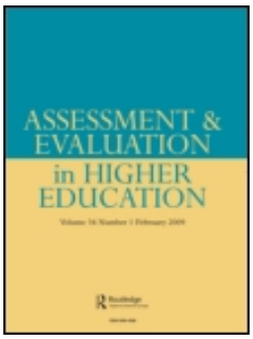

Assessment & Evaluation in Higher Education

ISSN: 0260-2938 (Print) 1469-297X (Online) Journal homepage: www.tandfonline.com/journals/caeh20

Influences on peer evaluation in a group project: an exploration of leadership, demographics and course performance

Molly Dingel & Wei Wei

To cite this article: Molly Dingel & Wei Wei (2014) Influences on peer evaluation in a group project: an exploration of leadership, demographics and course performance, Assessment & Evaluation in Higher Education, 39:6, 729-742, DOI: 10.1080/02602938.2013.867477

To link to this article: https://doi.org/10.1080/02602938.2013.867477

Published online: 16 Dec 2013.

Assessment & Evaluation in Higher Education, 2014
Vol. 39, No. 6, 729–742, http://dx.doi.org/10.1080/02602938.2013.867477

Routledge
Taylor & Francis Group

Influences on peer evaluation in a group project: an exploration of leadership, demographics and course performance

Molly Dingel[a]* and Wei Wei[b]

[a]*Center for Learning Innovation, University of Minnesota Rochester, Rochester, NY, USA;*
[b]*Mathematics Department, Metropolitan State University, Saint Paul, MN, USA*

Collaborative learning strategies are widely used in higher education to deepen learning, promote team-building skills and achieve course learning objectives. Using peer evaluation is an important strategy to ensure that engaged and active students are rewarded for their efforts, and to discourage loafing within groups. However, less is known about what biases may influence students' peer evaluations. In this paper, we investigate what variables students may (consciously or unconsciously) use to evaluate their peers. We explore the role of sex, race, course performance and group leadership on peer evaluation. We also investigate whether these variables correlate with students' final course grade. We found that students who reported being leaders in groups were evaluated higher than peers who reported being followers, and that course performance positively correlated with peer evaluations. White students received higher peer evaluations than students of colour. This difference reflects trends in group leadership and course performance, with more white students than students of colour reporting being leaders in groups and receiving higher grades.

Keywords: collaborative learning; peer ratings; peer evaluation; group work

Introduction

There is wide acceptance that collaborative/group learning strategies are useful in educational settings (Schroeder et al. 2007), with evidence indicating that these strategies increase student motivation, confidence and sense of responsibility (Caulfield and Caroline 2006; Bartle, Dook, and Mocerino 2011), as well as increase test scores and connection to classmates (Rau and Heyl 1990). There is also wide acceptance that peer evaluations provide a way for individuals to be held accountable for their efforts within the group, and provide a valid way to assign individual grades for group projects (Zhang and Ohland 2009). Evidence suggests that peer evaluations reduce frustration and disengagement from collaborative course activities, result in more equitable grades (Conway and Kember 1993; Kagan 1995; Cheng and Warren 2000) and are a predictor of positive student attitudes about group work (Pfaff and Huddleston 2003).

Though there is much interest in using collaborative learning techniques, and much effort aimed at identifying valid and effective methods of peer evaluation, there has been little work done to investigate what variables correlate with peer

*Corresponding author. Email: dinge016@umn.edu
The authors Molly Dingel and Wei Wei contributed equally to this manuscript.

evaluation and what types of biases influence evaluation. There exists only a small body of research on this topic and this paper investigates these questions.

Measures of peer rating

Researchers have explored the reliability and validity of various methods of evaluating peers in the context of collaborative learning. Kane and Lawler (1978) compare and contrast peer nomination, where group members individually designate one other group member as the highest performing on a particular trait or performance dimension; peer rating, where each group member rates every other member on a rating scale; and peer ranking, where members rate each other from best to worst on one or more factors. Kane and Lawler conclude that, despite suffering from lower validity and reliability, peer ratings are most useful for collecting feedback about specific behaviours and the way each team member performed. Importantly, most of the studies reviewed by Kane and Lawler collected data from the workplace or the military, sites where other researchers have observed relatively low inter-rater reliability, in part because subordinates and managers evaluate somewhat differently (Conway and Huffcutt 1997).

There has been far less research done on peer ratings in higher education. We might ask whether students, because of their common role as a student, might have higher inter-rater reliability. Other studies indicate that a variety of behaviourally anchored peer-rating methods has acceptable inter-rater reliability (Ohland et al. 2005; Baker 2008). Further, Ohland et al. (2005) show that the inter-rater reliability of a single-item behaviourally anchored instrument is as reliable as a 10-item instrument. Therefore, the literature seems to support the reliability of standard behaviourally anchored peer ratings in higher education. In this paper, we use 'peer ratings' interchangeably with 'peer evaluations'.

Inter-rater reliability is relatively straightforward. However, measuring the validity of evaluations presents additional challenges. Students who are similar by a variety of measures, including age, race and economic background, may hold similar biases that influence their peer ratings; therefore, ratings may be consistent but still include a set of biases. Little work has been done examining whether peer evaluations correlate with other variables – like leadership, sex, race and course performance – that may demonstrate either a weakness (in the case of sex and race bias) or a strength (in the case of leadership and course performance) of the validity of peer ratings.

Correlations with leadership

At least one study found that students who acted as leaders in a group received higher peer evaluations. In their qualitative study, Lee and Lim (2012) found that students rewarded those who exhibited 'managerial' skills with higher peer evaluations. In other words, engaging in task allocation, coordination, organisation and mediation correlated with higher peer evaluations, which would support the validity of using peer evaluations. There is a dearth of additional research, including quantitative studies, on this question.

Correlations with gender and race

There have been few studies that investigate whether gender bias emerges in student peer evaluation in higher educational settings, but Watson, BarNir, and Pavur (2010)

found no effect of gender on peer evaluation. This finding contrasts with several studies investigating gender differences in the evaluation of men and women in the workplace (Eagly, Makhijani, and Klonsky 1992). These latter studies indicate that gender does influence evaluation of leaders, though gender bias in evaluation is dependent both on context and upon whether a woman's leadership style conforms to or violates gender stereotypes and expectations. For example, researchers have found that women leaders are devalued if they break feminine stereotypes by engaging in autocratic or directive leadership styles (Eagly, Makhijani, and Klonsky 1992; Ridgeway 2001; Koenig et al. 2011). Because gender bias in evaluation is context-dependent, it can be difficult to anticipate how or when this bias will emerge.

There has been little research done on racial bias in peer evaluations in higher education. However, studies have found some evidence that peers may evaluate students of colour lower than white students (Kaufman, Felder, and Fuller 2000; Watson, BarNir, and Pavur 2010). This finding reflects research done in the workplace, with researchers identifying racial bias in performance reviews (Castilla 2008). As a whole, therefore, the degree to which race and gender influence peer evaluations in higher education remains unclear.

Correlations with course performance

Some scholars (Persons 1998; Watson, BarNir, and Pavur 2010), but not others (Dingel, Wei, and Huq 2013), found that peer evaluations correlate with performance. The positive correlations emerged even though the peer evaluation methods used in the research explicitly sought to measure cooperative efforts, or 'team citizenship', and not the quality of the work *per se*. It should be noted, however, that in Watson, BarNir, and Pavur's (2010) study exploring group testing, students had access to their teammates' individual test scores. It is plausible that students have a difficult time separating 'team citizenship' from the quality of the work, especially in cases when grades are visible. Alternatively, it is possible that the correlation is spurious, with higher performing students taking on leadership roles in the group, producing more work and/or being more conscientious, and therefore having higher performance as a team citizen. More work is needed to identify both whether students who perform better receive higher peer evaluations, and the causal factors in this relationship.

In this paper, we explore whether leadership, sex, race and course performance correlate with peer evaluation. We performed data analyses to compare students' peer evaluations with respect to sex, race, course grade and self-reported leadership in the group. Furthermore, we analysed whether the same pattern for peer evaluation holds for final course grades. That is, do we observe the same trend for course grades as for peer evaluations with respect to leadership, sex and race? If the trends for course grades are the same as for peer evaluation, we can conclude that peer evaluation reflects students' performance.

Methods

Group project overview

Students in an introductory sociology class were required to complete an interdisciplinary project using data from the US Census. Nearly, all students enrolled in this course are pursuing a Bachelor of Science in Health Sciences; in their first year,

these students enrol in a common set of courses, including statistics and sociology. This structure allows the mathematics and sociology faculty to work together to create a set of assignments that require students to use knowledge from both courses. Student consent was obtained at the beginning of the fall 2012 semester, consistent with a protocol approved by the Institutional Review Board. This study included two surveys on demographic data and students' roles in the project, final course grades, paper grades and peer evaluations from 113 (of 135) students enrolled in one of two sections of an entry-level sociology course (six students declined to participate in the study, and their teams were removed from analysis).

Instructors sought to create a set of assignments where students had to work together to actively ask questions, test a hypothesis and engage with course concepts. We organised students into teams of four to six, such that all team members were in the same section of sociology and statistics. A secondary consideration was race and sex, with instructors attempting to create diverse groups, while also not isolating minority students. Beyond these criteria, teams were randomly constituted. Our analysis draws from 22 teams formed for the project.

The project required students to analyse a subset of census data. Students were given the data-set and chose two variables about which to hypothesise a relationship. Over the course of three papers, written collaboratively, students: (1) articulated their null and alternative hypotheses and statistically described each variable; (2) tested the null hypothesis using skills learned in their statistics course; and (3) contextualised their statistical findings with sociological concepts. For every paper, each student evaluated themselves and each of their teammates. In the sociology course, these cooperative data papers made up 20% of students' course grades. Individual assessments (in-class examinations, essays, quizzes and participation) made up the remaining 80% of their grades.

Grading and peer evaluation

We calculated individual project grades based on a process described by Oakley et al. (2004). Instructors assessed each paper and assigned them a team grade. For each paper, instructors asked students to rate themselves and each of the peers on their team using a behaviourally anchored, nine-scale rating (excellent = 100, very good = 92, satisfactory = 84, ordinary = 76, marginal = 68, deficient = 60, unsatisfactory = 50, superficial = 25 and no show = 0). The scale is behaviourally anchored because, before assigning scores to their peers, students were instructed to think about the following questions: Has the student attended team meetings? Has the student made a serious effort at assigned work before the team meetings? Has the student made a serious effort to fulfil his/her team role responsibilities on assignments? Has the student notified a teammate if he/she would not be able to attend a meeting or fulfil a responsibility? Does the student attempt to make contributions in group meetings? Does the student listen to his/her teammates' ideas and opinions respectfully and give them careful consideration? Does the student cooperate with the group effort? In addition to behaviorally anchoring the scale, these questions push students to think about the 'team citizenship' of their peers.

This scale was converted into numbers, and instructors averaged the peer evaluations for each student and calculated the team peer evaluation average. Instructors then divided the individual average by the team average evaluation to calculate an

'adjustment factor' for each student, capped at 1.05. Individual students' paper grades were calculated by multiplying their adjustment factor by the team paper grade.

It should be noted that the nine-scale rating described above is different from that described by Oakley et al. (2004). We have modified the scale to make it more consistent with how instructors assign course grades, with more options (finer scaling) between 100 and 60%. In previous years, we used a scale that had even scaling between 0 and 100, and found students numerical evaluations did not match well with their verbal descriptions of their peers' work. Though anecdotal, it has been our experience that this finer scaling more accurately marries students' qualitative and quantitative assessment of their peers.

Survey

During the semester, the instructor administered two separate surveys to students, consistent with a protocol approved by the Institutional Review Board. During the first month of the course, students were asked, among other items, demographic variables, like sex, race and ethnicity (see Tables 1 and 2). Race and ethnicity were measured consistently with the US Census. In the last week of class, the instructor administered a survey where students were asked, among other items, to reflect on the three group papers and report whether they felt like a follower or a leader in the group project (Table 1). While students' leadership across and within each of the three papers may vary, asking about their leadership across the three papers pushes students to make a more holistic judgement about their leadership. This variable therefore mirrors the other variables we have chosen, like final course grade. Though students' grades on different assignments vary, final course grade provides a holistic vision of a students' performance over the entire course. In the analysis that follows, we will examine the correlation between students' peer evaluations and their final

Table 1. Relevant questions from beginning to end of semester surveys.

What is your sex?
- Male
- Female
- Other

Are you of Hispanic, Latino, or Spanish origin?
- No, not of Hispanic, Latino, or Spanish origin
- Yes, Hispanic, Latino, or Spanish origin

What is your Race? Check all that apply
- White
- Black or African American
- American Indian or Alaska Native
- Asian or Pacific Islander
- Some other race

Check all that apply*
- I felt like a follower in this group
- I felt like a leader in this group

*This question modified from (Pfaff and Huddleston 2003).

Table 2. Participant demographics ($N = 113$).

	N (%)
Sex	
Male	18(16)
Female	84(74)
No response	11(10)
Race	
White	93(83)
African American/Black	1(1)
Native American/American Indian	0(0)
Asian American/Pacific Islander	5(4)
Some other race	3(2)
No response	11(10)
Ethnicity	
Hispanic	3(3)
Not Hispanic	99(88)
No response	11(10)
Binary race	
White, not Hispanic	93(82)
Hispanic and/or students of colour	9(8)
No response	11(10)

course grades, sex, race and ethnicity, and reports of 'leadership' (feeling like a 'leader' or 'follower'). Because there were so few students of colour in the course, we converted race into a binary variable.

Statistical analysis

Peer evaluation

We performed a multiple regression analysis to detect the correlation between the dependent variable, peer evaluation and five independent variables: sex, race, being a leader, being a follower and course grade. The variables of sex, race, being a leader and being a follower are binary categorical variables; the variable course grade is a numeric variable. We used 0 for white and 1 for students of colour for the variable race, and 0 for male and 1 for female for the variable sex. We used 0 to represent a student who reported not feeling like a follower and 1 for feeling like a follower, and 0 for not feeling like a leader and 1 for feeling like a leader. Some students chose 1 for both of the two items while some students chose 0 for both of the two items. Thus, there are four types of leadership roles with which students could identify. The regression model is proposed as following:

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5$$

where α is the intercept and β_i is the coefficient associated with each variable. The variable y represents the peer evaluation score and x_i represents each of the five variables. We also conducted the analysis of multicollinearity among the independent variables and calculated the variance inflation factor (VIF) that measures multicollinearity in the regression model.

To follow up with the multiple regression analysis, we compared the peer evaluations among four student leadership roles: being a follower, being a leader, being both follower and leader, and being neither. The regression model does not

provide quantitative measurements on how the peer evaluations differ among these four leadership roles, since it only includes binary categorical variables. We observed uneven sample sizes among leaders, followers, students who report being both leaders and followers, and students who report being neither of them. The homogeneity of variance assumption and the normality assumption were both violated for using a one-way analysis of variance (ANOVA) to compare the average peer evaluations among the four leadership roles. Therefore, we used the non-parametric Kruskal–Wallis test to compare the median peer evaluations among the four leadership roles. We used multiple Mann–Whitney *U* tests, followed by the Kruskal–Wallis test, to detect pairwise differences.

Course grades

We performed a multiple regression analysis to explore the correlation between the dependent variable, final course grades and four independent variables: sex, race, reported leadership and reported following. The four variables are all binary categorical variables, with representations of 0s and 1s, and the same as those in the previous regression model for peer evaluation. The proposed model is

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$$

where α is the intercept and β_i is the coefficient associated with each variable. The variable y represents the final grade and x_i represents each of the four variables. We also conducted the analysis of multicollinearity among the independent variables and calculated the VIF that measures multicollinearity in the regression model.

To follow up with the multiple regression for the final course grades, we compared the final course grades among the four leadership roles. The test of normality showed that the data for each leadership role followed a normal distribution and the assumption of homogeneity of variances among leadership roles was satisfied. Therefore, we performed a one-way ANOVA to compare the mean final grades among leaders, followers, being both and being neither. A Tukey comparison was followed with the one-way ANOVA.

Leadership roles, race and sex

We observed that more women reported being leaders than men, and more white students reported being leaders than students of colour. We used a χ^2 test to examine whether the proportion of female leaders was significantly different from the proportion of male leaders. We used the same test to compare the proportion of white leaders and the proportion of leaders among students of colour.

We used the significance level of 0.05 for all the data analyses. All of the above analyses were performed using the IBM SPSS Statistics for Windows, Version 20.0.

Results

Peer evaluations and relevant variables

Relationship between peer evaluation and sex, race, follower, leader and course grade

Our multiple regression analysis showed that the five predictors explained 31% of the variance for peer evaluation score ($R^2 = 0.31$, $F(5, 299) = 27.36$, $p < 0.001$, 95%

CI [0.23, 0.39]). We found that the predictors race (β = −1.62, t = −1.98, p = 0.049, 95% CI [−3.24, −0.006]), reported following (β = −1.82, t = −3.71, p < 0.001, 95% CI [−2.78, −0.86]), reported leadership (β = 3.01, t = 5.38, p < 0.001, 95% CI [1.91, 4.10]) and course grade (β = 0.16, t = 3.82, p < 0.001, 95% CI [0.077, 0.24]) were significantly correlated with peer evaluation scores. Sex (β = 1.05, t = 1.75, p = 0.08, 95% CI [−0.13, 2.22]) was not significantly correlated with peer evaluation scores. The regression model was generated after removing one outlier based on the residual analysis.

Our results indicated that students of colour obtained significantly lower average peer evaluations than white students, and that course grades were significantly positively related to peer evaluation scores. We report the descriptive statistics of peer evaluations for white students and students of colour in Table 3. The results also suggested that followers obtained significantly lower average peer evaluations (M = 94.43, SD = 5.50) than non-followers (M = 97.65, SD = 3.09), and that leaders obtained significantly higher average peer evaluations (M = 97.41, SD = 3.41) than non-leaders (M = 92.72, SD = 5.79). Figure 1 shows the means and 95% confidence intervals for the four leadership roles. Sex was not significantly correlated with peer evaluation, though the average peer evaluation for women (M = 96.27, SD = 5.60) was higher than for men (M = 94.29, SD = 6.32).

Our multicollinearity analysis showed that there was an extremely low level of multicollinearity (VIF = 1.04 for sex, 1.05 for race, 1.19 for following, 1.26 for leading and 1.21 for course grade). This result indicated the five predictors were not interrelated with each other.

Comparison of peer evaluations among the leadership roles

The Kruskal–Wallis test showed a significant effect of leadership role on peer evaluations (χ^2[3, N = 294] = 81.80, p < 0.001, η^2 = 0.26). The Mann–Whitney U tests showed the median peer evaluation for leader was the highest and the median peer evaluation for follower was the lowest. The median peer evaluations for leaders (Mdn = 98.67) were significantly higher than those for: followers ([Mdn = 93.33], U = 1299.50, p < 0.001, r = 0.59), students who reported being both leaders and followers ([Mdn = 96.80], U = 4794.00, p = 0.007, r = 0.18), and students who reported being neither a follower nor a leader ([Mdn = 96.00], U = 883.00, p = 0.001, r = 0.25). The median peer evaluations for followers were significantly lower than those for: students who reported being both (U = 1170.50, p < 0.001, r = 0.50) and students who reported being neither (U = 321.50, p < 0.001, r = 0.40). The difference between the peer evaluations of students who reported being both and who reported being neither was not significant (U = 737.50, p = 0.24, r = 0.12).

Table 3. Descriptive statistics of peer evaluation for white students and students of colour.

			95% Confidence interval	
Race	Mean	SD	Lower limit	Upper limit
White students	96.39	4.45	95.86	96.91
Students of colour	91.11	12.39	86.21	96.01

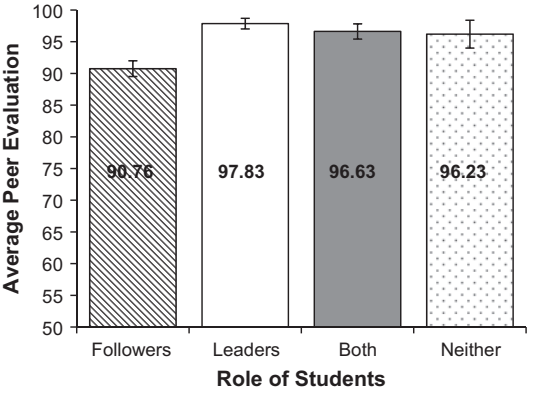

Figure 1. Average peer evaluation by 'leadership' role.
Note: The error bars represent the 95% confidence intervals.

Course grades and relevant variables

Relationship between final grade and sex, race, follower and leader

Our multiple regression analysis showed that the four predictors explained 18% of the variance for final course grade (R^2 = 0.18, $F(4,101) = 5.16$, p = 0.001, 95% CI [0.054, 0.30]). We found that the predictors following (β = −2.56, t = −2.19, p = 0.031, 95% CI [−4.88, −0.24]) and leading ($\beta = 2.99$, t = 2.24, p = 0.027, 95% CI [0.34, 5.63]) were significantly correlated with final course grade. Race (β = −2.41, t = − 1.23, p = 0.22, 95% CI [− 6.31, 1.48]) and sex (β = 1.39, t = 1.68, p = 0.094, 95% CI [− 0.24, 3.02]) were not significantly correlated with final grade.

Our regression model suggested that followers obtained a significantly lower average course grade than non-followers, and that leaders obtained a significantly higher average course grade than non-leaders. There was not a significant difference between the average final grade for students of colour and white students, but the average final grade for white students was higher than that for students of colour (see Table 4). There was not a significant difference between the average final grades by sex. The average final grade for women (M = 83.65, SD = 5.77) was higher than that for men (M = 81.39, SD = 6.65).

Our multicollinearity analysis showed that there was an extremely low level of multicollinearity (VIF = 1.03 for sex, 1.03 for race, 1.14 for reported following, 1.20 for reported leadership). This result indicated the predictors were not interrelated with each other.

Table 4. Descriptive statistics of final grade for white students and students of colour.

			95% Confidence interval	
Race	Mean	SD	Lower limit	Upper limit
White students	83.57	5.57	82.42	84.72
Students of colour	79.99	8.89	73.16	86.82

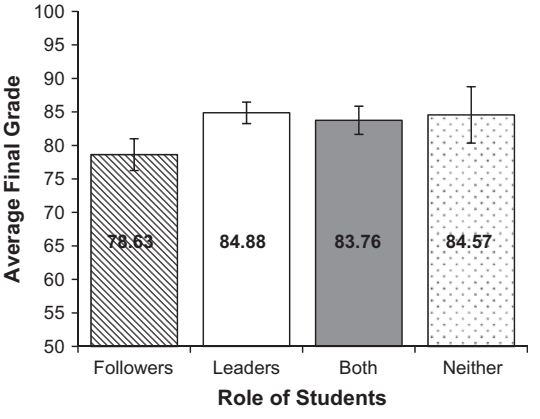

Figure 2. Average final grade by 'leadership' role.
Note: The error bars represent the 95% confidence interval.

Comparison of final grades among the leadership roles

The one-way ANOVA showed a significant effect of leadership role on final course grades, ($F(3, 101) = 6.51$, $p < 0.001$, $\eta^2 = 0.16$). The power of the test was 97%. The Tukey comparison showed that the average course grade for followers ($M = 78.63$, SD = 4.91) was significantly lower than that for leaders ($M = 84.88$, SD = 6.11), students who reported being both leaders and followers ($M = 83.76$, SD = 5.18), and students who reported being neither ($M = 84.56$, SD = 5.69) (Figure 2). There was not a significant difference in average final course grade among leaders, students who reported being both and students who reported being neither.

Differences in reported leadership roles

The percentage of women who identified themselves as leaders was higher (75%) than the percentage of men (55.6%). Similarly, the percentage of white students who identified themselves as leaders was higher (74.2%) than the percentage of students of colour (44.4%). However, the χ^2 test showed that the percentage of leaders did not significantly differ by sex ($\chi^2[1, N = 102] = 2.75$, $p = 0.097$, $r = 0.16$) or race, ($\chi^2[1, N = 102] = 3.57$, $p = 0.059$, $r = 0.19$).

Discussion

Peer evaluations are widely used in higher education to provide a way to hold students accountable for their behaviour and performance in group assignments, and to infuse validity into grades assigned to individuals for group projects. However, one concern is whether race or sex bias undermines the validity of peer evaluations. In this study, we sought to examine whether peer evaluations correlated with a variety of variables: personality or behavioural characteristics, like whether students reported that they acted like a leader or follower; demographic variables like race, ethnicity or sex; and course performance. Positive correlations with some of these variables (acting like a leader and course performance) would support the validity of

peer evaluations. Unless they are spurious correlations, positive correlations with other variables (race, ethnicity or sex) may undermine the validity of peer evaluations. There is scant research in this area, yet it remains an important question since it speaks to the validity of peer evaluations.

In support of the validity of peer evaluations, we found that people who reported taking on a 'leadership' role – that they 'felt like a leader' in the group – had significantly higher peer evaluations than students who reported that they were not leaders – that they 'felt like a follower' in the group. Leaders also had significantly higher course grades than followers. Regression analysis indicated that course performance significantly positively correlated with peer evaluations, despite students being asked to evaluate their peers' 'team citizenship' in a way distinct from the instructors' focus solely on content. We found no significant differences in peer evaluations by sex. Finally, potentially undermining the validity of peer evaluations, we found that white students received significantly higher peer evaluations than did students of colour. However, potentially in support of the validity of peer evaluations, we also found that white students had higher course performance, and were more likely to identify as leaders than students of colour, though the regression analysis showed no significant difference in average course grade between white students and students of colour. A χ^2 test indicated that, though white students were more likely to identify as leaders than are students of colour, these differences were not significant.

Our study supports the conclusions of others that peer evaluations positively correlate with both course performance (Persons 1998; Watson, BarNir, and Pavur 2010) and leadership (Lee and Lim 2012). Our study contributes to the literature by offering a quantitative analysis of the relationship between leadership and peer evaluations, and correcting for some weaknesses in previous studies. For example, because in our sample course grades were not divulged to group members, we can reject the notion that peer evaluations result only from *seeing* the grades of others, a question elicited by Watson, BarNir, and Pavur's (2010) study. The significant relationships among course performance, leadership and peer evaluations point to the likely relationship among these variables. Students who perform well in courses may be more confident in taking charge and assigning duties, and hence may be recognised by peers for both this leadership and the quality of their contributions to the group. These leaders, because of their higher course performance, may also have greater insight into what tasks need to be done to complete a successful project and how to complete those tasks, and are therefore in a better position to direct others. On the other hand, those students who struggle with course material may feel less confident regarding what tasks need to be completed to create a successful project, and therefore may be more likely to be told what to do by others. These relationships support the validity of peer evaluations.

Differences between this current and a past study (Dingel, Wei, and Huq 2013) with regard to the correlation between course performance and peer evaluation can be explained since the authors changed the peer evaluation scale between the studies. The scale used in this study has finer scaling between 60 and 100%, whereas the previous scale had equal scaling between 0 and 100%. Future research should explore what scales are most appropriate for use in peer review.

In addition to the proposed regression model for peer evaluation, we also incorporated the interaction terms between the two, three, four or five independent variables in the model. The adjusted R^2 value increased from 0.31 to 0.43. The significant variables are: being a follower; the interaction between race and being a

follower; the interaction between sex and being a leader; the interaction between final grade and being a follower; the interaction among final grade, sex and being a leader; and the interaction among final grade, sex, being a follower and being a leader. However, the standard errors and VIF values for most of the interaction terms were all extremely large. We also conducted different regression models of peer evaluations by adding the interaction terms each at one time, and found that our proposed model with the five independent variables was the best fit model to describe our data according to the adjusted R^2 values and the standard errors of each term. The regression model of final grade with the four individual variables was also the best fit model to our data.

Limitations

Our study had some limitations. First, there were only nine students of colour in our sample of 113 students. This small number of students of colour limits the power and generalisability of our conclusions. Students' peer evaluations with respect to race are consistent with the *trend* of course performance and reported leadership, with white students having a higher average grade and more likely to report being a leader than students of colour, though these latter differences were not statistically significant.

Conclusion

Since the trend in our sample is consistent with other research (Kaufman, Felder, and Fuller 2000, Watson, BarNir, and Pavur 2010), it is clear that the question of racial bias in peer reviews needs further exploration. Because of this trend, and because bias would negatively influence students' grades through unjustifiably low peer evaluations, we can conclude that instructors should carefully consider the effect of race when designing group assignments and constituting groups. Other scholars suggest that students in a demographic minority should not be isolated (Heller and Hollabaugh 1992, Oakley et al. 2004). We strongly agree with this recommendation, given that research has not identified harm in pairing minority students; given that effort should be made to mediate bias against students in the minority; given that the above research indicates that isolation may be harmful; and given that the trend in our sample is consistent with research showing the existence of bias.

Though it is not necessary (or preferable) that groups be homogenous with respect to race, minorities should be paired together and *not* distributed separately in different groups (Rosser 1998). In other words, even though our own study is inconclusive with regard to racial bias, the potential negative effect of bias is so damaging that, given the potential harms of isolating minority studies vs. the absence of benefit in not isolating them, it seems clear that minority students should not be isolated. In our sample, about half (four of nine) students of colour and/or Hispanic students were grouped with at least one other non-white and/or Hispanic student. The requirement that students be in both the same section of sociology and statistics, and the dearth of students of colour, presented barriers to creating groups that did not racially isolate individual students.

An alternate explanation is that, instead of individual bias on the part of student peer evaluators, race interacts in complex ways with leadership roles and course

performance. White students were more likely to identify themselves as group leaders than were students of colour. Since leaders had significantly higher average peer evaluations than followers, it is possible that students were accurately identifying and rewarding the leaders in their groups with higher peer evaluations. Leaders also had a significantly higher average course grade, therefore peer evaluations accurately reflect students' performance. Of course, there are complex reasons why a student may take on a leadership role in a group, with race possibly influencing a student's comfort level in the institution, the course and the group (Tinto 1975).

In conclusion, we find that leadership and course performance are important correlates to peer evaluation. That these characteristics positively correlate with peer evaluation can be explained by students accurately evaluating their peers' performance, and supports the validity of peer evaluations. However, potential discrepancies between students' course performance and peer evaluations with respect to race lead to more troubling implications. Do students hold biases against minority students? The relatively small number of students of colour in our sample prevents us from drawing solid conclusions, but warrants both further research, and awareness and consideration of these issues by instructors when forming groups.

Notes on contributors

Molly Dingel is a sociologist and an assistant professor at the University of Minnesota Rochester. She received her PhD in 2005 from the University of Kansas, and completed a postdoctoral fellowship at the Mayo Clinic in Rochester, MN in 2007. Her disciplinary research explores the bioethical dimensions of new genetic and medical technologies. She also engages in research on teaching and learning, with special interest in group work, interdisciplinary projects, and the experiences of diverse students in college.

Wei Wei is a statistician and an assistant professor at the Metropolitan State University. She received her PhD in 2008 from the University of Idaho. Her research interests include effects of classroom technology and group work on student learning and statistical analysis of clinical trial data.

References

Baker, Diane F. 2008. "Peer Assessment in Small Groups: A Comparison of Methods." *Journal of Management Education* 32 (2): 183–209.

Bartle, Emma K., Jan Dook, and Mauro Mocerino. 2011. "Attitudes of Tertiary Students Towards a Group Project in a Science Unit." *Chemistry Education Research and Practice* 12: 303–311.

Castilla, Emilio J. 2008. "Gender, Race, and Meritocracy in Organizational Careers." *American Journal of Sociology* 113 (6): 1479–1526.

Caulfield, Susan L., and Hodges Persell Caroline. 2006. "Teaching Social Science Reasoning and Quantitative Literacy: The Role of Collaborative Groups." *Teaching Sociology* 34: 39–53.

Cheng, W., and M. Warren. 2000. "Making a Difference: Using Peers to Assess Individual Students' Contributions to a Group Project." *Teaching in Higher Education* 5 (2): 243–255.

Conway, James M., and Allen I. Huffcutt. 1997. "Psychometric Properties of Multisource Performance Ratings: A Meta-analysis of Subordinate, Supervisor, Peer, and Self-ratings." *Human Performance* 10 (4): 331–360.

Conway, Robert, and David Kember. 1993. "Peer Assessment of an Individual's Contribution to a Group Project." *Assessment & Evaluation in Higher Education* 18 (1): 45–57.

Dingel, Molly J., Wei Wei, and Aminul Huq. 2013. "Cooperative Learning and Peer Evaluation: The Effect of Free Riders on Team Performance and the Relationship

between Course Performance and Peer Evaluation." *Journal of the Scholarship of Teaching and Learning* 31 (1): 45–56.
Eagly, Alice H., Mona G. Makhijani, and Bruce G. Klonsky. 1992. "Gender and the Evaluation of Leaders: A Meta-analysis." *Psychological Bulletin* 111 (1): 3–22.
Heller, Patricia, and Mark Hollabaugh. 1992. "Teaching Problem Solving through Cooperative Grouping. Part 2: Designing Problems and Structuring Groups." *American Journal of Physics* 60 (7): 637–644.
Kagan, Spencer. 1995. "Group Grades Miss the Mark." *Educational Leadership* 52 (8): 68–71.
Kane, Jeffrey S., and Edward E. Lawler III. 1978. "Methods of Peer Assessment." *Psychological Bulletin* 85 (3): 555–586.
Kaufman, Deborah B., Richard M. Felder, and High Fuller. 2000. "Accounting for Individual Effort in Cooperative Learning Teams." *Journal of Engineering Education* 89 (2): 133–140.
Koenig, Anne M., Alice H. Eagly, Abigail A. Mitchell, and Tina Ristikari. 2011. "Are Leader Stereotypes Masculine? A Meta-analysis of Three Research Paradigms." *Psychological Bulletin* 137 (4): 616–642.
Lee, Hye-Jung, and Cheolil Lim. 2012. "Peer Evaluation in Blended Team Project-based Learning: What do Students find Important?" *Educational Technology & Society* 15 (4): 214–224.
Oakley, Barbara, Richard M. Felder, Rebecca Brent, and Imad Elhajj. 2004. "Turning Student Groups into Effective Teams." *Journal of Student Centered Learning* 2 (1): 9–34.
Ohland, Matthew W., Richard A. Layton, Misty L. Loughry, and Amy G. Yuhasz. 2005. "Effects of Behavioral Anchors on Peer Evaluation Reliability." *Journal of Engineering Education* 94 (3): 319–326.
Persons, Obeua S. 1998. "Factors Influencing Students' Peer Evaluation in Cooperative Learning." *Journal of Education for Business* 73 (4): 225–229.
Pfaff, Elizabeth, and Patricia Huddleston. 2003. "Does it Matter if I Hate Teamwork? What Impacts Student Attitudes Toward Teamwork." *Journal of Marketing Education* 25: 37–45.
Rau, William, and Barbara Sherman Heyl. 1990. "Humanizing the College Classroom: Collaborative Learning and Social Organization among Students." *Teaching Sociology* 18 (2): 141–155.
Ridgeway, C. L. 2001. "Gender, Status, and Leadership." *Journal of Social Issues* 57 (4): 637–655.
Rosser, Sue V. 1998. "Group Work in Science, Engineering, and Mathematics: Consequences of Ignoring Gender and Race." *College Teaching* 46 (3): 82–88.
Schroeder, Carolyn, Timothy P. Scott, Homer Tolson, Tse-Yang Huang, and Yi-Hsuan Lee. 2007. "A Meta-analysis of National Research: Effects of Teaching Strategies on Student Achievement in Science in the United States." *Journal of Research in Science Teaching* 44 (10): 1436–1460.
Tinto, Vincent. 1975. "Dropout from Higher Education: A Theoretical Synthesis of Recent Research." *Review of Educational Research* 45 (1): 89–125.
Watson, Warren E., Anat BarNir, and Robert Pavur. 2010. "Elements Influencing Peer Evaluation: An Examination of Individual Characteristics, Academic Performance, and Collaborative Processes." *Journal of Applied Social Psychology* 40 (12): 2995–3019.
Zhang, Bo, and Matthew W. Ohland. 2009. "How to Assign Individualized Scores on a Group Project: An Empirical Evaluation." *Applied Measurement in Education* 22: 290–308.