Peer Evaluations in Teams of Predominantly Minority Students

Richard A. Layton, Matthew W. Ohland
North Carolina A&T State University / University of Florida

Abstract

This paper presents an analysis of student peer evaluations in project teams where the majority of the students are African-American. Peer evaluations were used to assign individual grades from group grades for design projects in a junior-level mechanical engineering course taught by Layton for three semesters in 1997-99. This study is similar to and complements a 1999 study by Kaufman, Felder, and Fuller. The results of the two 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/ethnicity. We concur with Kaufman et al. that, while racial prejudice cannot be ruled out, a more likely explanation of this result is that students tend to give low ratings to those who are weaker academically. Students seem to base ratings on perceived abilities instead of real contributions. To overcome this tendency, we suggest that instructors teach the behavioral characteristics of good teamwork and focus student peer evaluations on those characteristics.

I. Introduction

A well-known drawback of group projects in engineering 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 or "autorating" system described by Brown[1] 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 a recent study, Kaufman, Felder, and Fuller[2] 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 acronym "KFF" is used here to refer to this paper.) Their results include: 1) no gender bias in peer evaluations was detected; and 2) minority students on average receive lower ratings and give higher ratings than non-minority students, with the differences being statistically significant in one class but not in another class. They conclude that racial bias could be a factor, but alternative explanations are considered more likely. "Minorities" includes African-American, Hispanic, and Native American students; "non-minorities" includes students of all other ethnic backgrounds. Most of the students in the KFF study are non-minorities.

In this paper, we complement the KFF study by analyzing peer evaluations in classes where most of the students are minorities. In a mechanical engineering design course at North Carolina A&T State University, students were assigned to groups for term projects in design. Peer evaluations, based on Brown's autorating idea, were used to assign individual grades from group grades. This paper presents an analysis of these peer evaluations for three offerings of this course.

The main differences between KFF and this study are that the students in KFF are primarily non-minorities (89%) while our students are primarily African-American (87%); in KFF, group work is mainly homework while our group work is mainly design projects; and in our classes, the co-operative learning environment is not as developed as in KFF.

II. Class and team demographics

Peer evaluation data from three offerings of a single course taught by Layton in 1997, 1998, and 1999 are included in this study. The course is MEEN 440 Mechanism Design and Analysis, a required junior-level design course in Mechanical Engineering at North Carolina A&T State University. Student demographics are shown in Table 1, where "L&O" represents Layton & Ohland, the authors of this study. Data from KFF is shown for comparison.

Table 1: Demographic data.

Study	*N*	Men	Women	Non-minorities	Minorities
L&O	70	73%	27%	13%	87%
KFF	208	70%	30%	89%	11%

Here, *N* is the number of students receiving final course grades. In both studies, less than 1% of the students are Hispanic and Native American—the "minority" students are predominantly African American. Compared to KFF, we have essentially the same ratio of men to women but the inverse ratio of minorities to non-minorities.

On the first day of class, students are asked to fill out a questionnaire indicating their GPA, gender, course grade in a prerequisite course (MEEN 337 Dynamics), and whether they are repeating the course. They are required to sign a statement indicating that they have satisfied the prerequisites. In addition, a 7-day scheduling table is included on which students indicate times that they cannot meet for group work. All information is voluntary except the signature verifying prerequisites. The instructor uses this information to form project teams according to the following guidelines, c.f. Felder et al.[3]

- Groups selected by instructor.
- Groups of 3 or 4 members.
- Women and minorities are not outnumbered in a group.
- Heterogeneous ability level using GPA and grade in prerequisite course.
- Time available during the week for group work.

Final composition of the teams is shown in Table 2.

Table 2: Team composition.

Category	Number
Total	21
All female	1
All male	11
Mixed gender	9
All minorities	12
Mixed ethnicity	9

III. Team activities and the cooperative learning environment

Teams are formed in this course for design projects. Project deliverables are a final written report and an oral presentation. In a typical semester, the teams work together early in the semester on an introductory design project in mechanism kinematics. Peer evaluations are used to assign individual grades from group grades. Later in the semester, the teams do one or two computer-lab projects in mechanism simulation and an occasional group quiz in class. In the second half of the semester, the groups are given a second design project of greater scope. It is the peer evaluations from this second project that are analyzed here.

The cooperative learning environment is not fully developed in this course, although some 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 arbitrarily selecting students to give the oral presentations for both projects. Students are warned well in advance that the instructor will select speakers by pulling names out of a cup just before class the day of the presentations.
2. Individuals are held accountable by using peer evaluations to assign individual grades from a group project grade.
3. Face-to-face interaction is not always guaranteed.
4. Students receive some instruction in interpersonal skills. The instructor uses the remainder of the class period after the oral presentations to give a general critique of content and style and to facilitate a discussion of the difficulties of working in groups and the necessity of mastering these skills. Also, when the opportunity arises, visiting engineers from industry are invited to class to answer student questions about engineering practice. These sessions often include discussion of interpersonal skills.
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[1], 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. The verbal ratings are converted to a numerical equivalent and an individual’s weighting factor is the individual’s average rating divided by the group 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 peer evaluation instrument varied from semester to semester as improvements were tried. The results from the different instruments are normalized to a common 0-100 scale for comparison. (In a companion paper, it is shown that reliability coefficients for two of these instruments are similar enough to be normalized for comparison[4].) Our students, like those in KFF, never asked how these ratings are used to adjust their project grades.

Peer evaluations are performed at the conclusion of each of the two projects. The evaluations at the end of the first project accomplish several goals:

- Students learn about the peer evaluation procedure.
- Students reflect on the evaluation criteria.
- Students get feedback on how the group assesses each member’s work to date.
- The instructor is alerted if groups are not functioning.

This first evaluation 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 conflicts. Also, because the first project is weighted less than the second project in computing the final course grade, a poor peer evaluation after the first project has only a modest impact on a student’s course grade. Students are encouraged to view this first evaluation as a chance to identify areas of improvement.

V. Data analysis

All reported levels of significance are derived from a nonparametric Wilcoxon (Mann-Whitney) rank-sum test[5] unless noted otherwise, with “statistically significant” defined as $p<0.05$. The most noteworthy results are those relating to the peer ratings to gender, and race/ethnicity. Ratings given are summarized by gender in Table 3. Contrary to our expectations, there is no statistically significant gender-bias in the ratings. The ratings given by men to men compared to those given by men to women are the closest to statistical significance, but the true difference of only 3 rating points has no practical significance in any case.

Table 3: Ratings by gender

Average ratings given	*N*	Rating	p
By men	115	89.6	0.50
By women	50	87.2	
To men	117	89.4	0.19
To women	48	87.6	
By men to men	87	90.4	0.07
By men to women	28	87.1	
By women to men	30	86.6	0.93
By women to women	20	88.4	

Ratings given and received by minorities and non-minorities are summarized in Table 4. The differences here are statistically significant. Minorities give significantly higher ratings and receive significantly lower ratings. Non-minorities give significantly lower ratings and receive significantly higher ratings. The highest ratings are those given by minorities to non-minorities (96.5) and the lowest are those given by non-minorities to minorities (79.8).

Table 4: Ratings by race/ethnicity

Average ratings given	*N*	Rating	p
By minorities By non-minorities	145 20	90.2 79.6	0.0017
To minorities To non-minorities	145 20	87.8 96.4	0.012
By minorities to minorities By minorities to non-minorities	125 20	89.2 96.5	0.035
By non-minorities to minorities By non-minorities to non-minorities	20 0	79.8 -	-

These results are similar to those reported in KFF and we concur with their list of possible explanations and we've added one additional possibility (no. 6). To paraphrase KFF, students receive low ratings because:

1. they are less diligent or responsible, so low ratings were justified.
2. they are weaker academically than their teammates, and so were perceived as contributing less.
3. they tend to be passive in group sessions, and so were perceived as contributing less.
4. of statistical chance, and the observed results are not repeatable.
5. of racial prejudice.
6. of socioeconomic prejudice.

We concur with KFF, in that "we can neither confirm nor refute any of these hypothetical explanations on the basis of the available data…" If we consider student course grades, however, we find that in this study, the average non-minority course grade (84.0) is significantly (p=0.0003) higher than the average minority course grade (76.8). These results are shown in Table 5 with similar information (average test scores) from KFF shown for comparison.

Table 5: Grades by race/ethnicity.

Study	Course		Average grade	p
L&O	MEEN 440	non-minorities minorities	84.0 76.8	0.0003
KFF	CHE 205	non-minorities minorities	78.0 62.0	0.005
	CHE 225	non-minorities minorities	81.3 77.8	0.21

In KFF, non-minorities have significantly (p=0.005) higher test grades than minorities in one course (CHE 205) and higher, but not significantly higher (p=0.21), test grades in the other course (CHE 225). On average then, non-minority students are receiving higher grades. We assume that this grade differential is a result of differences in academic performance and not instructor bias, and we agree with KFF that this result supports explanation number 2.

Explanation 3 is neither supported nor refuted by our data. Just over half (12 out of 21) of our teams had all minority members. Presumably, minority students on these teams would be less likely to exhibit passive behavior than on a racially mixed team, reducing the likelihood that passivity is the cause of the lower peer ratings for minorities. And while 9 of the 21 teams were mixed-gender teams, we found no gender bias in the ratings. Thus the idea is not supported that women, as a minority culture in engineering, would exhibit passive behavior and therefore be rated lower.

Since our results generally support the findings of KFF, explanation 4 is less likely than the others are. In addition, we have no data to either support or refute explanations 5 and 6. Thus, we concur with KFF that, although biased peer ratings cannot be ruled out, the data support the conclusion that peer ratings are related to academic ability.

VI. Conclusions and recommendations

The results of this study are nearly identical to the results of Kaufman et al. First, both studies had comparable ratios of men to women and neither found evidence of gender bias in peer ratings. Second, the two studies found similar evidence of the effects of race/ethnicity on peer ratings. The highest ratings are those given by minorities to non-minorities and the lowest are those given by non-minorities to minorities. The racial composition of the two studies (KFF 88% non-minority, L&O 87% minority) had no apparent effect on this outcome. If we are to attribute this racial effect to racial bias, we would have to assert that sophomore-level chemical engineering students at NC State and junior-level mechanical engineering students at NC A&T, two universities with quite different histories, stakeholders, and missions, are racially biased in the same manner. It seems more likely that the significant differences between minority and non-minority ratings are due to the significant grade differential between the two groups. In both studies, minority students had average grades lower than average grades of non-minority students. We conclude that, although racial bias cannot be ruled out, students seem to base ratings on perceived abilities instead of real contributions.

We recommend, therefore, that instructors attempt to minimize the effects of academic ability in student peer evaluations. First, we agree with Kaufman et al. that the peer evaluation forms should include definitions of the qualitative ratings “excellent,” “very good,” “satisfactory” and the rest. Second, we suggest that instructors teach the behavioral characteristics of good teamwork and develop methods to focus the peer evaluations on those specific characteristics. A list of such characteristics might include, for example, “attends meetings regularly,” “contributes to discussions,” “listens effectively,” and “performs significant tasks.”

VII. Appendix

One of the peer evaluation instruments used in this study is shown on the following page. In recent classes, the form has been amended to include definitions of the terms “excellent,” “very good,” and so forth, as suggested in KFF.

Peer evaluation

Date:__________

Project no.__________

Group no.__________

Write down the names of all members of your group (including your own) and next to each person's name write the word from the following list that best describes that person's contribution to this project.

excellent
very good
satisfactory
ordinary
marginal
deficient
unsatisfactory
superficial
no show

Name	**Rating**

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. (1999). Peer ratings in cooperative learning teams. In proc. *ASEE Annual Conference*. ASEE, Charlotte, June 1999.
3. Felder, R.M., Stice, J.E. and Brent, R. (1999). Implementing formal cooperative learning. In handbook, National Effective Teaching Institute. ASEE, Charlotte, June 1999.
4. 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.
5. 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 and he is currently an Assistant Professor of Mechanical Engineering at North Carolina A&T State University. Prior to his academic career, Dr. Layton worked 12 years in consulting engineering, with the final five years as a group head and project manager. His technical research is in the area of dynamic systems and control. He is a registered Professional Engineer in NC.

MATTHEW W. OHLAND
Matt Ohland received his Ph.D. in Civil Engineering with a minor in Education from the University of Florida in 1996, and since then has been the Assistant Director of the NSF-sponsored SUCCEED Engineering Education Coalition. In 1998, he was awarded an NSF Postdoctoral Fellowship for Science, Mathematics, Engineering, and Technology Education, under which he is studying the impact of a wide variety of student programs on student success.