern recognition letters*, *27*(8), 861-874.

Fukazawa, M. (2010). Validity of peer assessment of speech performance. *ARELE: Annual Review of English Language Education in Japan*, *21*, 181-190.

Gabrilovich, E., & Markovitch, S. (2004, July). Text categorization with many redundant features: Using aggressive feature selection to make SVMs competitive with C4. 5. In *Proceedings of the twenty-first international conference on Machine learning* (p. 41).

Goolsarran, N., Hamo, C. E., & Lu, W. H. (2020). Using the jigsaw technique to teach patient safety. *Medical education online*, *25*(1), 1710325.
Gorham, T., Majumdar, R., & Ogata, H. (2023). Analyzing learner profiles in a microlearning app for training language learning peer feedback skills. *Journal of Computers in Education*, 1-26.
Horikoshi, I., Liang, C., Majumdar, R., & Ogata, H. (2022). Applicability and reproducibility of peer evaluation behavior analysis across systems and activity contexts. In *30th International Conference on Computers in Education Conference Proceedings* (Vol. 1, pp. 335-345).
Horikoshi, I., & Tamura, Y. (2021). How do students evaluate each other during peer assessments? An analysis using "evaluation behavior" log data. *Educational technology research*, *43*(1), 3-21.
Janssen, J., & Kirschner, P. A. (2020). Applying collaborative cognitive load theory to computer-supported collaborative learning: Towards a research agenda. *Educational Technology Research and Development*, *68*(2), 783-805.
Johnson, L. D. (2017). Exploring cloud computing tools to enhance team-based problem solving for challenging behavior. *Topics in Early Childhood Special Education*, *37*(3), 176-188.
Kim, Y., Dykema, J., Stevenson, J., Black, P., & Moberg, D. P. (2019). Straightlining: Overview of measurement, comparison of indicators, and effects in mail–web mixed-mode surveys. *Social Science Computer Review*, *37*(2), 214-233.
Kuromiya, H., Majumdar, R., & Ogata, H. (2020). Fostering evidence-based education with learning analytics. *Educational Technology & Society*, *23*(4), 14-29.
Liang, C., Gorham, T., Horikoshi, I., Majumdar, R., & Ogata, H. (2022a). Estimating peer evaluation potential by utilizing learner model during group work. In *Collaboration Technologies and Social Computing: 28th International Conference, CollabTech 2022* (pp. 287-294).
Liang, C., Majumdar, R., Horikoshi, I., Flanagan, B. & Ogata, H. (2022b). Exploring predictive indicators of reading-based online group work for group formation assistance. In *30th International Conference on Computers in Education Conference Proceedings* (Vol. 1, pp. 642-647).
Liang, C., Majumdar, R. & Ogata, H. (2021). Learning log-based automatic group formation: system design and classroom implementation study. *Research and Practice in Technology Enhanced Learning*, *16*(1), 1–22.
Michaelsen, L. K., Knight, A. B., & Fink, L. D. (Eds.). (2002). *Team-based learning: A transformative use of small groups*. Greenwood publishing group.
Ogata, H., Yin, C., Oi, M., Okubo, F., Shimada, A., Kojima, K., & Yamada, M. (2015). E-Book-based learning analytics in university education. In *International conference on computer in education (ICCE 2015)* (pp. 401-406).
Ohland, M. W., Loughry, M. L., Woehr, D. J., Bullard, L. G., Felder, R. M., Finelli, C. J., ... & Schmucker, D. G. (2012). The comprehensive assessment of team member effectiveness: Development of a behaviorally anchored rating scale for self-and peer evaluation. *Academy of Management Learning & Education*, *11*(4), 609-630.
Parmelee, D., Michaelsen, L. K., Cook, S., & Hudes, P. D. (2012). Team-based learning: a practical guide: AMEE guide no. 65. *Medical teacher*, *34*(5), e275-e287.
Patchan, M. M., Schunn, C. D., & Correnti, R. J. (2016). The nature of feedback: How peer feedback features affect students' implementation rate and quality of revisions. *Journal of Educational Psychology*, *108*(8), 1098.
Piech, C., Huang, J., Chen, Z., Do, C., Ng, A., & Koller, D. (2013). Tuned models of peer assessment in MOOCs. *arXiv preprint arXiv:1307.2579*.
Sánchez, O. R., Ordonez, C. A. C., Duque, M. A. R., & Pinto, I. I. B. S. (2021). Homogeneous group formation in collaborative learning scenarios: An approach based on personality traits and genetic algorithms. *IEEE Transactions on Learning Technologies*, *14*(4), 486-499.
Strauß, S., & Rummel, N. (2021). Promoting regulation of equal participation in online collaboration by combining a group awareness tool and adaptive prompts. But does it even matter?. *International Journal of Computer-Supported Collaborative Learning*, *16*, 67-104.
To, J., & Panadero, E. (2019). Peer assessment effects on the self-assessment process of first-year undergraduates. *Assessment & Evaluation in Higher Education*, *44*(6), 920-932.
Van Zundert, M., Sluijsmans, D., & Van Merriënboer, J. (2010). Effective peer assessment processes: Research findings and future directions. *Learning and instruction*, *20*(4), 270-279.
Yan, T., & Tourangeau, R. (2008). Fast times and easy questions: The effects of age, experience and question complexity on web survey response times. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, *22*(1), 51-68.
Zhang, C., & Conrad, F. (2014, July). Speeding in web surveys: The tendency to answer very fast and its association with straightlining. In *Survey research methods* (Vol. 8, No. 2, pp. 127-135).