

STRATEGIES AND MECHANISMS FOR ELECTRONIC PEER REVIEW

Edward F. Gehringer¹

Abstract — We have implemented a peer-grading system for review of student assignments over the World-Wide Web and used it in approximately eight courses. Students prepare their assignments and submit them to our Peer Grader (PG) system. Other students are then assigned to review and grade the assignments. The system allows authors and reviewers to communicate with authors being able to update their submissions. Unique features of our approach include the ability to submit arbitrary sets of Web pages for review, and mechanisms for encouraging careful review of submissions. Electronic peer review facilitates collaborative learning in several ways. First, there is the obvious fact that students can learn from their reviewers' comments. Second, students help each other to improve their communication skills. Third, in team projects, peer review allows team members to be assessed by each other. Finally, peer review makes it possible to break up a large project into small chunks. In fact, new releases of PG are being developed in exactly this way.

Index Terms — Collaborative Learning, Peer Grading, PG System.

INTRODUCTION

For generations, the academic community has relied on peer review as a way of enhancing the knowledge base and encouraging serious scholarship. Peer review can offer many of the same benefit to students. However, the mechanics of peer review have traditionally required too much paper-shuffling to make it practical as a classroom strategy. The era of networked computing—and the World-Wide Web, in particular—has changed all that. In recent years, electronic peer review has made its way into the classroom. Although peer review (students commenting on other students' work) and peer grading (students assigning grades for other students) have been used in many academic fields, the most common use has been in writing classes [Daed 97], to eliminate the need for instructors to read and grade hundreds of student essays.

But peer review has benefits far beyond improving writing. It encourages engineering instructors to assign more design problems, which are very important in an engineer's education, but very time consuming to grade adequately. In addition, the best work done by students can be turned into resources to help future classes learn. For example, students can be assigned to write research papers on various topics, with several students writing on the same topic. The most highly evaluated paper on each topic can then be presented to the next class of students as background

reading on that topic. The writers can be asked to include liberal doses of Web hyperlinks in their papers, so that later students can read not only their work, but also the analyses of experts. Finally, students can be asked to compose problems as well as papers; the best of these can then be assigned to later classes as homework or test questions.

PREVIOUS WORK

Dozens of studies report on different aspects of peer review, peer assessment, and peer grading in an academic setting. A comprehensive survey can be found in Topp 98. Experiments with peer assessment of writing go back more than 25 years [Ford 73]. Peer review has been used in a wide variety of disciplines, among them accounting [Pers 98], engineering [Maca 99], mathematics [Earl 86], mathematics education [LC 99], MBA programs [DW 99], and social science [Falc 94]. In recent years, cooperative learning has increased in prominence [MC 98], and a number of researchers have used peer assessment to gauge the contributions of individual members of a project team [CKSW 93, Math 94, EM 98, LC 99].

It is perhaps surprising that only since the early '90s have computers been used to mediate the interaction among peers. An early project in computer-science and nursing education was MUCH (Many Using and Creating Hypermedia) [RRR 93]. The first reported software program to support peer evaluation was evidently created at the University of Portsmouth [UP 95]. The software provided organizational and record-keeping functions, randomly allocating students to peer assessors, allowing peer assessors and instructors to enter grades, integrating peer- and staff-assessed grades, and generating feedback for students. That same year, Anderson and Michaels [AM 95] wrote a Macintosh application to "act as if it were the editor of a journal, and [send] all submissions out to two reviewers for anonymous comment." Because it required a network of Macintoshes, it did not spread as widely as it otherwise might have. One of the early Web-based peer-review experiments was described by Downing and Brown [DB 97]. Their psychology students collaborated to create hypertexts which were published in draft on the Web and peer reviewed via e-mail. Eschenbach and Mesmer [EM 98] have used the Web as a vehicle to gather peer assessment of the contributions of team members on engineering design projects.

¹ Edward F. Gehringer, North Carolina State University, Associate Professor, Department of Electrical Engineering and Computer Science, efg@ncsu.edu

THE PG SYSTEM

Our contribution to this field is PG, a portable, Web-based peer-evaluation system written in Java, which provides for peer review of homework over the Web. Students submit their work over the Web. Reviewers can be assigned using a variety of strategies (Section 5). Reviewers and authors communicate double-blindly via a shared Web page. At the end of the review process, the reviewer assigns a grade to each author whose work (s)he has reviewed. A student's grade is the average of the grades given by the reviewers, plus an incentive described below to encourage good reviews.

A student entering the PG system (Figure 1) has a choice of whether to submit a new page or review pages submitted by others. Upon choosing "submit", (s)he is presented with a screen describing how to submit and a browser to select a file. If more than one Web page is to be submitted, the student may either submit them sequentially, assigning different filenames to each, or submit a single Zip file, which PG will unpack into its components. Entire directory hierarchies may be submitted in this manner. The ability to submit directory hierarchies allows large projects to be submitted as easily as small ones.

The instructor assigns reviewers according to appropriate criteria, as detailed in Section 5. Reviewers communicate with their authors via a shared Web page. There is one such page for each author (Figure 2); the author can view the reviewers' comments and vice versa. If the instructor turns the "privacy" switch on, each reviewer can also see the other reviewers' comments and assigned grades (Figure 3). Otherwise, each reviewer is limited to seeing his/her comments and the author's responses. In either case, the author can post comments that will be seen by all the reviewers.

Early on, we discovered that sharing a Web page was not sufficient to stimulate give-and-take between authors and reviewers. Authors did not "poll" their reviewers' pages periodically to see when a new comment was submitted, nor would reviewers often notice when an author had posted a revised version of a submission. This led many students to ask the instructor to make sure their reviewers looked at their new version before the deadline for assigning grades. Now, when reviewers do reviews, their authors are notified by e-mail, and when authors submit new versions, their reviewers are e-mailed. Students can turn off e-mail selectively if the volume of e-mail gets annoying.

When an author revises a submission, PG creates a new version of that submission, and begins a new set of shared Web pages for author/reviewer dialog. Thus, each review is associated with a particular version—the version that is the most recent at the time it is submitted. If anyone wants to look at an earlier version with its reviews, a hyperlink will take them to it.

For PG-reviewed assignments where students in the class do different work (research different topics, for example), they sign up over the Web for their chosen topic.

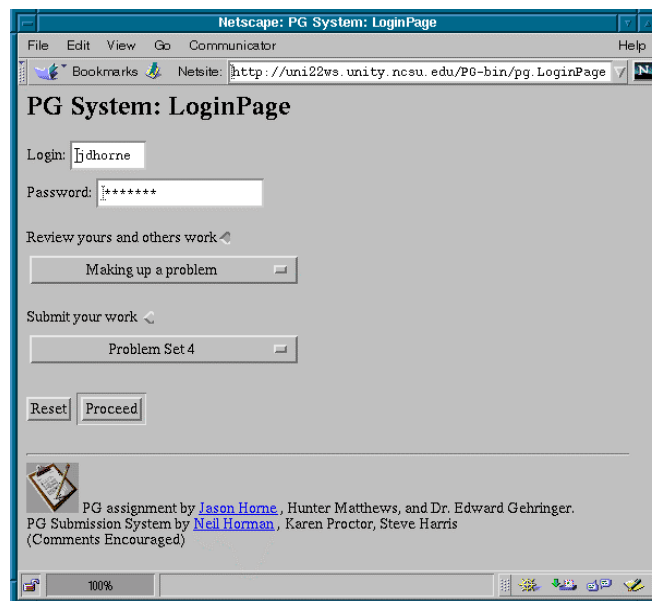


FIGURE 1
PG'S LOGIN PAGE

This allows the instructor to constrain their choices so that an equal number of students choose each topic. Signups are managed by another Java program called Shimmer; Shimmer and PG share the same database format so that the same password file can be used by both.

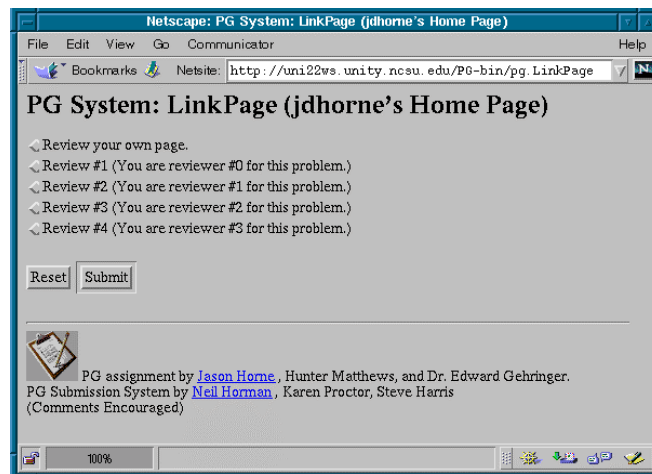


FIGURE 2
PAGE WITH LINKS TO SUBMISSIONS TO BE REVIEWED

HOW PEER REVIEW HAS BEEN USED

The kinds of assignments that can be peer-reviewed are almost unlimited. In the last two years, peer review has been used in these ways.

- *To review research papers.* In an operating systems class, for example, students were asked to research how a particular operating system (e.g., Linux) solved a particular problem (e.g., mutual exclusion). They were encouraged to include hyperlinks to documents describing the strategy in more detail.

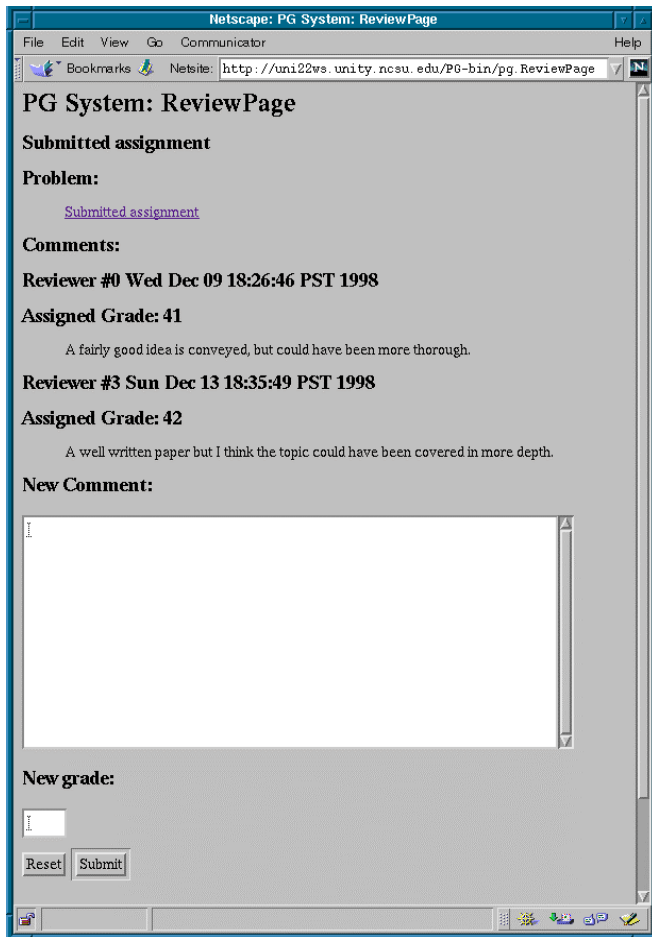


FIGURE 3
REVIEW PAGE

- *To research lecture material.* Students have been assigned to research one lecture on the Web. That is, they take the topics covered in the lecture and use search engines to find other treatments of the same material. The goal is to come up with five to ten good links to the material covered in class. A page of these links is compiled and submitted via PG. At the end of the semester, the best of these links (based on student evaluations) are compiled into a single page, which serves as a resource page the next time the course is taught.² PG is capable of automatically “publishing” a page containing links to student submissions.

² In one case, the author of the textbook used for the course made this compendium one of the resources listed on the textbook’s Website [http://www.belllabs.com/topic/books/os-book; see “Links to other OSC pages”].

- *To annotate lecture notes.* If lecture notes for a course are on line (as is the case in all of the author’s courses), students can be assigned to *annotate* those lectures—that is, to take the on-line notes and insert hyperlinks to explanatory documents at appropriate points. Later, these annotations will help fellow students to learn the lecture material if their background is weak or if they just want to explore the topic in more detail. Again, these annotations can be compiled at the end of the semester³ and serve as a resource for later generations of students. In some cases, instead of seeing large classes as a burden, an instructor may come to prefer them because they can create more formidable Web-based resources, and do so without burdening the instructor and with additional grading responsibility. This is an example of “education engineering” [Gehr 99b] — developing methodologies and tools to create educational materials more quickly and in greater volume, and disseminate them without loss of quality to the increasing numbers of students seeking a technologically up-to-date education.
- *To make up original problems.* Students have been assigned to make up a problem on some topic covered in the course. These problems have been reviewed via PG, and occasionally used on homework or tests in later semesters. The “yield” is generally 20%–25%; that is, about one-fifth to one-fourth of the problems designed by students will be usable on assignments in the future. For the instructor, being able to read the student evaluations and grades is a great help when looking for problems to assign.
- *To review other students’ designs.* In software-engineering and object-technology courses, as in many other areas of engineering, it is important for students to learn design principles. Generally, students will approach the same problem in slightly different ways and come up with different answers, since there is usually more than one “right” way to do the design. It is much more efficient for students to review each other’s designs than for the course staff to do it. In particular, PG has been used to have students review other teams’ designs for new PG modules, which have been assigned as semester projects in the author’s masters-level object-technology class.
- *To do weekly reviews in independent-study courses.* Each year, the author leads a team of students in developing and updating a large Web site.⁴ Instead of the instructor doing weekly reviews, the students are assigned to do them. This encourages the students to make steady progress, and the results figure heavily in the students’ grades.

³ See, for example http://www4.ncsu.edu/eos/users/e/efg/501/f98/course_locker/www/lectures/annotations.

⁴This is the NCSU Ethics in Computing Website, http://www2.ncsu.edu/eos/info/computer_ethics.

Peer review has also been found useful in reviewing individual contributions to team projects (Section 2). The ability to do this was recently incorporated into PG, but we have not yet had a chance to use it in a course.

REVIEWER-MAPPING STRATEGIES

When reviewers are to be assigned randomly, PG can automatically generate the reviewer mappings. But often it is better to constrain the assignment of reviewers. For example, if the class researches a large number of topics, with each student choosing one of them, it is a good idea to have some reviewers be students who have written on the same topic.

Sometimes individual students are assigned to review group projects. In one case, the author assigned a group project to his students (three students were assigned to research a topic and create a study guide and discussion questions). The reviews, however, were done individually. Each student was assigned to review one project, which meant that each team got three reviews. In another case, the group project was to design and implement a module of Java code. In this case, each two- or three-student group submitted its design to PG, and had it reviewed by individual students in other groups.

When students choose their topic from a list supplied by the instructor, it seems appropriate for each student to have some reviewers who have chosen the same topic, and some reviewers who have chosen others. For example, if each student reviews four other students, two of those students might have researched the same topic, and two might have written on another topic. This has the advantage that at least some of the reviewers will have a good idea of how hard it is to come up with material on that topic, while the other reviewers will be less than experts (and therefore better judges of the clarity of the writing). However, this assumes that students don't know which other students have chosen the same topic. This may not be true; students may sign up for the same topic in collusion with each other. To prevent this from undermining the review system, students are not assigned to review *all* the other students who have chosen the same topic, and if the instructor suspected collusion between two students, he specifically arranged the mapping so that they would *not* review each other.

- Because appropriate review strategies vary so widely, it is unlikely that we can ever anticipate all the review strategies that will be desired by instructors using the system. We plan to extend PG to handle the most popular constrained strategies for assigning reviewers. But there will always be a need to allow arbitrary user mappings. Thus, PG allows an instructor to upload a spreadsheet listing the user IDs of class members in the first column, and in subsequent columns, the IDs of students reviewing each student in the first column. This spreadsheet may be assembled using any desired strategy.

THE REVIEW PROCESS

In order for a peer-review system to work, of course, students must review the work they are assigned to. Students are given incentives to do their reviews, as described below (Section 7). But regardless of the incentive, some reviews will not be done because a reviewer has dropped the course. This is a major problem before the drop deadline, which occurs as late as two months after the beginning of the semester. We have investigated strategies to remap the remaining authors and reviewers, subject to the constraint that no reviewer will be remapped who has already reviewed an author's work. For random author-reviewer mappings, this is feasible, and can be done using either algebraic or iterative strategies. For non-random strategies specified by spreadsheets, automatic remapping is not feasible, because the remapping procedure would have to understand the original mapping strategy and follow it—when the original mapping was done “by hand.” Even if we could devise an automatic remapping strategy, that would not solve the problem, because a student may decide to drop a class (and thus stop doing homework for it) several weeks before (s)he actually drops. Consequently, at this point, no remapping is employed. If a piece of work does not receive “enough” reviews, the instructor and/or TAs review it themselves and factor the grade they assign into the grade awarded the author.

Perhaps the most frequent complaint from students was that their reviewers did not make comments in time for them to respond to them. If their reviewers waited until the last day before giving them feedback, they would have no time to improve their submission and earn a better grade. This problem was alleviated by assigning two review deadlines: By the first deadline (typically three or four days after the deadline for submitting work), reviewers were required to read the students' submissions and post some feedback on the shared Web page. By the second deadline (typically one week after the submission deadline), final grades needed to be assigned. In a follow-up survey (detailed below), most students agreed that the two-deadline system worked well.

GRADING STRATEGIES

Encouraging useful feedback. When PG was first introduced, most students wrote only cursory reviews. To encourage students to give better feedback, the grading formula for student x was changed to take into account the scores given to the students that x is reviewing—on the assumption that if x got some credit for the work (s)he is reviewing, (s)he would be more motivated to review it carefully. About three-quarters of the student x 's grade was based on the scores that x 's reviewers gave student x 's work. The other quarter was determined by the scores received by the authors x was reviewing (except for the scores given to these authors by x himself, which are not counted in

determining x 's grade). The reviews improved, but not enough.

Consequently, PG was extended to add another level of peer review: Each student can be assigned a set of reviews to evaluate, disjoint from the reviews (s)he had written and those that had been written on his/her work. The student rates each review on a scale of 1 to 10, according to how helpful (s)he thought it would prove to the author. These ratings are factored into the reviewer's grade for the assignment. Since this approach was adopted in Summer 1999, the volume of communication between students and their reviewers has increased by 15%–35% ($n = 733$, with 459 before Summer '99), though direct comparisons are difficult because the courses and assignments before the change were not exactly the same as after the change.

Combating grade inflation. One oft-cited problem with peer grading [KPD 95] is that students tend to give higher grades than the instructor. Their grades also cluster around the mean. Several antidotes come to mind, e.g., having the students *rank* their reviewees rather than give them numerical grades, or by giving each student a fixed number of *shares* to award to the other students [MG 98]. However, both of these techniques seem problematical when authors are allowed to improve their work during the review period, since one student's score can be raised only at the expense of other students. This might make reviewers reluctant to change grades in response to new submissions.

Two more promising strategies are these: MacAlpine [Maca 99] reports that having graders assign a specific weight to each of a set of characteristics, rather than give a single numerical grade, improves the correspondence between student-assigned and instructor-assigned grades. The latest version of PG supports this, but this feature has not yet been used in classes. Another promising technique is a slight modification of the shares approach: Have reviewers grade on a fixed scale, but award credit to authors based on the *fraction* of the points that each reviewer awarded to them (e.g., if a given reviewer awarded 280 points altogether and one author received 100, then that author received 35.7% of the points awarded by that reviewer).

STUDENT REACTION TO PG

Students in three classes, one undergraduate and two graduate, were surveyed in Fall '98 and Spring '99 to gauge their reactions to PG:

- CSC 210, a sophomore-level programming class, where students used PG for one assignment.
- CSC 501, Operating System Principles, where students used PG for 3 assignments.
- CSC 517, Object-Oriented Languages and Systems, where students used PG for one assignment.

Students had quite a positive reaction (Table 1) to peer review. When asked whether they agreed with the statement, "Peer review is helpful to the learning process," respondents' average response was 3.57 to 4.24 (on a scale

of 1 = "strongly disagree" to 5 = "strongly agree"). Notice that the students who used PG most (3 assignments) rated it most favorably. This may reflect the "learning curve" of getting accustomed to the system. Students responded at a mean of 3.26 to 3.88 to the statement, "I was satisfied with the reviews of my work."

In response to the question on satisfaction with reviews, one of the students noted that his work was not always reviewed by all four students who were assigned to review it. Anticipating this, the survey asked a question, "Not all students do the reviews they are assigned to do. Did this cause any problems for you?" The mean response was only 2.57 to 3.00, indicating that it was not a great problem. Perhaps more telling is the fact that 38 students (in all three classes) said it was not a problem, while only 17 said it was. Or maybe this says that missed reviews were a problem for a significant minority of students. Whatever the interpretation, it was clear that the reviews could use some improvement, and this motivated the decision to begin using reviews of reviews in Summer 1999.

The two-deadline approach seemed to help students respond to feedback from their reviewers. Students liked this approach by an average score of 3.52 to 4.12 on a scale of 1 (emphatically no) to 5 (emphatically yes), indicating that this strategy of encouraging timely reviews has been fairly successful.

CONCLUSION

The PG project is an effort to make peer grading practical. Students submit arbitrary hierarchies of Web pages, which are reviewed blindly by other students. For non-objective homework, e.g., design problems, it can provide better feedback to students than teaching assistants have time to produce. Not only does PG provide an alternative way of grading homework, but it also facilitates collaborative work. For example, it has been used to annotate all the lectures for a semester-long class with hyperlinks to related material from the Web. Student reaction to PG has been very positive, as demonstrated by post-semester surveys. PG has proven itself a valuable tool for enhancing the educational experience in courses as varied as Ethics in Computing and Advanced Object-Oriented Systems. Further information on PG may be found at <http://uni22ws.unity.ncsu.edu/PG>.

ACKNOWLEDGMENT

This work is supported by the NCSU Provost's office through an Instructional Grant. Many students have contributed to it through senior design projects and independent study projects, including Hunter Matthews, who wrote the original implementation, Neil Horman, Steve Harris, Karen Proctor, Mark Shaw, Daniel Walton, Nitin Dayhabhai, Kusay Rukieh, Will Whitaker, Phu Dinh, Jason Horne, Weigen Liang, Yuan Xu, Hassan Shehab, Archana Suthan, Ashwini Sidhaye, Chikka Rao, and Rick Flynn.

TABLE 1
AVERAGE RESPONSES TO SURVEY, BY CLASS, ON A SCALE OF 1 TO 5,
WITH ONE BEING "EMPHATICALLY NO" AND 5 BEING "EMPHATICALLY YES"

QUESTION		CSC210	CSC501	CSC517
		n=16	n=41	n=23
1	Peer review is helpful to the learning process.	4.06	4.24	3.57
2	I was satisfied with the reviews of my work.	3.88	3.68	3.26
3	Two review deadlines were imposed, one for the first review and another for the final grade. Did this provide an adequate opportunity for you as an author to respond to the comments of your reviewers?	4.06	4.12	3.52
4	Not all students do the reviews they are assigned to. Did this cause problems for you?	3.00	2.78	2.57
5	Should PG use HTML frames, so that you could see your author's work at the same time you are writing a review of it?	4.06	4.00	4.22

BIOGRAPHY

Edward Gehringer is an associate professor in the Department of Electrical and Computer Engineering and the Department of Computer Science at North Carolina State University. He has been a frequent presenter at education-based workshops in the areas of computer architecture and object-oriented systems. His research interests include architectural support for persistence and large object systems, memory management and memory-management visualization, and garbage collection. He received a B.S. from the University of Detroit(-Mercy) in 1972, a B.A. from Wayne State University, also in 1972, and the Ph.D. from Purdue University in 1979.

REFERENCES

- [Ford 73] Ford, B. W., *The effects of peer editing/grading on the grammar-usage and theme-composition ability of college freshmen*. Dissertation Abstracts International, 33, 6687.
- [Gehr 99a] Gehringer, Edward F. "Peer grading over the Web: Enhancing education in design courses," American Society for Engineering Education 1999 Annual Conference and Exposition, Session 2532.
- [Gehr 99b] Gehringer, Edward F., "A Web-Based Computer Architecture Course Database," American Society for Engineering Education 1999 Annual Conference and Exposition, Session 3232.
- [KPD 95] Kerr, Peter M., Park, Kang H., and Domazlicky, Bruce R., "Peer grading of essays in a principles of microeconomics course," *Journal of Education for Business* 70:6, July 1995, pp. 357 ff.
- [LC 99] Lopez-Real, F. and Chan, Y-P. R., "Peer assessment of a group project in a primary mathematics education course," *Assessment & Evaluation in Higher Education* 24:1, March 1999, pp. 67-79.
- [Maca 99] MacAlpine, J. M. K., "Improving and encouraging peer assessment of student presentations," *Assessment and Evaluation in Higher Education* 24:1, March 1999, pp. 15-25.
- [Math 94] Mathews, B. P., "Assessing individual contributions: Experience of peer evaluation in major group projects," *British Journal of Educational Technology* 25, 1994, pp. 19-28.
- [MC 98] Mills, Barbara J., and Cottell, Jr., Philip G., *Cooperative Learning for Higher Education Faculty*, Oryx Press, 1998.
- [MG 98] Maranto, Robert and Gresham, April, "Using 'World Series shares' to fight free riding in group projects" *PS, Political Science & Politics* 31:4, December 1998, pp. 789-791.
- [RRR 93] Rushton, C., Ramsey, P., and Rada, R., "Peer assessment in a collaborative hypermedia environment: A case-study" *Journal of Computer-Based Instruction* 20, 1993, pp. 75-80.
- [Topp 98] Topping, Keith, "Peer assessment between students in colleges and universities" *Review of Educational Research* 68:3, Fall 1998, pp. 249-276.
- [UP 95] University of Portsmouth, "Transferable peer assessment" in National Council for Educational Technology [ed.], *Using information technology for assessment, recording and reporting: Case study reports* (Vol. 1, pp. 73-78), 1995. Coventry, England: National Council for Educational Technology.
- [AM 95] Anderson, B., & Michaels, G., "Random anonymous peer review of undergraduate writing", http://www.ucsb.edu/detche/library/software/peer_review.html
- [CKSW 93] Conway, R., Kember, D., Sivan, A., and Wu, M. "Peer assessment of an individual's contribution to a group project". *Assessment and Evaluation in Higher Education* 18, 1993, pp. 45-56.
- [Daed 97] The Daedalus Group, Daedalus Integrated Writing Environment, <http://www.daedalus.com/info/overtext.html>
- [DB 97] Downing, T. and Brown, I., "Learning by cooperative publishing on the World-Wide Web," *Active Learning* 7, 1997, pp. 14-16.
- [DW 99] Druksat, Vanessa Urch and Wolff, Steven B., "Effects and timing of developmental peer appraisals in self-managing work groups," *Journal of Applied Psychology* 84:1, February 1999, pp. 58-74.
- [Earl 86] Earl, S. E., "Staff and peer assessment: Measuring an individual's contribution to group performance," *Assessment and Evaluation in Higher Education* 11, 1986, pp. 60-69.
- [EM 98] Eschenbach, Elizabeth A. and Mesmer, Marc A., "Web-based forms for design team peer evaluations," American Society for Engineering Education 1998 Annual Conference and Exposition, Session 2630.
- [Falc 94] Falchikov, N. "Learning from peer feedback marking: Student and teacher perspectives," in H. C. Foot, C. J. Howe, et al. [eds.], *Group and Interactive Learning*, Vol. 1, pp. 411-416.