Peer Ratings in Cooperative Learning Teams

Deborah B. Kaufman, Richard M. Felder, Hugh Fuller
North Carolina State University

Synopsis

A universal concern about cooperative learning is the possible existence of "hitchhikers," team members who fail to fulfill their team responsibilities but get the same high grade as their more responsible teammates. A common way to minimize hitchhiking is to use peer ratings to assess individual performance of team members and to adjust the team project grade for individual team members based on their average ratings. Peer ratings have potential drawbacks, however. Common concerns are that team members will agree to give one another identically high ratings, or give ratings based on gender or racial prejudice, or inflate their own ratings if self-ratings are collected. Some instructors also worry that many students will resent having their grades affected by their teammates' ratings. The objective of this study was to examine the validity of these concerns.

A peer rating system developed at the Royal Melbourne Institute of Technology was modified and used in two sophomore-level chemical engineering courses. The students completed their homework in instructor-formed teams in each course, and an average homework grade was computed for each team. At the end of each course the students confidentially rated how well they and each of their teammates fulfilled their team responsibilities, taking the ratings from a prescribed list of nine terms ranging from "excellent" to "no show." The instructor assigned numerical values to each rating and computed a weighting factor for each student as the student's individual average rating divided by the team average. The student's final homework grade was the product of the weighting factor and the team project grade. Correlations were computed between peer ratings and test grades, peer ratings and self-ratings, ratings given to teammates of the same sex and of the opposite sex, and ratings given to teammates of the same ethnic background and of different ethnic backgrounds.

Peer ratings correlated significantly with test grades, indicating that the more responsible students tended to be those who did best academically and/or that the academically stronger students were perceived as contributing most to the team effort. Self-ratings were remarkably consistent with peer ratings. Students rarely rated themselves higher than the rest of their teammates rated them; in fact, more (although still relatively few) gave themselves ratings lower than any they received from teammates. The incidence of identical ratings for all members of a team was also relatively low, on the order of 5–10% of all teams. No evidence of gender bias appeared in the data. Non-minority students gave lower ratings to minority students than to other non-minority students; racial prejudice could account in part for this result, but other explanations are equally likely or more so. Roughly 7% of the students were revealed as possible hitchhikers (as evidenced by their receiving less than satisfactory peer ratings from their teammates), but complaints about the system were almost non-existent. Most of the concerns frequently raised about peer ratings in cooperative learning were thus not borne out by the results of this study. Much additional research will be needed before the concerns can be definitively set aside.

Introduction

Cooperative learning (CL) is an instructional method in which teams of students work on structured tasks (e.g. homework assignments, laboratory experiments, or design projects) under conditions that meet five criteria: positive interdependence, individual accountability, face-to-face interaction, appropriate use of collaborative skills, and regular self-assessment of team functioning. Many studies have shown that when correctly implemented, cooperative learning improves information acquisition and retention, higher-level thinking skills, interpersonal and communication skills, and self-confidence (Johnson, Johnson, and Smith, 1998).

Holding each student individually accountable for work done in a team setting is a cornerstone of cooperative learning. One way to meet this goal is to adjust team project grades for all team members according to how well they fulfilled their responsibilities. A peer rating system designed for this purpose has been developed at the Royal Melbourne Institute of Technology (RMIT) by Professor Rob Brown (Brown, 1995). Team members confidentially rate themselves and one another, taking the ratings from a prescribed list of nine terms ranging from "excellent" to "no show." The instructor assigns numerical values to each rating and computes a weighting factor for each student as the student's individual average rating divided by the team average. The student's final project grade is the product of the weighting factor and the team project grade.

The potential of peer ratings to promote fairness in team project grading is evident, but their use gives rise to several concerns. Individuals could give themselves higher ratings than they deserve; team members could agree to give everyone identical ratings to avoid conflict; and personal prejudices—e.g., gender or racial bias—could influence the ratings. The objective of this study is to assess the likelihood of these occurrences.

Demographics

The RMIT peer rating system was used in two consecutive sophomore-level chemical engineering courses at North Carolina State University: CHE 205 (Chemical Process Principles, Fall 1997), and CHE 225 (Chemical Process Systems, Spring 1998). Table 1 reports demographic data for the students in each course.

Table 1
Demographic Data

Class	N	Men	Women	Non-minorities	Minorities
CHE 205	137	70%	30%	88%	12%
CHE 225	71	70%	30%	92%	8%

N is the number of students who received final course grades. "Minorities" includes African-American students (11% in CHE 205, 7% in CHE 225) and Native American students (<1% in CHE 205, 1% in CHE 225), and "non-minorities" includes Caucasian students and students of all other ethnic backgrounds enrolled in the course. (There were no students of Hispanic background in either course.)

On the first day of class the students filled out questionnaires that asked them to specify gender, ethnicity, grades in prerequisite courses (calculus, chemistry, and physics courses for CHE 205, advanced calculus and CHE 205 for CHE 225), outside interests, and times available for group work outside of class. The students were told that they could skip any questions that they felt intruded on their privacy, but only a few failed to respond to all questions. They were then grouped into teams of three or four by the instructor to assure as much as possible heterogeneity of academic ability (as measured by the prerequisite course grades), commonality of interests, and common blocks of time for meeting outside class.

In CHE 205, 39 three- and four-person groups were formed: 20 all-male, 6 all-female, and 13 mixed-gender groups; 26 ethnically homogeneous groups and 13 groups of mixed ethnicity. In CHE 225, 18 groups were formed: eight all-male, one all-female, and nine mixed-gender groups; 12 ethnically homogeneous groups and six groups of mixed ethnicity.

Cooperative Learning Procedures

Team members were assigned roles that rotated from assignment to assignment. The *coordinator* organized working sessions and made sure that all team members understood their responsibilities. The *recorder* prepared the final solution set. A *checker* (or two checkers in a team of four) proofread the final solution set, verified that all team members understood both the solutions and the problem-solving strategies used to obtain them, and took primary responsibility for submitting the solution set on its due date.

The teams were periodically asked to submit assessments of how well their team was functioning, and they were encouraged to see the course instructor if they were having problems of any sort. In some cases the course instructor sought out teams that reported having difficulties. Occasional mini-clinics were held to discuss ways of dealing with problems commonly encountered by cooperative learning teams. After the first six weeks, the students were told that their teams would be disbanded and reformed unless all members of a team indicated confidentially that they wished to remain together, in which case they would be permitted to do so. Of the 39 teams in CHE 205, only one elected to disband and so had to remain together. All of the teams in CHE 225 elected to remain together.

For more details about the cooperative learning model implemented in the two courses, see Felder (1995).

Peer Rating Procedure

The peer rating form used in the course is shown in Figure 1. In CHE 205, the students were specifically asked *not* to rate themselves, and the form they received differed from that shown in Figure 1 in only that respect.

Each student received a copy of the form on the first day of each course. The form was briefly explained, and the students were told that they would fill it out at the end of the semester and that their ratings would be used to adjust their average homework grade (which accounted for 15% of their final course grade). Midway through the semester, blank forms were handed out and the students were instructed to fill them out, show them to their teammates, and discuss reasons for ratings lower than "Very Good." In the last week of the semester, after the last

assignment had been turned in, they were given blank forms again and told to fill them out confidentially, sign them, and return them to the instructor. The explanations of the purpose of the form and the meaning of the ratings were repeated, and the students were cautioned that both fairness and self-interest dictated that they submit their ratings. The instructor logged in the forms and sent e-mail reminders to those who had not submitted them.

Each verbal rating was converted to a numerical equivalent, with "Excellent" = 100, "Very Good" = 87.5, and so on in 12.5-point decrements down to "No Show" = 0. The ratings were entered in a spreadsheet and analyzed in the manner explained in the introduction of this paper. The weighting factor used to determine each individual's homework grade was that individual's average rating divided by the team average. A maximum weighting factor of 1.10 was imposed, and calculated factors greater than this value were scaled down. This step was taken to preclude students receiving highly inflated homework grades by virtue of having a teammate with very low ratings.

Interestingly, no students in either course ever asked exactly how the descriptive ratings of their teammates would be used to adjust their homework grades. (Had one asked, the instructor would have explained it.) The students apparently assumed that the ratings would be used in some qualitative manner if they were used at all; it apparently never occured to them that the descriptive terms ("Excellent," "Very Good," etc.) would be converted to numbers and used to make quantitative adjustments to team grades.

Nomenclature and Data Analysis

The following nomenclature will be used in reporting results. IER (individual effort rating) denotes the average numerical peer rating a student received from his or her teammates, and GIER (group-average individual effort rating) denotes the average IER for all team members. Unless otherwise noted, ratings reported in this paper represent average peer ratings and do not include self-ratings. The "average test grade" is a weighted average of a student's three individual test grades (weighted at 20% per test) and final examination grade (weighted at 40%), all tests having been graded on a 0–100 basis.

PGPA (prior grade-point average) is a student's cumulative grade-point average scaled to a 0–100 basis for semesters up to but not including the one that included CHE 205 or CHE 225. The scaling formula is [GPA(0-100) = 12.5 x GPA(A=4) + 50]. The "normalized test grade" is the difference between a student's average test grade and his or her PGPA. Loosely speaking, the normalized test grade is a measure of performance relative to grades in prior courses: the higher the normalized grade, the better the performance relative to pre-course expectations. PGPAs for sixteen students in CHE 205 were unavailable for various reasons (e.g., because they were new transfer students), and so these students were omitted from statistical tests involving this variable.

In CHE 205, twelve of the 137 students did not submit peer ratings. Six of these students were male, six were female, and two were minorities. Six students did not receive ratings from their teammates. Since five of these students were in the same two groups and one was in a two-person group, three groups did not generate GIERs. These students and groups were excluded from analyses involving class IER and GIER averages. In CHE 225, all students submitted peer ratings for their teammates. One student did not submit a self-rating.

The remaining sections of this paper summarize the principal results. All reported levels of significance are derived from nonparametric Wilcoxon rank-sum tests unless otherwise noted, with "statistically significant" defined as $p<0.1$. Pearson correlations were used to test for association between average student ratings and student performance in the class.

Correlations between ratings and grades.

In CHE 205, peer ratings correlated positively with average test grades [R=0.54, p=0.0001]. The correlations between IER and average test grade were even stronger for women (R=0.76, p=0.001) and minority students (R=0.79, p=0.0005). In CHE 225 the correlation between peer ratings and test performance was weaker but still statistically significant (R=0.32, p=0.0008), and the correlations for female and minority populations were not statistically significant. These results indicate that the more responsible students tended to be those who did best academically and/or that the academically stronger students were perceived as contributing most to the team effort.

Of students entering CHE 205 with a PGPA less than 3.0, those with IER > 80 earned an average normalized grade of –12.3 and those with IER < 70 earned a normalized grade of –28.9. The difference is significant at the 0.06 level (1-tailed test). In other words, the performance relative to expectations of good team citizens exceeded that of poor team citizens. This result further supports the implication that responsible team performance has a beneficial effect on academic performance.

Correlations between self-ratings and peer ratings.

A common concern when self-ratings are included in a peer rating system is that students will inflate their own ratings to give themselves an advantage when the project grades are computed. The study results show that inflated self-ratings rarely occurred in CHE 225 (the only course in which self-ratings were collected). The average self-rating was 90.0 and the average peer rating was a statistically indistinguishable 89.1. Self-ratings of male, female, minority, and non-minority students were also not statistically different from ratings received from teammates. Roughly 6% of the CHE 225 students gave themselves at least one rating higher than any of the ratings they received from their teammates. None of them earned a higher course grade as a consequence of his or her self-rating.

A greater concern than inflation of self-ratings may be *deflation*. Fourteen percent of the CHE 225 students gave themselves lower ratings than they received from any of their teammates. One of them claimed that while his teammates believed he was well prepared and cooperative, he himself knew he could have done better. Fortunately, the course grades received by these students were not affected by their modest self-ratings. In short, as long as peer ratings are not given excessive weight in course grading, situations in which inflated or deflated self-ratings affect course grades are unlikely to arise.

Gender differences in ratings.

Ratings given by men and women to their teammates are summarized in Table 2. (Self-ratings are not included in these tabulations.)

Table 2
Gender Differences in Peer Ratings

	CHE 205			CHE 225		
Average ratings given	N	Rating	p	N	Rating	p
By men	244	87.6	.76	145	89.7	.81
By women	91	85.7		59	87.5	
To men	246	87.5	.67	145	89.7	.19
To women	89	86.0		59	87.7	
By men to men	205	87.1	.24	114	90.1	.48
By men to women	39	90.1		31	88.3	
By women to women	50	89.3	.07	28	87.1	.30
By women to men	41	82.8		31	87.9	

There were no statistically significant differences between the ratings men and women received from their teammates, in the ratings they gave to their teammates in either course, nor in the ratings men gave to women. (Self-ratings were not included in these tabulations.) The only marginally significant gender-related effect was that women gave lower ratings to other women than they gave to men in CHE 205, a result that reflected very low ratings within an all-female group that had considerable difficulty working together from the beginning of the semester. These data strongly suggest that gender bias was not a factor in the peer ratings collected in this study.

Effects of ethnicity on ratings.

Minority students entered CHE 205 and CHE 225 with slightly lower prior grade point averages than those of their non-minority counterparts (87.2 vs. 91.3 in CHE 205, 91.1 vs. 92.4 in CHE 225), and earned significantly lower test grades in CHE 205 (62.0 vs. 78.0, p=0.005) and slightly lower test grades in CHE 225 (77.8 vs. 81.3, not significant). Ratings given by minority and non-minority students to their teammates are summarized in Table 3.

Table 3
Ethnicity Differences in Peer Ratings

	CHE 205			CHE 225		
Average ratings given	N	Rating	p	N	Rating	p
By non-minorities	297	86.9	.81	186	94.4	.05
By minorities	38	88.5		18	88.6	
To non-minorities	294	87.7	.37	186	90.3	.0004
To minorities	41	82.9		18	77.1	
By non-minorities to non-minorities	26	87.6	.19	16	89.8	.0008

	2			8		
By non-minorities to minorities	35	81.4		18	77.1	
By minorities to non-minorities	32	87.9	.41	18	94.4	—
By minorities to minorities	6	91.7		—	—	

In CHE 205, minority students gave higher but received lower ratings than did non-minorities; in CHE 225, minority students both gave and received lower ratings. None of these differences was statistically significant. Non-minority students gave lower ratings to minority students than to other non-minority students, with the difference being highly significant in CHE 225. Minority students gave slightly higher but not significantly different ratings to other minority students than to non-minority students in CHE 205, and with no more than one minority student per group in CHE 225, minority students did not have the opportunity to rate other minority students in the class. Minority students also gave themselves lower self-ratings in CHE 225 than did non-minority students (87.5 versus 89.1, respectively), but the difference was not statistically significant.

The relatively low ratings given by non-minorities to minorities could have several explanations:

1. Students with lower ratings tended to have a lower mastery of the course material, and hence were seen as contributing less to the team problem-solving efforts.

2. Students with lower ratings tended to be relatively passive or reticent in group sessions, and so were perceived to be contributing less to the group their more vocal teammates contributed.

3. Students with lower ratings tended to approach teamwork with lower levels of commitment.

4. The ratings were influenced by racial bias.

Although we can neither confirm nor refute any of these hypothetical explanations on the basis of available data, we have reason to believe that the first two were likely contributors to the outcomes. Minorities tended to earn lower test grades than non-minorities, increasing the likelihood that their contributions would be perceived to have less value by their teammates, and the correlation between ratings and test grades was extremely high for minorities ($R = 0.79$). Also, research studies have shown that members of minority cultures tend to play more passive roles in mixed groups, especially if they are outnumbered in those groups. (See Felder *et al.*, 1995.) In short, while racial bias cannot be ruled out as a possible influence on peer ratings, other explanations for the observed results seem more likely.

Incidence of identical ratings.

Usually the first concern raised about peer rating methods is that many or most teams will agree among themselves to give everyone the same high rating. (With the RMIT system, it makes no difference which rating they settle on, since the grade adjustment factor will be 1.0 regardless of their choice.) There is nothing necessarily wrong with team members reaching such an

agreement. Their doing so would suggest that the team was functioning well, with everyone working responsibly throughout the semester, and so for every team member to receive the same project grade would be a perfectly reasonable result. In any event, such agreements were certainly not widespread in this study and there may well have been none of them. Identical peer ratings were submitted by only two CHE 205 groups (6% of all groups) and two CHE 225 groups (11%).

Use of ratings to identify hitchhikers.

"Hitchhikers" in cooperative learning terminology are team members who shirk their responsibilities to the team. Unless measures like peer ratings are instituted to assure individual accountability, hitchhikers receive the same grades as the more industrious group members who do the bulk of the work, and so get a "free ride." Educators who have reservations about cooperative learning often cite the possibility of successful hitchhiking as a drawback of the approach.

Peer rating provides a mechanism for identifying hitchhikers in a course. Many students are inclined to cover for teammates who occasionally miss team meetings or fail to contribute to problem solutions; however, they are unlikely to give good ratings to students who chronically fail to participate in team efforts. Granted, shirking responsibility is only one of several possible causes for low ratings: they may also be received by students who attempt to dominate their teammates or by bright students who do all of the work themselves and refuse to involve their teammates in the effort. However, our experience is that consistently low ratings are most likely to be given to students who are perceived as failing to pull their weight on the team.

In this study, we define hitchhikers to be students whose average peer ratings are less than 75—i.e., students whose citizenship is rated as less than satisfactory by their teammates. The incidence of such students in both classes was very low (see Table 4).

Table 4
Incidence of Probable Hitchhikers

	70 ≤ Test Average ≤ 100	Test Average < 70
CHE 205	< 1% (N=1)	6.1% (N=8)
CHE 225	4.3% (N=3)	2.9% (N=2)

Roughly 7% of the students in each class received less than satisfactory average ratings from their teammates. The average test scores for 10 of the 14 students in this category were below a "C" level (the level required to advance to the next course in the curriculum). The common concern that cooperative learning inevitably leads to widespread hitchhiking and that the hitchhikers earn undeservedly high grades was not borne out in this study. (We would speculate that the peer rating system minimized the incidence of hitchhiking, but we have no way to prove it.)

Student complaints

Almost anything an instructor can do in a class—lecture or require active student participation in class, assign groupwork or give only individual assignments, give unadjusted team grades or use peer ratings to adjust the grades for individual effort—is likely to lead to complaints from isolated students. Complaints become a matter of concern only if they are voiced by a significant fraction of the students taking a course.

We acknowledge that widespread objections to peer ratings (and to cooperative learning, for that matter) might occur in some circumstances, but they did not occur in the courses described in this study. Only one negative comment about the peer rating system appeared in mid-semester and final course evaluations in either CHE 205 or CHE 225. At the end of the first semester several students questioned their lowered homework grades, but they then admitted that they missed several group meetings and came to others unprepared, and acknowledged that they had been warned about the possible consequences of such behavior.

We would speculate that the use of peer ratings is likely to *reduce* the number of student complaints. When students know that hitchhikers will not receive the same grade as responsible team members, they are much less inclined to complain about the unfairness of cooperative learning. We are also inclined to believe that the low incidence of hitchhiking observed in this study might have been due in part to the knowledge that the hitchhikers would be penalized in some manner.

Conclusions and Recommendations

This study examined the application of a peer rating system in two sophomore chemical engineering courses in which the students completed homework assignments in cooperative learning teams. The students were instructed to rate how well their teammates and (in one of the courses) they themselves fulfilled their responsibilities to their teams. The ratings were used to determine individual grades for homework completed in teams.

We found the peer rating procedure used in this study to be easy to administer and to use for team grade adjustments. It provides a modest reward to students who go above and beyond the minimum required individual effort in teamwork, and effectively identifies hitchhikers and keeps them from getting full credit for work done primarily by their teammates.

Differences between peer ratings and self-ratings were insignificant. Only two teams in each class (out of 39 teams in one course and 18 teams in the other) submitted identical ratings for all team members. No evidence of gender bias in the ratings was detected. Non-minority students on average gave lower ratings to minority students than to other non-minority students, with the difference being statistically significant in the second course. Racial bias could have been a factor in the latter result, but alternative explanations suggested in the paper are considered more likely. Student complaints about having their grades influenced by peer ratings were almost non-existent. Many commonly expressed concerns about peer ratings in cooperative learning were thus not borne out by the results of the study, although many more studies with much larger populations would be required for the concerns to be definitively dismissed.

We believe that as successful as this experience was, more could be done to make peer ratings as effective as they could be. Above all, we recommend providing the students with more guidance and practice in assigning ratings than we provided in this study. In their excellent reference on cooperative learning in higher education, Millis and Cottell (1998) show a peer evaluation form that assigns numerical ratings to four different components of effective teamwork: attending meetings on a regular basis, making an effort at assigned work, attempting to make contributions and/or to seek help within the group when needed, and cooperating with the group effort. We are currently using results from a modified version of this form to derive quantitative definitions of the RMIT system terms ("Excellent"…"No Show") that should make peer ratings more objective and less subject to personal feelings. We also suggest taking some time in class to present several team scenarios, have the students fill out rating sheets for the hypothetical team members, and then discuss the ratings and reach consensus on what they should be.

PEER RATING OF TEAM MEMBERS

Name_____ Group #_____

Please write the names of all of your team members, INCLUDING YOURSELF, and rate the degree to which each member fulfilled his/her responsibilities in completing the homework assignments. The possible ratings are as follows:

Excellent	Consistently went above and beyond—tutored teammates, carried more than his/her fair share of the load
Very good	Consistently did what he/she was supposed to do, very well prepared and cooperative
Satisfactory	Usually did what he/she was supposed to do, acceptably prepared and cooperative
Ordinary	Often did what he/she was supposed to do, minimally prepared and cooperative
Marginal	Sometimes failed to show up or complete assignments, rarely prepared
Deficient	Often failed to show up or complete assignments, rarely prepared
Unsatisfactory	Consistently failed to show up or complete assignments, unprepared
Superficial	Practically no participation
No show	No participation at all

These ratings should reflect each individual's level of participation and effort and sense of responsibility, not his or her academic ability.

Name of team member	**Rating**
_____	_____
_____	_____

_____ _____

_____ _____

Your signature: _____

Figure 1. Peer rating form

References

Brown, R. W. (1995). Autorating: Getting individual marks from team marks and enhancing teamwork. *1995 Frontiers in Education Conference Proceedings*. Pittsburgh, IEEE/ASEE, November 1995. For a reprint, contact Rob Brown at rwb@rmit.edu.au.

Felder, R.M. (1995). A longitudinal study of engineering student performance and retention. IV. Instructional methods and student responses to them. *J. Engr. Education, 84*(4), 361–367. This paper may be viewed at < www2.ncsu.edu/unity/lockers/users/f/felder/public/Papers/long4.html >.

Felder, R.M. (1998). A Longitudinal Study of Engineering Student Performance and Retention. V. Comparisons with Traditionally-Taught Students. *J. Engr. Education, 87*(4), 469–480. This paper may be viewed at < www2.ncsu.edu/unity/lockers/users/f/felder/public/Papers/long5.html >.

Johnson, D.W., R.T. Johnson, and K.A. Smith. (1998). Active learning: Cooperation in the College Classroom. Edina, MN: Interaction Book Co.

Millis, B.J., and P.G. Cottell, Jr. (1998). *Cooperative learning for higher education faculty*, p. 198. Phoenix: Oryx Press.

DEBORAH KAUFMAN
Deborah Kaufman is currently a doctoral student in the Department of Chemical Engineering at North Carolina State University. She received her B.S. in Chemical Engineering from Cornell University. Her dissertation is in the area of bioseparations, and she is also participating in N.C. State's Preparing the Professoriate mentorship program. She expects to complete her degree requirements in August 1999 and is currently seeking an academic position.

RICHARD FELDER
Richard Felder is Hoechst Celanese Professor of Chemical Engineering at North Carolina State University and Faculty Development Codirector of the NSF-sponsored SUCCEED Coalition. He received his B.Ch.E. in Chemical Engineering from the City College of New York, and his M.A. and Ph.D. in Chemical Engineering from Princeton University. He is co-author of *Elementary Principles of Chemical Processes* (Wiley, 1978, 1986, 2000), a regular

contributor to engineering education journals, a Fellow Member of the ASEE, and codirector of the National Effective Teaching Institute.

HUGH FULLER
Hugh Fuller is Director of Educational Assessment for the North Carolina State University College of Engineering. Prior to taking this position, he was Director of Institutional Research and Director of the N.C. State Academic Skills Program. His current interests include assessing attitudes toward engineering and confidence levels of first-year engineering students.