# Experiences with Peer Evaluations of Student Papers

**Robert C. Sorensen, Department of Agronomy**
**Myra S. Wilhite, Department of Agricultural Leadership, Education, and Communication**
**University of Nebraska, Lincoln, NE 68583-0914**

## Abstract

A system for peer evaluation of student papers is described. A rationale for the system is proposed. Observations and results of the process are presented. Data for the 1998 and 1999 classes are given confirming the consistency of the ratings. No differences were observed in the consistency among the content, organization, interest, and presentation of the papers. Results from a questionnaire given to students after the scoring exercise demonstrated acceptance of the rating process by students, and suggested that it had positively affected their own writing.

## Introduction

In recent years considerable progress has been evident in higher education toward involving college students directly in their education through a variety of active learning teaching methods. However, an opportunity for student involvement that has received limited attention is using students as peer evaluators. A recent literature review of contemporary sources on grading identified only two published references to peer evaluation. Walvoord and Anderson (1998, p. 230) describe the use of peer evaluation as a way to make grading more time-efficient. Fry (1990) found that peer grades were positively correlated with teacher grades and that five advantages of peer evaluation had been achieved. He found also that students believed their work had been evaluated fairly and the scores should count toward final grades. A few unpublished accounts of peer evaluation in the context of active learning were found.

We propose that upper-class students may effectively judge student work in selected applications and that student ratings may have validity on the same order as teacher ratings. We describe a system using peer evaluations of student papers, provide a rationale for the system, and present evidence demonstrating that these peer evaluations are internally consistent.

## The Process

The course in which peer scoring has been and continues to be used is Soil Science 366, Soil Nutrient Relationships, a course which is taken primarily by juniors and seniors with majors in Agronomy, Agribusiness, Diversified Agriculture, Horticulture, Natural Resources, and a few other majors. There are usually 50 to 60 students in the class. The student paper whose rating is described herein constitutes about 5% of the student's course grade. Each student is asked to write a four to six page article on some subject he or she knows or is willing to learn about and which pertains to the course subject matter. Library research and citations are not required. The paper is intended to be a popular article, not a research paper. The objective for the students is to integrate concepts learned in the class with their own prior knowledge and experiences to produce a cohesively written article that will be interesting and provide useful information to their present peers. These peers are likely to be their future coworkers, employees, and employers.

Each paper is rated by six students. The review is double-blind. The rater does not know whose paper he or she is reading, and the author will not know the identity of the raters. The papers are rated on four categories: content, organization, interest, and presentation. Ratings are: 1 - excellent, 2 - average, and 3 - could use more work. The instruction sheet to raters is included as Appendix 1. Each student is asked to mark all errors, make comments throughout the papers, and add at least one general comment at the end of each paper being read. Unless he or she is the first rater, each subsequent student rater sees what corrections and comments others have placed on the paper. Raters will not see the previous ratings which are placed on a separate sheet.

Each student's rating process is itself evaluated by the course instructor to encourage a careful review of the papers by examining the distribution of ratings given. The raters are given a perfect score (10 points) unless the score distribution they gave other students' papers demonstrates a lack of discrimination among papers. The most frequent reason for an imperfect score is too many "1" ratings. For the

1999 class 56% of the students received a perfect rating score and 40% lost only one point.

The teacher tabulates and adds the ratings to yield the score for each paper. The lowest score represents the best paper. The best grade possible would be 24 (six raters x four categories x 1). The worst grade possible would be 72 (six raters x four categories x 3). For the 1998 class, the grades ranged from 28 to 54 with an average of 40.7 and a standard deviation of 5.2. For 1999, the range was from 32 to 57 with an average of 43.5 and a standard deviation of 5.9. The teacher arranges the papers in order of scores and checks selected papers with emphasis on those receiving the highest or lowest ratings. In five classes, the teacher has never found adequate cause to change the grade on a paper. Although only a few papers were read by the teacher, the student ratings appeared to be very appropriate. Using a linear scale with negative slope, the ratings were converted to traditional grades. At this point, the teacher could affect what grades the students as a group receive in general, but not influence individual student grades. The teacher offered to read any paper in detail and score it if a student does not believe his or her paper received fair treatment. This option has been requested only twice in five years. Apparently, most students believe they fare better with their peers than with the teacher.

**Consistency of Ratings**

This peer evaluation process is validated only if there is consistency in student ratings. Data from the 55 papers submitted for the 1998 class and 60 papers for the 1999 class were used to examine consistency of rating. Since how a person rates a paper can be heavily influenced by many factors, we developed a weighting system such that a t-test could be used to determine whether, on the average, the six peer evaluators were rating each paper similarly. A consistency index 'I' was computed for each paper:

$I = (A \times 4) + (B \times 3) + (C \times 2) + (D \times 1) + (E \times 0)$ where:

A = number of cases where all six students gave the same rating;
B = number of cases where five of the six students gave the same rating;
C = number of cases where four of the six students gave the same rating;
D = number of cases where three of the six students gave the same rating; and
E = number of cases where two of the six students gave the same rating (random rating).

Since there are four categories being rated (content, organization, interest, and presentation), $A + B + C + D + E = 4$. The maximum value of the index would be 16 if A = 4, and B, C, and D are zero. The minimum value would be zero if E = 4, and A, B, C, and D are zero.

If the ratings were random, the index for a paper should be zero. For 1998 the indices for 55 papers ranged from 1 to 12 with a mean of 7.62 and a standard deviation of 1.95. In 1999 the indices for 60 papers ranged from 4 to 10 with a mean of 7.47 and a standard deviation of 1.78. The computed t-values for the respective years were 32.5 and 29.0 based on the null hypothesis that the ratings were random, i.e. that the mean rating was zero. This is excellent evidence that the students rated with a great deal of consistency.

One may ask whether there were differences in consistency among the four components of the ratings (content, organization, interest, and presentation). Average indices for these components in 1998 were 2.1, 1.8, 2.0, and 1.8 respectively. In 1999 the respective values were 1.8, 1.8, 1.9, and 1.9. Thus consistency was essentially equal for all four components. As a matter of interest, these values indicate that, on the average, four of six students gave the papers the same ratings.

There are several advantages to this type of peer evaluation. Writing for their peers rather than the teacher is consistent with what students will be doing after graduation. It moves the focus from "what the teacher wants" to student values, often an unexplored area. Each paper is evaluated by six persons with diverse backgrounds, experiences, learning styles, academic abilities, and knowledge of the subject. Most had little reviewing experience. We have found no reason to believe that these student ratings are in any way inferior to ratings by a single teacher and, in theory at least, they are likely to be equivalent. Much is learned by students seeing what corrections and comments other students have put on the paper they are reading. In this process they can hone their own evaluative skills, which should ultimately improve their own writing. Having to read, evaluate, and score represents active learning at its best. The process also should reduce inconsistent scoring caused by fatigue of the teacher in a large course if all papers were read by him or her.

Although we have done the tabulations of ratings and computation of scores by hand, it would be a simple matter to collect the student ratings by mark-sense sheet and mechanize the entire tabulation and scoring process. This would provide opportunities for students in large classes to participate in writing experiences with high quality grading but with very minimal grading costs.

A question may be raised about bias in ratings caused by evaluators having seen previous comments placed on the paper. The presence of bias is difficult to assess. Also, any bias could be positive or negative. Observation of student comments by the teacher has shown that raters sometimes refer to other raters' comments, either

to agree or to disagree with them. Our judgment is that bias may be present, but the effect on final ratings is probably small.

Obviously this process is most effective when used with advanced and experienced students who have had writing courses, have a positive attitude, and are conscientious about their work. It should probably not be used for rating papers such as research reports which have many specific requirements for quality, unless a fairly detailed check sheet is used. The rating process needs to be done in a classroom environment without conversation where plenty of time is available for rating. Evaluators should not feel rushed. Papers should be assigned to raters at random.

**Student Responses**

In 1999 students were asked to complete a questionnaire regarding their views of the peer scoring exercise. A list of the questions is provided in Appendix 2 and the results presented in Table 1.

The students were comfortable scoring the papers (95%) and having other students score their papers (78%). A significant number (22%) found the papers to be more poorly written than they expected. Either they liked having the students score the papers (56%) or they did not care who scored them (31%). Thirty percent of the students reported that knowing other students would score their paper had a measurable effect on how they wrote the paper.

Table 1 Percentage of students responding to questions about peer scoring.[z]

| | Answers (%) | | | | |
|---|---|---|---|---|---|
| Question No. | a. | b. | c. | d. | e. |
| 1. | 53 | 42 | 3 | 2 | – |
| 2. | 11 | 60 | 22 | 7 | – |
| 3. | 13 | 36 | 51 | – | – |
| 4. | 7 | 35 | 58 | – | – |
| 5. | 78 | 16 | 6 | – | – |
| 6. | 56 | 13 | 31 | – | – |
| 7. | 45 | 53 | 2 | – | – |
| 8. | 47 | 21 | 28 | 2 | 2 |

[z] Question provided in Appendix 2

## Conclusions

Our conclusions are similar to those of Fry (1990). We have concluded that students in this class have not only had an opportunity to gain several educational benefits from their paper-scoring experience, but have probably received grades as appropriate and fair as those given by the course teacher. By the nature of the exercise, they also should have been given confirmation that their experiences and education have value and their ideas and values are important. In the consistency of their ratings, we are assured that scoring is based on commonly-held standards and that their work is reliable. From the responses to the questionnaire, we have found that students are pleased with the exercise, and it probably provides an incentive to improve their writing.

## Literature Cited

Fry, S.A. 1990. Implementation and evaluation of peer marking in higher education. Assessment and Evaluation in Higher Education 15:177-189.

Walvoord, B.E. and V.J. Anderson. 1998. Effective grading - a tool for learning and assessment. San Francisco, CA: Jossey-Bass Publishers.

Appendix 1. The instruction form given to the student raters.

**Paper Evaluation**

INSTRUCTIONS: During this period you will be asked to read 5-6 papers prepared by your classmates and evaluate them on a number of points. Please be critical but fair. Look for positive things as well as negative things about each paper.

A. Points for Evaluation.

a. Content - Does this paper provide adequate information so most readers would probably learn something?
b. Organization - Is the paper easy to follow from beginning to end? Is it organized logically?
c. Interest - Is the paper interesting? Did you enjoy reading it?
d. Presentation - Is the paper neat and free of grammar, spelling and typographic errors?

B. Rating Scale.

1 - Excellent
2 - Average
3 - Could use some more work

You should have some 1's, some 2's and some 3's. Be discriminating, but not sadistic.

C. Results

| Paper No. | Content | Organization | Interest | Presentation |
|---|---|---|---|---|
| 1. | | | | |
| 2. | | | | |
| 3. | | | | |
| 4. | | | | |
| 5. | | | | |
| 6. | | | | |

Appendix 2. Questionnaire given to students following the student rating exercise.

**Project Evaluation**

As you know, each paper written by students in AGRO 366 is scored by six other students. The rationale is that your audience for writing after graduation will be your present classmates, not the teacher. In addition, experience has shown that students do an excellent job of scoring other students' writing. Since you have just completed this scoring process, I would like to know how you felt about doing it.

Please put your answers on the mark-sense sheet. You don't need any other information on the sheet except your answers.

1. How comfortable were you with scoring other persons' papers?

a. I was very comfortable.

b. I was fairly comfortable.
c. I was somewhat uncomfortable.
d. I was very uncomfortable.

2. What was your general judgment of the papers you read?

a. They were better than I expected.
b. They were about as good as I expected.
c. They were worse than I expected.
d. I didn't know what to expect.

3. How would you rate your own approach to scoring?

a. I'm probably a more lenient grader than most others.
b. I'm probably a more critical grader than most others.
c. I think I grade about the same as most others.

4. How difficult was the process of scoring the papers for you?

a. Scoring was difficult for me.
b. Scoring was easy for me.
c. Scoring was not easy but not hard.

5. How do you feel about the scoring of your own paper?

a. I am confident it will be scored fairly.
b. I fear it will be scored lower than it should be.
c. I think it will be scored higher than it should be.

6. Considering the whole process:

a. I like the idea of students scoring the papers.
b. I would rather have the teacher score the papers.
c. It doesn't matter to me who scores the papers.

7. How much did you get out of the scoring exercise?

a. I learned a lot from reading others' papers.
b. I learned some things from reading others' papers.
c. Reading others' papers was just a job to me.

8. Did knowing that students would score your paper have any effect on how you wrote it?

a. No effect whatever.
b. It may have had a little effect.
c. It had some effect.
d. It had a lot of effect.
e. I didn't know it would be scored by students.

9. Please add any comments below you think would help me improve this scoring process: