Design of an Online Peer Assessment System for a Large Classroom Using Student-Defined Criteria

Panuakdet Suwannatat[1, a*], Chantima Patthamathamakul[1, b], Sakol Teeravarunyou[1, c], and Krittika Tanprasert[1, d]

[1]King Mongkut's University of Technology Thonburi, Bangkok, Thailand
[a*]panuakdet.mock@mail.kmutt.ac.th, [b]chantima.pat@kmutt.ac.th, [c]sakol.tee@kmutt.ac.th, [d]krittika.tan@kmutt.ac.th

Abstract

In a large classroom, providing formative assessment to the students in a timely manner is often not practicable due to time and resource constraints. The use of peer assessment, where the task of evaluation is crowdsourced to the students themselves, can address this challenge. In this paper, the design of an online peer assessment system is presented. The design's objective is to assist instructors in providing timely feedback to a large number of students. Using technology to supplement face-to-face communication faces its own challenges. In the first design iteration, each student was randomly assigned three other students whose work they would evaluate. Since each student rated more than one piece of work, the relative strictness of each rater can be calculated. In the second iteration, the students were divided into groups. Each group developed their own criteria on which their work would be evaluated. Each individual was assigned a group to evaluate — the assignments were random with fairness rules applied. The rater was given an opportunity in class to approach the group and ask questions. Using this approach, the feedback was found to be more informative in helping each group improve their work for the next submission. In this paper, the designs of the systems are discussed, as well as the major challenges including motivations of the raters, a fair distribution of raters, reliability of the quantitative rating, and usefulness of the qualitative feedback.

Keywords: formative assessment, peer feedback, large classroom, classroom technology

1. Introduction

1.1 Problem statement

Large class sizes pose significant teaching challenges in terms of the amount of feedback that must be provided to students. Instructors must provide high quality, individual feedback and fairly assessing a diverse mix of students. The OECD [1] defined *formative assessment* as an interactive assessment of student's progress and understanding to identify learning needs and adjust teaching appropriately. The goal of formative assessment is to monitor the student progress rather than giving specific grades to students.

Peer assessment is one of the techniques used in formative assessment. By letting students give feedback on other students' work, peer assessment gives students faster feedback and frees up teacher's time. The challenge of peer assessment is in the quality of feedback – due to a lack of knowledge or proper incentives for the evaluators. As a result, peer assessment needs to be used in conjunction with other teaching strategies.

For this study, an online peer assessment system was tested in two iterations. In the first iteration, the researchers focused on scores rather than comments. This focus was shifted to the quality of comments in the second iteration.

1.2 Background

A learning system based on technology differs from traditional learning. Bull et al [2] proposed an Open Learner Model that supports formative assessment and visual analytics. He used this model to accommodate a variety of learning styles. Other learning supporters such as friends, teachers and parents are included in order to enhance learning.

Chen el al [3] proposed using a data mining technique for formative assessment. He used *e-portfolio* and reflected the data analytic back to learners. As a result, the learners understand the content more clearly. Peer assessment also helps the evaluators develop a skill for appraising the work of others.

Several works on formative assessment employed mobile or internet technology such as a web-based system [4], [5], [6], [7], and [8]. An interesting work on peer assessment was done by Isabwe [8]. The students worked on a Mathematics assignment in groups of three. Each student assessed three assignments and gave feedback to their friends. At the end, each student received three peer feedback. With reference answers, the students were confident to provide a fair and responsible feedback. In contrast to [8], this study addresses the challenges of open-ended question and a large classroom.

1.3 Objectives

The objective is to develop an online tool, design a formative assessment method, and analyze the results. This tool should assist instructors in conducting peer evaluation for a large classroom, and provide students with timely feedback.

2. Methodology

Two iterations of peer evaluation were conducted as outlined in Figure 1, followed by a student survey.

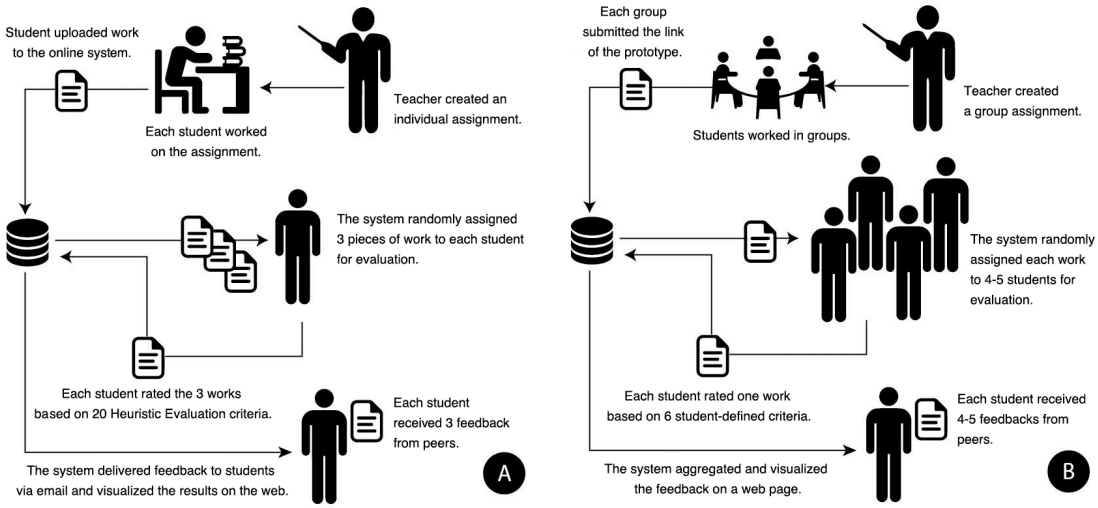

Figure 1: System Diagrams for Iterations 1 (A) and 2 (B)

2.1 The First Iteration

The main goal was to give rapid formative feedback for an individual assignment, Homework 3: Phone Menu. The design rationales and assumption are the following: (i) anonymity of the reviewers is important; (ii) a reasonable level of score reliability can be achieved by assigning multiple raters per work; (iii) raters and works must be matched by a system that ensures a fair distribution of work; and (iv) asking for an overall comment with an open-ended question can elicit useful responses from students.

2.1.1 The Design

Each student submitted his/her work by uploading it to an online system. The teacher monitored the submission progress on a dash board as shown in Figure 2 (a). After the deadline, the system randomly assigned each work to three raters. In turn, each rater was assigned three works to evaluate. Each student received an email with the following information: (i) a link to the peer evaluation form; (ii) the secret URLs of the three works they were asked to evaluate; and (iii) the secret, randomly-generated codes of the three friends, but not their names.

For each friend, the rater downloaded the work (a PowerPoint file implementing a phone menu prototype) and began evaluating. The rater also opened an evaluation form and entered his/her friend's secret code as an identifier for the work. The form asked for ratings on 20 Heuristic Evaluation questions [9], followed by a prompt for an overall comment.

The system collected all responses on a spreadsheet. For each piece of work, the system searched for the three ratings and presented the data on a new table. The result table, shown in Figure 2 (b), contained 70 rows, one per student, each row displaying the evaluation scores from the three friends, the mean, the SD, and the optional overall comments. Students were identified with their randomly-generated pseudonyms (fictional characters) both for the work owners and the raters. Each student also received a score report, along with the friend's comments, via email.

2.1.2 Lessons Learned

The raters may have felt overburdened because each was assigned three works with a lot of questions for rating. The overall comment box, lacking a specific prompt, failed to persuade most raters to give helpful comments. The technical nature of the work posed multiple hurdles such as large file sizes and incompatibility between versions. Insisting that the raters remained anonymous, along with allowing them to evaluate from home, created a situation where the raters could not ask the work's owner for clarifications, thereby missing an opportunity to give a more meaningful feedback. Despite the best effort by the system to distribute exactly three raters to each student, an unfair situation still emerged in which some work was evaluated by only one rater. This was because the matching was done in a batch – two raters who were assigned the same work might have failed to participate.

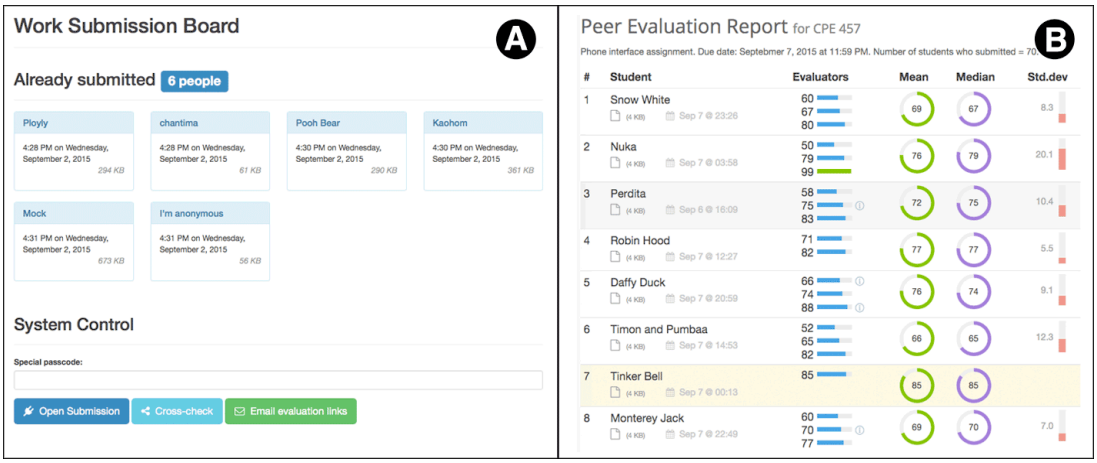

Figure 2: User interface for the first iteration: (A) the teacher's dashboard; (B) the results

2.2 The Second Iteration

The following strategies were designed to address the problems from the first iteration: (i) anonymity requirements were relaxed to make the raters more informed; (ii) students defined their own criteria, ensuring their relevance [10]; (iii) the teacher rated the quality of comments and gave points to raters who wrote good comments as an incentive; (iv) by evaluating only one group, each rater would be able to give more detailed comments; they were also given another tool to quickly express their overall feeling: the *reaction cards* [11]; (v) a fair distribution of raters was guaranteed by only matching students who participated as raters to groups; and (vi) peer evaluation was done in two rounds. After the first round, students improved their work based on formative feedback from peers before the final re-submission.

2.2.1 The Design

Prior to this assignment, each group performed a user test of a website and produced a list of six critical design issues. Their task for was to redesign the website using an online prototyping tool. For each of the two rounds of peer evaluation, each group submitted a link to the prototype. For the second round, each group also submitted a 10-minute video. The submission data were organized on a spreadsheet. In class, the students were asked to sit in groups to allow the raters to easily find and ask questions. Once logged in, the system assigned each rater a group to evaluate using a fair matching algorithm. The algorithm keeps track of the number of raters matched to each group, and always assigns a new rater to a group with the fewest raters; when there are multiple such groups, a choice is made randomly.

Once matched, each rater was presented with a link to the original website, the prototype (the work), and a video presentation. The rater then gave six ratings to how well the prototype addressed the six critical issues. For each issue, a comment was requested with an interactive prompt that changed depending on the score that he/she gave.

On the next screen, reaction cards [11] were laid out in a random order as shown in Figure 3 (a). The rater was asked to choose at least five cards that best described the work. At the bottom of the screen, a text box asked for an optional overall comment. The raters were asked to finish this evaluation process in class.

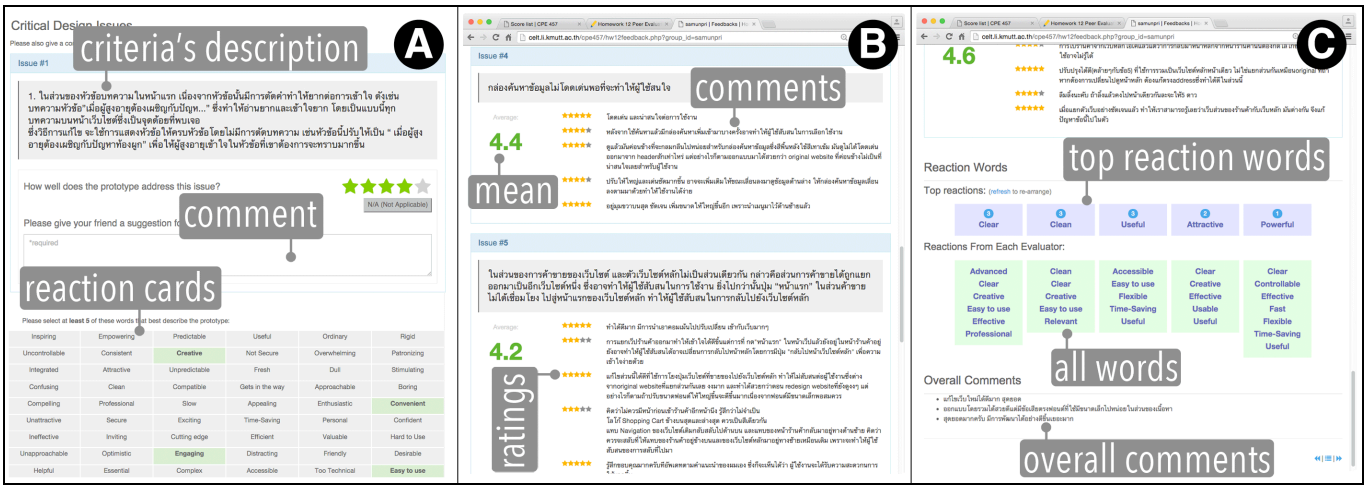

Figure 3: User interface for the second iteration: (A) the evaluation form; (B) a web-based report on issue-specific feedback; (C) reaction words and overall comments

Students received feedback in real time on a web page. For each group, the page aggregated the ratings, the comments, and the reaction words. For each of the issues, all individual comments, the star ratings, and the average were presented as shown in Figure 3 (b). At the bottom, top reaction words were reported in the order of their frequencies as shown in Figure 3 (c). The overall comments were listed as bullet points. Identities of the raters were not shown.

2.3 The Student Survey

At the end of the course, the students were asked to fill out an online survey. In order to ensure that the students consented to giving responses, they were informed at the beginning about the purpose of data collection and how their answers would be treated confidentially without affecting their grades. By design, the respondents remained anonymous; whether they chose to provide their identification numbers was completely voluntary. The instrument had two parts, including close-ended questions with five-point Likert scale (*strongly agree* to *strongly disagree*) and open-ended questions. The results of the former would be presented using descriptive statistics (percentage). In order to present the results more clearly and without distorting the meaning, the scales of *strongly agree* and *agree* were considered the same answer, so were *strongly disagree* and *disagree*. For the open-ended questions, a content analysis was conducted based on aspects of students' thinking, not by priori groupings. Two researchers independently coded the data and later discussed a few unclear statements in order to come to a consensus on the patterns.

3. Results

3.1 Results from the first iteration

3.1.1 Number of evaluators

By design, the work from each student was expected to be evaluated by three evaluators. In practice, out of the 70 students who submitted the work, one student did not get his/her work evaluated by anybody. Forty-one students (59%) received three evaluations as expected. Twenty-eight students (40%) received fewer than three evaluations, with seven students (10%) receiving only one evaluation each.

3.1.2 Agreements of Scores

In Figure 4, each column along the X axis represents a piece of work, each receiving scores from their peers as shown in green circles along the Y dimension. A single score from the teacher is shown as a red square. An average of the scores from peers was calculated and shown as a solid green disc. A disc may have a ring made of dashes, representing the total quality of comments.

There is a low agreement among student raters. The standard deviation of the peer scores for each work ranges from 0 to 29.80, with a mean of 7.80. The teacher gave a score to each work ranging from 20 to 100 with a mean of 72.59. The average peer score ranges from 20 to 94 with a mean of 72.03, strikingly similar to that from the teacher. However, the teacher scores have a higher standard deviation of 29.97, compared to 9.86 for the peer scores.

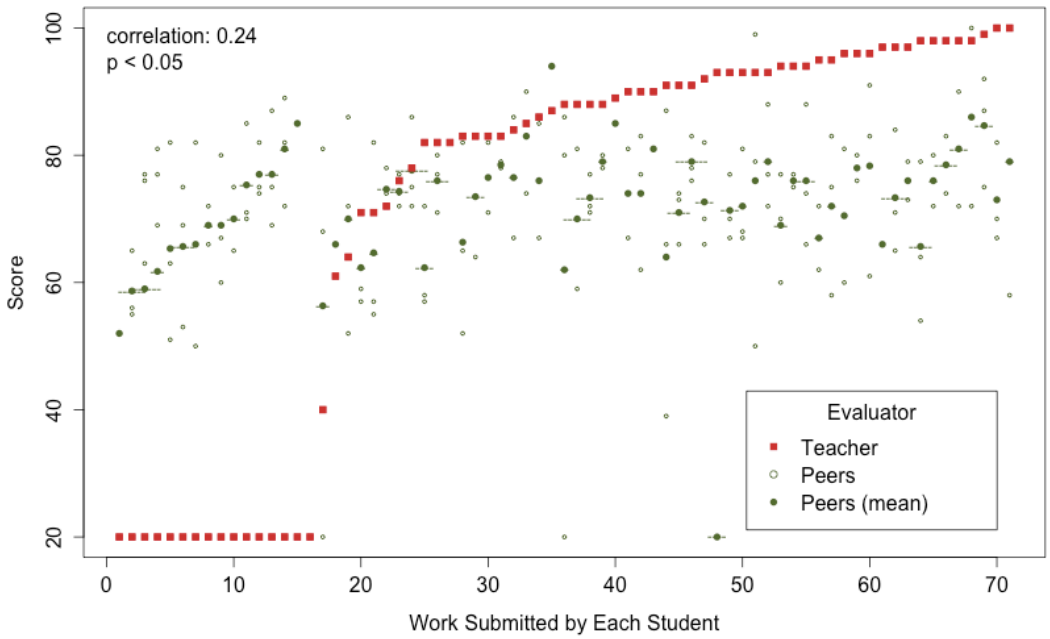

Figure 4: A comparison of scores from the teacher and the peers

Scores from the teacher are generally higher than those from the peers, except for a group of 17 students whose works were deemed incomplete and thus received very low scores. The teacher scores and the peer scores have a very small positive correlation of 0.235 ($p < 0.05$).

3.1.3 Quantity and Quality of Comments

The 70 students gave a total of 174 ratings (87% of 210 that was expected). There were 66 comments (37.9% of 174) with lengths ranging from 4 to 158, with a mean of 52.6 characters. The quality of each comment was rated by the teacher on a scale from 1 to 5. The comments were found to be generally unhelpful with a mean quality score of 2.4. Longer comments were more likely to be useful. The comment quality score was highly correlated with the length of comments with a correlation coefficient of 0.838 ($p < 0.01$).

Only one work received three comments as intended. The average number of comments per work is only 0.93, with as many as 25 works (36%) receiving no comments at all. For each work, the sum of the lengths of comments it received was calculated, along with the sum of the comments' quality scores. The distributions of these variables, shown in Figure 5, suggest that only a few students benefited from high-quality feedback during this iteration. The average total length of comments for a piece of work is 145.6 characters, with an average total quality of 3.48.

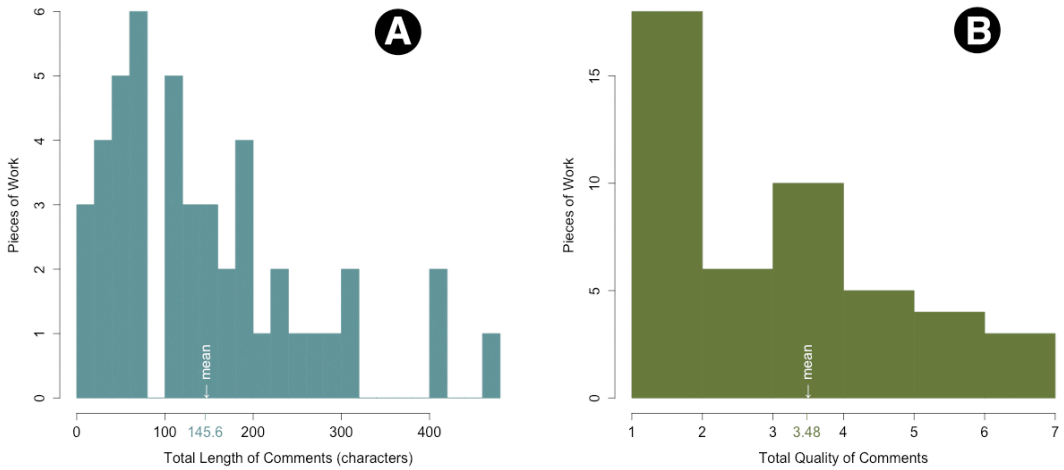

Figure 5: Histograms of comments by (A) the total length; (B) the total quality (first iteration)

In agreement with an earlier analysis, there is a correlation (0.79, $p < 0.01$) between the total quality and the total length of comments. The number of comments is slightly negatively correlated (-0.32, $p < 0.01$) with the average scores given by peers. The more comments a work received, the more likely that the work was rated poorly by peers.

3.2 Results from the second iteration

3.2.1 Group Sizes and Number of Evaluators

The 74 students formed 16 groups whose sizes ranged from two to seven members, with a mean of 4.62 and a median of 5 members. Seventy students participated as raters. Ten groups were assigned four evaluators. Six groups were assigned five evaluators. On average, 4.38 evaluators were assigned to each group.

The evaluation was done in two rounds. During each round, each group received 6 issue-specific comments, 5 or more reaction words, and an overall comment from each rater.

3.2.2 Distribution and Improvement of Scores

An average peer score for each group after the first round is shown in blue in Figure 6 (as bars with backward stripes) and Figure 7 (as triangles). The mean of this score across all groups was 76.05, with an SD of 8.3. The average peer score after the second round is shown in dark green in Figure 6 (as bars with forward stripes) and Figure 7 (as diamonds). The mean across all groups was 84.72, significantly higher than that from the first round, with an SD of 6.6.

The scores from the teacher are shown in Figure 6 as pink bars with vertical stripes. Compared to the peer scores, the teacher scores have a wider range (33.3 to 100), a lower mean (77.83), but a higher median (88.33). No significant correlation was found between the teacher scores and the peer scores.

Groups that received a high peer score in first round tended to also receive a high score in the second round. Scores from the two rounds have a positive correlation of 0.69 ($p < 0.01$). This relationship is visualized in Figure 7 where a pattern emerges that the second scores are almost always higher than the first – the amount of improvement indicated by the vertical distances between triangles and diamonds. Only one group received a lower score in the second round. The net improvement ranges from -1.56 to 21.97, with a mean of 8.67.

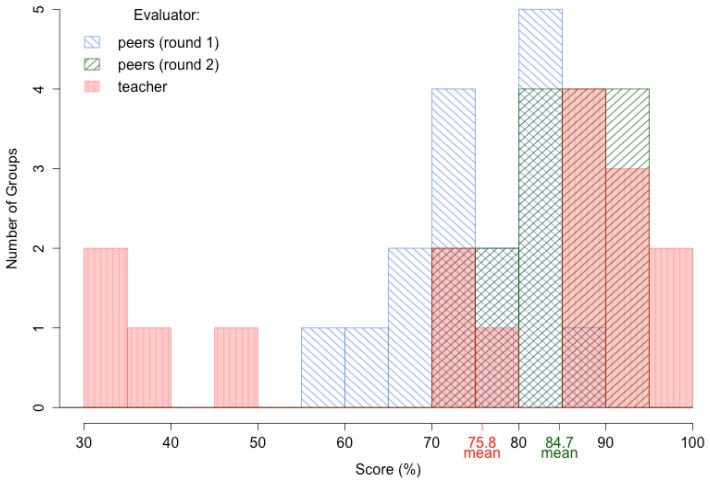

Figure 6: Histograms of scores from the teacher and the peers (second iteration)

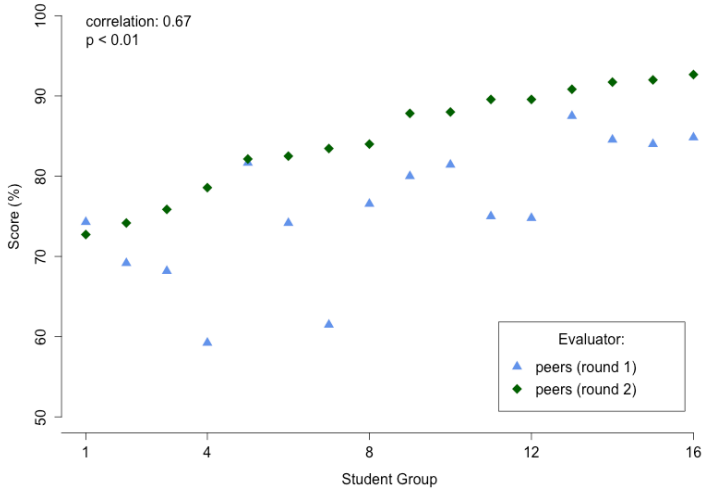

Figure 7: Score improvement after the first round of peer evaluation (second iteration)

3.2.3 Quantity and Quality of Issue-Specific Comments

Issue-specific comments are peers' responses to each group's six self-defined critical design issues. Each group was assigned 4-5 evaluators. Therefore, each group received between 24 to 30 comments per round of evaluation. The total length of all comments for each group were calculated. The histogram is shown in Figure 8 (a). An average group received 7,151 characters of comments from peers (compared with 146 characters from the first iteration and 1,527 characters from the teacher). Comments from the second round are, on average, shorter than those from the first round by 932 characters. That amount is still almost twice the length of comment that the teacher provided.

Each evaluator wrote six issue-specific comments. The aggregate quality of those comments was rated by the teacher on a scale from 1 (least helpful) to 5 (most helpful). Therefore, for each group, the total quality of issue-specific comments can be defined as a sum of these quality ratings. The histogram of the quality scores is shown in Figure 8 (b). The mean quality from the first round (19.2) is slightly higher than that from the second round (17.6). The total quality of comments from the second round is positively correlated with the total quality of comments from the first round with a correlation coefficient of 0.54 ($p < 0.03$).

Two factors were found to have significant effects on the length of comment: the length of student-defined critical design issues and the peer evaluation score from the first round. Groups that wrote longer, more elaborated and thereby less ambiguous criteria generally received longer comments from peers. Figure 9 illustrates this moderate but significant relationship, averaged across both rounds of peer evaluation, with a correlation coefficient of 0.58 ($p < 0.02$). For all groups, an average comment is always longer than an average student-defined criterion as indicated by the positions of blue dots above orange triangles in Figure 9.

During the first round, a student evaluator who gave his/her friend a lower score tended to write lengthier comments. The average score that each group received from peers is negatively correlated with the total length of issue-specific comments with a correlation coefficient of -0.53 ($p < 0.04$). This relationship has a few exceptions as shown in Figure 10. A similar pattern was not found during the second round.

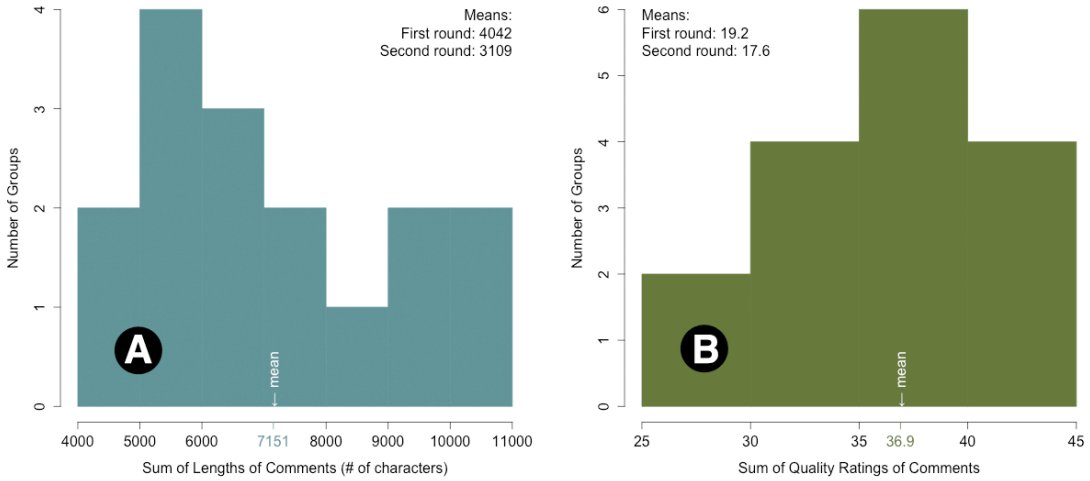

Figure 8: Histograms of comments by (A) the total length; (B) the total quality (second iteration)

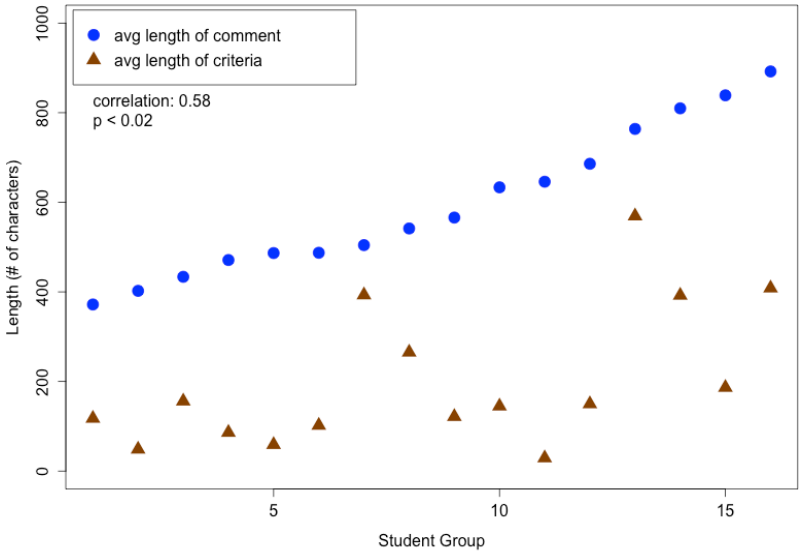

Figure 9: Relationship between the average length of a student-defined criteria and the average length of issue-specific comments (second iteration)

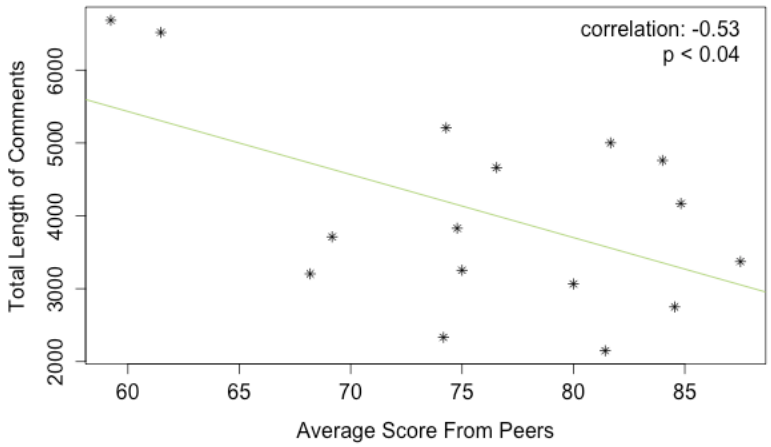

Figure 10: Score VS length of comment from peers during round (second iteration)

3.2.3 Overall Comments and Reactions Words

Each group may receive an overall comment from each peer. The total length of these overall comments ranges from 165 to 1165, with a mean of 573 characters. While a relationship between the length of issue-specific comments and the overall comments was not found, there is a positive correlation between the quality of issue-specific comments and the length of overall comments during the first round, with a correlation coefficient of 0.65 ($p < 0.01$).

The total number of reaction words that each group received (counting duplicates) is positively correlated with the average score from peers during the second round, with a correlation coefficient of 0.60 ($p < 0.02$). The number of words for both rounds ranges from 26 to 60 with a mean of 42.2 words.

Out of the 118 possible choices of reaction words, 96 were chosen, with 19 words being chosen only once. The frequency ranges between 1 and 62 with a mean of 7.03. The most popular words were shown in a word cloud [12] in Figure 11. The words *clear* and *clean* were applied to all groups. All the top-ten words have positive meanings. Nine of these words were applied to at least half of the groups.

Figure 11: A word cloud of the popular reaction words

To ensure the validity of the analysis, any words having a lower frequency than the third quartile (8) were considered noises and thus excluded (70 words). For each word, the scores given to groups associated with that word were aggregated – the min, max, count, and mean were calculated for both the teacher scores and the peer scores. With this information, it is possible to associate each word with the performance of the works it describes.

The words chosen to describe the best works according to the peer evaluation score are: *organized, convenient, accessible, comfortable, calm, powerful, useful, usable, time-saving,* and *understandable*. All words have positive meanings. The only word that may have an ambiguous meaning is *calm*.

According to the teacher's scores, the top words are: *complex, confusing, organized, time-saving, boring, easy to use, flexible, usable, stable,* and *useful*. This list highlights the difference in opinion between the students and the teacher. Works that were rated highly by the teacher were described with negative words such as *complex*, *confusing*, or *boring*.

Four words appear in both lists: *organized, useful, usable,* and *time-saving*. These positive words are associated with high scores from both the peers and the teacher.

3.3 Student Survey Results

3.3.1 Part One of the Student Survey

Out of the 71 respondents (96% of the class), 59.2% were male and 40.8% were female. Most were from the Faculty of Computer Engineering comprising of International Program (49.3%) and Thai Program (32.4%); the rest (18.3%) were from the School of Architecture and Design (International Program). Some students had a prior experience with peer assessment (39.4%), while 38% had never had such experience and 22.5% felt unsure.

As shown in Figure 12, most of the students perceived the positive value of peer assessment from both the aspects of feedback giver and receiver – an opportunity to improve one's own work as well as the others. Over 70% thought they felt encouraged to improve the work quality and over 50% thought they had put efforts into making comments so the work's owners would receive unbiased and helpful feedback. Interestingly, 45% of the respondents indicated that they felt uncomfortable to judge other's work even though their anonymity was secured. It was also found that only a small number of them viewed peer feedback as extra work that meant nothing to their work. The results for other relevant statements also indicated that most students viewed peer assessment favorably – most of them believed that their

understanding of the subject content increased with the help of such peer feedback and more than 70% felt that this system helped develop evaluative skills. However, 68% of the respondents thought that peer feedback should not be used as part of grading so that they would feel free to give comments.

Even though most students thought that peer assessment was useful, almost half felt neutral about whether the reiteration of the homework was more effective when incorporating peer feedback and, unsurprisingly, needed more guidelines from the teacher when doing peer feedback. While a few students thought that the teacher's feedback was enough, almost half of them acknowledged that such assessment system made the classroom environment collaborative.

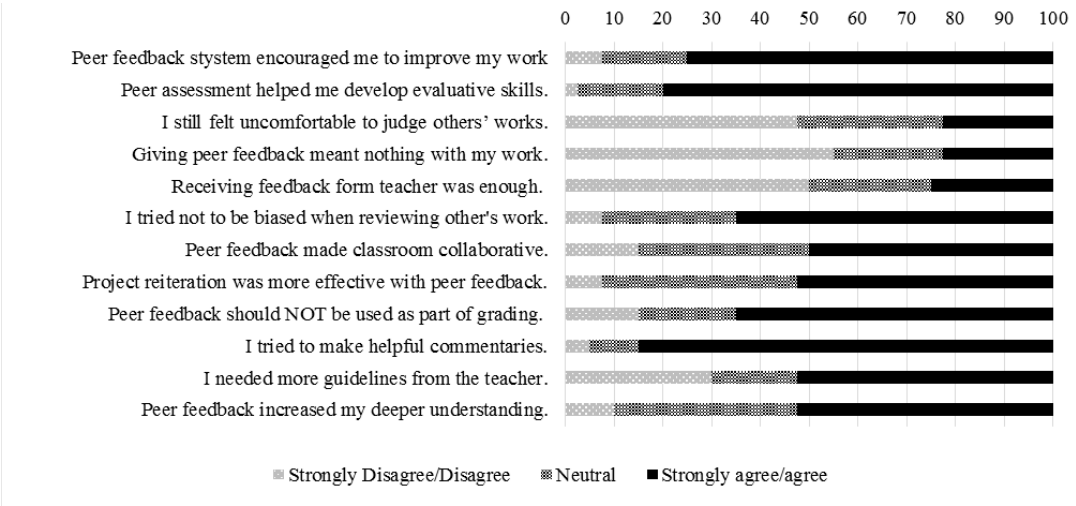

Figure 12: Percentage of students' attitude towards peer assessment

3.3.2 Part Two of the Student Survey

The findings of the qualitative data revealed how the students thought about peer assessment and its contribution to their learning. Five themes from the analysis of responses were identified according to the emergent patterns. Five emergent themes divided into two broad categories: positive and negative attitudes. The first three themes are related to positive attitudes including *work improvement*, *skill development through peer learning*, and *conditional benefit*; the other two themes were categorized as *negative attitudes* and *other*. The data was quantified as shown in Figure 13.

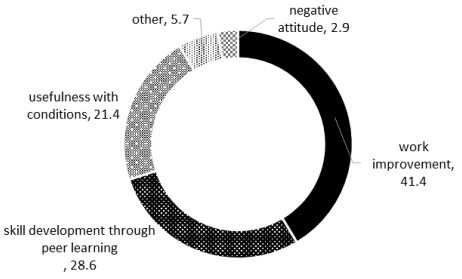

Figure 13: Emergent themes of students' attitudes towards the contribution of peer assessment to their learning

Out of the 70 comments, most students reported that peer assessment helped them learn. The students thought peer assessment encouraged them to improve their work (41.4%) while some reported that giving and taking feedback helped them develop critical thinking skills and a learning mindset (28.6%). Here are some selected responses:

> "I thought peer assessment gave me many comments to improve my work. Also, it came from students in other fields so it help me to recheck and develop my own work."
>
> "The student will have more various comment to improve their works. It also increases the understanding about the content of the subject."
>
> "This part allow me to comment other people work to give them suggestions. Also, I can get suggestions from other people as well which would be different from the teacher's."
>
> "Because we have to use our thinking skill to be able to critic friend's work and it would make us be more professional for deciding thing."

Some students thought that peer assessment was useful but would only be effective under certain conditions (21.4%). Students' concerns ranged from teacher's involvement in the feedback loop and the quality of feedback which depended on adequate knowledge and effort of peers. Here are some examples of the remarks:

> "Over all is fine but, I like the teacher to comment us too and we are able to talk with the teacher if we agree with the comments or not because sometimes people don't see things in the same way."
>
> "The comments from peer could be either helpful or destructive to the work development. The negative impact from comments is possible if the assessor do not take it serious."
>
> "It helps me improve my evaluative skill. But the other comments for my work that I received were not really useful. Because those who made the comments do not really understand my point of view in design and I cannot use them in improving my work at all."

A minority of students (2.9%) expressed negative attitudes towards the experiences of the peer assessment as they thought it was the teacher's responsibility when it came to assessment matters. Remaining statements were classified as *other* (5.7%) due to their ambiguous meanings. These include increasing collaboration in class and the impact of peer comments over one's own work.

4. Discussion

For a large classroom, an online peer assessment system enables the teacher to provide timely feedback to students while freeing up valuable teacher's time. In the second iteration of the proposed design, the students received substantially more comments than they did in the first iteration, or even from the teacher. Comparing Figure 5 to Figure 8, it can be observed that more comments are of higher quality. By asking more relevant questions using student-defined criteria, the raters were encouraged to write better comments. As illustrated in Figure 9, the length of the wording of a criterion has an *anchoring effect* [13] on the length of comments it receives. This suggests that the teacher should call for all student-defined criteria to be written out as clearly as possible within a reasonable length.

Raters who gave lower scores tended to write more or lengthier comments as explained in Section 3.1.3 and illustrated in Figure 10. On the contrary, those who gave higher scores did not feel as obligated to justify their ratings with detailed comments. This suggests that students should be encouraged to find ways to improve their friend's work even when they already like the work.

Whereas the comments were mainly used as justification for giving lower scores, reaction cards were mostly used to give compliments. The ten most frequently-used words were all positive. Peers who liked the work tended to select more words to describe it. An analysis of the association between reaction cards and scores also revealed interesting differences of opinion between novice and expert raters. Information gathered from reaction cards can be analyzed and visualized in a variety of ways. It can also serve as a guideline for a construction of student-defined criteria.

A fair distribution of raters is an important design goal of a peer evaluation system. In many situations, it is not known beforehand which students would be absent. It is important that they must not be matched to a group; otherwise, the group would be missing its rater. An on-demand matching algorithm solved this problem and ensured fairness.

A trade-off between being informed and remaining anonymous must be made when designing a peer assessment system. By enforcing full anonymity (the first iteration), the raters were left in the dark when they had questions about the work. By opening up the identity of the work's owner (the second iteration), the raters were able to ask questions as needed. Note that, by asking a question, they risked exposing their identity at their own choice. A future work may explore designs where questions from the raters can be made anonymously.

The survey results show that, by providing feedback to their friends, the students themselves became active learners and engaged in their own learning. Feedback is most effective in promoting learning if it involves them in the decision process, so they are not passive recipients of the teacher's judgments [14]. The students' attitudes towards this system aligned well with the rationale of using peer feedback as formative assessment. Students appreciated an opportunity to learn about their shortcomings as guided by peer critiques. In this class, to reiterate the work meant not only to refine their work for better scores; but they were also strategically induced to improve their learning skill.

Anonymity of evaluators and contribution of peer feedback to grades were the issues that concerned most students. Meaningful feedback is more likely to occur when the teacher lets the students know not only *what* to do, but also *why* and *how* to do peer assessment. This is especially important for students who are new to this learning approach. In order to elicit high-quality feedback from novice evaluators, the teacher's guidelines and feedback should also be timely managed. A peer feedback system can complement the teacher assessment but does not replace it.

5. Conclusion and Future Work

Peer feedback not only made the students aware of their own strengths and weaknesses but also gave the teacher pointers to adjust the learning process. Designed and implemented carefully, a peer assessment system can address the challenges of large classrooms: fairness, timeliness of feedback, limitation of the teacher's time, and student engagement. It also serves as a platform for students to practice self-regulated learning and critical thinking, which are important attributes of the *21st Century Skills* [15].

Future work should include developing the online peer evaluation tool to support general assignments, deploying it to the public, and real-time data collection and analytics from a variety of classrooms.

6. Acknowledgement

The authors would like to thank Dr Suporn Pongnumkul for her helpful insights and a pointer to the reaction cards.

7. References

[1] OECD. (2005). Formative Assessment: Improving Learning in Secondary Classrooms. *Assessment, 29* (November), 282. http://doi.org/www.oecd.org/dataoecd/19/31/35661078.pdf

[2] Bull, S., Johnson, M. D., Epp, C. D., Masci, D., Alotaibi, M., & Girard, S. (2014). Formative Assessment and Meaningful Learning Analytics. In *IEEE International Conference on Advanced Learning Technologies* (pp. 327–329). http://doi.org/10.1109/ICALT.2014.100

[3] Chen, C. M., & Chen, M. C. (2009). Mobile formative assessment tool based on data mining techniques for supporting web-based learning. *Computers & Education*, 52(1), 256–273. http://doi.org/10.1016/j.compedu.2008.08.005

[4] Wang, T. H. (2011). Developing Web-based assessment strategies for facilitating junior high school students to perform self-regulated learning in an e-Learning environment. *Computers and Education*, 57(2), 1801–1812. http://doi.org/10.1016/j.compedu.2011.01.003

[5] Mahroeian, H., & Chin, W. M. (2013). An Analysis of Web-Based Formative Assessment Systems Used in E-Learning Environment. In 2013 *IEEE 13th International Conference on Advanced Learning Technologies* (pp. 77–81). http://doi.org/10.1109/ICALT.2013.28

[6] Ng, H. Z., & Hussain, R. (2013). The use of Googlesites as a formative assessment tool [eportfolio] in higher education. In *63rd Annual Conference International Council for Educational Media* (pp. 1–9). http://doi.org/10.1109/CICEM.2013.6820157

[7] Kowalski, F. V., & Kowalski, S. E. (2012). Enhancing curiosity using interactive simulations combined with real-time formative assessment facilitated by open-format questions on Tablet computers. In *IEEE Frontiers in Education Conference Proceedings* (pp. 1–6). http://doi.org/10.1109/FIE.2012.6462282

[8] Isabwe, G. M. N. (2012). Investigating the usability of iPad mobile tablet in formative assessment of a mathematics course. *IEEE International Conference on Information Society (i-Society)* (pp. 39–44).

[9] Nielsen, J., & Molich, R. (1990). Heuristic evaluation of user interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems Empowering People - CHI '90* (pp. 249–256). http://doi.org/10.1145/97243.97281

[10] Orsmond, P., Merry, S., & Reiling, K. (2002). The Use of Exemplars and Formative Feedback when Using Student Derived Marking Criteria in Peer and Self-assessment. *Assessment & Evaluation in Higher Education*, 27(4), 309–323. http://doi.org/10.1080/0260293022000001337

[11] Benedek, J., & Miner, T. (2002). Measuring Desirability: New methods for evaluating desirability in a usability lab setting. *Proceedings of Usability Professionals Association*, *2003* (pp. 8-12).

[12] Heimerl, F., Lohmann, S., Lange, S., & Ertl, T. (2014). Word cloud explorer: Text analytics based on word clouds. In *Proceedings of the Annual Hawaii International Conference on System Sciences* (pp. 1833–1842). http://doi.org/10.1109/HICSS.2014.231

[13] Kahneman, D. (1992). Reference points, anchors, norms, and mixed feelings. *Organizational Behavior and Human Decision Processes, 51*, 296–312. http://doi.org/10.1016/0749-5978(92)90015-Y

[14] Harlen, W. (2006). On the relationship between assessment for formative and summative purposes. In Assessment and Learning, J. Gardner, ed. (London: Sage Publications), pp. 103–118.

[15] Bellanca, J., Eds, R. B., Barell, J., Darling-hammond, L., Dede, C., Dufour, R., … November, A. (2010). 21st Century Skills: Rethinking How Students Learn. *Solution Tree Press Study Guide*, 1–27.