EMPIRICAL EVIDENCE OF THE FAIRNESS AND QUALITY OF PEER EVALUATIONS

David Malone
Texas Tech University
Area of Accounting
Rawls College of Business Administration
Lubbock, TX 79416
Malone@ttu.edu
806-742-1545

ABSTRACT

This paper critically examines the use of peer evaluations in two semesters of a graduate level accounting class. While numerous authors have written on the use of peer evaluations, few have put to the test the issues of fairness and quality of those evaluations. Two primary research questions are asked: 1) Are there groups of students who systematically act in their own self-interest in evaluating their peers? 2) Are there characteristics in student peer evaluations that would suggest qualitative shortcomings to those evaluations? Preliminary evidence suggests that, with some qualification, these questions can be answered in the negative.

Introduction

A challenge to academics has long been the fair and rigorous evaluation of the performance of students in classes. During the 1990s, due to a surge in efforts directed toward making accounting education more participative (see, for example, AAA, 1986; AECC, 1990; Libby, 1991; Albrecht, et al., 1994; Lindquist, 1995,) pedagogical methods such as group work, case analysis, team projects, etc., have made evaluation of student performance more complex (Humphreys, et al. 1997).

Student peer evaluations offer a variety of benefits in supplementing the instructor’s task of evaluating students. First, when working in groups, fellow students have a unique perspective from which to evaluate the relative contributions of group members. Greguras, et al. (2001) observed that proximity of peers in performance of tasks make them uniquely positioned to observe level and quality of peer performance.

Second, if asked to assume partial ownership of the education process, students should be more engaged in that process. Thus, if expected to submit peer evaluations, students should be invested in paying attention to, being prepared for, and taking seriously work executed by their peers in order to compose a fair evaluation of their work.

An additional benefit to use of peer evaluations is their increasing use within firms (Greguras, et al. 2001.) Upon graduation, new employees often find themselves called upon to evaluate those with whom they work. Guidance on how students might formulate their evaluations while they are in school would invariably carry over into their ensuing professional lives.
Several problems associated with the use of peer evaluations present themselves, however. First, there exists the possibility of a prisoner’s dilemma when students are asked to evaluate each other. In a strictly competitive game (which, in a class using peer evaluations is almost certainly the case,) regardless of what another student does, and in the absence of signaling, a student’s best option (dominant strategy) will always be to assign a lower evaluation to the work of their peer. Thus, a concern of this study is that students, acting in their own self interest, will systematically grade their peers lower in an effort to make their own evaluations relatively better.

A second potential problem is that students may not have the capacity to judge the work of their peers. Technical courses in particular (e.g., accounting courses) present an environment in which, prior to the completion of the educational cycle, the student is not yet equipped to judge technical competency of a complex solution. How, for example, can a student evaluate the correctness of a solution to a cash flow problem if the student has not yet mastered the preparation of a cash flow statement?

### Course Environment and Peer Evaluations

The course in which peer evaluations were implemented and examined was a four-credit course covering introductory financial and managerial accounting offered at the graduate level for MBA students. Observations of behavior were made over two semesters and covered three sections of the course. The average enrollment was 35 students per section. Twenty-one Harvard Business School cases were used each semester, with students taking on team responsibilities in presenting the cases. In general, teams of two students were assigned one case each.

The case presentation counted for five percent of a student’s grade, and was earned by the team, rather than attempting to assess the individuals separately. Each student in the course, whether presenting or not, was expected to be thoroughly prepared for each case. Preparedness was monitored through a series of quizzes that were administered on a random basis. Participation was observed and graded to provide additional incentives for case preparation among class members.

A variety of benefits accrue from requiring student preparation and presentation of cases. Adler et al. (2004) argued that self-directed learning that emerges in student presentation of cases is more consistent with learning objectives intended in the case method, by comparison to a teacher-led case pedagogy. These include enhancement of communication skills, building of confidence, increased willingness to confront new experiences, promotion of self-directed learning, among others.

As the semester progressed, and cases were presented, students were asked to evaluate their peers on five dimensions (professionalism, technical quality, clarity and organization, identification of issues, and use of external resources), and on a scale of 0-5 on each of those dimensions. The five dimensions were provided on an evaluation form to which the students responded following each presentation. Evaluations were e-mailed to the professor, along with their assessment of degree of difficulty. Peer evaluations presented several challenges. Students in the first semester were not given specific instructions with respect to timeliness of their evaluations nor the importance of actually completing them. As a consequence, the response rate was only about 50%. By comparison, in the second semester, when asked to provide their
evaluations within two days of the presentation and told that their response rate may factor into their participation grade, response rate improved significantly, rising to over 80%.

Kilpatrick et al. (2001) identified several characteristics in peer evaluations that, according to students, are desirable. These include a structured evaluation form, allowance for additional comments, and that evaluators remain confidential. Each of these characteristics was incorporated into the peer evaluation process used here.

Research Questions and Hypotheses

Of significant concern is whether peer evaluations add or detract from a fair and impartial score. In MBA classes under a quasi-cohort system, one would be naïve to expect that peer evaluations would be completely impartial. One expects that both alliances and rivalries would develop over time – perhaps most obviously that friends would score friends highly; and, possibly, that rivalries or animosities may have emerged among students, having the opposite effect. There are, too, potential sources that arise from purely self-interested behavior. In its simplest form, a self-interested behavior might manifest itself as lower scores assigned by students hoping to gain a competitive advantage over their colleagues. The research question the study asks is:

**RI**: Are there groups of students who systematically act in their own self-interest in evaluating their peers?

A logical extension of this question is whether students who exhibit lower levels of moral development are more likely to use a peer evaluation system to put themselves at a systematic advantage to their classmates? To answer this question, the Defining Issues Test (DIT) was administered to each student in an effort to quantify various dimensions of the student’s moral reasoning. The most recent version of the DIT, the DIT-2, provides several measures that help identify progressively higher levels of moral reasoning. The N2SCORE is a developmental index that attempts to measure levels of sophistication in thinking about moral issues (Bebeau and Thoma, 2003, pp. 19-20). While it does not necessarily follow that more sophisticated thinking (and rejection of “simpler and biased” thinking) will produce ethical behavior, that there would be a systemic bias toward more moral behavior in the case of higher level thought does. Thus, the first hypothesis tested by this study is:

**H1**: Mean evaluations by students with a higher N2SCORE are higher than mean evaluations by students with a lower N2SCORE.

As results are discussed, whether the null is rejected or not, and its interpretation as a desirable outcome, or an undesirable one, will vary depending on the nature of the question. In this case, the regression results presented in Exhibit 1 do not support rejection of the null, suggesting that there is not a systematic, self-interested behavior exhibited during the peer evaluation process by students with a lower N2SCORE. Further, students with a higher N2SCORE (i.e., higher measured levels of moral development) are not at a systemic disadvantage to those with lower scores.
Another interesting question is whether students who are doing poorly in the course either consciously or subconsciously lower their evaluations to gain competitive advantage in an effort to improve their standing in the class. Two testable hypotheses were developed to address this question. They are:

H1₂: Change in mean evaluations by students from the first to the second half of the course is inversely related to their scores on the midterm exam.

H1₃: Student scores on the midterm exam are positively related to their mean evaluations in the second half of the course.

In the case of H1₂, upon receiving their score on the midterm exam, a student who has performed poorly may seek to obtain any competitive advantage they might be able to find. One possible source would be for that student to lower their peer evaluations for the duration of the semester. Since students are informed that grading is competitive in the course, this behavior would represent a dominant strategy if their goal is to raise their relative position in the class.

In much the same way, the population of students performing well on the midterm should be more confident (i.e. less insecure) about their grade and feel less pressure to lower their scores. In H1₃, mean peer evaluations prior to the midterm exam are assumed to be equivalent. This assumption was supported by an analysis of the data.

Results of the statistical tests of these hypotheses are presented in Exhibits 2 and 3. Again, in neither instance are these assertions supported; and, once again, this should be interpreted as a desirable outcome. Of course, there can be several explanations of why students appear to behave in a way true to the task of evaluating their peers fairly. The most optimistic interpretation is that students are behaving responsibly toward their peers, judging their work fairly, and acting in an altruistically consistent way with Kant’s first categorical imperative. It is also possible that students do not realize the marginal advantage to be gained by lowering their peer evaluations; or, that they do understand, but consider the probabilistic benefit to be so low that they do not wish to risk that their peers might discover the source of their low evaluations. In any event, there appears to be no evidence of gaming taking place in the peer process.

---

**Exhibit 1**: Average Evaluation = \( f(\text{N2SCORE}) \)
Another interesting question posed here is whether there is evidence that the peer evaluation system has characteristics that diminish the quality of the assessment. Although Kilpatrick et al. (2001) presented evidence that students favor student input into the evaluation process, there may be problems associated with the content of those evaluations. The research question suggested here is:

**R2: Are there characteristics in student peer evaluations that would suggest qualitative shortcomings to those evaluations?**

A variety of ways exist to approach answering this question. One interesting observation, for example, is the proportion of students who appeared to give uniform evaluations, offering very little discrimination among case presentations. Several examples illustrate this point. In one student presentation of Crystal Meadows of Tahoe, Inc. (HBS Case 192-150) requiring preparation of a cash flow statement, an income statement was presented instead. Because the error was so egregious, control of the presentation was temporarily assumed by the professor in order to correct any impression that the income statement might be a cash flow statement. Still, in the evaluations, under technical merit, several students assigned “5”, when a major technical flaw had been assertively pointed out. In several presentations, students would dress in shorts, wear baseball caps ( invariably turned backwards,) wear t-shirts and often suggested a lack of preparedness. Alternatively, other groups were dressed in business suits and had smoothly delivered, professional presentations. Still, a critical mass of students failed to discriminate between these two levels of apparent effort, assigning “5” in each instance to the “conducted in a professional manner” dimension. While this study did not attempt to measure these more subjective qualities, they exist as evidence that perhaps the marginal efforts made by some students was not rewarded in the peer evaluation process.
Another concern is that students who came to class unprepared may not have had a basis upon which to evaluate certain dimensions of the presentation. Question 2 on the evaluation form asked the reviewer to evaluate the presentation on its technical merits. Absent knowledge of the case and insight into viable solutions, a student may have given the presenter the benefit of the doubt and submitted a high evaluation. During both semesters, short quizzes were administered at the beginning of class periods, at random. These quizzes were used as a proxy for student preparedness, and were part of the grading mechanism serving that purpose. The hypothesis thus suggested is:

**H2:** Students performing poorly on daily quizzes submitted higher evaluations of technical merit for cases than students performing well on daily quizzes.

By a similar logic, students who performed well, and by extension are presumed to have been prepared each day, should have had more consistent insights into the technical merit of a presentation. The scores by those students, therefore, should be more narrowly distributed than scores assigned by students who were less well prepared. Regarding workload, preparing for an easier case will take less of a commitment on the part of students not assigned to present. With more difficult cases, one might expect that fewer students will have prepared for that case, and thus would be less informed in evaluating their peers whose responsibility it was to present. In those cases, too, one might expect that evaluations would be more widely dispersed than when the case assigned was less difficult. Based on these arguments, the following hypothesis was developed:

**H2:** Dispersion of evaluations of technical merit by students is inversely related to scores on the midterm exam.

Exhibits 4 and 5 provide the statistical results for the preceding two hypotheses. The results suggest no evidence that potential problems implied by either hypothesis exist. Again, failure to reject the null is a desirable outcome in each instance, indicating that lack of preparedness did not interfere with assessments when compared to those students who were more prepared. An interesting possibility is the “halo” effect that may accompany the presentation of more difficult cases. Anyone familiar with judging of diving understands this effect. Presumably, easier dives should be easier to execute and thus be accompanied by better scores. More difficult dives, however, seem to be those that will draw the 9s and 9.5s from the judges, while the easier dives will tend not to be scored as well. There thus
seems to be a subconscious awarding of additional credit for attempting the more difficult dives, even though the degree of difficulty system is intended to compensate automatically for this. In the same way, one expects that students executing easier cases should receive higher scores for their presentation. If the opposite were true, as seems to be the case in diving scores, rewards for cases would be distributed in a way other than intended. The following hypothesis, therefore, tests this notion:

\textbf{H2a: Unadjusted peer evaluations of cases are positively related to their degree of difficulty.}

Results reported in Exhibit 6 suggest there appears to be a strong statistical relationship between unadjusted peer evaluations and case difficulty, suggesting the aforementioned “halo” effect. The coefficient is positive, consistent with the hypothesized direction of the relationship. If there is solace to be found, one might find it in two places. First, the adjusted R-square is only 0.0567. That suggests that there are other, more important variables that would help explain better the variance among subjects. Second, this may be a “problem” that is acceptable. Students are taking on a risk and additional work by bidding aggressively on more difficult cases. The effect discussed here is simply a hidden reward associated with the extra risk taken on by those individuals.

Another indication of uninformed evaluations may be inconsistencies in distribution of evaluations on days when multiple cases were presented. When one case is assigned for a given day, the task of preparing adequately is more manageable than on days when multiple cases are assigned. Also true, perhaps, is that if evaluations of grouped cases are more widely distributed, a case could be made that students, in formulating their
R-square 0.0652
Adj R-square 0.0567

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>1</td>
<td>0.356</td>
<td>0.356</td>
<td>7.737</td>
<td>0.006</td>
</tr>
<tr>
<td>Error</td>
<td>111</td>
<td>5.109</td>
<td>0.046</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>112</td>
<td>5.465</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Exhibit 6: Unadjusted Peer Evaluations = f(Case Difficulty)

evaluations, are less focused because of the additional inputs. The fourth hypothesis for the second research question is thus suggested:

**H2:** Mean evaluations of cases presented alone are more narrowly distributed than of cases presented on days when multiple cases are presented.

Examining the output (F test for unequal variances) provided in Exhibit 7, the variance for these two samples was shown to be unequal at the 0.03 level of significance; however, the variance for the isolated cases is more narrowly distributed than that of the grouped cases. This result is opposite the relationship suggested in the hypothesis. The null, therefore, is not rejected.

<table>
<thead>
<tr>
<th></th>
<th>Grouped Cases</th>
<th>Isolated Cases</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>4.559</td>
<td>4.557</td>
</tr>
<tr>
<td>Variance</td>
<td>0.0085</td>
<td>0.0302</td>
</tr>
<tr>
<td>Observations</td>
<td>11</td>
<td>9</td>
</tr>
<tr>
<td>Degrees of Freedom</td>
<td>10</td>
<td>8</td>
</tr>
<tr>
<td>F</td>
<td>0.2813</td>
<td></td>
</tr>
<tr>
<td>p-value</td>
<td>0.0323</td>
<td></td>
</tr>
</tbody>
</table>

Exhibit 7: Mean Peer Evaluations = f(Case Isolation)

Conclusions and Recommendations

The purpose of this paper has been to explore the fairness and quality of student peer evaluations in accounting courses. Two questions were asked: 1) Did students exhibit self-interested behaviors in assessing the performance of their peers? 2) Were there qualitative shortcomings to peer evaluations?

In both questions 1 and 2, there seemed to be little evidence in the data gathered either that a) students behaved in a self-interested way; or, b) there were qualitative problems with peer evaluations.

On the subject of peer evaluations, guidance, perhaps in the form of specific instructions, should be offered to students on how to assign scores to the different dimensions of the peer evaluations. Knechel (1992) describes an interesting alternative to the method adopted here. Rather than
having students evaluate each case presentation, Knechel suggests having students, at the end of the semester, name the five best presentations. Students would then be rank-ordered according to the number of votes they received. There are obvious scaling issues that might be encountered with this problem (e.g., several or many groups receiving no votes, a recency effect, etc.) This method may, however, offer better discrimination.

One dimension that was not covered in the evaluations was intra-group evaluation. There were, of course, several confidential complaints by team members that they were “doing all the work.” The decision to assign grades equally to the team, rather than allowing intra-group allocations was done more for expediency than anything else. Since the grade component for the case was only 5% of the overall grade, the cost of administering an intra-group evaluation was judged to be greater than its benefits. Were the component higher, or if there were greater concern for the extent of free-rider problems, an intra-group evaluation might be advisable. Several citations exist on methods of incorporating such an evaluation (see, for example, Knechel, 1992; Stout, 1996; or, Greenstein and Hall, 1996.) Additional studies of those pedagogical models need to be made in order to assess the fairness of the evaluation processes related to those models.

This study was not an experiment, in the traditional sense. Rather, the study examined various characteristics associated with a particular pedagogy and its implementation in a real classroom. Obviously, the first priority in the class was to have the best possible pedagogy and associated evaluation system in place, such that learning potential was maximized. There were, therefore, no experimental manipulations among subjects. Future research may be well served by examining student behaviors within an experimental setting where variables similar to those examined in this study can be evaluated under more controlled circumstances.

REFERENCES


