

# Toward Better Training in Peer Assessment: Does Calibration Help?

Yang Song, Zhewei Hu,  
Edward F. Gehringer

Department of Computer Science  
North Carolina State University  
Raleigh, NC, U.S.

{ysong8, zhu6, efg}@ncsu.edu

Julia Morris, Jennifer Kidd

Darden College of Education  
Old Dominion University  
Norfolk, VA, U.S.

{jmorr005, jkidd}@odu.edu

Stacie Ringleb

Department of Mechanical &  
Aerospace Engineering  
Old Dominion University  
Norfolk, VA, U.S.

sringleb@odu.edu

## ABSTRACT

For peer assessments to be helpful, student reviewers need to submit reviews of good quality. This requires certain training or guidance from teaching staff, lest reviewers read each other's work uncritically, and assign good scores but offer few suggestions. One approach to improving the review quality is calibration. Calibration refers to comparing students' individual reviews to a standard—usually a review done by teaching staff on the same reviewed artifact. In this paper, we categorize two modes of calibration for peer assessment and discuss our experience with both of them in a pilot study with Expertiza system.

## Keywords

Educational peer review; peer assessment; calibration.

## 1. INTRODUCTION

Writing assignments are used across the curriculum because they hone communication skills and teach critical thinking. Unfortunately, they impose a considerable grading burden since it is time consuming to give good feedback on writing. Many instructors may turn to computer-supported peer-review systems for help; indeed, reviewing writing was the motivation behind long-lived peer-assessment systems like the Daedalus Integrated Writing Environment and Calibrated Peer Review™.

In educational peer-review systems, students submit their artifacts and other students rate and/or give comments on artifacts submitted by their peers. Previous research has shown that this process benefits both reviewers and reviewees. The reviewers benefit by seeing others' work and thinking metacognitively about how they can improve their own work. The reviewees profit from receiving comments and advice from their classmates. That feedback is both more timely and more copious than feedback from teaching staff [1].

The efficacy of peer assessment depends heavily on the quality of the reviewing. Left to their own devices, students tend to examine peers' work uncritically, and make few suggestions on how to

improve it. When asked to rate it on a Likert scale, they gravitate to the upper end of the scale, making little distinction between the various artifacts that they review [2].

One approach to improving the quality of peer review is to interpose a calibration phase before the actual peer-review task. "Calibration" refers to having students evaluate sample artifacts that have already been rated by teaching staff. Then the online peer-review system can use the comparison between students' reviews and those of the teaching staff to calculate review proficiency values for students. This approach was pioneered in Calibrated Peer Review™ [3], [4] and later adopted by other systems as well (such as Coursera [5], EduPCR5.8 [6], Expertiza [7], Mechanical TA [8], Peerceptiv [9] and Peergrade.io).

## 2. TWO MODES OF CALIBRATION

We can divide calibrations into two modes. The first mode separates the calibration from actual peer-review assignments, in which students rate and on comment each other's work. We call this *stand-alone calibration*. An example is Calibrated Peer Review™. A calibrated assignment has a separate calibration phase in which students need to rate three sample artifacts, one of which is exemplary, and the other two of which have known defects. The system uses their ratings to calculate the Reviewer Competency Index, which is a measure of the student's review proficiency [3], [4]. The motivation for this mode of calibration is to train students to become proficient reviewers first before they start to review each other's artifacts. The resultant peer-review grades should have greater validity and thereby, make grading easier for the teaching staff.

The other mode of calibration combines the calibration with ordinary peer-review activity. In the peer-review phase, students review both sample artifacts and artifacts submitted by their peers. Usually, they are not aware of whether the artifact is a sample for calibration or an actual peer submission. We call this approach *mixed calibration*. An example is the Coursera system [5]. In a calibrated assignment, the teaching staff grades only a small number of artifacts, which are then used as sample artifacts in the peer-review phase. When doing peer review, each student evaluates four random artifacts and one sample artifact that has already been graded by teaching staff. Just as in stand-alone calibration, the review proficiency is determined by agreement on the sample artifacts between students and teaching staff.

Comparing these two modes of calibration, we observe that stand-alone calibration requires more work for teaching staff: they need to locate sample artifacts (which they could take from earlier semesters) and set up a calibration phase in the assignment. Students are aware of the fact that they are rating some sample

artifacts, so they may pay more attention than they do in the actual peer-review tasks, which also makes it harder to test the efficacy of the calibration. However, stand-alone calibration fits in well with in-class lecture. Instructors can give students time to do the calibration in class as training. They can also explain how the rating was done on sample artifacts so that students may have a better understanding of the rating rubrics.

Mixed calibration does not emphasize training — to make students better peer-reviewers — but score aggregation — how to identify the good reviewers and use their peer-review responses to aggregate grades for each artifact. Therefore, students who did poorly on the peer-review do not receive any pedagogical intervention, though their identities are known. So the mixed calibration is used more often by classes of massive sizes, e.g. some courses in the Coursera system.

## 2.1 Calibration in Expertiza

Beginning in 2016, the Expertiza system has included a calibration feature, which supports both stand-alone calibration and mixed calibration. In setting up an assignment, an instructor can designate an assignment as a calibrated assignment, and submit sample artifacts and “expert” reviews. The instructor can give students the right to do reviews, but not submit work. This makes the assignment a stand-alone calibration assignment. (Ordinarily, students are permitted both to submit and to review.)

The review was done in double-blind style in Expertiza. In neither calibration mode did student reviewers see the expert review before they finished reviewing an artifact. But, after a student finishes reviewing an artifact that is a calibration sample that has been reviewed by the instructor, Expertiza shows a comparison between the student’s review and the expert review (see Figure 1 for an example). No update is allowed after the expert review is displayed.

### Your calibration results for Calibration\_CSC/ECE517\_S16

Question 1 [Criterion]: List the unfamiliar terms used in this wiki. Are those unfamiliar terms well defined or linked to proper references?		
	Answer	Comment
Expert review	3	Strategies - The term is well defined but links are not provided. The links provided map to the wiki page itself. A good resource would be the OmniAuth Strategy Contribution guide - <a href="https://github.com/intridea/omniauth/wiki/Strategy-Contribution-Guide">https://github.com/intridea/omniauth/wiki/Strategy-Contribution-Guide</a> which has not been included in the list of references.
Your review	3	In the Developer Straegies section, those listed terms should be linked to their corresponding wiki pages

Question 2 [Criterion]: Rate the overall readability of the article. Explain why you give this score.		
	Answer	Comment
Expert review	3	The article is readable and even somebody very new to the subject would feel comfortable reading it. But it does not cover the topic in much depth (there are only 10 sentences in the whole article) so the reader doesn't come out with a lot of information. Also, it would be better if some of the lists were changed to prose content. The reader won't remember the items in a long list without any explanation
Your review	3	Does not give a enough information about the topic.

Figure 1. Comparison page of between student’s review and expert review

## 3. ASSIGNMENT DESIGN

Three instructors at two universities set up a total of four calibration assignments using Expertiza. Those assignments used calibration feature in Spring 2016 but did not have calibration in Fall 2015. Other than the calibration, those four assignments were of the same settings including review rubrics.

- Assignment 1: *Course:* Foundations and Introduction to Assessment of Education; *Assignment:* Grade Sample Lessons. This assignment was a precursor to engage students in evaluating peers’ writing before they assessed each other’s work. Pre-service teachers were asked to grade two different example lesson plans with a five-item rubric

by ranking the (1) importance, (2) interest, (3) credibility, (4) effectiveness, and (5) writing quality of the lesson. They were asked to consider what was effective and ineffective in each lesson based on the strengths and weakness they identified from the rubric. The artifacts were lessons created by students of prior semesters whose lessons exemplified both noteworthy achievements and pitfalls. By evaluating these two lessons, students gain valuable insight into the act of evaluating peers’ writing and are provided with a model to guide their own submissions. The students’ completed the calibration assignment, ranking each of the rubric categories on a 1-5 scale. Their results were then compared with the “expert” review completed by the course instructor.

- Assignment 2: *Course:* Project Design and Management I; *Assignment:* Practice Introduction to Peer Review. This assignment was designed to expose students to writing an introduction for their senior project, to orient them to the peer review process, and to understand the instructor’s expectations for the peer review assignment. The calibration exercise had the students peer review two introductions from a previous class, one with a good grade and one that received a poor grade. The calibration exercise was performed before the introduction was drafted. The general introduction assignment included a draft with an in class peer review, a second draft peer review using Expertiza and the submission of a final draft.

- Assignment 3: *Course:* Object-Oriented Design and Development; *Assignment:* Calibration for reviewing Wikipedia pages. This assignment was to get the students ready to write and peer-review Wikipedia entries. The instructor provided a list of topics on recent software-development techniques, frameworks, and products. Some of these topics had pre-existing Wikipedia pages; some did not. Where the pages existed, they were stubs or otherwise in need of improvement. Students could choose one topic and create the corresponding page. Then students were required to review at least two others’ artifacts and provide both textual feedback and ratings.

We created a separate assignment for calibration. The sample artifacts were chosen from a previous semester. The instructor took two reviews done by good reviewers and made further changes in an effort to make the review of exemplary quality.

- Assignment 4: *Course:* Object-Oriented Design and Development; *Assignment:* create and review CRC (Class-responsibility-collaborator) cards. CRC cards are an approach to designing object-oriented software. The instructor’s students tended to make the same mistakes, semester after semester. The goals of this calibration assignment were to (1) allow students to submit their own CRC-card design and (2) review some CRC-card designs that contained common mistakes. In this assignment, each student reviewed one of their peers’ designs, and two designs arranged by the instructor to contain common mistakes. These designs were created by merging the errors made by previous students on an exam.

Unlike the other three calibration assignments, this assignment did not precede another assignment where the students submitted their own work. Rather, it was done as practice for the next exam.

We asked the instructors to identify a few good reviewers in the actual peer-review assignments of exemplary quality to compare the student performance on the calibration assignment and the actual assignments for which they received training. To test student performance on different assignments, we used the metrics below:

- Percentage of exact agreement on each criterion. All the rubrics used in our experiments were scored on either a 0-to-5 or a 1-to-5 scale. On each criterion, exact agreement was when instructor and student gave exactly the same score.
- Percentage of adjacent agreement on each criterion. On each criterion, adjacent agreement means that the score assigned by the student is within  $\pm 1$  of the instructor's score.
- Percentage of empty comment boxes. Some criteria asked students to give both a score and textual feedback. In the calibration, the instructors tried to give textual feedback on all these criteria. If the sample artifact was in good shape, the instructors commented why it was good; otherwise, if the sample artifact needed improvement, the instructors suggested changes for the author to consider. We hoped this would encourage students to comment on more of the criteria.
- Average non-empty comment length. We counted the words in the non-empty responses. In calibration, the expert reviews were usually longer than the average of students' review (see Figure 1 for example).
- Average of number constructive comments. We tried to measure how much constructive content was provided in the non-empty responses. We used the same constructive lexicon used by Hsiao and Naveed [10], [11]. This lexicon focuses mainly on assessment, emphasis, causation, generalization, and conditional sentence patterns.
- Readability. We used the Flesch-Kincaid readability index [12], which considers the length of sentences and the length

of words. The Flesch-Kincaid readability index rates work between 0 (difficult to read) and 100 (easy to read). Conversational English is usually between 80 and 90 on this index. Text is considered to be hard to read (usually requiring a college education or higher) if the index is lower than 50.

## 4. HOW CALIBRATION AFFECTS STUDENT PERFORMANCE

### 4.1 Results for stand-alone calibration

The first three calibration assignments (Assignment 1, 2 and 3) were followed by an actual assignment where the students carried out the same kind of review on which they were calibrated. We measured the percentage of empty comments, average comment length, and number of constructive comments in the response to each criterion, and the overall readability. In the following actual assignment, we also measured the students' agreement on exemplary reviews (done by students). The results are shown in Table 1.

In all three classes, we found there was a similar amount of exact agreement on calibration assignments and following assignment. But we observed increases in the adjacent agreement on the following assignment. The reason for that could be that the calibration phase led students to become more skilled and more polite as reviewers. The instructor of assignment 1 observed that her students were critical or even bullying, in their peer reviews at the very beginning of the semester. In the calibration phase, students were able to see how the instructor reacted to various issues and what the instructor grades were. This gave students guidance on how to rate artifacts that still needed improvement.

We also noted that the percentage of empty comments dropped between the calibration assignment and the assignment right after, indicating students were more willing to give comments after the calibration. Relative to the previous semester, two of the three classes had a lower empty-comment percentage on corresponding assignments.

**Table 1. Metrics for calibration assignments, the assignments following the calibration assignment, and the corresponding actual assignment in the previous semester**

	Assignment	Exact agreement %	Adjacent agreement %	Empty comment %	Avg. non-empty comment length	Avg. number of constructive comments	Readability
Calibration assignment	Assgt. 1	53.20%	83.80%	31.80%	17.4	0.35	58.9
	Assgt. 2	21.60%	32.10%	17.40%	22.1	0.31	49.8
	Assgt. 3	45.90%	85.80%	11.20%	18	0.27	54.4
Assignment right after the calibration assignment	Assgt. 1	48.00%	86.70%	26.80%	21.8	0.44	63.2
	Assgt. 2	26.70%	61.70%	13.20%	21.2	0.35	50.8
	Assgt. 3	49.10%	92.00%	8.50%	14.4	0.25	55.9
Corresponding actual assignment from former semester	Assgt. 1	N/A	N/A	20.80%	18.3	0.36	62.6
	Assgt. 2	N/A	N/A	15.10%	28	0.48	51.5
	Assgt. 3	N/A	N/A	46.10%	8.6	0.14	57.2

The comment length between the calibration assignment and the following assignment were almost the same. Two out of three classes had a higher average comment length after they did calibration, compared with corresponding assignments last semester.

From the amount of constructive content per response to each criterion, we found that the students tended to give as many or more constructive comments in the peer-review after the calibration. Two out of three classes made more constructive comments after calibration compared with corresponding assignments last semester.

In this study, we found that students tended to write more complicated sentences in calibration tasks, but in the assignments right after the calibration, their comments were a little easier to read but close to college level, which was acceptable to instructors.

## 4.2 Results for mixed calibration

Assignment 4 was our only experiment with the mixed calibration mode: each student reviewed two calibration submissions and one submission from their classmates. Unlike Assignments 1–3, which aimed to train students to become better reviewers on the actual peer assessment, Assignment 4 was not followed with an “actual” assignment on the same topic. Instead, Assignment 4 was designed to give students the opportunity to see common mistakes that others had made on a certain kind of question (on CRC-card design) on exams in earlier semesters.

On Assignment 4, the percentage of exact agreement was 52.2% and percentage of adjacent agreement was 91.3%, which were both very high. This was partially due to a review rubric that asked students to count the number of errors of certain types (e.g. the number of class names that are not singular nouns), instead of ordinary rubric criteria that ask students to rate the artifact on some aspect (e.g., the language usage of an article). This rubric design reduces ambiguity and thereby increased the agreements.

The percentage of the empty comment was 77.0%, the average of non-empty comment length was 5.4 and average of number constructive comments was 0.13, which are all lower than Assignment 1-3. The ostensible reason was that the review rubric was not designed to encourage students to give textual comments, but simply to count the errors. The review readability index was 60.1, which indicates that for those reviewers who gave textual feedback, the feedback was not short and simple as we expected.

We hypothesized that after this calibration, student's' average score on related questions on the exam would be higher. We compared the student performance on CRC-card related questions in exams of this semester (with calibration as training) and last semester (without training). However, we found that the students' average grade was 85.3% on those questions in this semester, and 85.4% on last semester. We did not find any significant change between this semester and last semester. Upon seeing those results, we surmised this calibration assignment was done several weeks before the next exam, and, without follow-up practice, students forgot the training they received.

## 5. WHAT SAMPLE ARTIFACTS WE SHOULD USE FOR CALIBRATION?

After students finish the calibration, the instructor can see the calibration reports for each artifact, as shown in Figure 2. Each table shows the students' grades on each question on a sample

artifact. The green color highlights the expert grade, and the bolded number was the plurality of students' grades.

Figure 2 shows a sample artifact where the calibration was quite successful, with exact agreement of more than 40% and adjacent agreement of almost 80%. However, it is still not clear that if it was related to the quality of the artifact. When we calculate the percentages of agreements for each sample artifacts, we found that the level of agreement is related to the quality of the artifact: the higher grade that a sample had, the higher agreement that students might achieve. This raises another question: what kind of artifacts work better as samples in calibration?

### Calibration 1

Question1: How IMPORTANT was the information included by the author?

Assigned Score	1	2	3	4	5
% of students	0.0%	2.17%	28.26%	<b>35.87%</b>	32.61%

Question2: How INTERESTING was the content created by the author?

Assigned Score	1	2	3	4	5
% of students	1.09%	5.43%	16.3%	35.87%	<b>40.22%</b>

Question3: How CREDIBLE was the lesson produced by the author?

Assigned Score	1	2	3	4	5
% of students	0.0%	4.35%	13.04%	31.52%	<b>50.0%</b>

Figure 2. A calibration report on Expertiza system

We put the percentages of agreement and grades for the artifacts together to compare the relationship between the agreement and the grades that the sample artifacts received. We used both the sample artifacts and the artifacts reviewed by the exemplary reviewers. The distribution and fit line are shown below.

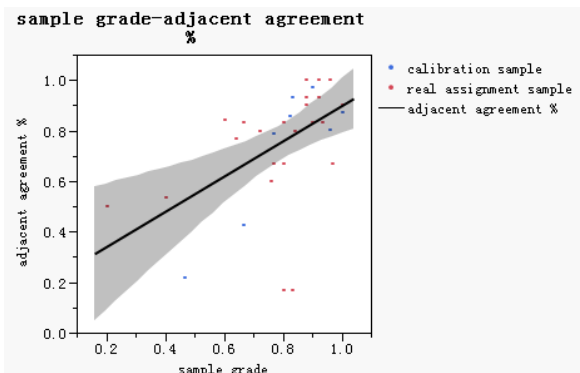


Figure 3. Relationship between adjacent agreement percentage and sample grade

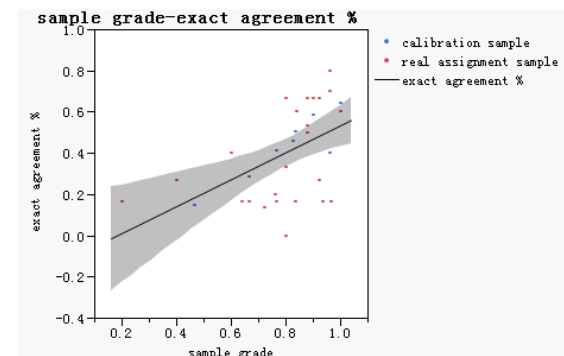


Figure 4. Relationship between the exact agreement percentage and sample grade

We find that the samples that received higher grades usually have higher levels of agreement (on both exact agreement and adjacent agreement). The lower quality a sample is, the lower agreement we observed between teaching staff and students.

We looked into the samples used in each assignment, and we found that usually it is harder for students to make the same judgment as teaching staff on an artifact of low quality. There could be multiple reasons. The first reason is that teaching staff has seen more artifacts, therefore they know the distribution of the quality of the artifacts and thereby they made better judgments. For student reviewers, they may be able to tell an artifact is of low quality based on one criterion, but they could be more critical than warranted since they have not seen even worse examples. From this perspective, it is important for instructors to use at least one or two low-quality sample artifact as a sample artifact to show students how to rate poor work.

Another factor that may lower the agreement between teaching staff and students is the reliability of the criterion: some of the criteria are not specific enough for the reviewers to make reliable judgments [2]. E.g. the criterion, “(On Likert scale) does the author provide enough examples in this article?” is not reliable, since “enough” is not well defined. To improve review rubrics, instructors can create “advice” for each level (sometimes known as an “anchored scale”). For example, “ $\frac{1}{5}$  - No example provided”, etc. From this perspective, the calibration can also be used to test the instructor’s review rubric.

## 6. CONCLUSION

In this paper, we have described our experience with the calibration in peer assessment in Expertiza. We first introduced two modes of calibration that have been used in online peer assessment systems, which are stand-alone calibration and mixed calibration. Stand-alone calibration trains students to become better reviewers, while mixed calibration finds credible reviewers in the course of performing peer assessment. We also discussed the pedagogical scenario in which each mode is suitable.

We calculated the agreement between students’ rating and teaching staff’s rating on the sample artifacts. We found that students in our assignments, on average agreed exactly with teaching staff on more than 40% of ratings. This means that on more than 40% of the ratings done by students during calibration gave exactly the same scores given by teaching staff. In addition, more than 70% of the ratings done by students gave the score within the  $\pm 1$  range to the scores given by teaching staff. To test if students still perform as well on the actual peer assessment after training, we asked the teaching staff to identify some good reviewers in each course. Using their reviews as exemplars, we found that, in the actual peer assessment phases, the agreement was similar to that on the calibration assignments, sometime even a little higher.

We compared the volume of textual feedback from the semester with calibration and the previous semester without calibration. We found that after calibration, students tend to give more extensive textual feedback, fill in more text boxes with comments, and give more constructive feedback.

We also found that the level of rating agreement between students and teaching staff is related to the quality of the artifact; namely students tended to agree less with teaching staff on artifacts of low quality. To improve agreement, we suggested: (1) on the calibration, an instructor can use both median-quality artifacts and

low-quality artifacts as samples and (2) the instructor can provide “advice” for each level of each criterion.

One future study we are interested in is to calibrate the textual feedback. In this paper, we have only calibrated the numerical scores. It is possible that both a student and the teaching staff gave a  $\frac{4}{5}$  on one criterion on a sample artifact, but may not see the same issue. This kind of agreement can only be measured by calibration of textual feedback.

## 7. REFERENCES

- [1] E. F. Gehringer, “A Survey of Methods for Improving Review Quality,” in *New Horizons in Web Based Learning*, Y. Cao, T. Våljataga, J. K. T. Tang, H. Leung, and M. Laanpere, Eds. Springer International Publishing, 2014, pp. 92–97.
- [2] 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.
- [3] R. Robinson, “Calibrated Peer Review™,” *Am. Biol. Teach.*, vol. 63, no. 7, pp. 474–480, Sep. 2001.
- [4] A. Russell, “Calibrated peer review-a writing and critical-thinking instructional tool,” in *Teaching Tips: Innovations in Undergraduate Science Instruction*, 2004, p. 54.
- [5] 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.
- [6] Y. Wang, Y. Jiang, M. Chen, and X. Hao, “E-learning-oriented incentive strategy: Taking EduPCR system as an example,” *World Trans. Eng. Technol. Educ.*, vol. 11, no. 3, pp. 174–179, Nov. 2013.
- [7] E. Gehringer, “Expertiza: information management for collaborative learning,” *Monit. Assess. Online Collab. Environ. Emergent Comput. Technol. E-Learn. Support*, pp. 143–159, 2009.
- [8] 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.
- [9] C. Schunn, A. Godley, and S. DeMartino, “The Reliability and Validity of Peer Review of Writing in High School AP English Classes,” *J. Adolesc. Adult Lit.*, p. n/a-n/a, Apr. 2016.
- [10] 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.
- [11] Y. Song, Z. Hu, Y. Guo, and E. Gehringer, “An Experiment with Separate Formative and Summative Rubrics in Educational Peer Assessment,” in *Submitted to IEEE Frontiers in Education Conference (FIE), 2016*, 2016.
- [12] J. P. Kincaid and A. Others, “Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel,” Feb. 1975.