Iyer, S. et al. (Eds.) (2022). Proceedings of the 30th International Conference on Computers in Education. Asia-Pacific Society for Computers in Education

Applicability and Reproducibility of Peer Evaluation Behavior Analysis Across Systems and Activity Contexts

Izumi HORIKOSHI[a*], Changhao LIANG[b], Rwitajit MAJUMDAR[a] & Hiroaki OGATA[a]

[a]*Academic Center for Computing and Media Studies, Kyoto University, Japan*
[b]*Graduate School of Informatics, Kyoto University, Japan*
* horikoshi.izumi.7f@kyoto-u.ac.jp

Abstract: Learning Analytics research provides findings and analytical methods. However, those have not been frequently shared and reused in other studies. The objective of this study is to clarify the possibilities and challenges in applying a behavioral analysis method to other system contexts. This study uses the Evaluation Behavior Analysis (EBA) as an example of an analytical method, and compared the applicability of the method and reproducibility of findings in two different studies: the original dataset (Study A), and which applied EBA (Study B). Not all methods were able to be applied as the data and activity were different. However, the reproducibility was far higher than we expected. This research contributes to expanding the reuse of analysis methods for small-scale systems such as EBA, which have not usually been reused, and further development in Learning Analytics.

Keywords: Peer Evaluation, Behavior, Learning Analytics, Applicability, Reproducibility

1. Introduction

Learning Analytics research has been conducted for a range of different purposes, and has focused on various learning activities (Lang, Siemens, Wise, & Gasevic, 2022). In fields where knowledge sharing and accumulation have begun, it is necessary to consider the reproducibility and replication of research findings. In education science research, the US National Science Foundation (NSF) and the Institute of Education Sciences (IES) jointly released the Companion Guidelines on Replication and Reproducibility in Education Research (NSF & IES, 2018). Based on these guidelines, McGill et al. (2019) organized problems and their solutions from the perspectives of reproducibility and replication for the maturity of the field of computer education. Learning Analytics research provides findings with contextual information and offers ideas for analytical methods for extracting behaviors. However, those findings and methods, especially the analytical process, have not been frequently shared and reused in other studies owing to technical constraints due to runnable necessities and context dependencies (Lebis, Lefevre, Luengo, & Guin, 2018). Lebis et al. (2018) proposed a framework called "Capitalization of Analysis Processes," and cited data standardization as related work.

If many people use the same system and do some analysis on that log, the same preprocessing and basic analyses are likely to be performed as the data structure is the same. For example, OpenLA provides common processing as an open-source library for an e-Book log analysis (Murata, Minematsu, & Shimada, 2020). Even if the system is not exactly the same, the general learning tool, such as e-Book or LMSs, have some common actions. Hence, xAPI Profiles have been proposed so that the data of the learning behavior that emerges there can be described in a common format (Usalearning). In other words, for systems such as e-Books or LMSs, there have been several proposals for applying methods and findings for extracting behavior from data from the original system or context to others.

However, Learning Analytics research does not always target behaviors on common systems that many people use, such as e-Books or LMSs. For example, the first author has focused on students' peer evaluation activities and has researched Evaluation Behavior Analysis (EBA), which visualizes students' behaviors during peer evaluation (Horikoshi & Tamura, 2021). EBA uses evaluation process data such as which evaluation items were evaluated by which student in what

order and with how much time. As a result, for example, we found students who evaluate from the evaluation items, or who evaluate in the order of evaluation items in a short time, and it has been clarified that the evaluation behavior varies depending on the student.

This study uses the EBA as an example of an analytical method on systems that are not in general use, unlike e-Books or LMSs. The objective of this study is to clarify the possibilities and challenges in applying a behavioral analysis method to other system contexts. To achieve this objective, we applied EBA to two datasets from different systems, using a situation where one of the authors involved in two different systems targeted the same behavior owing to the transfer of the institution. The research questions are: (1) whether EBA method can be applied to different datasets generated in different peer evaluation activity contexts and systems, and (2) whether the findings in the previous EBA study can be reproduced. We expect that this research will contribute to expanding the reuse of analysis methods for small-scale systems such as EBA, which have not usually been reused, and further development in Learning Analytics fields.

2. Methods

2.1 Research Design

To answer the research questions, we designed this research as shown in Figure 1. We compared the applicability of the method and reproducibility of findings in two different studies: Study A (Horikoshi & Tamura, 2021), which used the original system for EBA, and Study B (Liang, Majumdar & Ogata, 2022), which is a study this paper attempts to apply EBA. This study focuses on four basic steps in EBA: data preprocessing, behavior visualization, feature extraction and distribution, and correlation.

The structure of this paper is as follows. The following part of Section 2 summarizes the differences between the systems and activities in Studies A and B. In section 3, four basic steps in EBA are executed using the data from each system. The results are compared from the perspectives of the applicability of the method and reproducibility of the findings. Where the results in Study B did not match those in Study A (the original), we clarified whether it was because of the difference in the data format generated by the system or in the activity. Section 4 summarizes the results and discusses the challenges and solutions in applying an analysis method to the data of another system. The new data are only for Study B. The results in Study A for comparison were already reported in Horikoshi and Tamura (2021) and others.

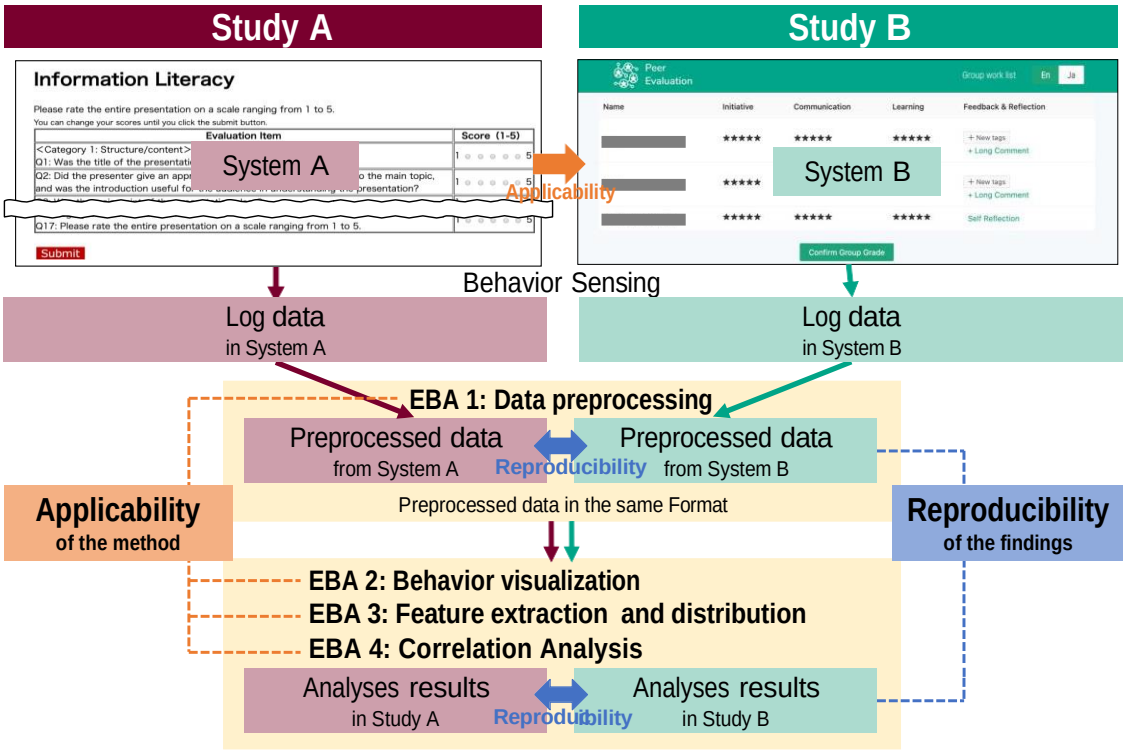


Figure 1. Research Design.

2.2 *Two Students' Peer Evaluation Studies Considered in this Paper*

We considered two contexts of peer evaluation activity. Here, we highlight the system, the peer evaluation activity, and the data collected in both contexts.

2.2.1 *Study A: The Original Evaluation Behavior Analysis (EBA) Study*

(1) System

Figure 2 presents the peer evaluation tool used in Study A. It is developed as a Web-based form to detect students' evaluation process data. The reviewer selects the score for each item and clicks the respective radio button to evaluate the evaluation target student. The reviewer can change their scores at any point in time before clicking the submit button. Evaluation process logs are sent to a server with a timestamp for when the reviewer clicks both the submit and radio buttons.

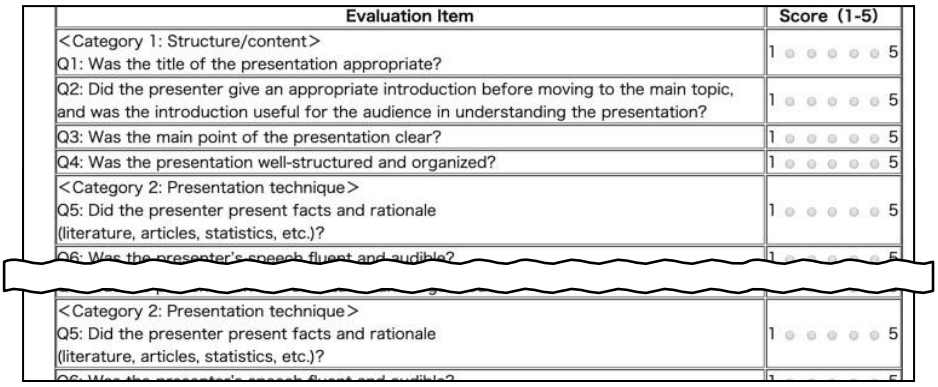


Figure 2. The Peer Evaluation Tool used in Study A (System A).

(2) Activity and Procedure

The target activity was one of the presentation classes at the end of a course at a university in Japan. The presentations were opinion speeches and the evaluations focused on the formal aspects of presentations. The students were divided into groups, each of which was given 15 minutes, which comprised a 10-minute presentation and a 4-minute question and answer (Q&A) session. Peer Assessments were conducted using the Peer Evaluation tool. The log data of the students' evaluations were acquired. The class was 90 minutes long. The students were instructed to submit an evaluation form during class.

(3) Scope of the Evaluation Behavior in Study A

The evaluation process log data used in Study A can only visualize "when the reviewer clicks a button on the evaluation form." As for the process that was not visualized in the log, the research field of questionnaire response behavior helped organize the cognitive process to provide the final response (Horikoshi & Tamura, 2021). This was called the "Cognitive Response Process Model" (Dillman, Smyth, & Christian 2009; Olson & Parkhurst 2013; Tourangeau, Rips, & Rasinski 2000). The components of this model differed slightly across the studies. Study A used the version created by Olson and Parkhurst (2013). Figure 3 illustrates the response process that operated when the respondents answer a form. Steps 1 to 5 and their descriptions were taken from the "Cognitive response process model" in Olson and Parkhurst (2013), and we added the component "Answering" as the sixth step.

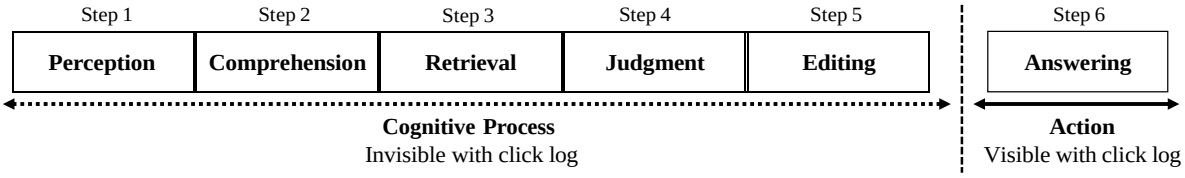


Figure 3. Response Process (Adapted from Fig.4 in Horikoshi and Tamura, 2021)

Only the sixth step can be visualized from the click log. However, as it is assumed that steps 1 to 5 are performed between each click, it is possible to infer the cognitive process until finally pressing submit. For example, if a student clicked from the top of the evaluation items, the cognitive process in steps 1 to 5 may be performed according to the evaluation items. On the other hand, if the student did not click in the order of the evaluation items, the cognitive process may have been performed according to what was noticed in the performance to be evaluated. If the time between clicks were extremely short, the cognitive process may have not taken enough time.

2.2.2 Study B: Where Evaluation Behavior Analysis (EBA) was Applied in this Paper

(1) System

Figure 4 shows the peer evaluation tool used in Study B. This system is one of the functions of the Group Learning Orchestration Based on Evidence (GLOBE: Liang, Toyokawa, Nakanishi, Majumdar, & Ogata, 2021) which constitutes the Learning and Evidence Analytics Framework (LEAF). In this study, we focused on the interaction log generated by the peer evaluation tool. This tool only sends the log when the submission button is clicked, and unlike System A, it does not send the log when the radio buttons (star-shaped) are clicked. In System A, the students had to press the submit button once for each evaluation target. The evaluation items were listed vertically, and the corresponding points were placed next to each item. In System B, evaluation items and targets were lined up horizontally and vertically, respectively. The evaluations of multiple targets were sent together in one submission. As with System A, students can change the scores as many times as they want until they press the submit button, but the process log is not recorded in System B. The log is sent only upon pressing the submit button.

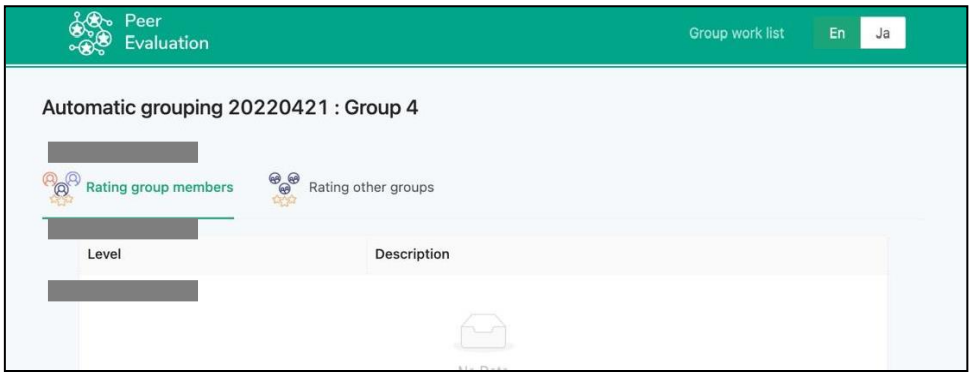


Figure 4. The Peer Evaluation Tool used in Study B (System B).

(2) Activity and Procedure

The target activity was a mini-presentation for sharing the status of assignments within a group at a university in Japan, which was different from Study A. The mini-presentation was instructed to include three elements pertaining to the assignment, and the students evaluated each element in the range of 1 to 5. The evaluation criteria for scoring each evaluation item were provided to the students. The instruction was to give each presentation in about 5 minutes, but as it was a group presentation, it was not possible to know how much time they actually took. The class length was about 90 minutes, of which the target activity was about 30 minutes, and the students were supposed to submit their evaluation by the end of the class.

3. **Results**

3.1 *Analysis 1: Data Preprocessing*

Comparing the log data of Systems A and B, the column names and their order were, not surprisingly, different between them. However, although the column names such as the timestamp when the log was stored, the reviewer, the evaluated target, the evaluation item number, and the

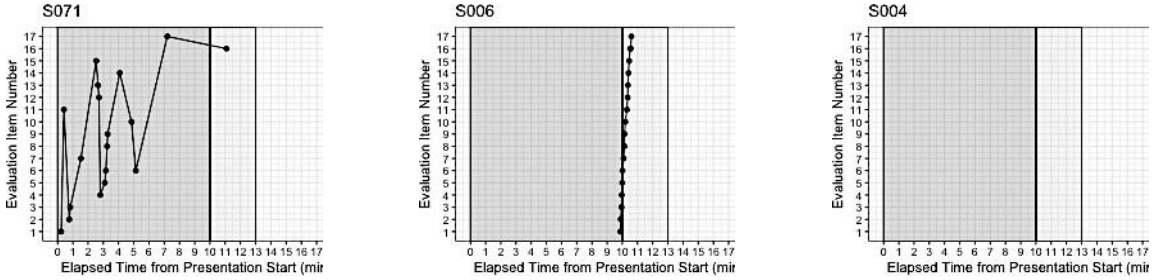
score were different, the contents of the stored information were almost the same. The timestamps in System A differed for each record as the system sent a log each time the radio button was clicked. In System B, the timestamps were the same for multiple lines because the logs were sent together when the submit button was clicked. Nevertheless, even in System B, the submit button can be pressed as many times as necessary, so some students submitted again with some changes.

In System A, these data were preprocessed and shaped into a form of when and who gave what points for what item to whom. Along with shaping the column, the record whose score had not changed was deleted from the previous log, leaving only the data of the change action and its timing behind. We tried to format the System B data into the same shape as the preprocessed System A data. As System B only comprised submission clicks, the number of records was small, and the timestamps of the lines extracted from the same submission became the same. Nevertheless, we could obtain the data in the same format as that in System A.

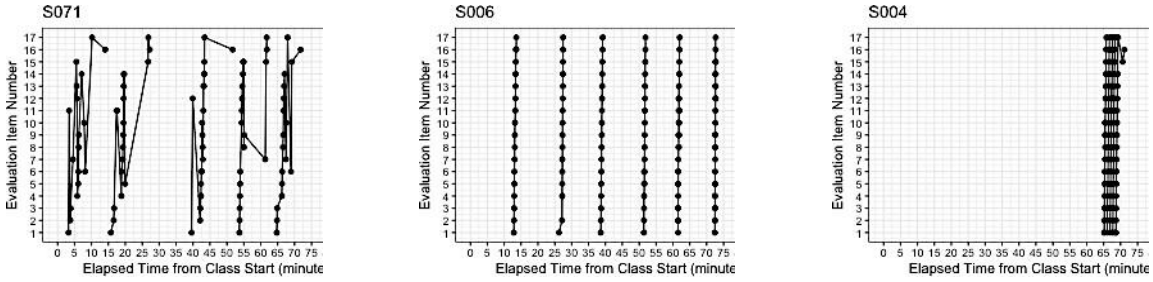
From this, it became clear that if the final goal of the preprocessing was to obtain the data form in basic experience expression of "when-who-did-what", even datasets with completely different columns from completely different systems can be shaped in the same shape. In other words, it can be said that the method of data preprocessing in EBA was applicable.

3.2 *Analysis 2: Behavior Visualization*

Figures 5 and 6 visualize the evaluation behavior from the data prepared with the preprocessing in Analysis 1. The vertical axis of the plots is the evaluation item number, and the horizontal axes are the elapsed time from the start time of the presentation, class, and activity respectively.



(a) Evaluation Behaviors for One Target



(b) Evaluation Behaviors for All Targets

Figure 5. Evaluation Behaviors from Study A's data. (Source: Fig.7 in Horikoshi and Tamura, 2021)

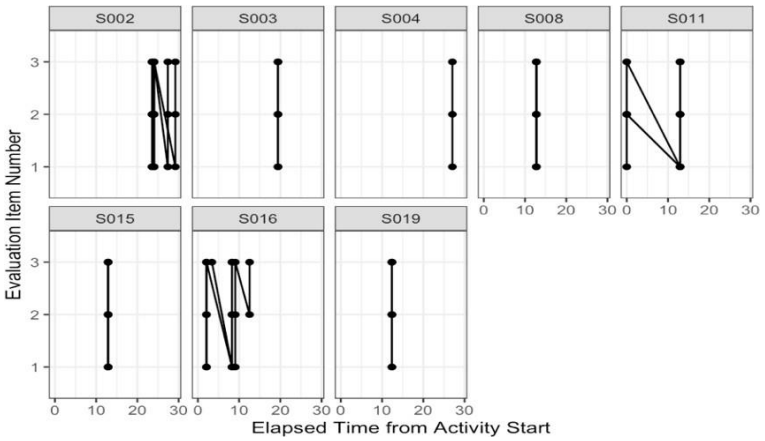


Figure 6. Evaluation Behaviors from Study B's data.

The evaluation behaviors in Study A, shown in Figure 5, can be divided into behaviors for one evaluation target (Figure 5 (a)), and for all evaluation targets performed in that class (Figure 5 (b)). Figure 6 shows the evaluation behaviors visualized from Study B's data. In System B, the radio button-click logs were not sent but the evaluations of multiple evaluation targets were sent together in one submission. Therefore, unlike Study A, it was not possible to separate the behavior for one evaluation target and that for all targets performed in that class.

First, it was possible to visualize evaluation behavior by placing Elapsed Time on the x-axis and Evaluation Item Number on the y-axis. In addition, it was also possible to visualize the timing (some evaluated at the beginning of the activity and some at the end) and modification of the evaluation. On the other hand, the evaluation timing for each evaluation target, evaluation item, and the order of evaluation could not be visualized. Moreover, modifications in evaluation without clicking the submit button could not be visualized. All of these differences are due to the difference in data transmission timing design that System B did not send the log when the radio button was clicked, and the evaluations of multiple evaluation targets were sent together in one submission. In other words, the applicability of the method was not sufficient for behavior visualization, and this was due to differences in data granularity.

In this way, many characteristic behaviors could not even be visualized, and therefore, findings based on behavior visualization were limited. However, the timing of the evaluation was able to be visualized, and the behaviors found in Study A, such as some submitted at the beginning and others at the end, were also visualized in Study B. In addition, the behavior that some students changed their evaluation was reproduced. However, only the submission level was visualized in Study B while this was the click level in the original method. Therefore, there might be more students who made modifications at the click level actually.

3.3 *Analysis 3: Feature Extraction and Distribution*

In Study A, the six feature variables shown in Table 1 were extracted and used as indicators to quantitatively capture the characteristics of evaluation behavior. Some of the definitions of the feature variables shown in Table 1 could not apply to Study B as they were, so the underlined parts were replaced, as shown in parentheses.

Table 1. Definition of Feature Variables of Evaluation Behaviors Proposed in Study A

Feature Variables	**Definition**
Evaluation Time (ET)	Time difference between clicking the radio button of the first evaluation item and the last evaluation item. (Replaced it with "the first record" and "the last record")
Click Count (CC)	Total number of <u>times the radio buttons for the evaluation items were clicked</u>. (Replaced it with "the record")

Mean of the Score (sM)	Average score for all the evaluation items scored by the reviewer.
SD of the Score (sSD)	Standard deviation for the scores of all evaluation items scored by the reviewer.
Mean of the Timestamp (tM)	Average elapsed time since the start of the presentation. (Replaced it with "the first record")
SD of the Timestamp (tSD)	Standard deviation of the timestamps for all evaluations.

Figure 7 compares the distribution of the feature variables extracted from the data in Studies A and B. The vertical axis of the plot is the value of Feature Variables; the range of which is different depending on each Feature Variable. As mentioned above, some of the definitions of feature variables could not apply to Study B as they were. ET and CC were because of system differences, and tM was because of activity differences. However, in fact, the program was applicable with almost no changes or modifications. This means that the main problem was that the expressions of the definitions of feature variables were specialized to the context (system or activity) in Study A, and were not sufficiently generalized.

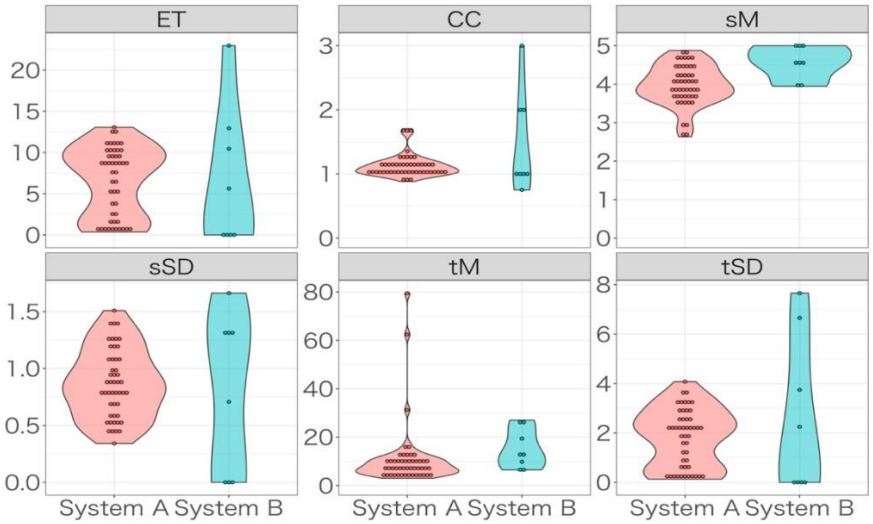


Figure 7. Distribution of Extracted Feature Variables.

All feature variables could be extracted after the above adjustment. However, the extracted variables did not necessarily express the same concept in Studies A and B. For example, score-related features (sM and sSD) were assumed to represent almost the same concept in Studies A and B. On the other hand, other variables related to time and clicks may have been significantly affected by differences in the timing of data transmission. This is because the System B log data did not record the action before the submission button was clicked; for example, ET and tSD will be 0 if there is only one submission, and the value of tM does not always reflect the timing of the evaluation. Also, for CC, the modifications before submission were not captured. Thus, the value may have been a little smaller than the original definition.

As for the visualization of the distribution of the extracted features, CC, sM, and tM showed almost the same distribution in Studies A and B, whereas the distributions in ET, sSD, and tSD were different. Most CCs were one time per evaluation item in both data. Also, regarding sSM, many students tended to use higher scores and sSM became high in both studies. For tM, most tMs were within presentation activity time. On the other hand, ET and tSD had two peaks (Speeder and non-speeder) in Study A, whereas most of them were 0 in Study B. This is because, as mentioned above, if there is only one submission, then ET becomes 0, and this is due to the difference in the system. For sSD, it was 0 to 1.5 and two peaks were observed in Study B while it was around 0.5 to 1.5 in Study A. Regarding sSD,

the extracted variables themselves were considered to express the same concept in Study A and B, but the distribution was different. The interpretation of the reason for this difference is that Study B clearly stated the criteria for each score, and for these reasons, it was easy for students to choose one.

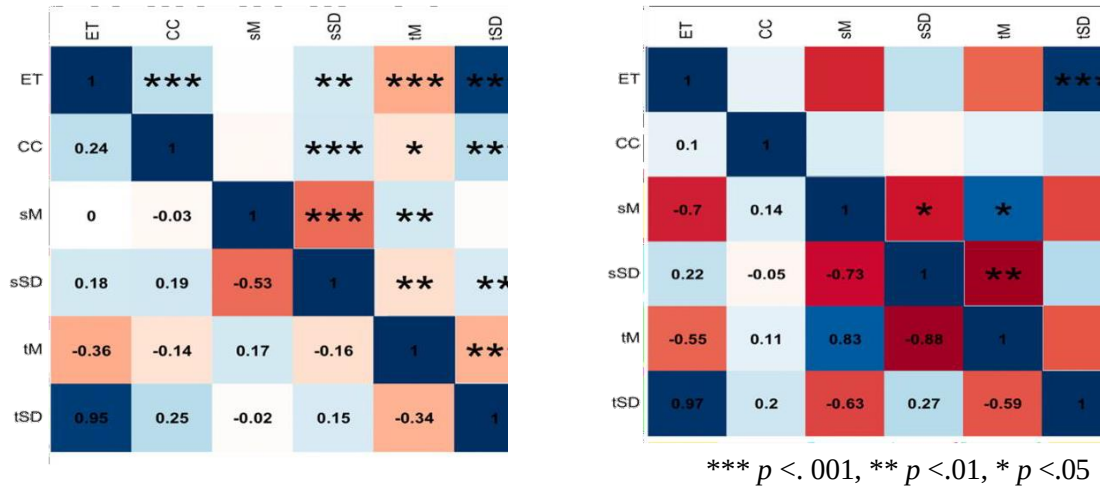
Thus, only three out of the six feature variables reproduced distribution tendencies. Two findings were not reproduced because of differences in data and one was because of differences in activity.

3.4 *Analysis 4: Feature Correlation*

As the last analysis, correlation analyses between the features were performed to compare the characteristics of the features extracted. In Study A, the correlation shown in Figure 8 (a) was found among the six types of feature variables. For this result, Study A interpreted that there was some common factor, and the factor made ET, CC, sSD, and tSD decreased, and sM and tM increased. This factor was interpreted as the students' motivational state in Study A. In other words, if the students' motivation for evaluation was low, the evaluation might be completed in a short time (small ET, tSD), not be changed (small CC), the same score might be used many times (small sSD), be given many full marks (large sM), and many evaluations might be made at the end of the class (large tM).

This interpretation was based on the findings of the answering behavior Web Survey research. In the Web Survey research field, response time has been used frequently as an indicator of the quality of a survey. According to Yan and Tourangeau (2008, p.64), "respondents tended to answer more quickly as they got closer to the end of the questionnaire." Therefore, we hypothesized that the proposed feature variables could also be influenced by repeatedly conducting peer evaluations in a single class.

As the feature variables were extracted in the same format, there was no problem in applying the correlation analysis. However, there were few significant correlations as the number of subjects and records in Study B were limited. Focusing on the value of the correlation coefficient, which reflects the tendency of the distribution of variables, the results in Studies A and B matched in 12 out of 15 pairs in the positive or negative direction. All three unmatched pairs were related to CC, perhaps because the data in Study B were not click-level ones, and therefore had smaller values than the other features.



*** p <. 001, ** p <.01, * p <.05

(a) Study A (b) Study B

Figure 8. Correlation among Feature Variables.

4. **Discussion**

The research questions were: (1) whether EBA method can be applied to different datasets generated in different peer evaluation activity contexts and systems, and (2) whether the findings

in the previous EBA study can be reproduced. To answer this question, we conducted four analyses from EBA, and compared the two studies conducted with two different systems that collected students' evaluation behavior data. In this section, we summarize the results and discuss the challenges and solutions in applying an analysis method to the data of another system.

4.1 *Applicability of the Method*

EBA 1 - Data Preprocessing: Even though the original data structures were completely different, the final structure became the same. This shows that if the final goal of the preprocessing was to obtain the data in basic experience expression of "when-who-did what," it is possible to convert datasets with completely different columns of completely different systems into the same shape. In other words, it can be said that the method of data preprocessing in EBA was applicable.

EBA 2 - Behavior visualization: It was possible to visualize the basic evaluation behavior plot with the same axes in both datasets. From these plots, the timing and modification behavior of the evaluation was visualized. The visualization of behavior was significantly affected by the difference in the design of the data transmission timing, and many behaviors such as the order of evaluation and change in evaluation without clicking the submit button could not be visualized. In other words, the applicability of the method was not sufficient for behavior visualization, and this was mainly due to differences in data granularity.

EBA 3 - Feature Extraction and distribution: Some of the definitions of feature variables could not apply to Study B as they were, because of both differences in the systems and activities. Nevertheless, the program was applicable with almost no change or modification. This means that the main problem was that the expressions of the definitions of feature variables were specialized to the context (system or activity) in Study A, and it was not sufficiently generalized. In addition, all feature variables could be extracted with the adjustment. However, the extracted variables did not necessarily express the same concept in Studies A and B. This was mainly caused by system differences.

***EBA 4 - Feature Correlation*:** There was no problem applying the correlation analysis, as the feature variables were extracted in the same format.

4.2 *Reproducibility of the Findings*

EBA 2 - Behavior Visualization: Many characteristic behaviors could not even be visualized. Therefore, findings based on behavior visualization were limited. However, the timing of the evaluation was visualized, and the behaviors found in Study A (e.g., some submitted at the beginning and others at the end, or some changed their evaluation) were reproduced in Study B.

EBA 3 - Feature Extraction and Distribution: Only three out of six feature variables reproduced distribution tendencies. Two findings were not reproduced because of differences in data and one was because of differences in activity.

EBA 4 - Feature Correlation: The results of Studies A and B matched in 12 out of 15 pairs in terms of the direction of correlation (positive or negative). All three unmatched pairs were related to CC, perhaps because the data in Study B were not click-level, in other words, due to system differences.

4.3 *Challenges and Solutions*

First, regarding the preprocessing of data, the data structure is usually completely different depending on the system, but it became clear that if preprocessing was performed appropriately and the data were brought into the same format, a considerable part of the later analysis method can be applied. This was considered possible because the result of the preprocessing in the EBA

was the structure of "when-who-did what," which consisted of the components of the description of the basic learning experience. This structure is also similar to the components of xAPI. However, preprocessing for the new dataset was complex and took a lot of time. One of the possible solutions for this is to standardize the data format, but it is unrealistic to standardize the data structure in the database for analysis. Because the data to be analyzed are not only stored just as log data but also usually working as a part of the system to operate. Therefore, storing additional data for analysis in the "when-who-did-what" format using xAPI or other standardized formats into Learning Record Store (LRS) is considered a practical solution for standardizing and reusing preprocessing.

It was also clarified that there are cases where the method can be applied but the findings cannot be reproduced because of the difference in activities. One solution for this problem is to separate the applicability of the method and reproducibility of the findings and the influence of data and activity as done in this paper. However, it was possible to discuss the effects of data and activity in detail because one of the authors was involved in both studies, but such cases are rare. Therefore, it is necessary to disclose data and activity information in papers and leave it thus so that similar discussions can be held even if the same researcher does not participate in subsequent studies.

Finally, the factor that had the most significant impact on both the applicability of the method and reproducibility of the findings in this analysis was the granularity of the stored data. No matter how much pre-processing or interpretation was devised, it was impossible to create unstored information that was supposed to be extracted from process data. The click-level process log is necessary for the behavior analysis related to the cognitive process shown in Figure 3.

As Learning Analytics shifts from the research level to the practical level in the future, it will become more common to apply the same methods and knowledge to log data from different systems. From the results clarified in this paper, it is necessary to acquire similar log data in similar granularity in order to increase the applicability of the method and the reproducibility of the findings. For this reason, when designing learning log data sensing systems, it is important to sense various types of data in as fine a granularity as possible, convert them into a standardized format, and store them with context information. This requires a more detailed discussion on the standardization of data format and stored information.

5. **Conclusion**

In this paper, we applied Evaluation Behavior Analysis (EBA) to two datasets from different systems and verified the applicability of the method. Not all methods were able to be applied as the data and activity were different, and the results showed that some findings were not reproduced. However, the reproducibility was far higher than we expected.

One limitation of this paper is that the datasets from the two compared systems differed in terms of data granularity and activity. Thus, in this research design, even though the applicability of the method had not been verified, the reproducibility of the findings was discussed. Therefore, in our future work, we would like to discuss the reproducibility of findings again with the same data granularity and in similar activity contexts, while ensuring that the applicability of the method remains satisfied. This time, we only used evaluation behavior data, which were used in Study A. However, the greatest advantage of System B's data is that it can be analyzed by combining behavioral data and other learning logs on the same LEAF platform. Thus, we would like to build further on this study, which was merely a confirmation of applicability and reproducibility, and clarify whether the new interpretation of behavior or knowledge that leads to class improvement can be obtained by combining the evaluation behavior data with other learning logs. Though this study is exploratory and has some limitations, we believe the results present possibilities, difficulties, and solutions with respect to applying similar behavior concepts and methods of behavior analysis to other systems.

Acknowledgments

This study was supported by the following grants: NEDO JPNP20006, JSPS KAKENHI 20H01722

References

Dillman, D. A., Smyth, J. D., & Christian, L. M. (2009). Internet, Mail and Mixed-Mode Surveys: The Tailored Design Method (3rd edition). Hoboken, NJ: Wiley and Sons, Inc.

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.

Lebis, A., Lefevre, M., Luengo, V., & Guin, N. (2018, March). Capitalisation of analysis processes: Enabling reproducibility, openness and adaptability thanks to narration. In Proceedings of the 8th international conference on learning analytics and knowledge (pp. 245-254).

Lang, C., Siemens, G., Wise, A., & Gasevic, D. (Eds.). (2022). Handbook of learning analytics – Second edition. New York: SOLAR, Society for Learning Analytics and Research.

Liang, C., Toyokawa, Y., Nakanishi, T., Majumdar, R., & Ogata, H. (2021, August). Supporting Peer Evaluation in a Data-Driven Group Learning Environment. In the proceedings of International Conference on Collaboration Technologies and Social Computing (pp. 93-100). Cham: Springer.

Liang, C., Majumdar, R., & Ogata, H. (2022, June). Continuous Data-Driven Group Learning Support: Case Study of an Asynchronous Online Course, In Proceedings of the 15th International Conference on Computer-Supported Collaborative Learning - CSCL 2022 (pp. 547-548).

McGill, M. M. (2019, November). Discovering empirically-based best practices in computing education through replication, reproducibility, and meta-analysis studies. *In Proceedings of the 19th Koli Calling International Conference on Computing Education Research* (pp. 1-5).

Murata, R., Minematsu, T., & Shimada, A. (2020, November). OpenLA: Library for Efficient E-book Log Analysis and Accelerating Learning Analytics. In 28th International Conference on Computers in Education, ICCE 2020 (pp. 301-306). Asia-Pacific Society for Computers in Education.

National Science Foundation and the Institute of Education Sciences. (2018). Companion Guidelines on Replication & Reproducibility in Education Research: A Supplement to the Common Guidelines for Education Research and Development.
Retrieved from https://www. nsf. gov/pubs/2019/nsf19022/nsf19022. pdf.

Olson, K., & Parkhurst, B. (2013). Collecting paradata for measurement error evaluations. In Kreuter, F. (Eds.), Improving Surveys with Paradata, 43-72. Hoboken, NJ: Wiley

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). The Psychology of Survey Response. Cambridge, UK: Cambridge University Press

USAlearning. (n.d.). Actionable Data Book (ADB) Profile
Retrieved from https://profiles.usalearning.net/profile/f24c8d44-4937-44a6-bef7-f5435f57492f