Analysis of "Evaluation Behavior" Using Students' Peer Assessment Process Data

Izumi HORIKOSHI [a*] & Yasuhisa TAMURA [b]
[a] *Graduate School of Science and Technology, Sophia University, Japan*
[b] *Dept. Information and Communication Sciences, Sophia University, Japan*
*izumihorikoshi@eagle.sophia.ac.jp

Abstract: In this study, we have focused on students' peer assessment and analyzed evaluation behavior using data from the evaluation process. Peer assessments by students are problematic in terms of reliability and validity. Many previous studies have discussed the reliability or validity of peer assessment, based on the evaluation scores of the peer assessment. However, the "Evaluation Process", that is, who, when, and which items were evaluated in what order, has not been studied. For this issue, in this research, we have proposed to acquire the "Evaluation Process" data in peer assessment and to analyze and visualize the students' "Evaluation Behavior". As a preliminary result, this study identified a range of characteristic evaluation behaviors, indicating the possibility that each student might have a unique style of evaluation and that there are students who do not take evaluation seriously. We expect that we will be able to estimate the students' motivation on the peer assessment or evaluation ability, and also to improve the design of the peer assessment form or the conditions of the peer assessment based on the "Evaluation Behavior".

Keywords: Peer assessment, Evaluation behavior, Learning analytics

1. Introduction

With the growing popularity of active learning in recent years, the practice of making presentations in class and evaluating these through peer assessment has increased. Peer assessment is defined as "an arrangement in which individuals consider the amount, level, value, worth, quality, or success of the products or outcomes of learning of peers of similar status" (Topping 1998, p. 250). According to Fukazawa (2010), the general benefits of peer assessment in learning situations are enhancing learner motivation (Orsmond, Merry & Reiling, 1996) and also reducing the teachers' burden (Brown 1998).

Although peer assessments are useful, these can also be seriously flawed in terms of reliability and validity when conducted by students. Several studies have been conducted in order to address these concerns. Previous studies have mostly discussed the reliability and validity of evaluation scores in peer assessments. In other words, the "Evaluation Process," that is, who, when, and which items were evaluated and in what order, has not been studied in terms of technology and significance.

In order to address this gap in research, in this study, we propose to acquire the "Evaluation Process" data in peer assessment and to analyze and visualize the students' "Evaluation Behavior".

2. Related Work and Our Study

2.1 Conventional Studies Discussing Quality of Peer Assessment

As mentioned in Chapter 1, peer assessments made by students are weak in terms of reliability and validity, and this issue has been studied extensively (Fujihara, Ohnishi & Kato, 2007). According to Fukazawa (2010), several studies have discussed reliability and validity using the correlation coefficient between scores given in evaluations by teachers and in peer assessments made by students. These studies can be placed into two categories: those that conclude that peer assessments made by students are as reliable as teacher assessments and those that question the reliability and validity of peer assessments made by students. Fukazawa cites three studies as examples of the former (Miller & Ng, 1996; Hughes & Large, 1993; Stefani, 1994) and also cites two studies as examples of the latter (Stefani, 1994 and Freeman, 1995).

We focused on the fact that any related research has discussed reliability and validity based on the score of peer assessments. However, we are considering it is difficult to conclude that the evaluations are "similar" just because the evaluation scores are similar. This is because there might be cases where the scores given by students are similar but their behaviors during the process of evaluation are different.

2.2 Paradata" Studies in Web Surveys

On the other hand, we found many previous studies focusing on behavior during the answering process in Web Survey research field. In the disciplines of social science and statistics, the number of computer-assisted interviews and Web surveys have increased. Computer-assisted interviews or Web surveys have advantages that traditional surveys do not have. For example, data related to the process of responding to the surveys are automatically generated. These log data related to response behavior, such as response time or interruption during answering the questionnaire, which is also called "paradata" (Couper, 1998), and these have been used to verify the quality of survey responses. No single formal definition exists for paradata. However, according to Olson and Parkhurst (2013), "examples of paradata collected automatically by many computerized survey software systems include timing data, keystroke data, mouse click data... (p.43)" and so on.

In studies on paradata, "response time" has frequently been the main focus. According to Couper and Kreuter (2013), response time is readily available in most computer-assisted interviewing systems. Based on the literature, "shorter response times" are associated with a "lack of motivation to answer accurately, caused by continuous survey (Yan & Tourangeau, 2008)" and "longer response times" are associated with "lower scores on knowledge items (Heerwegh, 2003)."

3. Expected Contribution of this Research

As described above, many previous studies on peer assessment have discussed the reliability and validity based on the score of peer assessments. On the other hand, we found many previous studies focusing on behavior during the answering process in Web Survey research field. Although the research fields are different, the analysis of "Evaluation Behavior" in peer assessment, which is the focus of this study, and that of "Response Behavior" or answering behavior in the field of social research are quite similar in terms of data and behavior. For this reason, we expect that there is a great possibility in visualizing and analyzing evaluation behavior to obtain valuable findings in peer assessment.

When this study is completed, it will be possible to understand in more detail the situation of the students during peer assessment. We believe that leads to visualizing issues that could not be seen with the conventional method based only on the scores. Furthermore, we expect that we will be able to estimate the students' motivation on the peer assessment or evaluation ability and also to improve the design of the peer assessment form or the conditions of the peer assessment based on the "Evaluation Behavior."

4. Preliminary studies and Results

We have already started preliminary studies and have reported some of the results in oral presentations [Horikoshi & Tamura, 2017a; 2017b; 2018a; 2018b; 2018c and 2018d]. In the preliminary studies, we aimed to achieve the following three objectives: (1) to investigate research on the evaluation process and behaviors, organize these, and compare these with our research, (2) to establish a method of acquiring the evaluation process data and visualize evaluation behavior, and (3) to extract characteristic evaluation behaviors.

First, we investigated similar studies and considered the relationship with our proposal. As a result, it became clear that there have been a lot of studies in social science, which acquire answering process data of surveys and questionnaires. We have already addressed this result in detail in Chapter 2.

Second, we developed a Web-based form as the peer assessment tool to detect students' evaluation process data using HTML, JavaScript, and PHP. Using this form, we conducted experiments to acquire "Evaluation Process" data of actual classes in which assessments were made and visualized the "Evaluation Behavior."

Finally, we extracted and discussed characteristic evaluation behaviors. As a result, this study identified a range of characteristic evaluation behaviors, indicating the possibility that each student might have a unique style of evaluation and that there are students who do not take evaluation seriously. For example, we found the following characteristic evaluation behaviors: "did not evaluate sequentially" or "evaluated in a short time". Furthermore, it was also suggested that each student showed similar evaluation behavior in every evaluation of all the presentations in the class. On the other hand, there are students who "complete evaluation before the start of presentation" or "incorrectly evaluate to a group not giving a presentation". Therefore, we need to be aware of the possibility that each student might have a unique style of evaluation behavior. However, at the same time, we also need to be aware that there are students who are not working seriously on evaluation, not on the learner's evaluation style, and that those students often make evaluations in a short time.

5. Future Research

The preliminary studies revealed the significance of conducting research on the evaluation process and evaluation behavior in peer assessment. As the next step, we will work on the following tasks:

First task: to compare the "Evaluation behavior" by our proposed method analyzed this time with the "Score" that has been treated by the conventional method

In the preliminary studies, it became clear that there are learners who spend a long time on evaluation and short learners. Therefore, for the next step, it is necessary to verify whether there is a difference in the degree of reliability and validity of the scores, depending upon the time taken for the evaluation or depending upon the evaluation behavior pattern (sequential or nonsequential).

Second task: to discuss evaluation behavior quantitatively

In the preliminary studies, we visualized and discussed the evaluation behavior qualitatively using graphs. Qualitative methods are suitable for discussing the characteristics of each student. However, qualitative methods have limitations in understanding or discussing the characteristics of the entire class and also features from the qualitative analysis are not machine-readable. Therefore, as the next step, a method to discuss the characteristics of evaluation behavior quantitatively needs to be developed. To be specific, we are planning to propose feature variables to quantitatively express the characteristics of evaluation behavior, such as short evaluation time or tendency to choose high scores. The feature variables will include , for example, "Evaluation Time", "Click Count", "Mean of the score", "Standard deviation of the score", "Mean of the evaluation timestamp", or "Standard deviation of the evaluation timestamp".

Third task: the establishment of extraction and interpretation of characteristic evaluation behaviors

It is necessary to construct an interpretation hypothesis for the characteristic evaluation behavior extracted in the preliminary studies and to verify whether our interpretation hypothesis is correct, using a questionnaire or an interview schedule. In addition, we expect that it would be effective to compare with the findings of the interpretation of the answering behavior of the paradata research mentioned earlier or to compare with the findings of the conventional method using the score.

Fourth task: to clarify the relationship between the above-mentioned characteristic evaluation behavior and "class design" or "evaluation form design".

In the preliminary studies, we discussed only the nature and style of individual reviewer as the cause of characteristic evaluation behavior. However, we have hypothesized that "class design," such as "how many times the teacher makes the students evaluate in each class," "how many minutes the presentation length is," or "How many evaluation items the students evaluate during a presentation" can influence students' motivation or seriousness and this might lead to behavior. We also hypothesized that "evaluation form design" such as "whether the form is multiple choice type or rubric type," "the wording of criteria," or "whether the selector is pre-selected or not" might influence students’ behavior. These hypotheses were formed by us because we found similar hypotheses and findings in research on paradata (For instance, Masuda, Sakagami, and Kitaoka, 2017).

References

Brown, J. D. (1998). *New ways of classroom assessment. new ways in TESOL series II. innovative classroom techniques.* ERIC.

Couper, M. (1998). Measuring survey quality in a CASIC environment. *Proceedings of the Survey Research Methods Section of the ASA*, 41-49.

Couper, M. P., and Kreuter, F. (2013). Using paradata to explore item level response times in surveys. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 176(1), 271-286.

Freeman, M. (1995). Peer assessment by groups of group work. *Assessment & Evaluation in Higher Education, 20(3), 289-300.*

Fujihara, Y., Ohnishi, H., Kato, H. (2007). Review of Research on Peer Evaluation. *Journal of Multimedia Aided Education Research*, 4(1), 77-85.

Fukazawa, M. (2010). Validity of Peer Assessment of Speech Performance. *Annual review of English language education in Japan*, 21, 181-190.

Heerwegh, D. (2003). Explaining response latencies and changing answers using client-side paradata from a web survey. *Social Science Computer Review*, 21(3), 360-373.

Horikoshi, I., and Tamura, Y. (2017a). Analysis of Students’ Peer Assessment Processes. In Yang, J. C. et al. (Eds.), (2017). *Proceedings of the 25th International Conference on Computers in Education*, 25-30.

Horikoshi, I., and Tamura, Y. (2017b). Timing Acquisition of Students' Peer Assessment. *Japanese Society for Learning Analytics report*.

Horikoshi, I., and Tamura, Y. (2018a). Evaluation Behavior Analysis in Peer Assessment using Evaluation Process Data: Comparison with Response Behavior Analysis using Paradata in Web Survey. *Proceedings of the 46th Annual Meeting of the Behaviormetric Society*, 46, 196-199.

Horikoshi, I., and Tamura, Y. (2018b). Comparison of Evaluation Behavior between Learner and Teaching Assistant during Peer Assessment of Oral Presentation. *Research report of JSET Conferences*. 18(5), 37-44.

Horikoshi, I., and Tamura, Y. (2018c). Analysis of “Evaluation Time” in Peer Assessment based on Evaluation Log. *Proceedings of the Annual Conference of Japanese Society for Information and Systems in Education*, 43, 325-326.

Horikoshi, I., & Tamura, Y. (2018d) Feature Extraction of Learners’ Motivation from Peer Assessment Process Logs. In Yang, J. C. et al. (Eds.) (2018). *Proceedings of the 26th International Conference on Computers in Education*, 352-354.

Hughes, I., & Large, B. (1993). Staff and peer-group assessment of oral communication skills. *Studies in Higher Education*, 18(3), 379-385.

Masuda, S., Sakagami, T., & Kitaoka, K. (2017). Responding to Many Items in a Questionnaire Increases Middle Category Response. The Japanese journal of behaviormetrics, 44(2), 117-128.

Miller, L., & NG, R. (1996). Autonomy in the classroom: Peer assessment. *Taking control: Autonomy in language learning*, 133-146.

Olson, K., and Parkhurst, B. (2013). Collecting Paradata for Measurement Error Evaluations. Kreuter (Eds.), *Improving Surveys with Paradata*, 43-72. F. Wiley.

Orsmond, P., Merry, S., & Reiling, K. (1996). The importance of marking criteria in the use of peer assessment. *Assessment & Evaluation in Higher Education*, 21(3), 239-250.

Stefani, L. A. (1994). Peer, self and tutor assessment: Relative reliabilities. Studies in Higher Education, 19(1), 69-75.

Topping, K. (1998). Peer assessment between students in colleges and universities. *Review of educational Research*, 68(3), 249-276.

Yan, T., and 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.