The role of initial input in reputation systems to generate accurate aggregated grades from peer assessment

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

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

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

ABSTRACT

High-quality peer assessment has many benefits. It can not only offer students a chance to learn from their peers but also help teaching staff to decide on grades of student work. There are many ways to do quality control for educational peer review, such as a calibration process, a reputation system, etc. Previous research has shown that reputation systems can help to produce more accurate aggregated grades by using peer assessors' reputations as weights when computing an average score. However, for certain kind of assignments, there still exist big gaps (larger than 10 points out of 100, on average) between expert grades (grades given by more than one expert markers) and aggregated grades (grades computed from peer assessment). In order to narrow down the gap and improve the accuracy of aggregated grades, we designed three experiments using different initial inputs (reputations) for reputation systems. These initial inputs came from calibration assignment, previous review rounds and previous assignments. Our experiments show that under certain conditions, the accuracy of aggregated grades can be significantly improved. Furthermore, for assignments not achieving the desired results, we analyzed that the reason can be the mediocre design of review rubrics and teaching staff's idiosyncratic grading style.

Keywords

Peer assessment; peer grading; educational peer review; reputation systems

1. INTRODUCTION

Peer assessment is commonly used in colleges, universities and in MOOCs. It can offer assessors a chance to learn from their peers and improve their understanding of the assignment requirements. Peer assessment can also help teaching staff to decide on grades for student work. However, in order to make the peer assessment process credible, we need a way to distinguish good peer assessors from bad ones. One solution is to use reputation systems [1]. In reputation systems, each peer assessor will have one or more reputation values. Reputation is a quantization measurement to judge how reliable each assessor is. Several reputation algorithms have already been created to calculate reputations from peer assessment grades [2, 3, 4]. Basically, each algorithm will consider one or more measurements, such as validity, reliability, spread[1], etc. The reputation can be used in different ways, such as to "… give credit to students for careful reviewing or to weight peer-assigned grades" [1]. According to previous research, reputation systems can play an important role in educational peer review systems. Moreover, using reputation algorithms to compute peer grades is indeed more effective than the naive average approach [5].

Although aggregated grades with reputations as weights can outperform naive averages, there is still much room for improvement. According to our previous research, for assignments based on writing, most of the time, the average absolute bias between expert grades and aggregated grades was less than 4 points out of 100. In this case, when teaching staff giving expert grades, they can use aggregated grades generated by reputation systems as references [5]. Since aggregated grades can be available immediately after peer assessment stage is finished, which is prior to the expert grading stage. The availability of aggregated grades can give teaching staff a general idea about the quality of each artifact based on assessors' point of views and help teaching staff to decide the expert grades. If we can narrow down the gap even further, we may be able to dispense with the expert grading of writing, and we can use aggregated grades instead. But we believe spot-checking is still necessary. However, for assignments based on both writing and programming, there are still big gaps between aggregated grades and expert grades (naive average bias larger than 10 points out of 100) even after applying reputation systems.

Hence, it is necessary to improve the accuracy of aggregated grades and reduce the burden on teaching staff. We designed three experiments and tried to answer these three questions:

1. Can reputation be taken from a formative review round to a summative review round?
2. Can reputation be taken from calibration to real assignments?
3. Can reputation be taken from one assignment to a different one?

In each experiment, we attempted a new set of data as initial input for reputation systems. In later parts of this paper, we will narrate

[1] Spread is a metric to measure the tendency of an assessor to assignment scores to different work. Generally speaking, a higher spread is better, because it indicates the peer assessor can distinguish good artifacts from bad ones.

the detail of experimental design and analyze the results of experiments.

2. REPUTATION SYSTEMS

In this paper, we focus on the performance of Hamer's and Lauw's algorithm [2, 3] since they are both iteration-based and comparable to each other.

To compute the reputation of each peer assessor, Hamer's algorithm first assigns the same weight to all of the assessors, which is 1 [2]. In each iteration, the algorithm calculates weighted average for each artifact based on peer assessors' reputations. Then Hamer's algorithm computes the difference between aggregated grade of each artifact and each peer assessment grade. The larger this difference is, the more inconsistent this peer assessor compares with others. After that, the algorithm updates the reputation of each peer assessor accordingly and calculate the aggregate grade for each artifact again till the grades converge. The steps of Lauw's algorithm are similar to Hamer's. But Lauw's algorithm applies different arithmetic formulas to calculate the differences and scale the final results.

3. DATA COLLECTION AND EXPERIMENTAL DESIGN

In this section, we provide an overview of the experimental design and validate the dataset we collected for later experiments.

3.1 Class Setting

We collected our data from two courses (CSC 517, Fall 2015 and Spring 2016 in NC State University) from Expertiza, a web-based educational peer review system [6]. For CSC 517, Fall 2015, 92 students enrolled the course (98.9% graduate level, 17.4% female). These students come from different countries but with a predominance from India (78.3% India, 9.8% China, 9.8% United States, 2.1% other countries). Moreover, most students major in Computer Science (88.9% Computer Science 3.7% Computer Engineering, 3.7% Electrical Engineering, 3.7% Computer Networking). For CSC 517, Spring 2016, 54 students enrolled the course (94.4% graduate level, 13% female). The majority of students also major in Computer Science (79.3% Computer Science 9.8% Computer Engineering, 7.6% Electrical Engineering, 3.3% Computer Networking). They come from different countries (74.1% India, 14.8% China, 9.3% United States, 1.8% other countries).

Each course contains four assignments, all of which are graded on a scale of 0 to 100. They are a Wikipedia contribution (writing a Wikipedia entry on a given topic), Program 1 (building an information management system with Ruby on Rails web application framework), OSS project (typically, refactoring an open-source software model) and the final project (adding new features to an open-source software project).

For each assignment, students have to write at least two peer assessments by completing different kinds of review rubric questions. Furthermore, they can do more peer assessments for extra credit. Several policies are proposed to avoid students playing the system. In summary, each Wikipedia contribution artifact received 9 peer assessments on average; each Program 1 artifact received 15 peer assessments on average; for OSS project, the number is 13. For final project, 20 assessors evaluated each artifact on average.

In Expertiza, there are three main types of review rubric questions, that is, choice question, text response and upload file. The choice question has two subtypes, scored question and unscored question. Criterion is scored question; conversely, dropdown, multiple-choice and checkbox are unscored questions. It is worth noting that in Expertiza, only scored questions are included in peer assessment grades. As a scored question type, criterion is the combination of dropdown and text area. It means that peer assessors can not only give a score to certain question, but also write some text comments. Moreover, criterion is one of the most frequently used question types in Expertiza.

Past research shows that the performance of Hamer's and Lauw's algorithm varies a lot on different kinds of assignments [5]. Hence, we tried to consider different assignment categories in our experimental design. According to different types of submissions, we classified five assignments (including one for calibration use, which will be mentioned in Section 4) into three categories: writing assignment, programming assignment and assignment combining writing with programming. We classified Wikipedia calibration assignment as a writing assignment because it helped students to improve their peer assessment skills on writing assignment. Wikipedia contribution is also classified as a writing assignment; Program 1 is a programming assignment; OSS project is considered as an assignment combining writing with programming. For final project, although it contains both writing section and programming section, we consider it as a writing assignment here. The reason is that students only peer assessed their peers' design documents due to the shortage of time.

3.2 Data Verification

We ran Hamer's and Lauw's algorithm on our dataset and checked whether aggregated grades with reputations as weights were more accurate than naive averages. Table I shows the comparison results between aggregated grades and naive averages. Two metrics are used to measure the accuracy of aggregated grades, namely, average absolute bias and root mean square error (RMSE). Average absolute bias indicates the average distance between aggregated grade and expert grade. RMSE is another frequently used measurement to present the differences between values. Lower average absolute bias and RMSE means better performance.

Table I shows that aggregated grades calculated by reputation algorithms perform better than naive averages in six out of eight assignments. In all these six assignments, Hamer's algorithm outperforms Lauw's algorithm, so we only used Hamer's algorithm for later experiments. Two assignments which violate our expectation are Wikipedia contribution, Spring 2015 and Program 1, Spring 2016. For Wikipedia contribution, Spring 2015, when comparing with expert grades, aggregated grades produced by Hamer's algorithm show less validity than naive averages. One potential reason is that Hamer's algorithm uses the square to amplify the differences between peer assessment grades and aggregated grades during each iteration, which makes assessors' reputations have larger variance and degrades the performance of Hamer's algorithm. For Program 1, Spring 2016, it is very likely that the instructor-defined test cases in this semester are not as elaborated as those in last semester (the average absolute bias of Program 1 in Fall 2015 is much less than those in Spring 2016).

Table I Comparison of differences among aggregated grades from Hamer's and Lauw's algorithm and naive averages

Assgt. name	Metric	Hamer's alg.	Lauw's alg.	Naive average	Assgt. name	Metric	Hamer's alg.	Lauw's alg.	Naive average
Wikipedia contrib., Fall 2015	Avg. abs. bias	4.62	3.49	3.50	Wikipedia contrib., Spring 2016	Avg. abs. bias	2.91	3.15	3.17
	RMSE	6.13	4.71	4.72		RMSE	3.61	3.89	3.94
Prog. 1, Fall 2015	Avg. abs. bias	4.32	5.58	6.21	Prog. 1, Spring 2016	Avg. abs. bias	11.46	10.77	10.59
	RMSE	5.84	7.59	8.19		RMSE	13.06	12.46	12.36
OSS project, Fall 2015	Avg. abs. bias	5.30	6.55	7.29	OSS project, Spring 2016	Avg. abs. bias	5.22	6.90	7.00
	RMSE	6.49	7.46	8.06		RMSE	6.12	8.47	8.57
Final project, Fall 2015	Avg. abs. bias	4.64	5.93	6.27	Final project, Spring 2016	Avg. abs. bias	4.65	5.91	6.07
	RMSE	7.52	8.70	9.03		RMSE	5.91	7.48	7.61

The instructor-defined test case is more like a test case in software engineering, is a set of conditions to check whether an application is working as it was originally designed [7]. The purpose of instructor-defined test cases is to help students to understand the requirements of the certain assignment and also help teaching staff to grade the students' artifacts. For instance, "Can an admin delete other admins other than himself and the preconfigured admin?" is an instructor-defined test case. This question was used in both review rubric and expert grading stage. By manually testing a series of instructor-defined test cases, teaching staff and peer assessors are able to decide the grades. Instructor-defined test cases are used a lot in Program 1. Because Program 1 is not a topic-based assignment and all students are required to build web applications with same functionalities, it is easier for the instructor to create such test cases comparing with assignments with different topics.

Hamer's algorithm is iteration-based, which means the algorithm will take several iterations before a solution (fixed point) is reached. However, results generated by these two algorithms can be locally optimal solutions, instead of a globally optimal solution [2]. It means the result of each algorithm can be optimal within a neighboring set of candidate solutions, instead of the optimal solution among all possible solutions [8]. One reason is that the initial reputation assigned to each peer assessor is always equal to 1, which mandatorily sets each peer assessor's ability the same at the very beginning.

In order to verify that different initial inputs will lead to different fixed points, we assembled a very small set of peer assessment records shown in Table II. There are four peer assessors (a, b, c, d) who assessed four artifacts (1, 2, 3, 4). To make the dataset more similar to a real scenario, we assumed that assessor b did not assess artifact 3.

We used two sets of data as initial inputs for Hamer's algorithm, whose reputation range is $[0, \infty)$. The first set of initial input is the same as the default setting of Hamer's algorithm, 1 for all assessors; the second set of initial input is arbitrarily chosen, that is, 0.2 for assessor 3, and 1 for the rest. The final reputations are shown in Table III. As you can see, we obtained two sets of data with totally different results. It is obvious that different initial inputs affect the final results.

This test shows that there are different fixed points in this dataset. If we use 1 as initial input, it will often lead to a "reasonable" fixed point, but not always [2]. Instead, if we have prior knowledge about which assessor might be credible or not, we should make use of this prior knowledge and the algorithm may converge to a more reasonable fixed point accordingly.

Thus, we tried to use different initial inputs to obtain more accurate aggregated grades. Instead of assigning a random initial reputation to each student, considering other available data, such as reputations from another review round, calibration results [9] or reputations from former assignments can be better ways to achieve more reasonable results.

Table II. Scores assigned by four peer assessments to four artifacts

Peer assessment grade		Assessor			
		a	b	c	d
Artifact	1	10	9	10	8
	2	7	6	8	7
	3	7	-	2	4
	4	6	7	3	3

Table III. Reputations with initial input equal to 1 and other values

Assessor	Rep. values with init. rep. all eq. to 1	Rep. values with init. rep. not all eq. to 1
1	0.50	2.66
2	0.77	2.67
3	2.00	0.42
4	2.59	0.79

3.3 Research Questions

This part presents three research questions. By answering these questions, we can figure out whether replacing the initial input of Hamer's algorithm from 1 to some other available data in the same course will obtain more accurate results.

3.3.1 Can reputation be taken from a formative review round to a summative review round?

Since Fall 2015, Expertiza has allowed different rubrics to be used in each round of review. For each assignment with this feature, students were encouraged to finish two rounds of peer assessments - a formative review round and a summative review round. During the formative review round, the teaching staff presented an elaborate formative rubric to peer assessors. Two questions asked in formative rubric are presented below. "Rate how logical and clear the organization is. Point out any places where you think that the organization of this article needs to be improved.". "List any related terms or concepts for which the writer failed to give adequate citations and links. Rate the helpfulness of the citations." The purpose of these questions is to encourage peer assessors to look into the artifact, point out the problems and offer insightful suggestions [10]. After one assessor submitted formative feedback, Expertiza calculated the assessment grade based on scored questions in the formative rubric.

After that, authors will have a chance to modify their work according to information given by their peers. In the summative review round, teaching staff offered a summative rubric which is designed to guide peer assessors to evaluate the overall quality of artifacts and check whether authors made the changes they suggested in the formative review round. Below are two questions used in the summative rubric. "Coverage: does the artifact cover all the important aspects that readers need to know about this topic? Are all the aspects discussed at about the same level of detail?". "Clarity: Are the sentences clear, and non-duplicative? Does the language used in this artifact simple and basic to be understood?" After assessors submitted their summative feedback, Expertiza calculated the assessment grades again for each artifact received new feedback.

We hypothesized that the assess credibility of the same assessor on formative and summative review round are related and reputations calculated from the formative review round, if used as initial input of the summative review round, can produce more accurate aggregated grades.

3.3.2 Can reputation be taken from calibration to real assignments?

At the beginning of the Spring 2016 semester, we created a Wikipedia calibration assignment before real assignments. The instructor selected several representative artifacts from former semesters. These artifacts had major differences in quality. Then the instructor submitted an expert peer assessment based on the same review rubric that students would use for each artifact. During the class, students assessed those artifacts on Expertiza. After that, Expertiza generated the report for both the instructor and students. According to the report, the instructor analyzed the results and helped students enhance their peer assessment skills.

We hypothesized that the assess credibility of the same assessor on the calibration assignment and the real assignment are related and reputations calculated from the calibration assignment, if used as initial input of the subsequent real assignment, can produce more accurate aggregated grades.

3.3.3 Can reputation be taken from one assignment to a different one?

In our dataset, there are four real assignments in a fixed order in each semester. We hypothesized that the assess credibility of the same assessor on one assignment and a subsequent one are related and reputations calculated from one assignment, if used as initial input of the subsequent assignment, can produce more accurate aggregated grades.

What's more, since we have already classified all assignments into three categories, we also assumed that using reputations from one assignment as the initial input of a subsequent assignment of the same category can also produce more accurate aggregated grades.

4. EXPERIMENTS AND ANALYSIS

According to three questions listed in the last section, we did corresponding experiments to verify derived hypotheses. Since in data verification section, Hamer's algorithm outperforms Lauw's algorithm for six out of eight assignments, we only displayed the reputation results from Hamer's algorithm in experiments.

4.1 Can reputation be taken from a formative review round to a summative review round?

Table IV shows the differences among aggregated grades, naive averages and expert grades on assignments in CSC 517, Spring 2016 by using two metrics (average absolute bias and RMSE). The reason why we chose CSC 517, Spring 2016 is because all assignments in this course support two rounds of peer assessments.

We found that using reputation results from the formative review round as initial input does not work well in all assignments. Among four assignments, two of them (Program 1, Spring 2016 and final project, Spring 2016) saw improvement by using this method. One of them (Wikipedia contribution, Spring 2016) converged to the same fixed point as using 1 as initial input. Since Hamer's algorithm is iteration-based, it is possible that different initial inputs converge to the same fixed point. The last one (OSS project, Spring 2016) fared even worse with alternative initial input. Overall, we were not able to draw the conclusion that whether initial input from formative review round is a good input option of Hamer's algorithm.

One potential reason is that according to first author's master's thesis, peer assessment records from formative review round would generate less accurate aggregated grades comparing with peer assessment records from summative review round. It is because during the formative review round, peer assessors were encouraged to offer suggestions, and authors might make changes before the summative review round. Hence, it is possible that peer assessments based on the initial version of products were not accurate. However, during the summative review round, artifacts were unchangeable, and the same version as the one teaching staff graded. Therefore, it is reasonable that when comparing with expert grades, peer assessments during the summative review round could have higher validity than those during the formative review round. And initial input from formative review round cannot help to improve the accuracy of aggregated grades.

Table IV Comparison of differences between aggregated grades from Hamer's algorithm with initial input equal to 1 and from formative review round

Assgt. name	Metric	Initial input equal to 1	Initial input from formative review round	Naive average
Wikipedia contribution, Spring 2016	Avg. abs. bias	2.91	2.91	3.17
	RMSE	3.61	3.61	3.94
Program1, Spring 2016	Avg. abs. bias	11.46	11.35	10.59
	RMSE	13.06	12.96	12.36
OSS project, Spring 2016	Avg. abs. bias	5.22	5.29	7.00
	RMSE	6.12	6.16	8.57
Final project, Spring 2016	Avg. abs. bias	4.65	4.54	6.07
	RMSE	5.91	5.77	7.61

Table V Comparison of differences between naive averages and aggregated grades from Hamer's algorithm with initial input equal to 1 and from calibration assignment

Wikipedia contribution, Spring 2016		
Different sets of aggregated grades	**Avg. abs. bias**	**RMSE**
Hamer's alg. with initial input equal to 1	2.91	3.61
Hamer's alg. with initial input from calibration	2.80	3.51
Naive averages	3.17	3.94

4.2 Can reputation be taken from calibration to real assignments?

In this experiment, we further tested whether the aggregated grades can be improved by using calibration results as initial input. Since we trialed calibration assignment for Wikipedia contribution only in Spring 2016 semester, for this hypothesis we did the experiment based on data from Wikipedia calibration, Spring 2016 and Wikipedia contribution, Spring 2016.

There was only one round of peer assessment in Wikipedia calibration, Spring 2016. And we used the formative review rubric in this assignment just the same one used in Wikipedia contribution, Spring 2016. After assessors submitted their feedback, Expertiza computed the assessment grades for representative artifacts. Then the instructor submitted expert peer assessments based on the same formative review rubric. After that, we calculated each assessor's reputation value based on their assessment grade and expert grade. When Wikipedia contribution, Spring 2016 finished, we used reputation values produced from calibration assignment as the initial input to compute a new set of reputation values. We compared this new set of reputation values with reputation values calculated based on initial reputation equal to 1.

Table V shows that both average absolute bias and RMSE are decreased by using Hamer's algorithm with calibration results as initial input. However, data used in this experiment is quite limited. If we want to further verify the efficacy of the calibration process, more data and more experiments are needed.

4.3 Can reputation be taken from one assignment to a different one?

In this experiment, we tried to test the hypothesis that using reputations from former assignments as initial input will get more accurate aggregated grades. Both course CSC 517, Fall 2015 and CSC 517, Spring 2016 have Wikipedia contribution, Program 1, OSS project and final project. And these four assignments are in fixed order. To verify this hypothesis, we designed three sub-experiments separately on these two courses. The first sub-experiment is based on Wikipedia contribution and Program 1. In this sub-experiment, we used initial input from the Wikipedia contribution assignment and peer assessment records from Program 1 to compute the aggregated grades. We compared these results with aggregated grades that based on the initial input equal to 1. The second sub-experiment was designed between Program 1 and OSS project. The same process as the first sub-experiment, we produced aggregated grades with the initial input from the Program 1 and peer assessment records from OSS project. Then we compared these aggregated grades with those grades calculated with the initial reputation equal to 1. The third sub-experiment was between OSS project and final project. The results are shown in Table VI.

Table VI shows that in Fall 2015, aggregated grades with initial input from former assignments have higher validity than those grades produced by initial input equal to 1. For Spring 2016, we found that among three sub-experiments, one of them (sub-experiment between Wikipedia contribution, Spring 2016 and Program 1, Spring 2016) converged to the same fixed point as using 1 as initial input, and another one (sub-experiment between Program 1, Spring 2016 and OSS project, Spring 2016) became

Table VI Comparison of differences between aggregated grades from Hamer's algorithm with initial input equal to 1 and from former assignments

Method	Metric	Hamer's alg.	Method	Metric	Hamer's alg.	Naive average
Wiki → Prog 1, Fall 2015	Avg. abs. bias	4.13	Prog 1, Fall 2015	Avg. abs. bias	4.32	6.21
	RMSE	5.76		RMSE	5.84	8.19
Wiki → Prog 1, Spring 2016	Avg. abs. bias	11.46	Prog 1, Spring 2016	Avg. abs. bias	11.46	10.59
	RMSE	13.06		RMSE	13.06	12.36
Prog 1 → OSS, Fall 2015	Avg. abs. bias	5.08	OSS, Fall 2015	Avg. abs. bias	5.30	7.29
	RMSE	6.31		RMSE	6.49	8.06
Prog 1 → OSS, Spring 2016	Avg. abs. bias	5.41	OSS, Spring 2016	Avg. abs. bias	5.22	7.00
	RMSE	6.36		RMSE	6.12	8.57
OSS → Final, Fall 2015	Avg. abs. bias	4.52	Final, Fall 2015	Avg. abs. bias	4.64	6.27
	RMSE	7.46		RMSE	7.52	9.03
OSS → Final, Spring 2016	Avg. abs. bias	4.55	Final, Spring 2016	Avg. abs. bias	4.65	6.07
	RMSE	5.81		RMSE	5.91	7.61

Table VII Comparison of differences among aggregated grades from Hamer's algorithm with initial input equal to 1 and from Wikipedia contribution and Program 1

OSS project, Fall 2105		
Different sets of aggregated grades	**Avg. abs. bias**	**RMSE**
Hamer's alg. with initial input equal to 1	5.30	6.49
Hamer's alg. with initial input of writing section from Wikipedia contribution and initial input of programming section from Program 1.	3.32	4.44
Naive averages	7.29	8.01
OSS project, Spring 2016		
Different sets of aggregated grades	**Avg. abs. bias**	**RMSE**
Hamer's alg. with initial input equal to 1	5.22	6.12
Hamer's alg. with initial input of writing section from Wikipedia contribution and initial input of programming section from Program 1.	4.45	5.12
Naive averages	7.00	8.57

worse by using initial input from former assignments. The last sub-experiment (between OSS project, Spring 2016 and final project, Spring 2016) supported our hypothesis. In general, initial input from former assignments obtained equal or better results than initial input equal to 1 in five out of six experiments. Therefore, we believe that initial input from former assignments can increase the accuracy of aggregated grades.

Although the new method introduced above (initial input from former assignments) can have better performance in most cases, the improvement is limited. Most of the time, it only improved by less than 0.5 points in average absolute bias. What's more, one sub-experiment (between Program 1, Spring 2016 and OSS project, Spring 2016) obtained even worse results by using this new method. One potential explanation is that Program 1 is a programming assignment, but the OSS project combines writing section with programming section. During the grading of OSS project, teaching staff gave scores to writing section and programming section separately. These two scores are related, but not always direct proportional with each other. If one team did a good job on programming and wrote the writing section perfunctorily, they would get a low score of writing section regardless of their high score on programming section. However, if another team did not accomplish the programming section well, they would not receive a high score on writing section in most of the time. Both writing and programming scores are on a scale of 0

to 100. The final score of OSS project is the combination of these two scores with corresponding weights defined by teaching staff.

In order to verify the effect of assignment categories and try to obtain more improvement, we designed a new experiment by using initial input from both the Wikipedia contribution and Program 1 acting on peer assessment records from OSS project. That's to say, the initial input of OSS project writing section came from Wikipedia contribution and the initial input of OSS project programming section came from Program 1. Furthermore, we also combined the aggregated grades of writing section and programming section with the same weights used for producing the final expert grades of OSS project. Table VII presents the experiment results in both Fall 2015 and Spring 2016. Comparing with the results produced from initial input equal to 1, average absolute bias is decreased by more than 1.3 points on average by using this new method. It is a big improvement, which indicates that assignment categories should be made into consideration in future work.

5. DISCUSSION

After three experiments, we found that average absolute biases of some assignments are still high even using our new method, which means that there still exist some obvious differences between expert grades and aggregated grades. And there must be some other issues also affecting the aggregated grades and not being considered into these algorithms, such as mediocrely-designed rubrics, insufficient peer-review training, etc.

Then we figured out that the mediocre design of review rubrics and teaching staff's idiosyncratic grading style may help to explain these high biases. For instance, OSS project, Fall 2015 is an assignment with both formative review round and summative review round. Its summative rubric has seven questions. Each question in this rubric has the same weight and Expertiza uses the naive average as the final grade, which means that each question will affect more than 14% of final grade.

One OSS artifact got 91 for the expert grade but only got approximately 75 for aggregated grade. The final comments given by teaching staff is

"Well, from the video they did the thing we expect them to do, but their tests are failing, and they should have fixed them."

In summative rubric, there is a test-related question

"IF it is an Expertiza project, check the pull request. Did the build pass in Travis CI? Was there any conflict that must be resolved? You can check those on the pull request on GitHub. Ignore this question if it is not an Expertiza project."

According to 13 valid peer-review records, most peer reviewers were able to figure out this problem. And the average score of this question is 2.16 out of 5, which means that on average more than 8 points will be deducted from total score since the code did not pass the TravisCI.

And during grading, teaching staff almost did not consider another question in this rubric. That is

"Check the commits. Was new code committed during the 2nd round?"

Since this team did not commit new code or did not commit promptly, the average of this question is 3.58 out of 5, which means on average more than 4 points will be taken off from the total score. Only these two questions have already deducted more than 12 points from the total score.

What's more, only 3 out of 31 artifacts got the grades lower than 90 and this one got 91. It is obvious that teaching staff also considered it is not a quite successful artifact. However, a relatively tolerant grade is still assigned to this team. So it can be the reason why there are large differences between expert grades and aggregated grades. A new grading method or newly-designed rubric may help to solve this problem.

6. CONCLUSIONS

In this paper, we propose several novel methods to improve the accuracy of aggregated grades generated by reputation algorithms. Since Hamer's and Lauw's algorithms are iteration-based, we tried different sets of initial inputs in order to get aggregated grades with least biases.

We designed three experiments. The first one was to use reputations from the formative review round as initial input into summative review round peer assessment records. Comparing with the initial input equal to 1, this method cannot help us to obtain aggregated grades with higher accuracy. Since after the formative peer assessment stage, authors have a chance to modify their work, which makes the peer assessments in formative review round inaccurate.

The second experiment was to use the reputations from calibration assignment as initial input. The result shows that this method can help us to get more accurate aggregated grades. However, lots of questions are needed to answer to further verify the efficacy of the calibration process. For example, when should we let assessors perform calibration process, at the beginning of the course or before each real peer assessment stage? How many calibration processes do we need, just one or one for each assignment category? What content should be included in calibration? By answering these questions, we can have a deeper understanding of calibration process and help to improve the quality of peer assessment.

The last experiment focused on initial input taken from former assignments. The results supported our hypothesis that aggregated grades calculated in this way can outperform naive averages. We also verified that under certain circumstances, considering assignment categories will improve the accuracy of aggregated grades a lot.

Our new methods can help to improve the performance, but the absolute averages biases of some assignments are still high. After looking into this, we figure out that there is still room for us to improve our review rubric to resolve ambiguity and provide more guidance to students (e.g. training or calibration). What's more, both Hamer's and Lauw's algorithm are rating-based. Some other educational peer review systems, such as Critviz[2] and Mobius SLIP[3], measure the qualities of peer assessments based on ranking. A different set of results might be found if we use ranking-based algorithms. We hope we can solve these issues, use

[2] https://critviz.com/

[3] http://www.mobiusslip.com/

different kinds of algorithms and obtain aggregated grades with even higher accuracy in the future.

7. ACKNOWLEDGMENTS

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

8. REFERENCES

[1] E. F. Gehringer, "A Survey of Methods for Improving Review Quality," in *New Horizons in Web Based Learning*, Y. Cao, T. Väljataga, J. K. T. Tang, H. Leung, and M. Laanpere, Eds. Springer International Publishing, 2014, pp. 92–97.

[2] J. Hamer, K. T. K. Ma, H. H. F. Kwong, K. T. K, M. Hugh, and H. F. Kwong, "A Method of Automatic Grade Calibration in Peer Assessment," in *of Conferences in Research and Practice in Information Technology, Australian Computer Society*, 2005, pp. 67–72.

[3] H. Lauw, E. Lim, and K. Wang, "Summarizing Review Scores of 'Unequal' Reviewers," in *Proceedings of the 2007 SIAM International Conference on Data Mining*, 0 vols., Society for Industrial and Applied Mathematics, 2007, pp. 539–544.

[4] K. Cho, C. D. Schunn, and R. W. Wilson, "Validity and Reliability of Scaffolded Peer Assessment of Writing From Instructor and Student Perspectives," *J. Educ. Psychol.*, vol. 98, no. 4, pp. 891–901, 2006.

[5] 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.

[6] E. F. Gehringer, L. M. Ehresman, S. G. Conger, and P. A. Wagle, "Work in Progress: Reusable Learning Objects Through Peer Review: The Expertiza Approach," in *Proceedings. Frontiers in Education. 36th Annual Conference*, 2006, pp. 1–2.

[7] Wikipedia. (2016). Test case. [online] Available: https://en.wikipedia.org/wiki/Test_case.

[8] Wikipedia. (2016). Local optimum. [online] Available: https://en.wikipedia.org/wiki/Local_optimum.

[9] Y. Song, E. F. Gehringer, J. Morris, J. Kid, and S. Ringleb, "Toward Better Training in Peer Assessment: Does Calibration Help?," presented at the EDM 2016, CSPRED workshop, 2016.

[10] S. Yang, Z. Hu, Y. Guo, and E. F. Gehringer, "An Experiment with Separate Formative and Summative Rubrics in Educational Peer Assessment," *IEEE Front. Educ. Conf. FIE 2016*, 2016.