Assignment and Quality Control of Peer Reviewers

Edward F. Gehringer
North Carolina State University
efg@ncsu.edu

Abstract

Much work has been performed on assessing the validity and usefulness of peer assessment in the classroom. Much less effort has been invested in enumerating and classifying strategies for assigning reviewers, encouraging good feedback, and preventing clustering of grades. This study reviews the different approaches that have been taken to those problems. Several of these strategies are employed in PG, our Web-based application for peer review and peer grading.

Usually, students are assigned randomly to review other students' work. Often, students work in teams, with each member of the team reviewing the other members. Or, students or teams may choose from a list of topics to work on. In this case, it is helpful to assign students to review others who have chosen the same topic. To encourage students to provide adequate feedback to their reviewees, several approaches can be taken. Students can be denied credit for the assignment unless they do the required reviews. Or, they can be prevented from seeing feedback on their work until they provide feedback to others. Multiple review periods may be employed, with students required to give some feedback in each period. A formula may be devised to allow reviewers to share in good grades received by their reviewees. Or students may be assigned to review each other's reviews.

To improve the accuracy of grading, students can be required to pass a pre-certification test before being allowed to serve as peer graders. The instructor can supply a set of grading criteria, and discuss it with the students, either in advance or after the students complete their first round of review. Reviewer mappings can be constrained to assure that each student will review one paper from each quartile (etc.) of the class.

1. Introduction

Peer review in the classroom is a technique that is becoming increasingly popular, with over 100 papers published on the topic in the past ten years. Much work has been performed on assessing usefulness of the technique (students generally like it, and learn well from it) and its validity (students do in general rate better work more highly, though some effort needs to be invested in the assessment procedure to assure this). However, very few published reports discuss appropriate strategies for matching reviewers with reviewees, how students can be induced to give good feedback, or how student-assigned grades can be prevented from clustering closely around the mean. These topics are the focus of this review paper.

2. Strategies for assigning reviewers

In his 1998 survey paper [Topp 98], Topping says, "How peer assessors and assessees should best be matched ... is discussed surprisingly little in the literature." In most cases, he says, a

single assessor was matched with an assessee. The matching has been done along two dimensions: blindly or non-blindly, electronically or non-electronically.

Often, reviews are done blindly, e.g., by collecting student assignments in one class, and passing them out to other students in the next class period, using an instructor-assigned ID number* to identify the students [KPD 95]. However, some projects use face-to-face interaction, frequently called "peer revision" [Stys 98, Stys 99]. In this case, of course, review is not blind. In cases where electronic review is done, usually the review is done blindly, via an application such as the Daedalus Integrated Writing Environment [Daed 97], but sometimes it is done non-blindly by e-mail [DB 97].

Most papers omit entirely any indication of how reviewers are chosen; others just say they are matched "randomly." When matching is done randomly, it may be truly randomly, meaning that each time the mapping is done, a different result (set of author-reviewer pairings) will be produced (let us call this Strategy R), or pseudo-randomly, meaning that if the mapping is rerun with the same group of students, the same pairing will be be produced (Strategy P). Generally, Strategy R would be preferred to P, since it is not good for students to tailor their writing to a particular reviewer, but strategy P has advantages where one assignment builds on the the previous one. If a particular student has reviewed the design document for a project, for example, there are advantages in having him (her) also review the finished project. If groups of students work on a single project, students may be randomly assigned to review other student(s) within the same group [Topp 00]; let us call this Strategy G.

One variant of Strategy G was used by Henderson and Buising [HB00]. They had groups of 3-5 students select topics from a list of 13. The groups then subdivided the topics, assigning part to each member of the group. The groups then exchanged their papers with another group, "preferably" one working on the same topic.

Strategy G also makes sense in a situation where students individually choose projects from a list supplied by the instructor; then students should be assigned to review students who have chosen the same project [Gehr 00]. However, in this situation, it is best also to have work reviewed by students who have not chosen the same project, so that a student author is evaluated on how well (s)he writes for non-experts; this suggests that multiple reviewers be assigned to each submission, some of whom are "experts" and some who aren't. Let us call this Strategy G+. Another example of G+ may be a project [Siva 00] where students did presentations in groups, with intra- and inter-group assessment.

In advanced classes where the students have different backgrounds and experience and students work on different projects, some students may be particularly well qualified to review certain projects. In this situation, self-selection of reviewers (Strategy S) makes sense. Of course, whether students choose review work on paper or electronically, limits have to be imposed so that each submission gets some review, and a disproportionate number of reviewers don't choose certain assignments. This is a refinement of the approach used by peer-reviewed journals that ask reviewers to list their areas of specialty. Again, to assure that submissions are reviewed for readability, it is also good if they are read by non-experts. This yields a strategy where a student

* Similarly, students could choose their own IDs (or "handles") and register them with the instructor.

gets to choose some, but not all, of the documents (s)he reviews; let us call this Strategy S+.

If the peer-review process is undertaken in two equal-size sections of the same course, a good way to make sure no student gets his (her) own paper to review is to have one section review the work of the other [KPD 95, SR 97]. This is really just a pragmatic variant of the P or R strategy, and so will not be given a separate name.

Sometimes the students are given class time to revise their submissions. In these situations, they are usually divided into groups, with each member of the group giving feedback to the others. There are advantages in making these groups "academically diverse," as Styslinger [Stys 98] did; the stronger students reinforce their understanding by explaining concepts to weaker students, and weaker students benefit from more individual attention than the instructor could possibly give. She found that students preferred either to self-select group members, or not to know them at all. However, if a project concerns assessment rather than revision, Topping [Topp 98] notes that most studies have paid little attention to having more expert students assess the less expert. He also notes that "the Piagetian model of cognitive conflict" seems to prefer "reciprocal same-ability peer assessment, between partners who are equally but differently competent."

Table 1. Review-mapping strategies

Strategy R	Random; each time the algorithm is run, a different mapping is obtained.
Strategy P	Pseudo-random; reviewers assigned at random, the algorithm will assign the same reviewers each time.
Strategy G	Students arranged into groups; assignment of reviewers to authors is random within the group.
Strategy G+	Students are assigned some reviewers from within their own group and some from other groups.
Strategy S	Self-selection; students choose which submission(s) to review (blindly).
Strategy S+	Students choose some of the submissions they will review, but are also assigned randomly to review other submissions.
Strategy V	Students review all of the other members of their group (with each member doing a different project).
Strategy T	Students review all of the other members of their team (which is cooperating to produce a single project).

One of the better developed experiments in diversification of peer-revision groups is that reported by Nelson [Nels 00]. She reports—

> "To accomplish the goal of improving communication and teamwork skills in my class, I put my students into teams of four at the beginning of the term. I use a writing skills pre-test to assess current competence, and then diversify my teams in terms of writing competency, engineering major, gender, and ethnicity. These teams sit together and collaborate on writing tasks as well as review one another's writing throughout the term. While I do not grade teams on group writing tasks, I point out continually that they are responsible for improving each other's

> grades through the peer evaluation process. In addition, in several group writing exercises they compete against other writing teams for extra credit points. Importantly, a component of their grade is based on evaluation by their team members on their performance as a peer reviewer and their performance as a collaborative task contributor."

Let us call peer revision in (non-blind) groups Strategy V. It differs from Strategy G, where the students evaluate authors who are not in their own group.

In Strategy V, all of the students in a particular group are working on their own assignment. If we assume instead that they are all doing parts of a single assignment, we get team peer review (Strategy T). This is one of the most frequently used kinds of peer review, being used in team projects at many schools in all branches of engineering [Esch 97, EM 98, DVFF 00]. Projects are graded by the course staff, but team members assess each other's contributions, and these assessments are used to avoid the need to give all members of the group the same grade, regardless of their contributions. Table 1 summarizes the different reviewer mappings we have considered.

3. Encouraging good reviews

Peer review can only be successful if students take the task of reviewing seriously. But, given the other demands on their time, they won't do this unless given an incentive to do so. Unfortunately, even less has been published on incentives for good reviewing than on how to select reviewers.

Kerr, Park, and Domazlicky [KPD 95] took the approach of giving a zero to any student who did not complete the assigned review. Their assignment sheet stated,

> ....your work is not done once you have completed your own paper. At the next class meeting you will be given another student's paper to read and grade. If you fail to complete this task, your own paper will receive a score of zero and the other student will receive the instructor's score.

(In their scheme, the instructor grades all the papers, and the students are given the average of their reviewer's and the instructor's score.) They report that the severity of punishment induces almost all of the students to complete their reviews. Let us call this Scheme 0.

A less drastic penalty is not to allow students to see comments on their own paper until they give feedback to all the students they have been assigned to review. Similarly, they can be denied seeing their own grade until they give grades to the students they are reviewing. Let us call this Scheme 1. A refinement—implementable by software for electronic reviewing—would be to let students see only the same volume of comments on their papers as they provided to other students. However, in the absence of some other scheme for rating reviews, this might encourage students to pad their reviews with marginal or even irrelevant information.

When we began the PG project [Gehr 99], the grading formula for student *x* took 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). This, unfortunately, did not seem to give enough of an incentive to provide thoughtful reviews. This may be because it was a difficult scheme to explain to students; or it may have been because their reviews could have only a small effect on their reviewees' work (this is especially true if a student gave high grades initially; then the reviewee has little incentive to improve). Let us call this Scheme 2.

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. In a study published last year [Gehr 00], we found that 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. Let us call this Scheme 3. The schemes for ensuring careful reviewing are listed in Table 2.

Table 2. Schemes for ensuring careful reviews

Scheme 0	Give a student a 0 on his own work if (s)he doesn't do the assigned review(s).
Scheme 1	Students are allowed to see comments (grades) for their own work only after they have provided comments (grades) to their reviewees.
Scheme 2	Students share in the grades of the students they review by being awarded a percentage of the grades their reviewees achieve.
Scheme 3	In an additional round of review, students review *reviews* submitted by other students; the score for these reviews are one component of their grades.

4. Preventing clustering of grades

Several researchers [KPD 95, Maca 99] have noted that students tend to give higher grades than instructors. This is a problem because it leads to grade inflation, where grades less clearly differentiate the performance of different students. It may make it politically difficult for an instructor to assign final grades lower than the students received from their peer reviewers. Furthermore, if multiple reviewers are used, the central tendency of averaged grades [Maca 99] leads to a still narrower range of grades. This decreases the incentive students have to do a good job on their peer-reviewed work.

To combat this problem, the instructor has two alternatives: Change the rating scheme, or help students to improve the quality of their reviews. Changing the rating scheme consumes less of the students' time, and is therefore the method of choice when peer review is only a small component of the homework assigned in a class. Attempting to improve the quality of reviews helps students improve their critical thinking skills, and thus is a good investment when time permits.

The rating scheme could be changed by having 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 and resubmit 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.

A refinement of the ranking approach would have the final grade depend not only on the ranks received by the student author, but also on the strength of the competition with which the student was paired. In other words, the student's average rank would be scaled by the average rank of the students with whom (s)he was ranked against. This would produce a *relative-strength* scoring system, akin to those used in computerized football and basketball rankings.

The most common approach based on rating schemes is *category-based* rating, where students are required to rate their reviewees on several categories (e.g., insightfulness, technical correctness, clarity of presentation). MacAlpine [Maca 99] found that requiring students to rate each other in 4 categories markedly diminished the tendency of grades to cluster around a mean. Orsmond, Merry, and Reiling [OMR 00] found that grading using the instructor's criteria evidently made students think more critically about their grading. Kerr et al. [KPD 95] found that giving the categories for grading used by the instructor and having students develop a scheme that resembled it resulted in a decrease in average grade given by students. Finally, Topping [Topp 98] notes that students "prefer specific performance criteria to vague ratings."

Although category-based rating helps students improve the quality of their reviews, it can be differentiated from training programs that have as their explicit goal improving the quality of reviews. One such scheme was reported by Allain [Alla 00]. Essays are placed in "bins" by the course staff, based on quality to prevent students from receiving two essays of similar quality. A bin can contain an "exemplar" essay that all students will receive for grading. (This is to give the student an idea of what a good essay should be like.) Students submit their rankings of the essays, and receive a grade based on the correlation between their ranking and the instructors ranking.

This approach is similar to that employed by the Calibrated Peer Review project [CPR 00]. After a student submits a document, CPR presents three "calibration" documents in random order. One is written by an expert; the other two are typical student submissions. The process proceeds as follows.

> For each calibration document, the student answers a series of content questions and then a series of style questions. The style questions may range from very simple questions to questions that require detailed grammatical analysis. After answering the questions for each document, the student rates each document on a scale of 1-10, with 10 best. When the student has completed the calibration, CPR prepares and presents the student calibration report. The report shows the student answers to the questions and the official answers and compares the student and official scores for each document. ... Extensive feedback in the assessment of the calibrations, clarifies students' understanding of the issues, and corrects misconceptions that they might have. In the background, CPR assesses the student's performance on the calibration and either directs the student to instruction or permits the student to proceed to peer review.

This approach, obviously, is more feasible for assignments that can be reused from semester to semester; otherwise, calibration documents would have to be created and collected anew for each assignment. The various approaches to preventing clustering of grades are summarized in Table 3.

Table 3. Approaches to improving grading

ranking	Have students rank their reviewees 1 to *n* instead of giving them grades
shares	Give each student a fixed number of shares to divide up among reviewees.
relative strength	Ranking, with a student's average rank scaled by the average rank of the students against whom (s)he is ranked.
category-based rating	Students are asked to rate reviewees in several categories, with their grade being the sum of ratings in all the categories.
calibrated peer review	Students are given "calibration" documents to grade, and only allowed to do peer reviewing if they grade the calibration documents well enough.

5. Conclusion

It is surprising that strategies for assigning peer reviewers and controlling the quality of peer reviews have received so little attention in the literature. As long as most peer review was conducted in the classroom, in small classes, these issues could perhaps be dealt with on an *ad hoc* basis. But as peer review begins to be used for larger numbers of students, and as software tools are developed to manage it, more formal strategies are needed. Of the eight reviewer-mapping strategies mentioned, each one is appropriate in some circumstances, and software should support most or all of them. Of the four schemes for ensuring careful reviews, research is necessary to determine the most effective scheme. Of the five schemes for improving grades given by students, category-based rating is the most widely used and best evaluated, but calibrated peer reviewing is a promising innovation, especially for improving student grading of writing.

Bibliography

[Alla 00] Allain, Rhett, "Peer ranking with WebAssign," WebAssign User's Group meeting, North Carolina State University, May 5-6, 2000.

[CPR 00] Calibrated Peer Review, http://cpr.molsci.ucla.edu.

[Daed 97] The Deadalus Group, Daedalus Integrated Writing Environment, http://www.daedalus.com/info/overtext.html

[DVFF 00] DeJong, Nicole C., Van Treuren, Kenneth W., Farris, Donald R, and Fry, Cynthia C., "Using design to teach freshman engineering," *Proc. American Society for Engineering Education 2000 Annual Conference and Exposition*, Session 2253.

[DB 97] Downing, T., and Brown, I., "Learning by cooperative publishing on the World Wide Web," *Active Learning* 7 (1997), pp. 14-16.

[Esch 97] Eschenbach, Elizabeth A., "Using Peer Evaluations for Design Team Effectiveness" *Proc. American Society for Engineering Education 1997 Annual Conference and Exposition*, Session 2553.

[EM 98] Eschenbach, Elizabeth A. and Mesmer, Marc A., "Web-based forms for design team peer evaluations," *Proc. American Society for Engineering Education 1998 Annual Conference and Exposition*, Session 2630.

[Gehr 99] Gehringer, Edward F., "Peer grading over the Web: enhancing education in design courses," *Proc. American Society for Engineering Education 1999 Annual Conference and Exposition*, Session 2532.

[Gehr 00] Gehringer, Edward F., "Strategies and mechanisms for electronic peer review," *Proc. Frontiers in Education 2000*, Session F1B.

[HB 00] Henderson, LaRhee, and Buising, Charisse, "A peer-reviewed research assignment for large classes," *Journal of College Science Teaching* 30:2 (October 2000), pp. 109-113.

[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.

[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.

[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.

[Nels 00] Stephanie Nelson, "Teaching collaborative writing and peer review techniques to engineering and technology undergraduates," *Proc. Frontiers in Education 2000*, Session S2B.

[OMR 00] Orsmond, Paul, Merry, Stephen, and Reiling, Kevin, "The use of student derived marking criteria in peer and self-assessment," *Assessment and Evaluation in Higher Education* 25:1 (Mar. 2000), pp. 23-38

[Siva 00] Sivan, Atara, "The implementation of peer assessment: An action research approach," *Assessment in Education 7:2* (July 2000), pp. 193-213.

[Stys 98] Styslinger, Mary E., "Some milk, a song, and a set of keys: Students respond to peer revision," *Teaching and Change* 5:2 (Winter 1998), pp. 116-138.

[Stys 99] Styslinger, Mary E., "Mars and Venus in my classroom: Men go to their caves and women talk during peer revision," *English Journal* 88:3 (Jan. 1999), pp. 50-55.

[Topp 98] Topping, Keith, "Peer assessment between students in colleges and universities," *Review of Educational Research* 68:3 (Fall 1998), pp. 249-276.

[TSSE 00] Topping, Keith, Smith, E.F., Swanson, I., and Elliot, A., "Formative peer assessment of academic writing between postgraduate students," *Assessment and Evaluation in Higher Education* 25:2 (June 2000), pp. 149-169.

EDWARD F. GEHRINGER
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.