Online Student Peer Reviews
William J. Wolfe
Professor of Computer Science
California State University Channel Islands
One University Drive
Camarillo, CA 93012
805-437-8985
william.wolfe@csuci.edu

ABSTRACT
This paper describes an online method of implementing student peer reviews. The results of applying the method to several courses is also discussed. Each week students accessed each others' assignments, as posted on individual web pages, and submitted scores and comments on the course web site. The course web site did all the bookkeeping, giving students instant access to the reviews they received. Statistics for a particular software engineering class of 34 students are presented.

Categories and Subject Descriptors
K.3.1 [Computers in Education]: Using an online system to assist students in the learning process.

General Terms
Measurement, Documentation, Experimentation, Human Factors.

Keywords
Online, Peer Review, Web, Computer Based Instruction.

1. INTRODUCTION
Peer reviews are very common in college classrooms ([3],[4],[5],[6],[7],[9]). Exchanging papers with the student next to you is a primitive form of it, and many English Composition classes have students break into groups and evaluate each other's writing [2]. But, without technology, it would not be possible to implement a complete set of peer reviews once a week in a class of, say, 30 students. To do so for one assignment would require making 900 copies (of possibly multi-page documents), distributing them, and then keeping track of the results as they were returned. Using web technology such a process can be envisioned, not just for one assignment but for weekly assignments.

I implemented a Web based system of peer review in several classes. The system is based on weekly assignments. When students are working on assignment N they are reviewing assignment N-1. After a student completes an assignment they post it on their web site. They submit the URL to the course web site (my own creation) which keeps track of student names, their URLs, and the reviews they have given and received. The course web site also has the syllabus, lecture material, quizzes, and assignments. To do a review the student logs on to the course web site, accesses the list of URLs and reviews the posted assignment. Then the student submits a score (1-10) and a comment. This score and comment are immediately available to the receiving student. The process is anonymous in that the receiving student does not know who submitted the review, but the reviewer knows who they are reviewing.

I have implemented the peer review process in upper division and first year graduate classes, and primarily in computer science classes.

2. RESULTS
The process worked exceedingly well. The peer reviews dominated the course activity. The peer review process supported the teacher in the roles of “coach” and “resource” as opposed to “lecturer” and “enforcer”.

In summary, the reasons for success are:
1. Students learned from their peers.
2. Students received more feedback than any one teacher could provide.
3. Students received almost instantaneous feedback.
4. Students got to know their peers as professional colleagues.
5. Students felt comfortable in the role of critical reviewer.

The students liked reviewing each other, getting fast feedback, knowing where they stood with respect to their peers, and getting to know each other. The teacher also found the process very rewarding.

3. ADVANTAGES OF PEER REVIEWS
With peer reviews:
1. Students see things from the teacher's perspective (In particular, they get to see all the other students' work);
2. Students exercise and refine their ability to be critical reviewers, often gaining a better understanding of the grading process and making it easier for them to accept criticism [1];

3. Students interact with each other in ways that help to build community, and to view their classmates as professional colleagues;

4. Students get quick and plentiful feedback (i.e.: removes the bottleneck of waiting for the teacher to grade all the papers);

5. Students get to see the distribution of the quality of the work submitted by their classmates, which, among other things, helps them understand what's required of them;

6. Students with superior knowledge of the subject can help their classmates;

7. The teacher can take on a supervisory role, and pay more attention to designing the assignments and keeping the course moving in the right direction.

There is yet another effect that I noticed. Students appeared to be working a lot harder to impress their classmates, possibly because they had a lot of respect for their classmates or simply because there were more of them (i.e.: 25 classmates to impress rather than a single teacher), I am not sure. However, in [8] there is a description of an English teacher who made it a point to read student papers aloud, and who described the same phenomenon: students did better work when they knew their work would be made public. Reciprocally, the poorer performing students seemed to accept the feedback from their peers even if it was sometimes harsh, and to view their teacher as a resource who could help them. This makes for a much nicer teacher-student relationship with more teaching, and less judging.

2. PROBLEMS

2.1 Harsh Language

Although the experience was hugely positive, there were a few problems. For example, there was an occasional problem with the wording submitted in some reviews. It was rare, but over 4 semesters and thousands of reviews, there were 2 or 3 times that a reviewer chose words that were deemed offensive. When confronted, the reviewer immediately apologized and all was well (at least I did not hear any more about it). For example, one reviewer wrote "This page sucks". When these events happened I contacted the reviewer and corrected the problem, and then sent out an email to the whole class thanking them for all the hard work they were putting into the reviews and tactfully reminding them of the difference between constructive/professional reviews and useless/insulting reviews. Overall, this was not a big problem, in fact I am amazed at how small a problem it was, but I would caution anyone who implements peer reviews to be very aggressive in checking the comments and responding to complaints. The offended students were placated, I think, because the problem was immediately acknowledged and addressed. I hesitate to think of what might have happened had these sentiments been allowed to fester over a 16 week period.

With all the comments stored in a database it was easy to select a student and review all his or her comments. I found myself forming mental images of the students as I read their individual comments. And, it caused me to reflect on my own grading, and how my comments might be perceived. It was as if I had attended a workshop for teachers, and we all graded some papers and then shared our comments and reflected on how we approached the task, our choice of words, etc.

2.2 Potential for Cheating

With assignments posted on publicly accessible web pages there was an obvious potential for cheating. However, there was a natural force working against copying homework: there were a lot of "eyes" on these assignments. That is, a small similarity that might get past the teacher was unlikely to get past all the reviewers. There were 2 cases of outright plagiarism in the 2 years of implementation. In both cases the plagiarism was detected by other students and reported to me before I noticed it, and the cheaters dropped the class the moment I confronted them. Despite this, students were told they could use information or knowledge gained from reviews to modify their own work, but to give appropriate credit to the source. This worked fine as several students pointed out things they learned from other students and made formal references to other students, often thanking them for their help. It was gratifying to see the high degree of respect they had for each other. It was also clear that the peer reviews provided a "mini-laboratory" for developing ethical practices, an unintended but valuable side effect.

2.3 Late Assignments

A potential problem in any class is late or missing work. One way to motivate a student to be on time is to threaten a reduced grade, but this puts the teacher in the role of "enforcer", which can lead to a very negative student-teacher relationship, especially if the student knows the subject matter fairly well and feels that participation is a waste of time. The peer reviews addressed this issue by putting students on notice that not only does the teacher expect timely work but so does the rest of the class. A late assignment will get several scores of "1" submitted, as opposed to just the teacher's feedback, thereby distributing some to the burden of enforcement.

2.4 Too Much Work

In most classes (e.g.: more than 15 students) it was too much work for the students to do a meaningful review every other student each week (although that's what a teacher is expected to do -- but we get paid to do it!). It depends a lot on the particular assignment but I found that the students could do about 10 reviews a week without complaint. I told students that 10 reviews was enough, and that they could do as many reviews as they wished, but not to sacrifice quality for quantity. It was not hard to evaluate the quality of a review. Roughly 70% of the reviews appeared to be superficial (such as: "9, nice work"), but many reviews were reasonably detailed and quite useful, I would think, to the receiver.
I noticed that some students did many more reviews than were required. I guess those students either saw the process as fruitful (i.e.: helped resolve technical issues), or were very curious.

To make sure that each student got roughly the same number of reviews, the web site indicated how many reviews were already submitted for each student. The reviewers were directed to review the student with the fewest reviews first. Although there was nothing on the web site that enforced this rule, it was obeyed most of the time.

2.5 The Need for Computer Literacy

At the heart of this peer review method is the ability of students to post their assignments on web pages. This requires a reasonable knowledge of web servers, HTML, and FTP. This was not a problem for most computer science majors but for other majors this can be a significant hurdle. For example, I tried the system in an upper division mathematics class [9] and was slowed down by two things: 1. mathematics assignments have a lot of special characters and figures that make it difficult to construct a web page, and 2. mathematics students are not necessarily adept at web technology. To overcome these hurdles I held special sessions to teach them the basics of web technology and demonstrated the use of mathematical symbols and figures in web documents. Most of the students were excited to learn about web technology but I was surprised that there were a few hard core resisters. They seemed to see this as extra drudgery, taking precious moments away from their first love, mathematics.

In addition to Computer Science and Mathematics classes, I also tried the system out in a Marketing class (I collaborated with CB Claiborne, Professor of Marketing, CSUCI). In this case the students broke into teams with at least one web master (adept at web technology) on each team. I had to modify my course web site to account for teams but it was well worth the effort. The students were very excited to post a marketing plan, with flashy graphics, on a web site and compete with the other teams for higher review scores. In this case the reviews were done only 3 times, corresponding to 3 phases of building a marketing plan for a fictitious company.

Finally, the peer review system was tried in a sophomore Psychology class (I collaborated with Beatrice de Oca, Professor of Psychology, CSUCI). In this case we did not expect students to post their own web pages, but instead we asked them to answer questions online. That is, they did their assignment (which was a technical reading) and then logged on to the web site and answered 5 questions about the assignment (i.e.: entered text into a web form and hit submit). After the assignment was due, the students could then log on, read answers submitted by other students, and then enter a peer review by submitting a score and comments. The system seemed to work quite well but we only did it for one assignment so it was difficult to draw too many conclusions. Using the database, it was interesting to be able to list out all the responses (from 40 students) to the same question, and to list out all the comments submitted by a student. That is, the fact that all student responses were recorded in a database gave us the ability to evaluate student performance in many ways, especially after the semester was over.

3. STATISTICAL RESULTS

Despite the problems mentioned above I believe the system was a great success. In one software engineering class (34 students) there were 5,212 reviews in the system after 15 assignments (the maximum possible number of reviews was 34x33x15= 16,830). The average score received by each student is shown in chart 1. Chart 1 shows that the top 16 ranking students had an average review score greater than 8.00. After that (moving to the right), student performance drops off steadily. The steep decline is traceable to having one or more incomplete assignment (and there were quality problems as well, but the dominant effect was missing work). The students in the upper half of the class did nearly all the assignments on time, whereas students in the lower half had varying degrees of missing and incomplete assignments. The steady drop off in student performance was almost certainly caused by the demanding schedule.

Chart 1: The average score received by each student.

Having weekly assignments, with peer reviews, proved to be to be an arduous schedule. I compensated for this by allowing students more time to post their solutions, helping with technical content, and giving full credit to many late assignments. So, some of the peer review scores do not reflect the official grades (for example, there were only 2 F’s). Half the class proved they could meet the schedule, but in the future I will scale things back a tad.

With over 5000 reviews in the system a "large-number" effect seemed to emerge. From Chart 1 it is easy to see that the student with rank=1 had an average score just above 9.0 while the next highest score was about 8.8. This seemingly small difference of 2 tenths of a point appeared to be completely justified as I reviewed all the work from those 2 students. The amazing thing, however, was that even differences of a tenth of a point appeared to be significant. That is, I actually started to believe that fractions of a point separating students was significant. This seemed to be true at each step as I walked down the chart. That is, even the smallest distinctions appeared accurate based on my own evaluation of their work. Based on this data it is easy to hypothesize that a large number of novice graders, acting over many assignments, can produce a relatively accurate assessment.
Chart 2 shows the number of reviews received by each student (keeping them in the same order, or ranking, as in Chart 1). It shows that the lower ranking students tended to get slightly fewer reviews. This was primarily due to late or missing assignments (students resisted submitting a review for a missing assignment, lacking a "killer" instinct I suppose), but I think it was also due to the fact that there was more "reward" for reviewing the better papers (and the source of the better papers became more apparent as the semester rolled on).

Chart 3 shows the average review score given by each student. It shows no obvious trend in the harshness or leniency of the reviewers. That is, the higher ranking students tended to be just as harsh or lenient as the lower ranking students. It also shows that 3 of the 34 students did not submit any reviews.

Chart 4 shows the number of reviews given by each student. It shows that the higher ranking students embraced the process while most of the lower ranking students did not participate as much. It shows that most of the 5000+ reviews were given by the top third of the class. There is an obvious positive correlation between participation in the peer review process and good performance in the class.

Chart 5 shows the distribution of scores submitted over the 15 assignments. It shows a large spike for the score of "1". This tended to skew the statistics but I can think of no better way to discourage late assignments. Also, notice that there is a dip for the score of "10". The students were instructed to give only one 10 for a given assignment. They were told to give a score of 9 to a very good assignment, but to reserve the score of 10 for the very best that they had seen. My hope was to expand the scale a little bit and compensate for the compression at 10. Again, this was not enforced by the web site, but most students obeyed this rule. The result was that there were many fewer 10's than 9's, and students appeared to be singling out especially good work for the score of 10.

4. CONCLUSIONS
Future work should focus on more refined statistical analysis of the results, a more formal experimental design based on randomization of the assignment of peer reviewers with peers to be reviewed, as well as the anonymization of all relationships, and a more careful design of each assignment. I feel there is enough evidence to support a few hypotheses, such as a strong
correlation between participation and performance, a strong incentive to do better work, and a high level of student satisfaction with the process (a student survey was conducted, and described in [9], that supports these hypotheses). Also, teacher-centered hypotheses should not be ignored, such as a high level of teacher satisfaction derived from being separated from, or supported in, the grading process. Halls of Yearning [10] pointed out in 1971 that the roles of “teacher” and “judge” are mutually exclusive. The academic system has essentially fused these two roles, emphasizing judging over teaching. The peer review system I have implemented partially restores the teacher-student relationship, de-emphasizing the judging aspect.

5. ACKNOWLEDGMENTS

Thanks to Professor Carol Holder for her help and support during the development of this peer review process and for finding many references to similar systems, and to Professors Harley Baker, Beatrice De Oca and CB Claiborne for their help. Thanks also to the anonymous ACM SIGITE reviewers for their helpful comments and to Michael Cook for helping with a major reorganization of the paper (and for noting the connection to "Halls of Yearning"), and to Cheryl Dwyer Wolfe for her assistance in reviewing this paper.

6. REFERENCES


