Research and Practice in
Technology Enhanced Learning

RESEARCH Open Access

Rater behaviors in peer evaluation: Patterns and early detection with learner model

Changhao Liang [1]*, Izumi Horikoshi [1], Rwitajit Majumdar [2] and Hiroaki Ogata [1]

*Correspondence:
liang.changhao.8h@kyoto-u.ac.jp
Academic Center for Computing and Media Studies,
Kyoto University, Japan
Full list of author information is available at the end of the article

Abstract

Peer evaluation is a common practice in team-based learning (TBL) designs, which can cover the assessment of individual or group work. However, the integrity of peer evaluation can be compromised by unserious raters—individuals who do not earnestly engage in the evaluation process. These raters may exhibit behaviors like consistently assigning the same score, rushing through evaluations, or evaluating before or long after the target presentations. This study delves into the issue of unserious peer evaluation in group presentations, with a specific focus on understanding the behavior patterns in the digital system. Utilizing evaluation behavior analysis (EBA) indicators, we identify patterns linked to unserious raters during the peer evaluation process. Meanwhile, we also connect these patterns to rating consistency and actual course performance, underscoring the significance of behavior patterns. Further, we conduct a preliminary analysis to explore the application of learner model data available before the peer evaluation starts for the early detection of unserious raters. This finding can assist teachers in providing personalized prompts and interventions before the peer evaluation stage, hence enhancing the evaluation quality through targeted interventions in a timely manner.

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

Introduction

Evaluation is an essential aspect in collaborative learning, but teachers may struggle to supervise every small group and properly evaluate each student (Amarasinghe et al., 2021; Yoon et al., 2018). Peer evaluation offers formative feedback that encourages reflection and overcomes the limitations of traditional evaluation such as social loafing and free ridings (Kasch et al., 2021; Ohland et al., 2012; Strijbos, 2010). It has become widely adopted in online settings where student-centered learning is prevalent and can enhance both learning and interpersonal skills (Chen et al., 2021; Rohmah et al, 2021). However, some raters may not take the evaluation process seriously, as Horikoshi and Tamura (2021) discovered. Such evaluations involve consistently assigning the same score, rushing through evaluations, and evaluating before or long after target presentations. 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 performance (Piech et al., 2013), afford group awareness information (Strauß & Rummel, 2021), 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. With the accumulation of online learning logs and the scaffold of learning analytics, we find an opportunity to model such capabilities in a data-driven environment.

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. Subsequently, we connected these behavior patterns with consistency indicators and actual course performance to understand the significance of evaluating behaviors. 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.

Research background

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 (Gorham et al., 2023; Ohland et al., 2012). 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 involving multiple rounds of group work with peer evaluation, initially introduced in medical education (Michaelsen et al., 2002). In each TBL round, students begin by exploring the learning topic individually before collaborating in teams to accomplish tasks (Parmelee et al., 2012). Collaborative learning processes, such as group discussions and presentations, form a crucial part of teamwork applications, while peer evaluation serves as the conclusion for each round. Peer evaluation is a crucial stage in TBL, which can ensure accountability for group work (Yoon et al., 2018). During the peer evaluation stage, students assess their peers' learning outcomes and engage in reflective practices as part of a formative process (Topping, 1998). Additionally, in the data-driven environment, previous rounds' learning log data empowers teachers to implement targeted interventions (Johnson, 2017).

Nowadays, digital systems facilitate in-class peer evaluation activities promptly and anonymously (Cleynen et al., 2020). This provides teachers with greater flexibility in integrating peer evaluation into their class design. The implementation of peer evaluation not only assigns a course grade to assess the quality of the learning outcome but also serves as a learning process. Through feedback, learners can identify their strengths and weaknesses, fostering improvement through critical thinking and self-reflection (Horikoshi & Tamura, 2021).

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). Therefore, the behavior paradata in web surveys deserve further attention when filtering invalid responses, as it can reflect the quality of answers.

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 from 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)

In addition to the behavior perspective that highlights the evaluation process, there is another consistency perspective in peer evaluation that predominantly focuses on scores and conformity with others (Cho & Schunn, 2007; Fukazawa, 2010). Two constructs define consistency: validity and reliability. Validity is assessed by comparing the student evaluation with a standard, such as a teacher's score (Kulkarni et al., 2013). Validity can be calculated only when the instructor assigns grades, which typically measure knowledge while neglecting student contribution (Yoon et al., 2018). Reliability is determined by the consistency among students' evaluations, estimated through the deviation from the average scores given by all raters. Reliability dynamically changes during the assessment session with constant submissions from raters.

Compared to the conventional perspective of peer evaluation consistency as consequences, the EBA indicators focus on the process of the peer evaluation. The initial objectives of EBA are to instruct students for appropriate evaluations thus enhancing the overall effectiveness of class design. These evaluation behavior indicators go beyond mere consistency, offering insights into different aspects of peer evaluation performance (see the proposed construct in Table 1). They furnish valuable insights for guiding appropriate score selection, determining evaluation timing, and improving the design of rating forms. These insights from EBA can unveil the reasons behind unexpected rating behaviors, thereby informing instructional design and enabling targeted feedback. By identifying strengths and weaknesses in specific behavior indicators, EBA allows for the recognition of areas requiring intervention, hence promoting the development of peer evaluation skills in line with the goal of formative assessment and enhancement of class design. Concurrently, learners can benefit from peer evaluation activities in TBL to refine their presentations, deepen their understanding of the learning topic, and in turn achieve the learning goals of the lecture.

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 (Cleynen et al., 2020; Ismail et al., 2019). 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. Within the GLOBE framework, data-driven support in group learning comprises four phases: formation, orchestration, evaluation, and reflection. The framework is operationalized through an algorithmic group formation system, a forum discussion dashboard, and a peer evaluation system, all utilizing learning logs. Peer evaluation plays a significant role in GLOBE, serving as a data sensor for collecting peer ratings and feedback (Liang et al., 2022), 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 and 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. (2022) 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.

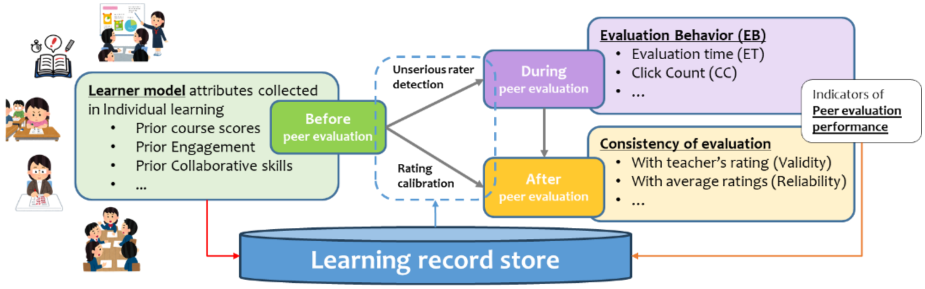

Learner model attributes collected in Individual learning
Prior course scores
Prior Engagement
Prior Collaborative skills
...
Before peer evaluation
Unserious rater detection
Rating calibration
During peer evaluation
After peer evaluation
Evaluation Behavior (EB)
Evaluation time (ET)
Click Count (CC)
...
Consistency of evaluation
With teacher's rating (Validity)
With average ratings (Reliability)
...
Indicators of Peer evaluation performance
Learning record store

Fig. 1 Data-driven design of peer evaluation related study

In the context of peer evaluation activity, data related to peer evaluation attributes can be categorized based on the phase they target. As depicted in Figure 1, antecedent indicators before peer evaluation, derived from data in individual learning activities, have the predictive capacity for the peer evaluation performance of raters, referred to as the learner model. The phases during and after peer evaluation can serve as reflections of the dynamics within the ongoing peer evaluation activity, and they can also be reused as antecedents for subsequent rounds of evaluation tasks. All this log data is recorded as learner model attributes that can be utilized for learning analytics models, including performance prediction and capability estimation.

Method

In this study, we perform several analyses to address our research questions. First, we examined the behavior patterns of unserious raters using clustering analysis based on evaluation behavior indicators. Then, statistical comparisons on the rating consistency with average level and actual course performance were implemented between clusters with different behavior patterns. Subsequently, we implement statistical comparisons on rating consistency with the average level and actual course performance across clusters exhibiting different behavior patterns.

Further, to explore the feasibility of utilizing learner model data from learning logs for early detection, we conduct a preliminary classification analysis. The research questions guiding our investigation are outlined below.

RQ1: What are the distinct behavior patterns exhibited by unserious raters during peer evaluations?

RQ2: What significance do behavior patterns hold for reflecting rating consistency and actual performance?

RQ3: How can the learner model data be used to early detect unserious raters before peer evaluations?

Participant and context

The data of this study comes from a course of a Japanese university with a four-week experiment. The course is for students beyond sophomore in computer science, with 35 students enrolled this year. It should be noted that one student withdrew from the course midway, so the course grade for this student is not available for the subsequent data analysis, despite their participation in the four-week learning activities. The experiment employed an adapted TBL and jigsaw design (Goolsarran et al., 2020), which is shown in Figure 2.

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. The system enables the formation of homogeneous or heterogeneous groups based on diverse learner attributes extracted from learning logs. It employs a genetic algorithm, with the squared difference within each group serving as the fitness function (high for heterogeneous and low for homogeneous). To ensure balanced compositions among groups, we employed the heterogeneous strategy, including both high- and low-engaged students in lecture slide reading and forum discussions within each group.

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.

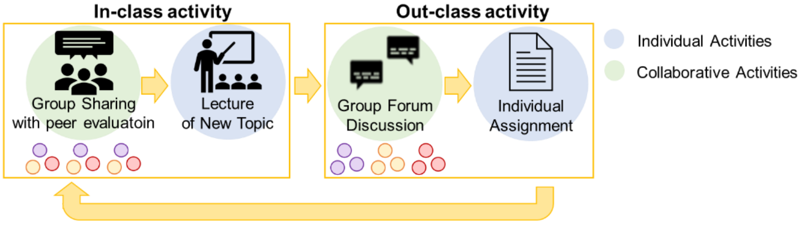
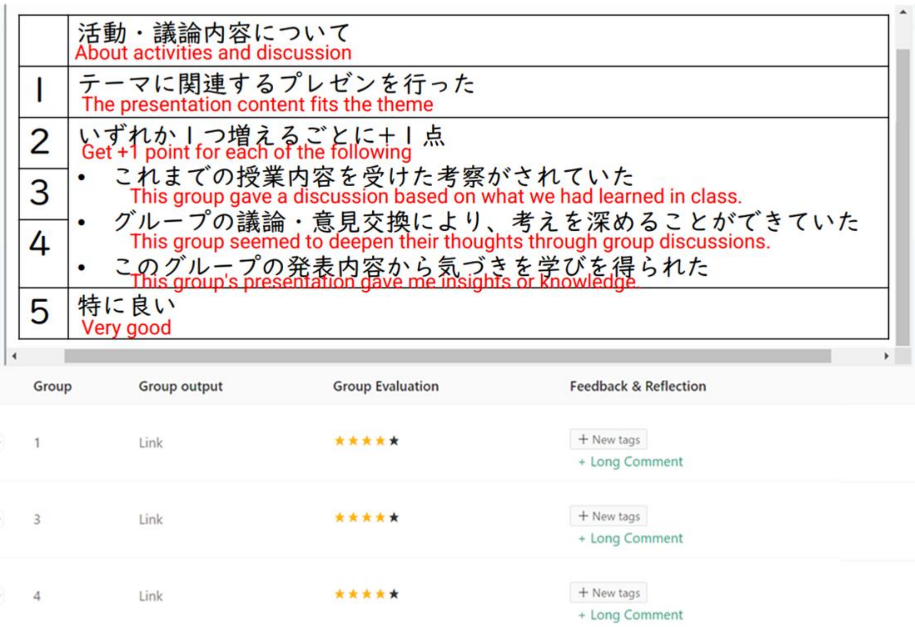


Fig. 2 Workflow of the class



	活動・議論内容について About activities and discussion
1	テーマに関連するプレゼンを行った The presentation content fits the theme
2	いずれか１つ増えるごとに＋１点 Get +1 point for each of the following
3	• これまでの授業内容を受けた考察がされていた This group gave a discussion based on what we had learned in class.
4	• グループの議論・意見交換により、考えを深めることができていた This group seemed to deepen their thoughts through group discussions. • このグループの発表内容から気づきを学びを得られた This group's presentation gave me insights or knowledge.
5	特に良い Very good

Group	Group output	Group Evaluation	Feedback & Reflection
1	Link	★★★★★	+ New tags + Long Comment
3	Link	★★★★★	+ New tags + Long Comment
4	Link	★★★★★	+ New tags + Long Comment

Fig. 3 Peer evaluation system for group evaluation with rubrics

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., 2022). The rubric was displayed at the top of the rating section in the peer evaluation system for reference (see Figure 3). 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.

Data collection and preprocessing

To detect unserious behavior prior to the peer evaluation (RQ3), 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. For each indicator, the percentage rank is calculated by (1). To scale various indicators, the percentage rank indicates the relative position of the value compared to others in the dataset, ranging from 0% to 100%.

$$\text{Percentage rank} = \frac{rank\ of\ current\ value - 1}{total\ number\ of\ values - 1} \qquad (1)$$

- **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 for the first two rounds were collected as the input indicators for classification.

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. The primary goal of this dimension reduction was to minimize the number of inputs for the classification and enhance the interpretability of the educational implications associated with these input indicators, going beyond pure behaviors. 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 absolute 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 utilized eight prior behavior indicators (four for each round) for the classification modeling, distinguishing indicators from different rounds of peer evaluation by adding a suffix number.

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 performed statistical examinations to assess the disparity in terms of the deviation from average peer rating scores and the course grades of this four-week section between two clusters of raters. Since the Shapiro-Wilk test indicated non-normality for the dependent variables, we employed the Mann-Whitney test.

For RQ3, 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. Next, we conducted a feature ablation analysis (Gabrilovich & Markovitch,

2004) based on the information gain (IG) of ten input indicators as discussed in the “Data collection and preprocessing” section, to figure out the predictive indicators for the classification. Further, not restricted to evaluation behaviors, to inspect the impact the learner model indicators on the consistency, we adopted correlation analysis to figure out potential input for the consequence of the peer evaluation.

Result

Behavior patterns clustering

Figure 4 illustrates the distribution of EBA indicators for each cluster and Table 2 shows the statistics of Mann-Whitney test. It is evident that students in cluster C1 possess longer

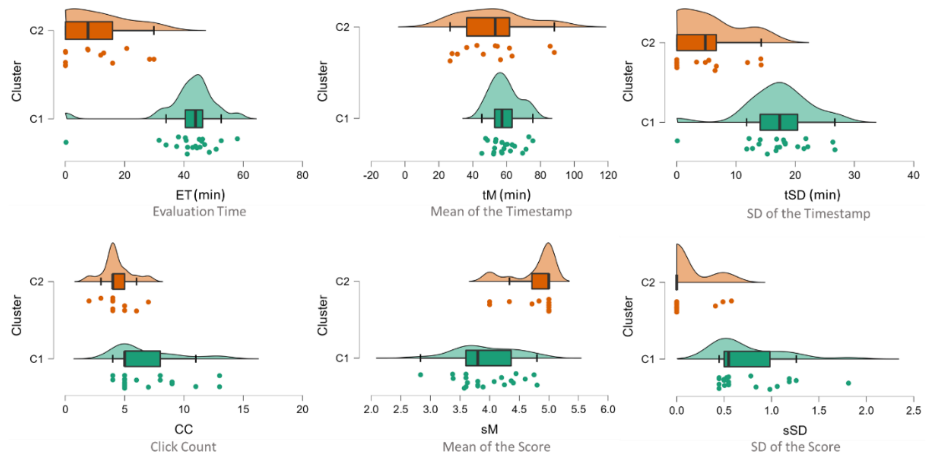


Fig. 4 Distribution of EBA indicators of the two clusters

Table 2 Descriptive statistics and Mann-Whitney test of EBA indicators between two clusters

	Cluster	N	Mean	SD	Range	Mean of $\lvert z\rvert$	p
ET	C1	21	41.716	11.259	57.800	0.817	< .001***
	C2	14	9.658	10.748	29.867	1.018	
tM	C1	21	58.870	8.416	30.170	0.491	.077
	C2	14	50.775	19.494	61.604	1.130	
tSD	C1	21	17.137	5.670	26.607	0.756	< .001***
	C2	14	4.804	5.370	14.250	0.985	
CC	C1	21	6.762	2.791	9	0.774	< .001***
	C2	14	4.071	1.492	6	0.697	
sM	C1	21	3.941	0.508	1.967	0.819	< .001***
	C2	14	4.777	0.378	1.000	0.957	
sSD	C1	21	0.746	0.370	1.365	0.620	< .001***
	C2	14	0.105	0.212	0.577	0.903	

***p < .001.

ET, more CC, and give a wider range of scores with lower sM and higher sSD. This may indicate that cluster C1 is an active and conscientious evaluation group. Although tM does not show a significant difference between the two clusters, the deviation (absolute *z*-score) differs (n = 35, p = .044 < .05 for Mann-Whitney test). This may suggest that the evaluation behaviors of cluster C1 are more consistent. 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.

Consistency and course grade of unserious raters

Table 3 presents the outcomes of the Mann-Whitney test concerning the significance of behavior patterns. The results indicate no significant difference in the deviation from average peer rating scores between the two types of raters, underscoring the independence of behavior patterns from indicators related to the consistency perspective (n = 35, p = .933). Concurrently, an observation was made that the course scores of unserious raters in the four-week learning task were significantly lower than those of serious raters (n = 34, p = .032 < .05, one student's course score is not available due to midway withdrawal).

Early detection of unserious raters

Figure 5 presents a performance comparison of various classification methods when using the top N input indicators ranked by IG, and Table 4 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 (tDEV) 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. This

Table 3 Mann-Whitney test on significance of behavior patterns

	Group	N	Mean	SD	p
Deviation from average	C1	21	0.564	0.310	.933
	C2	14	0.549	0.221	
Course score	C1	21	8.008	1.421	.032*
	C2	13	6.423	2.626	

*p < .05.

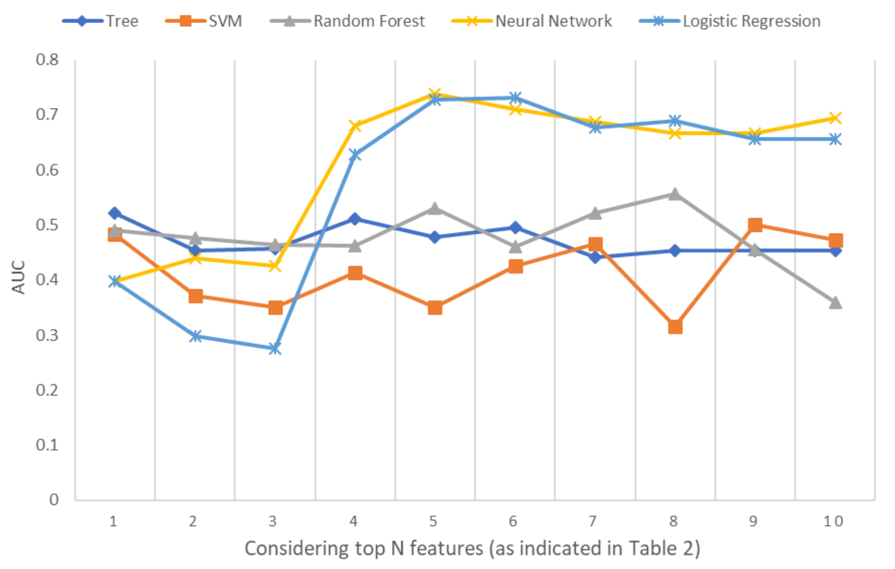


Fig. 5 Prediction accuracy of classification based on learner model

Table 4 Input indicators for the classification modeling 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.

suggests that in the initial peer evaluation, the serious raters in the C1 group paid more attention to scoring and aimed to assign variant scores to candidates based on the rubrics. As they became familiar with the rubrics and evaluation process, they tended to allocate time more evenly for each presentation in the second round.

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.

Table 5 Correlation between deviation from the average score and various learner model indicators

Indicator with IG rank	Pearson's r	p
3. RE	-0.318	0.031*
5. CC-2	0.244	0.903
2. SF-2	-0.232	0.109
7. tDEV-1	-0.076	0.347
1. tDEV-2	-0.048	0.400
6. TF-2	0.039	0.582
4. SF-1	0.027	0.555

*p < .05

Further, we conducted a correlation analysis for indicators with top-rank information gains to determine whether these indicators are also associated with consistency. Table 5 shows the correlation analysis results. We found that reading engagement indicators exhibit a significant negative relationship with the deviation from the average score (n = 35, r = -0.318, p = .031 < .05). While other input learner model attributes do not exhibit significant association.

Discussion

The findings of this study emphasize the significance of integrating learning model data in peer evaluation of TBL. Three research questions aim to address "what", "why" and "how" of this issue respectively. 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 and 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.

Typical behaviors of unserious raters (RQ1)

To provide a clearer understanding of the behavior pattern, Figure 6 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 (Kim et al., 2019; Zhang & Conrad, 2014), 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,

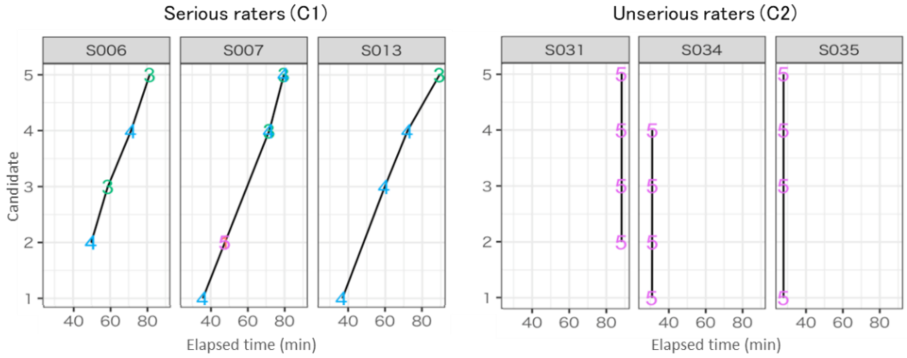


Fig. 6 Visualization evaluation behavior of typical raters in two clusters

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. Based on the constructs of the EBA indicators in Table 1, we can label the two clusters as serious (C1) and unserious (C2) raters.

Significance of underscoring behavior patterns (RQ2)

Besides behavior patterns of unserious raters, we also examine the significance of focusing on the evaluation behavior. Though we found no significant difference in their consistency, this result aligns with previous studies suggesting that behavior patterns and consistency indicators illustrate different aspects of peer evaluation activity. Therefore, the results support the earlier argument in Horikoshi and Tamura (2021) that these aspects represent independent facets of peer evaluation. It proves that focusing on the evaluation behaviors in addition to the scores has additional significance. Concurrently, the observed difference in course grades suggests that unserious raters in peer evaluation also tend to invest less effort in the overall learning task. This connection resonates with the implications of web survey responses for filtering low-quality answer (Couper & Kreuter, 2013), validating hypotheses with shared mechanisms, and the appropriateness of leveraging web survey theories in peer evaluation behaviors. In addition, the results also unveil a mutual association between rater ability and overall course performance. Not only can learning performance predict rater seriousness (Piech et al., 2013), but serious raters also tend to perform well in the overall course, demonstrating a rigorous attitude toward their academic responsibilities. Moreover, this finding aligns with the original goals of EBA – improving presentation quality and overall effectiveness of course design – as articulated in Horikoshi and Tamura (2021), which is evident in their connection to course grades.

Predictive learner model indicators for detection (RQ3)

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 (SF) in each round of TBL play a significant role in the detection model. Time features (TF), 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 on forum activities may be attributed to forum posts being compulsory and formatted as part of the course grade, resulting in minimal variation among learners. Further analysis on consistency indicators suggests that individuals with higher reading engagement in individual reading tend to demonstrate higher reliability in peer evaluation, as they provide scores closer to the average level. This result also reflects the significance of the individual preparation phase on the performance of subsequent TBL phases, as shown in Lyu et al. (2023). 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). Additionally, given that the final group presentations were not assessed by the teacher, further investigation is needed to explore whether the evaluation behavior patterns of conscientious raters can uphold higher validity, evidenced by the alignment of their scoring consistency with teachers' expectations.

Contributions and limitations

First, the study exemplifies the implementation of the data-driven design in Computer-Supported Collaborative Learning (CSCL). Learner model attributes, which involve all data collected during the individual learning phase of TBL, portray learning-associated characteristics with the potential to enhance peer rating quality. This data can be synthesized into rater modeling, with specific weights fine-tuned for each learning attribute. During the group learning phase, the dynamics of the evaluation process from peer evaluation behaviors logged in online systems can also be incorporated, along with consistency measures indicating the deviation between rating scores from instructors and from the average level. Thereby, the study presents a feasible solution to the gap of data interoperability in existing studies.

The study also contributes to iterative TBL design with multiple rounds of group learning, where learning log data from previous rounds can be utilized for various learning analytics purposes. This study deliberates one example of the early detection of unserious raters in peer evaluation, which is intended to improve the quality of peer assessment and the group presentation. Furthermore, continuous data support has broader applications in group learning. Not only peer evaluation, but other phases of group learning can also apply the design, as learning logs from past activities can provide data for creating groups (Liang et al., 2023) and calibrating peer rating scores (Piech et al., 2013). These accumulated data can also be useful for data visualization platforms for reflecting on teaching interventions (Kuromiya et al., 2020). The potential for data-driven TBL design can also extend beyond higher education context, overcoming the cold start problem for lack of learning logs in traditional classrooms (Pliakos et al., 2019). With rich data sensors ready before and during the TBL in this design, data-driven learning can be launched in manifold contexts, facilitated by the prevalence of digital devices in this mobile era.

However, there are several limitations to this study. The sample size of learners was relatively small, which might limit the generalizability of the findings. Hence, while the study's contribution may not be entirely conclusive, there is still potential in the current results for future exploration. 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 as antecedents. 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.

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. We also found these behavior patterns of raters are independent of consistency indicators, but associated with their course performance. 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.

Abbreviations

TBL: Team-Based Learning; EBA: Evaluation Behavior Analysis; LMS: Learning Management System; GLOBE: Group Learning Orchestration Based-on Evidence; AUC: Area Under Curve; IG: Information Gain; CSCL: Computer-Supported Collaborative Learning.

Acknowledgements

The author would like to acknowledge all of the contributors to this work and the reviewer comments and invitation from the ICCE2023 committee.

Authors' contributions

CL drafted the initial manuscript and performed data analysis. CL and IH designed and implemented the empirical study. RM, IH, HO provided insight and editing of the manuscript. HO provided supervision of the research. All authors read and approved the final manuscript.

Authors' information

Changhao Liang is a program-specific researcher at the Academic Center for Computing and Media Studies and the Graduate School of Informatics, Kyoto University. His research focuses on algorithmic group formation system and group learning support in data-driven environments with learning analytics.

Izumi Horikoshi is an assistant professor at the Academic Center for Computing and Media Studies and the Graduate School of Informatics, Kyoto University. Her research interests include learning analytics and classroom visualization for formative assessment and reflection.

Rwitajit Majumdar is an associate professor at the Research and Educational Institute for Semiconductors and Informatics, and the Graduate School of Social and Cultural Sciences, Division of Instructional System Studies at Kumamoto University. His research focuses on learning analytics and human-data interactions in educational platforms.

Hiroaki Ogata is a full Professor at the Academic Center for Computing and Media Studies, Kyoto University. His research interests include learning analytics, evidence-based education, educational data mining, educational data science, computer supported ubiquitous and mobile learning, and CSCL.

Funding

This research was partly supported by CSTI SIP Program Grant Number JPJ012347, JSPS KAKENHI Grant Numbers JP22K20246 and JP24K16751, and JSPS KAKENHI Grant-in-Aid for Scientific Research (B) 22H03902.

Availability of data and materials

Not applicable.

Declarations

Competing interests

The authors declare that they have no competing interests.

Author details

[1] Academic Center for Computing and Media Studies, Kyoto University, Japan

[2] Research and Educational Institute for Semiconductors and Informatics, Kumamoto University, Japan

Received: 27 December 2023 Accepted: 30 May 2024

Published online: 1 January 2025 (Online First: 3 June 2024)

References

Amarasinghe, I., Hernández-Leo, D., & Hoppe, H. U. (2021). Deconstructing orchestration load: Comparing teacher support through mirroring and guiding. *International Journal of Computer-Supported Collaborative Learning*, *16*(3), 307–338. https://doi.org/10.1007/s11412-021-09351-9

Boticki, I., Akçapınar, G., & Ogata, H. (2019). E-book user modelling through learning analytics: The case of learner engagement and reading styles. *Interactive Learning Environments*, *27*(5-6), 754–765. https://doi.org/10.1080/10494820.2019.1610459

Chen, W., Tan, J. S., & Pi, Z. (2021). The spiral model of collaborative knowledge improvement: An exploratory study of a networked collaborative classroom. *International Journal of Computer-Supported Collaborative Learning*, *16*(1), 7–35. https://doi.org/10.1007/s11412-021-09338-6

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. https://doi.org/10.1016/j.compedu.2005.02.004

Cleynen, O., Santa-Maria, G., Magdowski, M., & Thévenin, D. (2020). Peer-graded individualised student homework in a single-instructor undergraduate engineering course. *Research in Learning Technology*, *28*. https://doi.org/10.25304/rlt.v28.2339

Damgaard, M. T., & Nielsen, H. S. (2018). Nudging in education. *Economics of Education Review*, *64*, 313–342. https://doi.org/10.1016/j.econedurev.2018.03.008

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, *27*(8), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010

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 C. Brodley (Ed.), *Proceedings of the Twenty-First International Conference on Machine Learning* (p. 41). ACM. https://doi.org/10.1145/1015330.1015388

Goolsarran, N., Hamo, C. E., & Lu, W. H. (2020). Using the jigsaw technique to teach patient safety. *Medical Education Online*, *25*(1), 1710325. https://doi.org/10.1080/10872981.2019.1710325

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*, *10*(3), 549–574. https://doi.org/10.1007/s40692-023-00264-0

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. https://doi.org/10.15077/etr.43114

Horikoshi, I., Liang, C., Majumdar, R., & Ogata, H. (2022). Applicability and reproducibility of peer evaluation behavior analysis across systems and activity contexts. In S. Iyer et al. (Eds.), *Proceedings of the 30th International Conference on Computers in Education* (Vol. 1, pp. 335–345). Asia-Pacific Society for Computers in Education.

Ismail, N. A., Mohd, T., Othman, N. A. A., Abdullah, M. N., & Nasrudin, N. H. (2019). Development of an online peer assessment system in teamwork skills—A preliminary survey. In M. Mohamad Noor, B. Ahmad, M. Ismail, H. Hashim & M. Abdullah Baharum (Eds.), *Proceedings of the Regional Conference on Science, Technology and Social Sciences (RCSTSS 2016)* (pp. 175–185). Springer, Singapore. https://doi.org/10.1007/978-981-13-0203-9_17

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. https://doi.org/10.1007/s11423-019-09729-5

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. https://doi.org/10.1177/0271121417715318

Kasch, J., van Rosmalen, P., Löhr, A., Klemke, R., Antonaci, A., & Kalz, M. (2021). Students' perceptions of the peer-feedback experience in MOOCs. *Distance Education*, *42*(1), 145–163. https://doi.org/10.1080/01587919.2020.1869522

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. https://doi.org/10.1177/0894439317752406

Kulkarni, C., Wei, K. P., Le, H., Chia, D., Papadopoulos, K., Cheng, J., Koller, D., & Klemmer, S. R. (2013). Peer and self assessment in massive online classes. *ACM Transactions on Computer-Human Interaction*, *20*(6), 1–31. https://doi.org/10.1145/2505057

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. (2022). Estimating peer evaluation potential by utilizing learner model during group work. In L. H. Wong, Y. Hayashi, C. A. Collazos, C. Alvarez, G. Zurita & N. Baloian (Eds.), *Collaboration Technologies and Social Computing. CollabTech 2022. Lecture Notes in Computer Science, vol 13632* (pp. 287–294). Springer, Cham. https://doi.org/10.1007/978-3-031-20218-6_20

Liang, C., Horikoshi, I., Majumdar, R., Flanagan, B., & Ogata, H. (2023). Towards predictable process and consequence attributes of data-driven group work. *Educational Technology & Society*, *26*(4), 90–103.

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*, 14. https://doi.org/10.1186/s41039-021-00156-w

Lyu, Q., Chen, W., Su, J., & Heng, K. H. (2023). Collaborate like expert designers: An exploratory study of the role of individual preparation activity on students' collaborative learning. *The Internet and Higher Education*, *59*, 100920. https://doi.org/10.1016/j.iheduc.2023.100920

Michaelsen, L. K., Knight, A. B., & Fink, L. D. (Eds.). (2002). *Team-based learning: A transformative use of small groups*. Greenwood Publishing Group.

Nicol, D., Thomson, A., & Breslin, C. (2014). Rethinking feedback practices in higher education: A peer review perspective. *Assessment & Evaluation in Higher Education*, *39*(1), 102–122. https://doi.org/10.1080/02602938.2013.795518

Ogata, H., Yin, C., Oi, M., Okubo, F., Shimada, A., Kojima, K., & Yamada, M. (2015). E-Book-based learning analytics in university education. In H. Ogata et al. (Eds.), *Proceedings of the 23rd International Conference on Computers in Education, ICCE 2015* (pp. 401–406). Asia-Pacific Society for Computers in Education.

Ohland, M. W., Loughry, M. L., Woehr, D. J., Bullard, L. G., Felder, R. M., Finelli, C. J., Layton, R. A., Pomeranz, H. R., & 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. https://doi.org/10.5465/amle.2010.0177

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. https://doi.org/10.3109/0142159X.2012.651179

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), 1098–1120. https://doi.org/10.1037/edu0000103

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.

Pliakos, K., Joo, S. H., Park, J. Y., Cornillie, F., Vens, C., & Van den Noortgate, W. (2019). Integrating machine learning into item response theory for addressing the cold start problem in adaptive learning systems. *Computers & Education*, *137*, 91–103. https://doi.org/10.1016/j.compedu.2019.04.009

Rohmah, K., Priyatni, E. T., & Suwignyo, H. (2021, January). Assessment of learning development to improve student's appreciative and critical thinking abilities in drama appreciation course. In Meilinda, N. L. Pitaloka, Saparini, E. Nurdiansyah, M. R. Pahlevi, M. Ariska & D. Kurniawan (Eds.), *Proceedings of the 4th Sriwijaya University Learning and Education International Conference (SULE-IC 2020)* (pp. 495–502). Atlantis Press. https://doi.org/10.2991/assehr.k.201230.153

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. https://doi.org/10.1109/TLT.2021.3105008

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*(1), 67–104. https://doi.org/10.1007/s11412-021-09340-y

Strijbos, J. W. (2010). Assessment of (computer-supported) collaborative learning. *IEEE Transactions on Learning Technologies*, *4*(1), 59–73. https://doi.org/10.1109/TLT.2010.37

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. https://doi.org/10.1080/02602938.2018.1548559

Topping, K. (1998). Peer assessment between students in colleges and universities. *Review of Educational Research*, *68*(3), 249–276. https://doi.org/10.3102/0034654306800324

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. https://doi.org/10.1016/j.learninstruc.2009.08.004

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*, *22*(1), 51–68. https://doi.org/10.1002/acp.1331

Yoon, H. B., Park, W. B., Myung, S. J., Moon, S. H., & Park, J. B. (2018). Validity and reliability assessment of a peer evaluation method in team-based learning classes. *Korean Journal of Medical Education*, *30*(1), 23.

Zhang, C., & Conrad, F. (2014, July). Speeding in Web Surveys: The tendency to answer very fast and its association with straightlining. *Survey Research Methods*, *8*(2), 127–135. https://doi.org/10.18148/srm/2014.v8i2.5453

Publisher’s Note

The Asia-Pacific Society for Computers in Education (APSCE) remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.