An Experiment with Separate Formative and Summative Rubrics in Educational Peer Assessment

Yang Song, Zhewei Hu, Yifan Guo, Edward F. Gehringer
Department of Computer Science
NC State University
Raleigh, U.S.
{ysong8, zhu6, yguo14, efg}@ncsu.edu

***Abstract*—Educational peer assessment has proven to be a powerful approach for providing students timely feedback and allowing them to help and learn from each other. In an educational setting, most peer assessment consists of a single round. The problem with this setting is that, either the authors do not have a chance to update their work, which makes the suggestions from their peers useless, or the author can make changes after receiving the peer reviews, which forecloses using peer review to help assign grades. To address these issues, in our classes we now use two rounds of online review, with a different rubric for each. Our Expertiza peer-review system allows the evaluation rubric to vary by rounds. In the first review round, we present a formative review rubric to the peer reviewers. In the formative rubric, we try to encourage student reviewers to look into details, point out the problems they can find in the author's work, and offer insightful suggestions. After the formative review round, authors have the opportunity to submit an updated version of their artifacts. Next comes a summative peer-review round, using a summative rubric. A summative peer-review rubric focused more on evaluating the quality of the artifact by comparing it against specific benchmarks. In this paper, we discuss the design of the two-round peer-review assignments in a computer-science course and present our observations on student peer-review activity. An analysis of students' peer-assessment responses confirms the effectiveness of this design of peer-review activity.**

Keywords—educational peer review; peer assessment; student-generated content; formative assessment; summative assessment; reliability; validity

I. INTRODUCTION

Traditionally, feedback on student work is provided mainly by teaching staff—they mark up the hard copies of students' artifacts, give comments on the issues, and return the hard copies to students. This approach is generally inefficient, as teaching staff only have an opportunity to do this once for each assignment. Thus, students receive the comments and grades as final decisions. Though some formative comments might be provided, the students have no chance to improve their work on the current assignment.

Online peer-review[1] systems are now in common use in higher education. They free the teaching staff from having to provide so much feedback to each student; instead, their student peers can furnish feedback in a more formative and timely manner. Numerical grades, if obtained from the peer-review process, can also be used in computing the final expert grade for each artifact [1]. Student reviewers benefit from having a chance to see their peers' work and learning what constitutes a good piece of work at the same time [2].

Though educational peer assessment has proven to be helpful for both teaching staff and students, some details are still unclear. One of them is whether revision should be allowed, and a second round review provided. In other words, is it good practice to use a single round of review which conflates formative and summative feedback? Ideally, the first round of review should be focused on helping students to improve their work, with a later round evaluating the finished product. Doing a review in a single round means that if the reviewers give any feedback, it's based on the first version. That means either it does not help the students authors to improve their work (if they have no chance to revise their work), or it does not provide much guidance to the instructor in assigning a final grade (if revision is allowed). We are not saying that students should assign final grades, but it is often helpful for instructors to look at grades they have recommended before making their final decision.

Having both summative and formative assessments performed by peers can help students take control of their own learning [3], or improve their earning [4]. Summative review is the phase when students review each other mainly to measure success, or what they have and have not done correctly. Summative peer reviews should happen further down the learning path when the students have produced a finalized or polished version of the artifact, benefiting from feedback provided in the formative stage. By contrast, formative peer review is an instructional and peer-learning phase which should happen long before the final due date. In the formative peer-review phase, student reviewers are encouraged to provide suggestions on their peers' artifacts so

[1]"Peer review" is the term usually employed for academic research, but "peer assessment" is more common when talking about work done by students in a course. We will use the terms interchangeably.

that the authors can improve their work. In this phase, it is not necessary for students to evaluate their peers' work on any numeric scale.

To guarantee the credibility of the peer-review activity, teaching staff usually provide peer-review rubrics, ideally tailoring rubrics for each assignment. There is not much literature on the difference between formative and summative review rubrics. However, past research has shown that:

- topic-specific rubrics are more likely to produce dependable scores than unconstrained scoring or generic rubrics [5];
- analytical scoring can lower the differences in judgments by raters [6];
- shorter prose comments will be provided when a rubric criterion can be answered yes/no, or if it only requires the reviewer to verify something [7];
- the more levels used by each criterion, the less reliable the scoring will be [8, 9].

II. An Overview of "Varying-Rubric-by-Round" Peer-Review Setting

Our data set was generated by students in an Object-Oriented Design and Development course using Expertiza, which is a web-based educational peer-review system [10]. Expertiza started to support varying rubrics by round at the beginning of fall semester 2015. This feature allows instructors to create different versions of review rubrics and assign different review rubrics to each review round. In the review phase, Expertiza delivers different review rubrics, depending on what round of review the assignment is in. This feature allows instructors to provide different rubrics as guidance to trigger the expected kinds of responses.

At the start of each Expertiza assignment, students may create files, Github repositories, or wiki pages, and then submit their artifacts, as files or hyperlinks to web objects.

When doing peer reviews, students are asked to fill out a rubric (Fig. 1 is the user interface in doing peer-review on Expertiza). Before fall semester 2015, students used the same review rubric in both formative and summative rounds. Starting in Fall 2015, the "varying-rubric-by-round" feature was available, and the course staff devised both a formative and a summative review rubric for each assignment (but in some cases only formative review was done due to time constraints). To deliver peer-review feedback in a timely manner, the peer-review phase was usually only two or three days long. This helps get feedback back to students while they are still thinking about what they have submitted, and before they have moved on to another task or lost their interest in making further changes to improve the artifact.

The peer-review responses, together with the peer-grading results, are displayed to the authors. After the first round of review, the authors have a few days to make changes to their artifacts and resubmit. Those artifacts will be reviewed by the same reviewers again in the second review round. Each student was required to do two peer reviews for each assignment, but was permitted to do two to five additional reviews for extra credit [11] (depending on the semester).

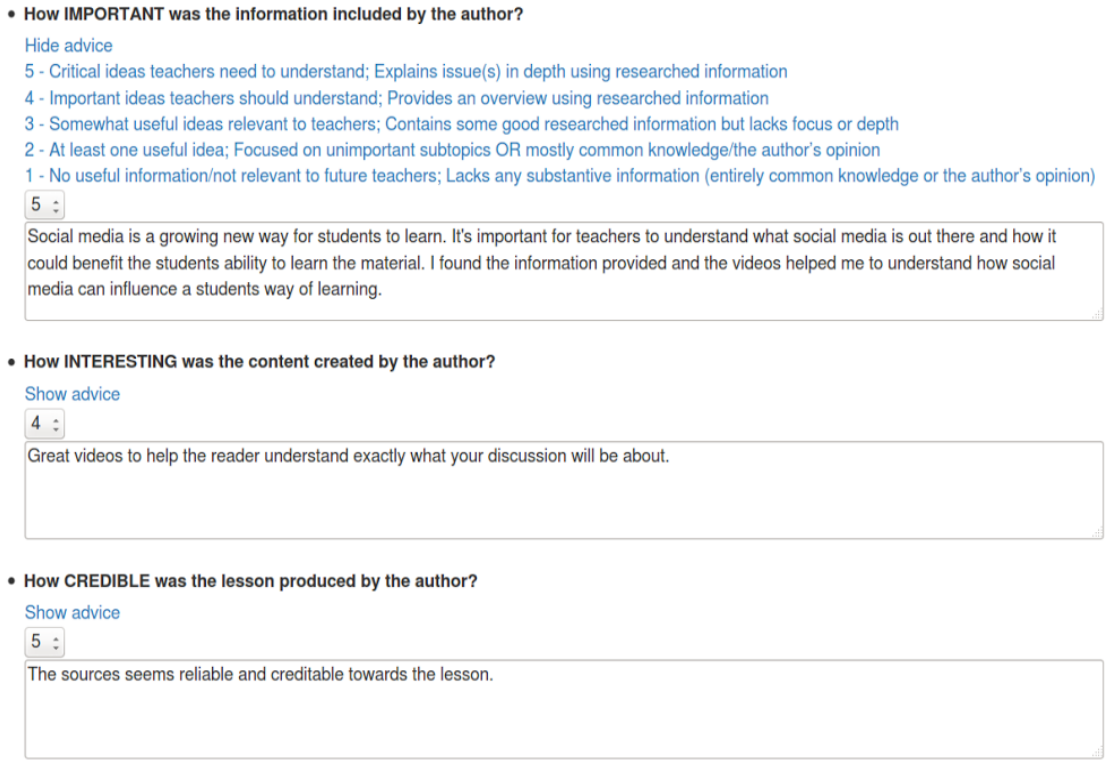

Fig. 1. Screenshot on peer-review user-interface from Expertiza

Below are a few aspects we paid attention to when we constructed the formative rubrics:

- We tried to avoid creating questions that can be answered by yes/no.
- If the question could be answered by yes/no, ask why it is good or how to improve.
- Ask for explanations and suggestions.
- Remind students of some common mistakes that might be mentioned in answering this question.
- Merge several shorter questions into a larger one, if that could be done gracefully.

Below are a few aspects we paid attention to when constructing summative rubrics (see Fig. 2 for examples):

- Criteria should cover all the important dimensions but not overlap.
- Make sure the questions can be answered in a binary way, or can be evaluated on a Likert scale.
- Ask reviewers whether the authors have made improvements based on their previous suggestions.

Sixteen assignments in four categories were used in this research. All of them were submitted to and peer-reviewed in Expertiza. Each category contains three or four similar assignments, some of which used separate formative and summative rubrics with the "varying-rubric-by-round" feature. We call them the *treatment group*. Assignments that did not use the "varying-rubric-by-round" feature comprise the *baseline case*. All of them were from a same 500- (masters) level course taught by the same instructor each semester during the past two years (Fall 2014–Spring 2016) in the College of Engineering at NC State University. This course is offered in both spring and fall semesters, but student populations vary between spring and fall semesters. Therefore, when we display the experimental results, we put the results from spring semester side by side and the results from the fall semester side by side. Table I shows the assignments in each category. Please note that Assignment 4 and 8 only used specially-designed formative review rubric (without a summative rubric), but they are still considered part of the treatment group.

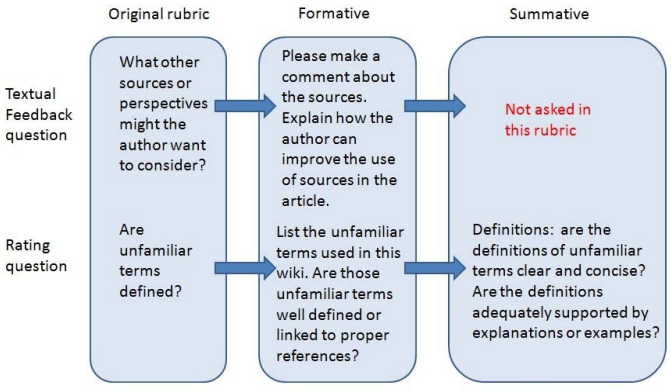

Fig. 2. Examples on devising formative/summative review questions

TABLE I. ASSIGNMENTS IN DIFFERENT CATEGORIES

Assignment Category	Treatment Group (formative)	Treatment Group(summative)	Baseline Case
Writing a wiki page	Assgt. 1 (S16) Assgt. 2 (F15)	Assgt. 1 (S16) Assgt. 2 (F15)	Assgt. 9 (S15) Assgt. 10 (F14)
Ruby on Rails app.	Assgt. 3 (S16) Assgt. 4 (F15)	Assgt. 3 (S16)	Assgt. 11 (S15) Assgt. 12 (F14)
OSS project	Assgt. 5 (S16) Assgt. 6 (F15)	Assgt. 5 (S16) Assgt. 6 (F15)	Assgt. 13 (S15) Assgt. 14 (F14)
Final project	Assgt. 7 (S16) Assgt. 8 (F15)	Assgt. 7 (S16)	Assgt. 15 (S15) Assgt. 16 (F14)

III. EXPERIMENTAL OVERVIEW

The goal of our experiments was to test the impact of separate formative and summative rounds in peer assessment. We attempted to contrast the difference in guidance, inter-rater reliability and rating validity between the classes that used the "varying-rubric-by-round" setting and the ones that did not use it.

A. Guidance

In this class, the teaching staff always encourages students to give descriptive feedback. Descriptive feedback provides student authors with information on "what they are doing well ... and gives specific input on how to reach the next step in the learning progression" [12]. In the training phase before students started to do peer-reviews, the instructor asked students to give descriptive feedback on both strong and weak artifacts: if the artifact is good, the reviewer should tell the authors where they did particularly well, so that they can continue doing this; if the artifact is not good enough, reviewers should tell the authors specifically where they went wrong and give suggestions on how to improve.

Our measures of guidance in peer-review responses are—

- Percentage of empty comment boxes. The comment boxes are the main source of formative feedback. However, we did observe that, if we used the same review rubric in both the formative and summative rounds, a large number of the comment boxes were left empty. We hoped our "varying-rubric-by-round" design would encourage students to leave fewer comment boxes empty (where the reviewer just selected a Likert rating but gave no text to explain it) than the baseline-case assignments.
- Average non-empty comment length. We also counted the words in the non-empty responses. We hoped our "varying-rubric-by-round" design would also result in a higher average non-empty comment length on the formative rubrics than the baseline-case assignments did.
- Average of number constructive feedback comments. We also tried to find how much constructive content was contained in the non-empty peer-review responses. We used the same constructive lexicon used by Hsiao and Naveed [13]. This lexicon focuses mainly on assessment, emphasis, causation, generalization, and conditional sentence patterns. We hoped our "varying-

rubric-by-round" design would trigger more constructive comments in response to the formative rubrics than the baseline-case assignments did.

- Readability. In this research, we used the Flesch-Kincaid readability index which considers both the words per sentence and syllables per word. The scale of Flesch-Kincaid readability index is usually between 0 (difficult to read) and 100 (easy to read). Conversational English is usually between 80 and 90 on the Flesch-Kincaid index. Text is considered to be hard to read (usually requiring a college education or higher) if the value is lower than 50.

B. Inter-rater reliability

Inter-rater reliability refers to the consistency of ratings across different reviews of the same artifact [7, 9]. Our measure of reliability of peer-review responses are:

- Student perspective on reliability (SPR). The SPR for one student author is the root-mean-square error between the peer-review grades received by an author and their median on this artifact. The range of SPR is $[0, \infty)$. The higher the SPR is, the lower the level of agreement that the reviewers have reached. SPR is usually higher when student reviewers are more critical, which is not a bad phenomenon, especially in the formative review phase. We hypothesize that the SPR is higher in the formative review rounds and drops in the summative review rounds.
- Instructor perspective on reliability (IPR). For each reviewer, we calculate the individual reviewer reliability as Pearson product-moment correlation between the peer-review grades assigned by this reviewer and the median of the others' on the same artifacts. We then calculate the average of individual reviewer reliability of all the peer-reviewers and use this as the IPR for this peer-review task. IPR can range between $[-1, 1]$. A positive IPR means, generally speaking, the student reviewers agree with each other about the artifacts they have reviewed. We hypothesize that the IPR on the review rubrics that we designed specially to be formative or summative should be higher than on the rubrics which are not designed with those purposes in mind.

C. Validity

Though higher IPR (which means student reviewers agree with each other) is what instructors desire, it may not always be a good phenomenon, since it could be caused by students colluding to "game" the peer assessment. Therefore, we also measured the validity of peer reviews, so that the instructor could tell whether the peer-review grades are close enough to expert grades (assigned by the course staff).

Our measures of validity of peer-review responses include:

- Student perspective on validity (SPV). The SPV for one student author is the root-mean-square error between the peer-review grades (s)he receives and the expert grade on this artifact. The range of SPV is $[0, \infty)$. The higher the SPV is, the lower the agreement between the reviewers and the teaching staff. SPV is usually higher when student reviewers interpret the peer-review rubric differently than the teaching staff, or one party does not follow the rubric to give grades. We hypothesize that the average SPV in summative rounds in treatment group assignments should be lower than the SPV in the baseline cases.
- Instructor perspective on validity (IPV). The IPV is the Pearson product-moment correlation of expert grades and the median of the peer-review grades on each artifact. The range of IPV is $[-1, 1]$. A negative IPV means that the instructors and student reviewers think about (some of) the artifacts in an opposite way (what instructors think is better, students think is worse, and vice versa), which is not what we expect to see. If the IPV value is positive, the higher the IPV is, the more correlated the peer-review grades to the expert grades. We hypothesize that, compared with the baseline-case assignments, the summative rounds of treatment-group assignments should have higher IPV.

IV. Analysis of Response Guidance

We calculated four measures for guidance to capture whether the specially designed formative and summative rubrics will bring more guidance. Those four measures are percentage of empty comments, average non-empty comment length, average constructive feedback in non-empty comments, and overall comment readability on the Flesch-Kincaid readability index. Please note we calculated average non-empty comment length and average constructive feedback based on each peer-review response to each single rubric criterion. The full result is reported in Table II.

From Table II, we found that for eight review tasks with formative rubrics, five of them received a lower percentage of empty feedback than corresponding baseline-case assignments, which means students were less likely to leave the textual feedback boxes blank. However, three formative rubrics received higher percentages of empty comments than corresponding baseline-case assignments (two of them were on Ruby on Rails application projects). After looking into the rubric, we found a possible explanation that there were more criteria for verification purpose than in other formative rubrics. Though we encountered the same issue in our earlier research [7], we still had to use verification criteria for Ruby on Rails application projects to help us verify the existence of expected features. We also found that in six assignments that used both formative and summative review rubrics, summative rounds usually had an equal or higher percentage of empty comments in the two rounds of review, which is acceptable because after receiving summative reviews, authors did not have a chance to make further changes.

We also found that of eight formative review rubrics, seven of them elicited longer textual feedback than the corresponding baseline-case assignments. This result confirms that our approach to designing formative review rubrics has been able to elicit more comments. In addition, five summative rubrics out of six also had more textual feedback,

which indicates that the reviewers also tried to justify their peer grades when they knew clearly that this was a summative round and the teaching staff might use their peer assessments to decide the grades for the artifacts.

On the constructiveness of comments, we found that 6 formative review rubrics out of 8 had increased in constructiveness from the corresponding baseline-case assignments, and there was an average of 0.10–0.27 keywords from the constructive lexicon in each response to a criterion. By comparison, research done on StackOverflow with the same lexicon reported that the median number of constructive keywords in accepted answer was 0.827 and the median length of an accepted answer was 84.47 [13]. This suggests that the peer reviews done by our students might be even more constructive. Taking the formative round of Assignment 1 for example, in every 100 words of students' peer-review comments, there were 1.18 constructive keywords (100/18.56×0.22), whereas the number in accepted answers on StackOverflow in 100 words was 0.97 (100/84.7×0.827). In addition, five summative comments out of six had more constructiveness than in the baseline-case assignments. This again shows that even in the summative review rounds, student reviewers were still willing to explain and justify their grades.

On readability, we found that six of eight formative review rubrics and five of six summative review rubrics had lower readability than in corresponding baseline-case assignments, with a readability between 51.38 and 74.37 (which is between plain English and college-level English). This was in our expected range. We used to see students giving too-short comments such as "Good job" or "Could be improved", which were readable, but not helpful to authors. The lower readability index indicates that, if a review rubric is designed explicitly to be either formative or summative, the textual feedback could be composed more formally with more complex words than the feedback to rubrics designed for both purposes.

TABLE II. COMPARISON OF GUIDANCE

Assignment Category	Treatment Group	% of Empty Comments	Avg. Non-empty Comment Length	Avg. Number of Constructive Comments	Readability	Baseline Case	% of Empty Comments	Avg. Constructive Feedback	Avg. Number of Constructive Comments	Readability
Wiki writing	Assgt. 1 (S16)-formative	5.32%	18.56	0.22	51.38	Assgt. 9 (S15)	19.46%	10.1	0.15	61.31
	Assgt. 1 (S16)-summative	8.51%	14.37	0.25	55.92					
	Assgt. 2 (F15)-formative	18.22%	13.43	0.17	51.74	Assgt. 10 (F14)	40.95%	10.96	0.14	54.24
	Assgt. 2 (F15)-summative	46.11%	8.62	0.15	57.20					
Ruby on Rails app.	Assgt. 3 (S16)-formative	35.19%	16.38	0.22	62.71	Assgt. 11 (S15)	14.09%	5.27	0.08	76.34
	Assgt. 3 (S16)-summative	10.56%	19.33	0.24	62.88					
	Assgt. 4 (F15)-formative	41.32%	8.98	0.10	65.35	Assgt. 12 (F14)	30.25%	11.09	0.11	66.36
OSS project	Assgt. 5 (S16)-formative	10.96%	19.39	0.27	69.18	Assgt. 13 (S15)	12.39%	8.01	0.13	74.01
	Assgt. 5 (S16)-summative	12.85%	15.44	0.27	74.37					
	Assgt. 6 (F15)-formative	40.45%	11.14	0.17	73.38	Assgt 14 (F14)	53.36%	6.15	0.09	65.52
	Assgt. 6 (F15)-summative	39.86%	10.05	0.21	74.14					
Final project	Assgt. 7 (S16)-formative	4.28%	17.74	0.24	66.08	Assgt. 15 (S15)	1.26%	15.6	0.35	66.78
	Assgt. 7 (S16)-summative	2.98%	21.48	0.39	63.80					
	Assgt. 8 (F15)-formative	26.02%	15.43	0.29	60.16	Assgt. 16 (F14)	29.07%	9.73	0.21	53.24

V. ANALYSIS OF PEER-REVIEW RELIABILITY

A second question that this study tries to answer is whether specially designed formative/summative rubrics can improve peer-review reliability. Though higher peer-review reliability is what teaching staff desire, it may be brought about by collusion. Theoretically speaking, if a large group of students starts to game the peer reviewing and gives each other perfect peer-review scores, the reliability will be higher since all the peer-review scores on each submission are in perfect

agreement [14]. Consequently, lower reliability, plus some clues (e.g. constructiveness) that students are taking the peer-review activity seriously may not be bad news, especially in formative review rounds.

In Table III we report the average student perspective on reliability (avg. SPR) and instructor perspective on reliability (IPR) in the dataset. IPR is the overall correlation between each peer-reviewer's rating and the rest of the reviewers. We found higher IPR in most treatment-group assignments than in corresponding baseline-case assignments. This indicates that peer-reviewers may have higher agreement with each other when they are using specially designed formative/summative review rubrics. From four out of six assignments which used both formative and summative review rubric, we also found that the reliability levels in summative review rounds were even higher than the reliability levels in formative rounds. This indicates that after the resubmission of the artifacts, reviewers had higher agreement in the second round.

TABLE III. COMPARISON OF RELIABILITY

Assignment Category	Treatment Group	IPR	Avg. SPR	Baseline Case	IPR	Avg. SPR
Wiki writing	Assgt. 1 (S16)-formative	0.42	7.77	Assgt. 9 (S15)	0.25	12.56
	Assgt. 1 (S16)-summative	0.51	5.37			
	Assgt. 2 (F15)-formative	0.35	10.26	Assgt. 10 (F14)	0.30	9.05
	Assgt. 2 (F15)-summative	0.37	5.49			
Ruby on Rails app.	Assgt. 3 (S16)-formative	0.76	13.61	Assgt. 11 (S15)	0.70	14.14
	Assgt. 3 (S16)-summative	0.63	9.28			
	Assgt. 4 (F15)-formative	0.67	19.55	Assgt. 12 (F14)	0.67	13.93
OSS project	Assgt. 5 (S16)-formative	0.41	11.71	Assgt. 13 (S15)	0.53	10.35
	Assgt. 5 (S16)-summative	0.81	8.82			
	Assgt. 6 (F15)-formative	0.57	12.42	Assgt. 14 (F14)	0.29	10.00
	Assgt. 6 (F15)-summative	0.47	10.41			
Final project	Assgt. 7 (S16)-formative	0.37	8.54	Assgt. 15 (S15)	0.14	13.21
	Assgt. 7 (S16)-summative	0.64	10.36			
	Assgt. 8 (F15)-formative	0.67	9.25	Assgt. 16 (F14)	0.34	8.55

Average student perspective on reliability (avg. SPR) is a measure of the variance of the peer grades received by student authors. Higher SPR may be salutary, since previous research showed that it takes multiple individual reviewers to cover most of the issues in an artifact [15], and each reviewer may catch different issues, thereby causing the peer-review grades to vary. In eight assignments with formative rubrics, five had higher average SPR than corresponding baseline-case assignments, indicating that reviewers for those assignments were more critical and found more issues. We also notice that there was a drop in average SPR from formative round to summative round in most of the assignments in the treatment group. This supports our assumption that, after the formative review phase and revision, the summative review grades in the second round can be more reliable.

VI. ANALYSIS OF PEER-REVIEW VALIDITY

The last analysis attempted to examine the peer-review validity of this separate formative and summative rubrics setting. In Table IV we report the average student perspective on validity (avg. SPV) and instructor perspective on validity (IPV) in the dataset. We used only the summative rounds to test the validity since they are based on the final versions of artifacts. Note that, for Assignments 4 and 8, we did not do summative review; therefore, the data was not included in validity tests. For most assignment pairs (four out of five) between treatment group and baseline cases, we found that the IPV improved after we used separate formative and summative rubrics. Before we used this strategy, the IPV varied significantly, and one observation was even negative, which means students' peer-reviews disagreed with the grades assigned by teaching staff. On Assignment 11, a Ruby on Rails application, we achieved an IPV of 0.94, which may be because it was a small class and most reviewers were accurate. However, we achieved reasonable IPV from the corresponding Assignment 3 (0.52) (also a Rails application).

Average student perspective on validity (avg. SPV) is a measure of the variance between the peer grades and expert grades received by student authors. From Table IV we found that, in the separate formative/summative review setting, the average SPV was lower than in the corresponding baseline-case assignments in most cases.

TABLE IV. COMPARISON OF VALIDITY

Assignment Category	Treatment Group	IPV	Avg. SPV	Baseline Case	IPV	Avg. SPV
Wiki writing	Assgt. 1 (S16)	0.58	6.18	Assgt. 9 (S15)	-0.31	14.06
	Assgt. 2 (F15)	0.49	7.04	Assgt. 10 (F14)	0.45	9.44
Ruby on Rails app.	Assgt. 3 (S16)	0.52	14.29	Assgt. 11 (S15)	0.94	18.71
	Assgt. 4 (F15)	-	-	Assgt. 12 (F14)	0.73	20.42
OSS project	Assgt. 5 (S16)	0.38	10.74	Assgt. 13 (S15)	0.17	11.56
	Assgt. 6 (F15)	0.49	12.58	Assgt. 14 (F14)	0.48	10.03
Final project	Assgt. 7 (S16)	-*	-*	Assgt. 15 (S15)	0.19	16.49
	Assgt. 8 (F15)	-	-	Assgt. 16 (F14)	0.34	9.69

* The grading of Assgt. 7 will be finished by early this May. Data will be available in the final version.

VII. Conclusion and Future Work

This paper reports on our efforts to apply a formative and a summative peer-review round in peer assessment in a graduate course. We analyzed the effectiveness of this separate formative and summative review rubric setting in three aspects: guidance, review reliability and review validity.

Our analysis of guidance showed that comments on the reworked review rubrics, especially the formative review rubrics, had a lower empty rate and higher volume. They also had higher constructiveness, which may improve constructive learning on both the reviewer and reviewee side. More composed and formally phrased comments were also given, which lowered the readability, but left it still between plain English and college-level English. These finding confirmed that our approaches on reworking formative review rubrics could increase the volume and helpfulness of peer-review comments

The second analysis involved instructor perspective on reliability and student perspective on reliability. We found that comparing with baseline-case assignments, the instructor perceived higher review reliability in both formative and summative rounds in most treatment-group assignments. In addition, the reliability was usually higher in summative rounds than in formative rounds. The student authors encountered more variance in peer-review grades in formative review rounds, but this variance usually dropped in summative review rounds. This means that, though in the formative review rounds, peer-reviewers could be critical and not agree with each other, the agreement level usually rose in the summative review rounds after revisions.

The third analysis investigated the instructors' and students' perspective on validity. We found that in most treatment-group assignments, the validity of peer-review grades was higher than in baseline-case assignments. This could mean two things: the formative rubrics helped authors find and fix more issues in their artifacts, and/or the summative rubrics helped reviewers do the assessment better than when they were using rubrics which mixed formative and summative questions.

In terms of future work, we hope to collect more data to derive better approaches to designing formative and summative peer-review rubrics. We are working on designing a data-sharing portal and a data repository that can collect data from multiple educational peer-review systems [16]. We believe that this big data repository with larger numbers of students at different educational levels will help us gain more knowledge about peer-review rubric design.

References

[1] 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.

[2] 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.

[3] D. J. Nicol and D. Macfarlane‐Dick, "Formative assessment and self‐regulated learning: a model and seven principles of good feedback practice," *Studies in Higher Education*, vol. 31, no. 2, pp. 199–218, Apr. 2006.

[4] R. M. O'Connell, "On the effectiveness of using midterm examinations strictly for formative feedback," in *IEEE Frontiers in Education Conference (FIE), 2015. 32614 2015*, 2015, pp. 1–3.

[5] R. J. Marzano, "A Comparison of Selected Methods of Scoring Classroom Assessments," *Applied Measurement in Education*, vol. 15, no. 3, pp. 249–267, Jan. 2002.

[6] R. Johnson and J. Penny, "The Relation Between Score Resolution Methods and Interrater Reliability: An Empirical Study of an Analytic Scoring Rubric," *Applied Measurement in Education - APPL MEAS EDUC*, vol. 13, no. 2, pp. 121–138, 2000.

[7] 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.

[8] L. Williams and J. Rink, "Chapter 5: Teacher Competency Using Observational Scoring Rubrics," *Journal of Teaching in Physical Education*, vol. 22, no. 5, pp. 552–572, Oct. 2003.

[9] A. Jonsson and G. Svingby, "The use of scoring rubrics: Reliability, validity and educational consequences," *Educational Research Review*, vol. 2, no. 2, pp. 130–144, 2007.

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

[11] E. Gehringer, Z. Hu, and Y. Song, "Five Years of Extra Credit in a Studio-Based Course: An Effort to Incentivize Socially Useful Behavior," *IEEE Frontiers in Education Conference (FIE) 2016*, 2016.

[12] C. Garrison and M. Ehringhaus, "Formative and summative assessments in the classroom," 2007.

[13] I. H. Hsiao and F. Naveed, "Identifying learning-inductive content in programming discussion forums," in *IEEE Frontiers in Education Conference (FIE), 2015. 32614 2015*, 2015, pp. 1–8.

[14] L. de Alfaro, M. Shavlovsky, and V. Polychronopoulos, "Incentives for Truthful Peer Grading," *arXiv:1604.03178 [cs]*, Apr. 2016.

[15] K. Cho, C. D. Schunn, and R. W. Wilson, "Validity and reliability of scaffolded peer assessment of writing from instructor and student perspectives," *Journal of Educational Psychology*, vol. 98, no. 4, pp. 891–901, 2006.

[16] Y. Song, F. Pramudianto, and E. Gehringer, "A Peer-Review Markup Language: One Step Toward Building a Data Warehouse for the Educational Peer-Assessment Research Community," *Submitted to Computer-Supported Peer Review in Education workshop (CSPRED 2016)*, 2016.