Who Took Peer Review Seriously: Another Perspective on Student-Generated Quizzes

Yang Song
Department of Computer Science
North Carolina State University
Raleigh, United States
ysong8@ncsu.edu

Zhewei Hu
Department of Computer Science
North Carolina State University
Raleigh, United States
zhu6@ncsu.edu

Edward F. Gehringer
Department of Computer Science
North Carolina State University
Raleigh, United States
efg@ncsu.edu

ABSTRACT

In educational peer-review activities, one challenge is to tell which peer reviews are credible. Reputation systems are one approach. However, they work best in a scenario where (i) most of the peer reviewers can do a decent job, or (ii) reviewers tend to do peer-reviews in fixed styles─e.g., negative reviewers are negative on all artifacts they review; accurate reviewers grade all work accurately. We argue that, those hypotheses may not hold for many cases for educational peer review. For this reason, we invited student authors in one class to create quiz questions based on their own artifacts. After their reviewers finished peer-review responses, they could take the corresponding quizzes. The mere existence of those quizzes may encourage reviewers to take the peer-review activity more seriously. In addition, their quiz scores can be used as another source of reputation. Since those quizzes are generated by authors to "test" the reviewers, the quiz scores give teaching staff another input on deciding whether a peer review is credible. Our experiments show that a reputation generated from quiz scores provided more accurate estimates of final grades for students' artifacts than some existing reputation systems.

Keywords

Educational peer review; student-generated quiz; contributing student approach; student-generated content; reputation system.

1. INTRODUCTION

Educational peer assessment has proven to be a useful approach that can provide students timely feedback and opportunities to help and learn from each other. Reviewers are often expected to provide both formative feedback─textual feedback telling the authors where and how to improve the artifacts─and summative review─peer grading telling authors how good their artifacts are─at the same time. Formative feedback is important for the authors when the assignments are still ongoing because timely and insightful feedback can help authors improve their artifacts. In a large class or MOOC, when the teaching staff is stretched thin in providing help, formative feedback from peers is the best help that the course participants may receive. Summative feedback can also aid the teaching staff, as it provides more input to help determine final grades. In a large class or MOOC, the teaching staff is almost forced to base final grades on the summative reviews that students have received [1].

1.1 Reputation Systems

After peer reviews have been completed, they can simply be shown to the student authors. However, they may not tell the authors much because there is no guarantee that the peer reviews are insightful or helpful. A peer-review done without dedication or conscientiousness may not have much impact on the author's learning process [2]. If teaching staff want to assure that all students receive competent and insightful feedback, or even use the peer-grading scores to generate final grades, an approach is needed to help the instructor identify the credible reviewers.

A reputation system [1, 3, 4] is one of the solutions for this. Reputation systems may take many different inputs:

- Are the scores assigned by the reviewer close to the scores assigned by the instructor, on work that they both review?
- Are the scores assigned by the reviewer close to the scores assigned by the other reviewers, on work they both review?
- Is the reviewer habitually stingy, or generous, to different kinds of work?
- How competent has the reviewer been on other work in class?

Any or all of the above information may be factored into a reviewer's reputation. These reputation scores can help teaching staff come up with the final grades for each artifact. If students are allowed to see their reputation scores, they can also learn how effective their reviewing has been.

There is a common assumption in some existing reputation systems that a reviewer's reputation on one task carries over to other review tasks [3, 4]. In other words, if one is deemed to be a credible reviewer, reputation systems assume that all his/her past peer reviews are credible and that future peer reviews will be credible as well. This "holistic reputation" hypothesis might hold in some cases, e.g., where all artifacts are reasonably similar to reach other, or even, on the same topic. However, if the artifacts are on different topics, or some of the artifacts are beyond the knowledge of the reviewer, this hypothesis cannot hold.

There are other issues related to existing reputation systems. First, the reputation scores should be calculated after the peer-review phase is finished [3, 4]. Though the teaching staff can apply a reputation system in the middle of the peer-review phase, the reputation scores for reviewers will keep changing as more reviews are done. In addition, some reputation systems calculate scores iteratively: they calculate the reputation scores, use the reputation scores to calculate the grades for each artifact, then re-calculate the reputations, till the results converge. The convergence, however, may not be at the global optimal fixed point [3]. A third problem is that reputation systems are "black boxes" to students, therefore, letting them see reputation scores may raise more questions about their validity.

1.2 Student-Generated Questions

Though unrelated to peer review, student-generated questions have long been considered a helpful process which triggers critical thinking [5, 6]. The process of authoring questions and answers helps students to retain prior knowledge and relate new knowledge to it. Question authoring usually starts by retrieving information from one's memory, which is a process that students are not motivated to undertake. As they author questions, they also need to figure out which part of the learning content is worth testing. While doing this, they clarify their own understanding, which is another aspect of self-directed learning. Finally question authors also need to construct answers for the questions. On the whole, the question-authoring process includes retaining learning, identifying key points, connecting knowledge, self-clarifying knowledge and self-assessment [7]. Previous research has also proved that students, though are not usually place in the role as question creator, can create practice questions of high quality: with well written question stems and good distracters [8].

In our research, we invited student authors to create quiz questions on their own artifacts (student-authored Wikipedia pages on different topics). After that, in the peer-review phase, student reviewers were encouraged to take those quizzes after giving peer reviews. Quiz scores served not only to help teaching staff tell whether student reviewers read the artifacts carefully enough, but also served as another means of computing reputations.

2. EXPERIMENTAL DESIGN AND DATA COLLECTION

In this section we provide an overview of our experimental design, class policies for our later experiments and our approach to checking and validating the data set we collected.

2.1 Class Setting

Our data is collected from students using Expertiza, a web-based system developed to facilitate peer-review activities on student-generated course content [9]. In 2014 we added a student-generated quiz feature into this system so student authors could create quiz questions after they submitted their artifacts. In the peer-review phase, after a reviewer finishes reviewing an artifact, (s)he can also take the quiz written by the author of the artifact. The question types supported are multiple-choice questions, checkbox questions (similar to multiple-choice questions, but potentially with several correct choices) and true/false questions. In our study, each quiz contained five questions. Each quiz could only be taken once by each quiz taker.

The assignment on which we applied our experiments was a wiki-writing assignment over two semesters in a graduate-level course in the College of Engineering in NC State University. Students were required to write or edit a Wikipedia page on a recent technique, product, programming framework, etc., related to object-oriented design or development.[1] Students were allowed to work in teams with maximum of two members per team.

The peer-review task for this assignment asked about a variety of aspects which related to the quality of the artifacts, including originality, structure, use of language and examples, etc. Most of the criteria in the review rubric allowed reviewers to give both textual feedback (as formative feedback) and Likert-scale scores (as summative feedback). According to the syllabus, each student was required to do only two peer reviews, but could do extra reviews for extra credit. The assignment consisted of two rounds of submission and review. After the first round, authors could change their artifacts based on the reviews they received. After the second review phase, the reviewers could take the quizzes on the artifacts they had reviewed.

Almost 200 students completed those assignments. We collected 74 sets of quiz questions. On average, each artifact was reviewed 9.1 times, and each quiz was taken 4.8 times. Quiz-authoring and quiz-taking was not required in the syllabus but students were given extra credit for them.[2] Quiz takers needed to score 80% or higher to pass the quiz. The overall passing rate was 81.6%.

2.2 Data Verification

A member of the teaching staff reviewed the student-generated quizzes to make sure that they were well formed (e.g., each choice of a multiple-choice question should have some content) before the quizzes were taken by the reviewers. We found three cases of incomplete questions and asked the authors to change them.

After this assignment was finished, we did further checks on the quality of student-generated questions. Figure 1 shows the distribution of passing rates of student-generated quizzes against the average quiz scores from quiz takers. Most of the quizzes had a passing rate of almost 80% (*x*-axis of Figure 1). The average quiz score on each set of the quiz was also high; most of them were higher than 80% (*y*-axis of Figure 1).

We suspected that the quizzes which had both a high average score and high pass rate might be too easy, and the ones which had both low average score and low passing rate might be poorly designed. Therefore, we examined all the quizzes from the upper-right corner and lower-left corner of the graph within a [10%, 10] range. We also examined some of the quizzes not in those 2 ranges. Among 14 samples, we found that two of them were so easy that the quiz takers could answer most of the questions correctly without fully understanding the artifacts. The quizzes in the lower left-hand corner are harder to answer because the authors created more checkbox questions. We did not identify any question that was invalid, irrelevant or given the wrong answer.

[1] The Wikipedia did not have that page, or only had very limited content when students started to add the content. The writing was done in the Wikipedia sandbox.

[2] We provided a online instruction document to students: goo.gl/7Eud42

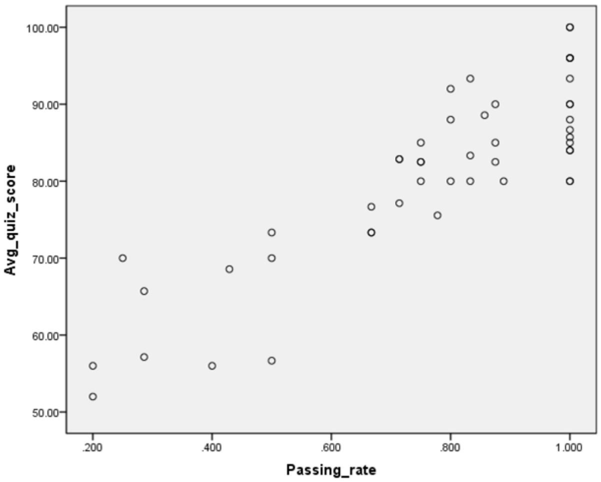

Figure 1: Distribution of the passing rates and average scores of student-generated quizzes

3. REPUTATION SYSTEM BASED ON STUDENT-GENERATED QUIZZES

Inputs of reputation systems vary, but a common way to represent peer-review scores is to use an adjacency matrix. In this matrix, each row stands for an artifact and each column represents a reviewer. The values in the matrix are the scores given by each reviewer to each artifact (if the reviewer reviewed that artifact). Reputation systems use the matrix to generate two quantities: the reputation scores for each reviewer and the weighted grades for each artifact. Since we used scores of student-generated quizzes as reputations, we have another adjacency matrix for quiz scores. In this matrix, again, each row stands for an artifact and each column stands for a reviewer. The values in the matrix are the quiz scores obtained by each reviewer on each quiz from the artifact author. Please note that a reviewer could only take a quiz if that reviewer had reviewed the artifact created by the same author.

The reputation can be calculated by comparing the peer-review scores and instructor-assigned grades. Since in practice, the reputation systems runs before the grading is done by teaching staff, we made use of two existing reputation algorithms which do not use instructor-assigned grades [3, 4].

To give precise definitions of the algorithms in this paper, let *a* be an artifact; *A* be the set of all the artifacts; *r* be a reviewer; *R* be the set of all the reviewers; g_a^r be the grade that *r* assigned to *a*; q_a^r be the quiz score that *r* earned from the quiz associated with *a*; R_r be the set of artifacts reviewed by *r*; A_a be the set of reviewers who have reviewed *a*;W_r be the weight for reviewer *r* (the weight could be temporary depending on which algorithm is used); $G_{aggregated}^a$ be the grade that the reputation algorithm aggregated for artifact *a* based on current reputations; and G^a be the temporal grade for artifact *a* in the algorithm.

We assume that the quiz score q_a^r can be used as reputation:

$$W_r = \begin{cases} q_a^r & q_a^r \neq \text{NULL} \\ \sigma & q_a^r = \text{NULL} \end{cases} \quad (1)$$

in which σ is an small constant for the reviewers who did not take the quiz. The idea behind this is that, those reviewers who did not take the quiz still did the peer reviews, so they should have a reputation, though modest.

$G_{aggregated}^a$ is the grade aggregated from the weights of the reviewers who have reviewed *a*:

$$G_{aggregated}^a = \sum_{r \in A_a} g_a^r \times W_r \Big/ \sum_{r \in A_a} W_r \quad (2)$$

We compare our algorithm with Hamer's algorithm and Lauw's algorithm, which are both reputation algorithms based on only peer-review scores (without quiz scores).

Both Hamer's [3] and Lauw's [4] algorithms are iterative. In each round the weights are calculated and then used to calculate the aggregated grades. Then, for the next round, the aggregated grades are used as temporary grades to calculate the new reputations: $G^a = G_{aggregated}^a$.

In Hamer's algorithm the grading variance of *r* is calculated from the difference between the reviewer's grading and the aggregated grades from last round:

$$\Delta_r = \sum_{a \in R_r} (G^a - g_a^r)^2 \Big/ |R_r| \quad (3)$$

After calculating the variances from all the reviewers, the weight is defined as:

$$W_r' = \overline{\Delta_r} \big/ \Delta_r \quad (4)$$

So the higher W_r' is for $r \in R$, the more "reliable" reviewer *r* is. The range for W_r' is $(0,\infty)$.

Since some of the reviewers may have very high weights and thereby "dominate" the aggregated grades, Hamer proposed a "log-damping" process:

$$W_r = \begin{cases} 2 + \log(W_r' - 1) & W_r' > 2 \\ W_r' & W_r' \leq 2 \end{cases} \quad (5)$$

Lauw's algorithm is similar to Hamer's algorithm. The difference is that Lauw's algorithm keeps track of the leniency for each reviewer. The leniency can either be positive or negative, depending on whether the reviewer tends to grade artifacts higher or lower compared with aggregated grades.

Let l_r be the leniency of *r*:

$$l_r = \frac{\sum_{a \in R_r} (g_a^r - G^a) / g_a^r}{|R_r|} \quad (6)$$

Given l_r for all $r \in R$, $G_{aggregated}^a$ can be calculated by:

$$G_{aggregated}^a = \frac{\sum_{r \in A_a} g_a^r \times (1 - \alpha \times l_r)}{|A_a|} \quad (7)$$

a $(\alpha \in [0,1])$ is a scaling factor, it controls "the extent to which the scores may be adjusted to compensate for leniency".

The weight from Lauw's algorithm can be defined as:

$W_r = 1 - |l_r|$ (8)

4. EXPERIMENTS AND ANALYSIS

4.1 Experiment Results

We used students' quiz scores as reputations to aggregate the grades for all the artifacts and compare them with the final grades assigned by teaching staff ("expert" grades). Similar to our approach, Hamer's algorithm and Lauw's algorithm are also used to calculate the reputation-aggregated grades. The comparison results are shown in Table 1 (please note that all the grades are 100 based in this paper).

Table 1. Comparison of difference between aggregated grades and expert grades

	Quiz-based	Hamer's algorithm	Lauw's algorithm
Bias range	[–8.4, 12.5]	[–9.6, 19.1]	[–10.3, 12.5]
RMSE on bias	4.4	5.3	4.4
Avg. absolute bias	3.4	3.9	3.4

The bias range is the range of difference between aggregated grades and expert grades. The quiz-based reputation approach has the smallest bias range, which means that even in the extreme case, the aggregated grades by quiz-based reputation algorithm may make smaller mistakes than other reputation algorithms. The root-mean-square error on quiz-based reputation approach was 4.4 and average of absolute bias is 3.4, which means that if teaching staff applies a similar approach and do not grade all by themselves, the average absolute bias between aggregated grades and expert grades is 3.4, which is still acceptable in most cases.

Not all the peer-reviewers took the quiz since it is not a required activity. We next tested if our quiz-based reputation approach works for the artifacts with smaller number of quiz takers. We divided the artifacts into two groups with roughly the same sizes: Group 1 is formed by the artifacts with higher numbers of quiz takers and Group 2 is formed by the artifacts with lower numbers of quiz takers. We did the comparison again, and the results are shown in Table 2.

From Table 2 we find that, for the artifacts in Group 2 which all have less than 4 quiz takers, the aggregated grades are still quite close on average to the expert grades. There is also no significant difference on the biases between Group 1 and Group 2 (the *p*-value between the biases from Group 1 and Group 2 from the quiz-based algorithm is 0.81). This indicates that, even in the extreme cases that some artifacts only have a few quiz takers, our approach still works. However, comparing the biases from Group 1 and Group 2, quiz-based algorithm works better when there are more quiz takers.

Another perspective we considered was whether our approach still works on the artifacts that received limited numbers of peer reviews. For the reputations calculated from reputation systems, the reputation of a reviewer was not based on whether his/her peer-review grades agree with others' on one artifact, but on all the artifacts that (s)he has reviewed. Therefore, even if an artifact is reviewed by a limited number of reviewers, the reputations and aggregated grades should still be reliable since the reviewers may also review other artifacts. However, if the reputations are the reviewers' quiz scores on the artifacts, will smaller numbers of reviewers lead to lower validities of aggregated grades using our approach? We again divided the artifacts into Group 3 and Group 4 which almost had the same size. Group 3 contains the artifacts which have higher numbers of reviewers; Group 4 contains the artifacts with fewer reviewers. We did the comparison again and the results are shown in Table 3.

Table 2. Comparison of difference between aggregated grades and expert grades between Group 1 and Group 2

		Quiz-based	Hamer's algorithm	Lauw's algorithm
Group 1 (more quiz takers)	Bias range	[–7.3, 11.7]	[–6.0, 12.6]	[–7.1, 11.8]
	RMSE on bias	3.6	4.0	3.7
	Avg. absolute bias	2.5	3.1	2.7
Group 2 (less quiz takers)	Bias range	[–8.4, 12.5]	[–9.6, 19.1]	[–10.3, 12.5]
	RMSE on bias	5.1	6.3	5.0
	Avg. absolute bias	4.2	4.7	4.0

Table 3. Comparison of difference between aggregated grades and expert grades between Group 3 and Group 4

		Quiz-based	Hamer's algorithm	Lauw's algorithm
Group 3 (more reviewers)	Bias range	[–8.0, 11.7]	[–8.1, 12.6]	[–7.5, 11.8]
	RMSE on bias	4.3	4.7	4.4
	Avg. absolute bias	3.2	3.5	3.3
Group 4 (less reviewers)	Bias range	[–8.4, 12.5]	[–9.6, 19.1]	[–10.3, 12.5]
	RMSE on bias	4.5	6.0	4.4
	Avg. absolute bias	3.6	4.5	3.3

From Table 3 we find that, all three approaches work better on Group 3 which had more reviewers, and our approach performed the best. On Group 4, our approach performed slightly worse than Lauw's algorithm, but the difference was small (0.3 on average bias and 0.1 on RMSE). We also found that the results are similar to Table 2 because in our experiment the quiz takers have to be reviewers of an artifact first. Therefore, the artifacts from Group 1 and Group 3 are mostly the same, similarly, a large portion of the artifacts in Group 2 and Group 4 are also the same.

4.2 Discussion

From the results above, we found that using quiz scores as reputations worked as well as, or sometimes even better than some existing reputation systems.

To look more closely into the reason for this, we used one artifact from our data set as an example. This artifact received 4 peer reviews, and all those 4 reviewers took the quiz created by the author.

Figure 2 shows the bubble graph of the peer-review scores and the reputations calculated by Hamer's algorithm. The *x*-axis is for different reviewers; the *y*-axis shows the peer-review scores; the size of bubbles demonstrates the reputation of each reviewer from Hamer's algorithm. The bubbles are bigger for the reviewers with higher reputations. We found that Reviewer 1 earned a very high reputation score (2.6) while the rest of the reviewers had low reputations (0.5, 0.1 and 0.1). Considering that the range of reputations from Hamer's algorithm is $(0,\infty)$, and a reviewer with average review skill should receive a 1 for reputation, Reviewer 2, Reviewer 3 and Reviewer 4 are considered to be very poor in peer-reviewing. Since Reviewer 1 gave this artifact a 100, the aggregated grade for this artifact from Hamer's algorithm was also very close to 100, which is much higher than the expert grade (by 11.4 points).

Figure 3 shows another bubble graph of the peer-review scores and reputations calculated by Lauw's algorithm on the same artifact. Lauw's algorithm usually assigns reputations more leniently than Hamer's algorithm. The *x*-axis is for different reviewers; the *y*-axis shows the peer-review scores; the size of bubbles demonstrates the reputation of each reviewer from Lauw's algorithm. We found that Reviewer 1 and Reviewer 2 both received higher reputation scores (0.93 and 0.96) from Lauw's algorithm and Reviewer 3 and Reviewer 4 received low reputations (0.51 and 0.65) (the reputation range of Lauw's algorithm is $[0,1]$). The aggregated grade for this artifact with Lauw's algorithm, in this case, is not dominated by the one(s) with high reputation(s), therefore is lower than the aggregated grade from Hamer's algorithm, but still higher than expert grade (by 6.4 points).

Comparing Figure 2 with Figure 3, there is a disagreement about Reviewer 2's reputation between Hamer's algorithm and Lauw's algorithm. We sometimes observe this kind of disagreement, but it does not result in a large performance difference between the two reputation algorithms [10].

Figure 4 shows bubble graph of peer-review scores and reviewers' quiz scores. The *x*-axis is for different reviewers; the *y*-axis shows the peer-review scores; the sizes of bubbles demonstrate quiz scores for reviewers. Reviewer 1, 2 and 4 all passed the quiz but reviewer 3 did not (actually, only answered 40% of the quiz questions correctly). The aggregated grade, in this case, is lower because Reviewer 3 and 4 gave lower peer-review scores. Even though considered as unreliable reviewers by other reputation algorithms, Reviewer 3 and Reviewer 4 were actually correct on this artifact─the expert grade was closest to aggregated grades from this approach (the difference is 4 points).

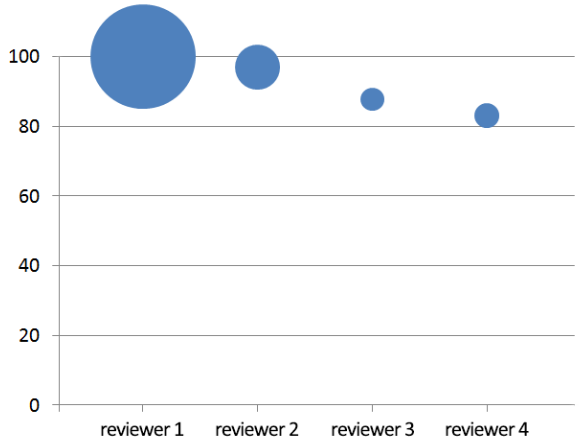

Figure 2: Peer-review scores and reputations from Hamer's algorithm

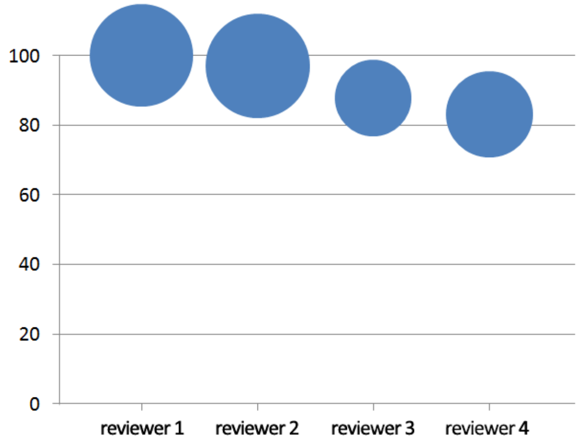

Figure 3: Peer-review scores and reputations from Lauw's algorithm

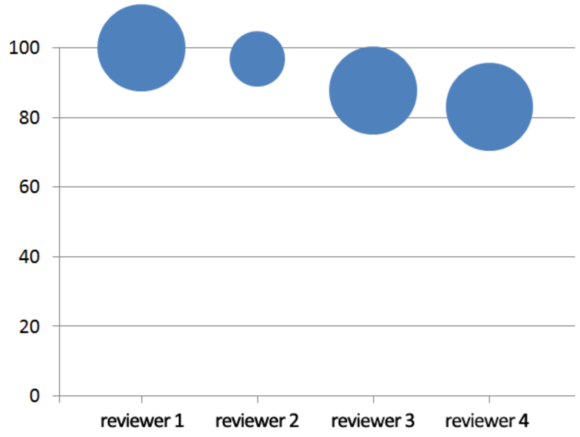

Figure 4: Peer-review scores and quiz scores as reputation

This case suggests that the “holistic reputation” assumption may not hold in this assignment where the authors worked on different topics. Some reviewers with lower reputations may not be able to perform credible peer review all the time, but their reviews should still be considered as credible if there is proof that they understand the artifact well enough. By contrast, the good peer reviewers with higher reputations, even though their peer reviews are reliable most of the time, may still assign grades which are very far from expert grades (in our case, this could cause by that they did not have a sufficient understanding of the topic to which the artifact related). In this case, letting those who have higher holistic reputation scores “dominate” the grade aggregation will lead to mistakes, such as the case in Figure 2.

5. CONCLUSION

The lesson of this paper is that reputation systems for peer review, and, by extension, the reliability of peer reviewing and peer grading, can be enhanced by having student authors write quizzes on the artifacts that they have submitted, and using reviewers’ scores on these quizzes as evidence for the validity of their reviews.

In our experimental design, after authors submit their artifacts, they can create a set of quiz questions for their peer reviewers. The reputations of peer reviewers are based on how well they did on quizzes written by the authors of the artifacts they have reviewed. The philosophy behind this is, the reviewer's formative feedback (suggestions) and summative feedback (peer-grading) are credible if they are capable of passing the quiz created by the authors.

We compared our approach with two existing reputation algorithms. We found that using the quiz scores as reputations can help teaching staff to aggregate final grades and this approach provide similar or even better estimates of expert grades than the reputation algorithms. In section 4.2 we used examples to explain the reason for this finding.

We found that using quiz scores as reputations has several advantages. First, this approach is not based on “holistic reputation” hypothesis since in the assignment that we used for experiment students worked on different topics separately. We argue that, if a student peer reviewer is qualified to evaluate some artifacts, then this reviewer also has higher chances of passing the quizzes created by the authors. On the contrary, for the topics that this reviewer is not competent to review, the reviewer has a lower chance of passing the corresponding quizzes. If so, the peer reviews may not be qualified to help the authors to improve the artifacts, nor to the teaching staff aggregating the final grades. Another advantage for using student-generated quizzes in the peer-review task is that we do not need to display the reputation scores calculated from reputation algorithms to students. We have been hesitating to show those reputation scores because they come from “black boxes” from the students’ perspective. But students have no trouble understanding their own quiz scores.

Despite our success, there are still more perspectives we can investigate. Our work suggests that the “holistic reputation” hypothesis (which assumes that the credibility of a reviewer is stable) of some existing reputation algorithms may not hold for an assignment where students are working on different topics. However, it remains unclear whether this hypothesis holds for assignments where students are working on the same topic or very similar topics. Some other recent findings of our team also show that the reliability of peer grading is related to the design of peer-review rubrics [11]. It is another challenging task to establish common understandings between teaching staff and student reviewers on peer-grading standards [12]. In the future we will also try to provide students more guidance or training before they start peer-reviewing each other.

6. ACKNOWLEDGMENTS

The Peerlogic project is funded by the National Science Foundation under grants 1432347, 1431856, 1432580, 1432690, and 1431975.

7. REFERENCES

[1] C. Piech, J. Huang, Z. Chen, C. Do, A. Ng, and D. Koller, “Tuned Models of Peer Assessment in MOOCs,” *ArXiv13072579 Cs Stat*, Jul. 2013.

[2] J. A. Kulik and C.-L. C. Kulik, “Timing of Feedback and Verbal Learning,” *Rev. Educ. Res.*, vol. 58, no. 1, pp. 79–97, Mar. 1988.

[3] J. Hamer, K. T. K. Ma, and H. H. F. Kwong, “A Method of Automatic Grade Calibration in Peer Assessment,” in *Proceedings of the 7th Australasian Conference on Computing Education - Volume 42*, Darlinghurst, Australia, Australia, 2005, pp. 67–72.

[4] H. W. Lauw, E. Lim, and K. Wang, “Summarizing review scores of ‘unequal’ reviewers,” in *In Proceedings of the 2007 SIAM International Conference on Data Mining*, 2007.

[5] J. T. Dillon, “Questioning the use of questions,” *J. Educ. Psychol.*, vol. 83, no. 1, pp. 163–164, 1991.

[6] C. Gillespie, “Questions about Student-Generated Questions,” *J. Read.*, vol. 34, no. 4, pp. 250–257, Dec. 1990.

[7] S.-B. Chang, H.-M. Huang, K.-J. Tung, and T.-W. Chan, “AGQ: A Model of Student Question Generation Supported by One-on-one Educational Computing,” in *Proceedings of th 2005 Conference on Computer Support for Collaborative Learning: Learning 2005: The Next 10 Years!*, Taipei, Taiwan, 2005, pp. 28–32.

[8] P. Denny, A. Luxton-Reilly, and B. Simon, “Quality of Student Contributed Questions Using PeerWise,” in *Proceedings of the Eleventh Australasian Conference on Computing Education - Volume 95*, Darlinghurst, Australia, Australia, 2009, pp. 55–63.

[9] E. Gehringer, L. Ehresman, S. G. Conger, and P. Wagle, *Reusable Learning Objects Through Peer Review: The Expertiza Approach*. .

[10] Y. Song, Z. Hu, and E. F. Gehringer, “Pluggable reputation systems for peer review: A web-service approach,” in *IEEE Frontiers in Education Conference (FIE), 2015. 32614 2015*, 2015, pp. 1–5.

[11] Y. Song, Z. Hu, and E. F. Gehringer, “Closing the Circle: Use of Students’ Responses for Peer-Assessment Rubric Improvement,” in *Advances in Web-Based Learning -- ICWL 2015*, F. W. B. Li, R. Klamma, M. Laanpere, J. Zhang, B. F. Manjón, and R. W. H. Lau, Eds. Springer International Publishing, 2015, pp. 27–36.

[12] J. R. Wright, C. Thornton, and K. Leyton-Brown, “Mechanical TA: Partially Automated High-Stakes Peer Grading,” in *Proceedings of the 46th ACM Technical Symposium on Computer Science Education*, New York, NY, USA, 2015, pp. 96–101.