

## Peer Ratings Revisited: Focus on Teamwork, Not Ability

Richard A. Layton, Matthew W. Ohland  
Rose-Hulman Institute of Technology / Clemson University

### Abstract

In a previous study, we determined that student peer ratings used to assign individual grades from group grades showed no effects relating to gender but significant effects relating to race. A likely explanation of this result was that students seem to base ratings on perceived ability instead of real contribution to the group effort. To overcome this tendency, we modified the peer-rating instrument, instructed students on the behavioral characteristics of good teamwork, and focused the student peer ratings on those characteristics. This paper presents an analysis of peer ratings using the modified instrument and compares these results to the results of the previous study. Using the new approach, student peer ratings showed no effects relating to race, suggesting that the new approach has the desired effect. However, contrary to prior studies, a gender bias is observed. In this study, women rate other women much lower than they rate men, although men's and women's average course grades are not significantly different. We conclude that our efforts to focus student peer ratings on real contributions to the group effort have been only partially successful.

### I. Introduction

A well-known drawback of student group work is the possibility that one or two students will do most of the work and that one or more students will “hitchhike,” that is, fail to do their share of the work but get the same high grade as the rest of the group. This problem can be addressed in part by using the peer-evaluation system described by Brown<sup>1</sup> for assigning individual grades based on a group grade. For this technique to be effective, groups should be assigned and coached by the instructor according to the established practices of cooperative learning.

In recent studies, Kaufman, Felder, and Fuller<sup>2</sup> (KFF) and Layton and Ohland<sup>3</sup> (L&O I) examine the incidence of hitchhiking and other aspects of group work with the aim of addressing common concerns about the validity of peer evaluations. The main difference between the two studies are that the students in KFF are primarily non-minorities (89%) while the students in L&O I are primarily African-American (87%). “Minorities” includes African-American, Hispanic, and Native American students; “non-minorities” includes students of all other ethnic backgrounds.

The results of these two prior studies—one at a majority-black institution (NC A&T) and the other at a majority-white institution (NC State)—are consistent, showing no effects relating to gender but significant effects relating to race. While racial prejudice could not be ruled out, a more likely explanation of this result was that students tend to give low ratings to those who are

weaker academically. Students seem to base ratings on perceived ability instead of real contribution to the group effort. To overcome this tendency, we modified the peer-rating instrument, instructed students on the behavioral characteristics of good teamwork, and focused the student peer ratings on those characteristics. This paper presents an analysis of peer ratings using the modified instrument and compares these results to the results of the two previous studies.

## II. Class and team demographics

Peer evaluation data from two sections of a single course taught by Layton in Spring 2000 are included in this study. The course is MEEN 337 Dynamics, a core course for most engineering majors at North Carolina A&T State University. Student demographics are shown in Table 1, where “L&O II” represents this study. Data from L&O I and KFF are shown for comparison.

Table 1: Demographic data.

Study	<i>N</i>	Men	Women	Non-minorities	Minorities	Institution
L&O II	70	85%	15%	10%	90%	NC A&T
L&O I	70	73%	27%	13%	87%	NC A&T
KFF	208	70%	30%	89%	11%	NC State

Here, *N* is the number of students receiving final course grades. In all three studies, less than 1% of the students are Hispanic and Native American—the “minority” students are predominantly African American. This study (L&O II) has about half as many women as our previous effort (L&O I), and the inverse ratio of minorities to non-minorities compared to KFF.

On the first day of class, students are asked to fill out a questionnaire indicating their GPA, gender, course grade in a prerequisite course, and whether they are repeating the course. A 7-day scheduling table is included on which students indicate times that they cannot meet for group work. All information is voluntary except a signature verifying that prerequisites have been satisfied. The instructor uses this information to form teams according to the following guidelines, based on Felder et al.<sup>4</sup> Final composition of the teams is shown in Table 2.

- Groups of 3 or 4, selected by instructor.
- Women and minorities are not outnumbered in a group.
- Heterogeneous ability level using GPA and grade in prerequisite course.
- Heterogeneity of major discipline: mechanical, electrical, civil, and so forth.
- Time available during the week for group work.

Table 2: Team composition.

	L&O I	L&O II
Category	Number	Number
Total	21	17
All female	1	
All male	11	11
Mixed gender	9	6
All minorities	12	11
Mixed ethnicity	9	6

### III. Team activities and the cooperative learning environment

Teams are formed in this course for homework. (Teams in KFF were formed primarily for homework and teams in L&O I were formed for design projects.) Peer ratings are performed twice in the semester, one at approximately midterm, the second near the end of the semester. The second set of peer ratings is analyzed here.

The cooperative learning environment is not fully developed in this course, although some of its elements are present. The five criteria of cooperative learning—positive interdependence, individual accountability, face-to-face interaction, appropriate use of interpersonal skills, and regular self-assessment of group functioning—are developed in this course to the degree described below.

1. Positive interdependence is promoted by awarding extra points on exams based on team performance. For the first exam, if a team average test score is greater than 75, each team member's exam score is increased by 5 points. On the second exam, if any member of the team scores at least 10 points higher than his or her raw score on the first exam, each team member's exam score is increased by 5 points.
2. Individuals are held accountable by using peer ratings to assign individual grades from a group homework grade.
3. Face-to-face interaction is not always guaranteed.
4. Students receive some instruction in interpersonal skills whenever the peer rating system is discussed.
5. Regular assessment of group functioning is accomplished by formal evaluations twice in the semester and informal discussions outside of class as required.

### IV. Peer rating procedures

The peer evaluation system is an adaptation of the one advocated by Brown<sup>1</sup>, in which students use a prescribed list of terms such as “excellent,” “very good,” “satisfactory,” and so forth to evaluate one another's contributions to the team's deliverables. In administering the peer ratings, the instructor spends about 5 minutes of class time discussing the meaning and purpose of the ratings. Students are given 10-15 minutes to complete the rating forms, which are folded when collected to maintain confidentiality. Students that need more time are allowed to turn them in later that day.

The verbal ratings are converted to a numerical equivalent and an individual's weighting factor is the individual's average rating divided by their group's average. An individual student's grade is the group grade multiplied by this weighting factor. A maximum factor of 1.07 was imposed to prevent students from receiving artificially high grades due to having a teammate with very low ratings. (KFF used a maximum factor of 1.10.)

The first evaluation (at approximately midterm) allows the groups to identify both “hitchhikers” and “overachievers,” that is, group members that are contributing either too little or too much to the group effort. The instructor meets with such groups outside of class to help them find ways to more evenly distribute the work load and to help resolve interpersonal difficulties and time con-

flicts. Students are encouraged to view the first evaluation as a chance to identify areas of improvement.

## V. What we did differently

Prior studies have noted that students tend to give low ratings to those of their peers who are weaker academically. Students seem to base ratings on perceived ability instead of real contribution to the group effort. To try to overcome this tendency, we modified the peer rating instrument and procedure.

The first change to the peer rating instrument was the listing of behavioral characteristics of good teamwork such as attending scheduled meetings, contributing to discussions, attempting to communicate clearly and with civility, and so forth. The complete list is shown on the peer rating instrument included at the end of this paper. Incidentally, as we reported in a related prior study,<sup>5</sup> the use of such behavioral characteristics can lead to a more reliable instrument, that is, one that measures some characteristic in a more consistent and repeatable way. Reliability, however, is not the same as validity, which investigates the question whether we are measuring what we seek to measure. So while there is evidence that focusing students' attention on the behavioral characteristics of good teamwork helps us measure *something* more consistently, we cannot claim with certainty that we are measuring what we seek to measure—individual contributions to the team effort.

The second change to the peer rating instrument, following the suggestion of KFF, is a description of what the ratings (excellent, very good, satisfactory, etc.) mean in terms of individual performance as well as an admonition that the ratings should reflect “each individual's level of participation, effort, and sense of responsibility, not his or her academic ability.” The rating descriptions and the focus on participation instead of ability are discussed with the students just before the peer-rating instrument is administered. The purpose of the discussion is to alert students to their responsibilities as evaluators and to point out that because they will be doing similar evaluations of coworkers in industry, they should treat the exercise seriously, as a component of their preparation for professional practice.

## VI. Data analysis

All reported levels of significance are derived from a nonparametric Wilcoxon (Mann-Whitney) rank-sum test,<sup>6</sup> with statistical significance defined by  $p < 0.05$ . Ratings are reported on a 100-point scale. The most noteworthy results are those relating to race and gender.

Ratings given and received by minorities and non-minorities are summarized in Table 3. None of the differences in ratings are statistically significant; thus, the racial bias in peer ratings observed in the two previous studies is not observed in this study. Note that minority students gave non-minority students the same average rating (89.2) as non-minority students gave minority students. In L&O I, minority students gave non-minority students an average rating of 96.5 while non-minority students gave minority students an average rating of 79.8—a 16-point difference related to race. Using the new peer rating instrument and procedure, this difference did not appear.

Table 3: Ratings by race/ethnicity

Average ratings given	<i>N</i>	Rating	<i>p</i>
By minorities	137	84.1	0.18
By non-minorities	15	89.2	
To minorities	137	84.1	0.18
To non-minorities	15	89.2	
By minorities to minorities	122	83.5	0.15
By minorities to non-minorities	15	89.2	
By non-minorities to minorities	15	89.2	-
By non-minorities to non-minorities	-	-	

To further illuminate this finding, student course grades with comparable results from L&O I (course grades) and from KFF (average test scores) are shown in Table 4. In all cases, non-minority students have higher grades than minority students do and in all but one class (KFF's CHE 225) the difference is significant. The prior studies, in which racial bias was observed, concluded that students tend to give low ratings to those who are weaker academically. In this study, a significant difference in average grades still exists between minority and non-minority students, but the previously observed racial bias is absent.

Table 4: Grades by race/ethnicity.

Study	Course		Average grade	<i>p</i>
L&O II	MEEN 337	non-minorities	80.3	0.02
		minorities	64.9	
L&O I	MEEN 440	non-minorities	84.0	0.0003
		minorities	76.8	
KFF	CHE 205	non-minorities	78.0	0.005
		minorities	62.0	
	CHE 225	non-minorities	81.3	0.21
		minorities	77.8	

Ratings are summarized by gender in Table 5. Contrary to both KFF and to L&O I, a significant effect relating to gender appears. Ratings given to men are 9 points higher than ratings given to women, but this difference is neither practically nor statistically significant ( $p=0.09$ ). More strikingly, the difference between the ratings given by women to women and those given by women to men is not only statistically significant ( $p=0.02$ ), but is a gap of 20 points. Neither we nor Kaufman et al. observed such an effect in previous studies.

Table 5: Ratings by gender

Average ratings given	<i>N</i>	Rating	<i>p</i>
By men	126	85.1	0.89
By women	26	82.2	
To men	126	86.1	0.09
To women	26	77.4	
By men to men	110	85.6	0.58
By men to women	16	82.0	
By women to men	16	89.8	0.02
By women to women	10	70.0	

To further illuminate this finding, we again examine apparent academic ability indicated by student grades. Table 6 shows student course grades by gender. Here, the 6-point difference in grades between men and women is not significant ( $p=0.49$ ).

Table 6: Grades by gender.

Study	Course		Average grade	p
L&O II	MEEN 337	men	67.7	0.49
		women	61.3	

Thus we find a significant gender-based bias in peer ratings that is apparently unrelated to academic ability. Men and women receive on average the same grades, yet the peer ratings given by women show a 20-point bias in favor of men.

Possible reasons for women rating other women so poorly are:

1. Gender-based prejudice in women.
2. Perception of ability based on discipline. Students in L&O I were all mechanical engineering majors and students in KFF were all chemical engineering majors. Students in this study are from several engineering disciplines (mechanical, electrical, civil, and so forth) and we know that some students and faculty deem certain engineering disciplines more demanding than others. Peer ratings may have been influenced by such prejudices.
3. Perception of ability based on age and experience. Students in L&O I were all at essentially the same point in their academic career. Students in this study ranged from first-semester juniors to graduating seniors (and one graduate student as well). Peer ratings may have been influenced by prejudices based on academic level.

The previous studies tend to refute explanation 1. Peer-rating gender-bias was not observed in KFF and L&O I. Explanations 2 and 3 are proposed because the basic difference between the students in this study and the students in L&O I and KFF is that the students in both prior studies were homogeneous with respect to major discipline and academic level while the students of this study are from different disciplines and are at different academic levels. This does not explain, however, why a bias appears in women and not in men. There were half as many women in this study compared to the two prior studies (see Table 1) but we do not know how this might have affected the results. It is likely that existing research in the broader area of gender studies can illuminate this result.

## VII. Conclusions

Peer ratings are used to assign individual grades from a group grade given to instructor-assigned homework teams. To overcome an apparent student tendency to give low ratings to those who are weaker academically, we modified our peer-rating instrument, and focused the students' attention on the behavioral characteristics of good teamwork and on the relevance of such instruments to professional practice. The results, in contrast to prior studies, show no racial bias in peer rating. We would like to conclude that our modifications to the process brought about this desirable result. However, a gender bias appeared in this study that was not observed previously. In this study, women rate other women much lower than they rate men, although men's and women's average course grades are not significantly different. Possible explanations for the appearance of this gender bias where none was observed before are perceptions of ability based on

discipline or academic level, and that this class had half as many women as in the prior studies. Additional work is required to determine which of these possibilities, if any, explain the observed results. We conclude that our efforts have been partially successful in addressing the issue of perceived ability in peer ratings.

#### Bibliography

1. Brown, R.W. (1995). Autorating: Getting individual marks from team marks and enhancing teamwork. In proc. *Frontiers in Education Conference*. IEEE/ASEE, Pittsburgh, November 1995.
2. Kaufman, D.B., Felder, R.M., and Fuller, H. (2000). Accounting for individual effort in cooperative learning teams. *J. Engineering Education*, Vol. 89, No. 2, pp. 133-140.
3. Layton, R.A. and Ohland, M.W. (2000). Peer evaluations in teams of predominantly minority students. In proc. *ASEE Annual Conference*. ASEE, St. Louis, June 2000.
4. Felder, R.M., Stice, J.E. and Brent, R. (1999). Implementing formal cooperative learning. In handbook, National Effective Teaching Institute. ASEE, Charlotte, June 1999.
5. Ohland, M.W. and Layton, R.A. (2000). Comparing the reliability of two peer evaluation instruments. In proc. *ASEE Annual Conference*. ASEE, St. Louis, June 2000.
6. Shavelson, R.J. (1988). *Statistic Reasoning in the Behavioral Sciences*, 2nd ed. Allyn and Bacon, Inc., Boston, p. 608.

#### RICHARD A. LAYTON

Richard A. Layton received his Ph.D. from the University of Washington in 1995, taught for four years at North Carolina A&T State University, and is currently an Assistant Professor of Mechanical Engineering at Rose-Hulman Institute of Technology with a specialty in system dynamics. Previously, Dr. Layton worked for 12 years in industry, culminating as a technical group head and project manager. He is a registered Professional Engineer.

#### MATTHEW W. OHLAND

Matt Ohland received his Ph.D. in Civil Engineering with a minor in Education from the University of Florida in 1996. He began as an Assistant Professor in Clemson University's General Engineering program in Fall 2000 after serving as the Assistant Director of the NSF-sponsored SUCCEED Engineering Education Coalition and as an NSF postdoctoral fellow. His research is primarily in freshman programs and educational assessment.

**Peer ratings of team members**

Name \_\_\_\_\_

ME 337 Dynamics  
R. Layton

Date: \_\_\_\_\_

*Please write the names of all the members of your team, INCLUDING YOURSELF, and rate the degree to which each member fulfilled his or her responsibilities. Such responsibilities include:*

1. Attending scheduled meetings.
2. Contributing to discussions.
3. Attempting to communicate clearly and with civility.
4. Listening effectively.
5. Accepting criticism gracefully.
6. Completing tasks fully and on time.

*Your responses are used to assign individual grades from the group grades. Your responses are confidential. The possible ratings are:*

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

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

Name of team members (including yourself)	Rating
_____	_____
_____	_____
_____	_____
_____	_____

Your signature \_\_\_\_\_

**Ref.**

Kaufman, D.B., Felder, R.M., and Fuller H., Peer ratings in cooperative learning teams, in proc. *ASEE Annual Conf.*, Charlotte (Jun 1999).

Ohland, M.W. and Layton, R.A., Comparing the reliability of two peer evaluation instruments, in proc. *ASEE Annual Conf.*, St. Louis (Jun 2000).