Case Study: Peer Assessment among Business School Finance Students

James L. Slaydon, Lynn Godkin

(Lamar University, Beaumont, TX 77710, USA)

Abstract: An exploratory study related to the use of peer assessment involving 36 undergraduate Finance students' course generating 657 observations from 24 different groupings was under taken. The literature on PA in higher education seems to cluster around two issues; the extent to which PA is appropriate for formal evaluation, and assessing the extent to which individual students have contributed to group activities. In turn, this paper provides insight into two research questions; (1) How appropriate is the formal PA used to the college Finance class as an evaluation tool? and (2) Does the PA assess identify the stronger performing students as measured by test scores, semester grade and overall grade point average (GPA)?

Key words: peer assessment; group evaluations; finance students

JEL codes: A2, G3

1. Introduction

Peer assessment (PA) practices have been around for over 50 years (Sluijsmans, Brand-Gruwel, & van Merrienboer, 2002). During that time, evidence has accumulated that educators need a variety of assessment methods (Matsuno, 2009) and that students along with faculty benefit when peer review is among those used (K. Topping, 1998). Though some reject the efficacy of PA for formal class evaluation (Goldfinch & Raeside, 1990), it is attracting renewed interest in higher education of late (Bouzidi & Jaillet, 2009; Chen & Tsai, 2009; Ljungman & Silen, 2008; van den Berg, Admiraal, & Pilot, 2006) where researchers are seeking ways to enhance the process (Chen & Tsai, 2009; van den Berg et al., 2006). This circumstance led these authors to question their use of peer assessment in Finance classes. The authors considered PA where students evaluated the individual contributions of their class peers and the overall performance of teams within those classes (K. J. Topping, 2009).

2. Research Questions

The literature on PA in higher education seems to cluster around two issues (Hanrahan & Isaacs, 2001); the extent to which PA is appropriate for formal evaluation (Hanrahan & Isaacs, 2001) and assessing the extent to which individual students have contributed to group activities (Falchikov, 1986; Falchikov, 1986; Stefani, 1992, 1994). Therefore, our first research questions were:

(1) How appropriate is the formal PA used to the college Finance class as an evaluation tool?

(2) Does the PA assess identify the stronger performing students as measured by test scores, semester grade

---
James Slaydon, Associate Professor of Finance, Lamar University; research areas/interests: finance and economics. E-mail: james.slaydon@lamar.edu.

Lynn Godkin, Professor of Management, Lamar University; research area/interests: management. E-mail: lynn.godkin@lamar.edu.

and overall grade point average (GPA)?

3. Background

Peer assessment takes place between equal-status learners (K. J. Topping, 2009). It is a platform from which students consider the value or quality of work or the effort expended and participation of others in a class. PA crops up in every work situation people encounter throughout their careers. Assessment skills used in the university are readily transferable to the world of work (e.g., Blair, Cline & Bowen, 2007). Similarly, the quality of peer feedback is important to student learning (Davies, 2000) and provides a number of other benefits (Ljungman & Silen, 2008). Negatively, students find the PA process difficult and can exhibit outright hostility toward it (Hanrahan & Isaacs, 2001). Their attitudes soften when the evaluator is anonymous to those being evaluated (Davies, 2002). Nonstudents have raised concerns as well (Chen & Tsai, 2009; Ljungman & Silen, 2008).

3.1 Benefits of Peer Assessment

A number of benefits are purported to be associated with PA. For example, both self-evaluation and PA increase student engagement in their learning (Anderson, Howe, Soden, Halliday & Low, 2001; K. J. Topping, 2005). Students and their peers reportedly benefit as they explain and defend their ideas before one another (Anderson et al., 2001; Wu, 2003). PA provides an opportunity for and a platform from which to gain independent judgment and increase the ability to learn autonomously (Ljungman & Silen, 2008; K. J. Topping, 2005). Students appear to benefit from being either assessor or assessee (K. J. Topping, 2009).

Studies undertaken in writing and science classes have also reveal context specific benefits. Matsuno (2009) from a study of writing students concluded that the PA process was more beneficial when raters were oriented to methodology and rater bias. Trautmann (2009) reported improvements in writing skill following PA and students credited the process with giving key insights into their work. In science classes, PA benefits have included increased critical thinking skill among students (Gratz, 1990; Towns et al., 2001), improved motivation (Towns et al., 2001), and enhanced ability to understand higher order concepts (Trautmann, 2009).

3.2 Concerns about Peer Assessment

Despite the potential benefits available from PA, a number of concerns have been raised about the process (Chen & Tsai, 2009; Ljungman & Silen, 2008) including the validity and fairness of PA (Falchikov, 1995; Orsmond et al., 1996), general acceptance of responsibility for PA by participants (Falchikov, 1995), rater bias based on social relationships (Magin, 2001; Ozogul & Sullivan, 2009) and student attitudes toward PA (Sluijsmans et al., 2002). Ljungman and Silen (2008) aptly provide a review of the literature outlining other key concerns, namely: (1) is the efficacy of PA, per se, considered in the learning context rather than alone, (2) is the accuracy of PA a function of the learning context and training, (3) to what degree are students involved in the creation and understanding of rating criteria, (4) to what degree do students accept the PA process, and (5) what is the extent of student exposure to PA (i.e., is the PA process incorporated into an entire program). Dominant concerns include the reliability of PA and student attitudes toward PA.

Peer examiners take on responsibility similar to that of faculty and they must handle the task in a similar manner (Ljungman & Silen, 2008). Unfortunately, only a handful of related studies (e.g., Falchikov & Goldfinch, 2000; Haaga, 1993; Mowl & Pain, 1995) have been undertaken (Bouzidi & Jaillet, 2009) and the sample sizes

have been small (Bouzidi & Jaillet, 2009; Cho, Schunn, & Wilson, 2006). Related empirical studies have clustered around reliability, validity, and bias in peer grading (e.g., Falchikov & Goldfinch, 2000; Ghorpade & Lackritz, 2001; Trautmann, 2009). Trautmann (2009) suggests that studies to date addressing learning outcomes are limited. Similarly, Zhang, Johnston, and Kilic (2008) report that research on the reliability of peer rating in group work is limited. Most has been related to the agreement between student and teacher ratings of course work. Examination of inter-rater reliability among student assessors is rare.

3.3 Training for Peer Assessment

Successful use of PA requires that faculty and student work responsibility and together. As has been noted, students come to the process with both anticipation and trepidation. They doubt the efficacy of the process and express the need for training in the process (Sluijsmans et al., 2002). At a minimum, ratings must accurately reflect the contributions of each individual to be valid and be fairly consistent across groups to be reliable (Zhang, et al., 2008). The reticence expressed by students and issues surrounding validity and reliability may be addressed through practice and training (Hanrahan & Isaacs, 2001). Matsuno (2009) has found from a study of writing classes that PA improves with orientation to methodology and potential rater bias.

Training for and orientation to PA is important (Ozogul & Sullivan, 2009) and should include several subjects. For example, the literature suggested that students should get an idea of what constitutes good and bad work (Ljungman & Silen, 2008) with supporting examples. Students should be provided or guided in the development of appropriate rubrics from which to operate (Ozogul & Sullivan, 2009). These would contribute to student understanding of the curriculum and contribute to the validity and reliability of the PA. One study found that raters who received qualitative assessment reports from peers in conjunction with other training outperformed those in the control group not benefiting by such exposure (Sluijsmans et al., 2002). Tseng and Tsai (2007), analyzing 184 high school students, peer feedback given with reinforcement, encouragement, and friendly suggestions helpful (Chen & Tsai, 2009).

Therefore,

H_1 Peer assessors who are trained for peer assessment will provide evaluations that identify higher performing students as measured by average test scores, semester final grade, and overall GPA.

4. Methodology

Few studies related to PA have been undertaken (Bouzidi & Jaillet, 2009) and the sample sizes have been small (Bouzidi & Jaillet, 2009; Cho et al., 2006). Related empirical studies have clustered around reliability, validity, and bias in peer grading (e.g., Falchikov & Goldfinch, 2000; Ghorpade & Lackritz, 2001; Trautmann, 2009). For these reasons and our concerns about PA in the context of Finance courses, the authors undertook this study.

A sample of 36 students in Cases in Financial Management in spring 2008 used peer assessment to assess their peers when working on four separate group projects in groups of six which were randomly rotated after completing two projects. Each participant completed the Peer Evaluation Form from The Business Strategy Game. The Peer Evaluation Form was used by more than 300 schools. A total of 657 usable student evaluation forms for the 24 different groups were collected. Additionally, a ranking table was added to the form.

The Peer Evaluation form has 12 individual questions. Questions 1-11 use a Likert scale to evaluate each

individual member of each group project by every other member of the group. For every group project, there were at least 5 separate student evaluations of each student. Using the Likert scale, seven out of the eleven questions were worth 6 points. Highest response was worth 6 points and the lowest response was worth 1 point. The other 4 questions were worth 12 points. The highest response was worth 12 points. The lowest response was worth 2 points. These 11 questions totaled to 90 points. Question 12 was an overall evaluation worth 10 points. Highest response was worth 10 points. The lowest response was worth 1 point. The total possible number of points was 100. The lowest possible number of points was 16.

Ordinary least squares regression was used to see if a student's average group of peer evaluations from other group members be statistically significant in identifying who's performance was higher on the four semester tests covering the material in each case. A regression was run using only questions 1-11 to predict performance. A second regression was run using all 12 questions to predict performance.

Ordinary least squares regression was used to see if a student's average group of peer evaluations from other group members be statistically significant in identifying who's performance was higher on the final semester grade. A regression was run using only questions 1-11 to predict performance. A second regression was run using all 12 questions to predict performance.

Ordinary least squares regression was used to see if a student's average group of peer evaluations from other group members be statistically significant in identifying who's performance was higher on the students overall GPA. A regression was run using only questions 1-11 to predict performance. A second regression was run using all 12 questions to predict performance.

A secondary inquiry was does question 12 (overall performance) match with the assessments made in questions 1-11. A t-test was run to see if there was a statistical difference between the two assessments. The ranges given in the overall question 12 were: 0-50, 50-59, 60-64, 65-69, 70-74, 75-79, 80-84, 85-89, 90-95, and 96-100. I used 25, 55, 62, 76, 72, 77, 82, 87, 92.5, and 98 to represent the ranges in the T-test for the corresponding ranges. Since there is a difference, I calculated the weight average difference to use for the hypothesized mean of 5.96.

The ranking within the group was added to give a measurable way to see if their perceived work rankings matched with the peer assessment numerical rankings. A Wilcoxson-Signed rank test was used to evaluate the difference between rankings given within the group at the end of the peer assessment form and the rankings in questions 1-11.

5. Results

Table 1 reports the regression results findings for the average test scores for the semester. The intercept is statistically significant at the 10% level for questions 1-11 but this is not surprise given the minimum point total is 16 because of the Likert scale. The average score for questions 1-11 is statically significant t the 1% level. This shows that high performance of peer evaluations seems to predict better scores on tests. The intercept is not statistically significant at the 10% level for questions 1-12 but this is a surprise given the minimum point total is 16 because of the Likert scale. The average score for questions 1-12 is statically significant at the 1% level. This shows that high performance of peer evaluations seems to predict better scores on tests.

Table 2 reports the regression results findings for the final semester grade for the semester. The intercept is statistically significant at the 1% level for questions 1-11 but this is not surprise given the minimum point total is

16 because of the Likert scale. The average score for questions 1-11 is statically significant at the 1% level. This shows that high performance of peer evaluations seems to predict better scores on tests. The intercept is statistically significant at the 1% level for questions 1-12 but this is not a surprise given the minimum point total is 16 because of the Likert scale. The average score for questions 1-12 is statically significant at the 1% level. This shows that high performance of peer evaluations seems to predict better scores on tests.

Table 1   Regression of Average Peer Evaluation Scores against Average Test Scores

	Intercept	Average Scores	R^2
Questions 1-11	16.761	0.821	0.5475
	(1.962)*	(6.584)***	
Questions 1-12	13.760	0.846	0.5471
	(1.530)	(6.579)***	

Table 2   Regression of Average Peer Evaluation Scores against Final Semester Grade

	Intercept	Average Scores	R^2
Questions 1-11	50.110	0.472	0.6944
	(13.920)***	(8.973)***	
Questions 1-12	48.313	0.487	0.6971
	(12.806)***	(9.031)***	

Table 3 reports the regression results findings for the overall grade point average (GPA). The intercept is statistically significant at the 1% level for questions 1-11 but this is not surprise given the minimum point total is 16 because of the Likert scale. The average score for questions 1-11 is statically significant at the 1% level. This shows that high performance of peer evaluations seems to predict better scores on tests. The intercept is statistically significant at the 5% level for questions 1-12 but this is not a surprise given the minimum point total is 16 because of the Likert scale. The average score for questions 1-12 is statically significant at the 1% level. This shows that high performance of peer evaluations seems to predict better scores on tests.

Table 3   Regression of Average Peer Evaluation Scores against Grade Point Average

	Intercept	Average Scores	R^2
Questions 1-11	1.021	0.028	0.4833
	(3.141)***	(5.808)***	
Questions 1-12	0.929	0.0.28	0.4783
	(2.702)**	(5.752)***	

Table 4 reports the t-test results for comparing the overall evaluation (question 12) and the average score on questions 1-11. The means are hypothesized to be different at the 1% level. I correct for using a Likert scale with a 5.96 expected difference. This table shows that there is a statically difference when after correcting for the Likert scale problem. The overall question was higher for students than the accumulated score given in questions 1-11. This table shows that some upward bias might be included when students give an overall evaluation of another student's performance. This upward bias was demonstrated as 97.4% (640 out of 657) were given a higher overall score (question 12) than they got on the accumulated questions 1-11.

Table 4   T-test: Paired Two Sample for Means

	Overall (12)	1 thru 11
Mean	82.60	67.39
Variance	184.14	281.67
Observations	657	657
Pearson Correlation	0.87	
Hypothesized Mean Difference	5.96	
df	656	
t Stat	28.35	
P (T <= t) one-tail	2.3501E-116	
t Critical one-tail	1.65	
P (T <= t) two-tail	4.7001E-116	
t Critical two-tail	1.96	

Table 5 helps shows if this upward bias cause's then ranking of a student's contribution to the groups work was affected by tendency to give higher overall evaluations. This shows that 46% of rankings had the exact matching ranking. Testing the other 54%, the authors find that there is not a statistically significant difference in the ones that are not a perfect match. Therefore, while differences might happen, these differences do not seem to be an important for the work contributed by the students.

Table 5   Matching Ranking between Students on Questionnaire and Questions 1-11

Number of Group Rankings	Number with Exact Matching Rankings	Percent with Exact Matching Rankings	Number with Ranking Differences	Number Statistically Significantly Different Rankings (Wilcoxon Signed Ranks Test)
120	55	45.83%	65	0

References:

Anderson T., Howe C., Soden R., Halliday J. and Low J. (2001). "Peer interaction and the learning of critical thinking skills in further education students", *Instructional Science*, Vol. 29, No. 1, pp. 1-32.

Billington H. L. (1997). "Poster presentations and peer assessment: Novel forms of evaluation and assessment", *Journal of Biological Education*, Vol. 31, No. 3, pp. 218-220.

Blair B., Cline G. R. and Bowen W. R. (2007). "NSF-style peer review for teaching undergraduate grant-writing", *American Biology Teacher*, Vol. 69, No. 1, pp. 34-37.

Bouzidi L. H. and Jaillet A. (2009). "Can online peer assessment be trusted?", *Educational Technology & Society*, Vol. 12, No. 4, pp. 257-268.

Chen Y. C. and Tsai C. C. (2009). "An educational research course facilitated by online peer assessment", *Innovations in Education and Teaching International*, Vol. 46, No. 1, pp. 105-117.

Cho K., Schunn C. D. and Wilson R. W. (2006). "Validity and reliability of scaffolded peer assessment of writing from instructor and student perspectives", *Journal of Educational Psychology*, Vol. 98, No. 4, pp. 891-901.

Davies P. (2000). "Computerized peer assessment", *Innovations in Education and Training International*, Vol. 37, No. 4, pp. 346-355.

Davies P. (2002). "Using student reflective self-assessment for awarding degree classifications", *Innovations in Education and Teaching International*, Vol. 39, No. 4, pp. 307-319.

Falchikov N. (1995). "Peer feedback marking: Developing peer assessment", *Innovations in Education and Training International*, Vol. 32, No. 2, pp. 175-187.

Falchikov N. and Goldfinch J. (2000). "Student peer assessment in higher education: A meta-analysis comparing peer and teacher marks", *Review of Educational Research*, Vol. 70, No. 3, pp. 287-322.

Ghorpade J. and Lackritz J. R. (2001). "Peer evaluation in the classroom: A check for sex and race/ethnicity effects", *Journal of Education for Business*, Vol. 76, No. 5, pp. 274-281.

Goldfinch J. and Raeside R. (1990). "Development of a Peer Assessment Technique for Obtaining Individual Marks on a Group Project", *Assessment and Evaluation in Higher Education*, Vol. 15, No. 3, pp. 210-231.

Gratz R. K. (1990). "Improving lab report quality by model analysis, peer review, and revision", *Journal of College Science Teaching*, Vol. 19, No. 5, pp. 292-295.

Haaga D. A. F. (1993). "Peer review of term papers in graduate psychology courses", *Teaching of Psychology*, Vol. 20, No. 1, pp. 28-32.

Hanrahan S. J. and Isaacs G. (2001). "Assessing self- and peer-assessment: The students' views", *Higher Education Research & Development*, Vol. 20, No. 1, pp. 53-70.

Ljungman A. G. and Silen C. (2008). "Examination involving students as peer examiners", *Assessment & Evaluation in Higher Education*, Vol. 33, No. 3, pp. 289-300.

Magin D. (2001). "Reciprocity as a source of bias in multiple peer assessment of group work", *Studies in Higher Education*, Vol. 26, No. 1, pp. 53-63.

Matsuno S. (2009). "Self-, peer-, and teacher-assessments in Japanese University EFL writing classrooms", *Language Testing*, Vol. 26, No. 1, pp. 75-100.

Mowl G. and Pain R. (1995). "Using self and peer assessment to improve students' essay writing: A case study from geography", *Innovations in Education and Training International*, Vol. 32, No. 4, pp. 324-335.

Orsmond P. et al. (1996). "The importance of marking criteria in the use of peer assessment", *Assessment & Evaluation in Higher Education*, Vol. 21, No. 3, pp. 239-250.

Ozogul G. and Sullivan H. (2009). "Student performance and attitudes under formative evaluation by teacher, self and peer evaluators", *Educational Technology Research and Development*, Vol. 57, No. 3, pp. 393-410.

Sluijsmans D. M. A., Brand-Gruwel S. and van Merrienboer J. J. G. (2002). "Peer assessment training in teacher education: Effects on performance and perceptions", *Assessment & Evaluation in Higher Education*, Vol. 27, No. 5, pp. 443-454.

Topping K. (1998). "Peer assessment between students in colleges and universities", *Review of Educational Research*, Vol. 68, No. 3, pp. 249-276.

Topping K. J. (2005). "Trends in peer learning", *Educational Psychology*, Vol. 25, No. 6, pp. 631-645.

Topping K. J. (2009). "Peer assessment", *Theory into Practice*, Vol. 48, No. 1, pp. 20-27.

Towns M. H., Marden K., Sauder D., Stout R., Long G. and Waxman M. et al. (2001). "Interinstitutional peer review on the internet: Crossing boundaries electronically in a student-refereed assignment", *Journal of College Science Teaching*, Vol. 30, No. 4, pp. 256-260.

Trautmann N. M. (2009). "Interactive learning through web-mediated peer review of student science reports", *Educational Technology Research and Development*, Vol. 57, No. 5, pp. 685-704.

Tseng S. C. and Tsai C. C. (2007). "On-line peer assessment and the role of the peer feedback: A study of high school computer course", *Computers & Education*, Vol. 49, No. 4, pp. 1161-1174.

Van den Berg I., Admiraal W. and Pilot A. (2006). "Design principles and outcomes of peer assessment in higher education", *Studies in Higher Education*, Vol. 31, No. 3, pp. 341-356.

Wu A. (2003). "Supporting electronic discourse: Principles of design from a social constructivist perspective", *Journal of Interactive Learning Research*, Vol. 14, No. 2, pp. 167-184.

Zhang B., Johnston L. and Kilic G. B. (2008). "Assessing the reliability of self- and peer rating in student group work", *Assessment & Evaluation in Higher Education*, Vol. 33, No. 3, pp. 329-340.

Peer Evaluation Form

Instructions: Fill an evaluation form for each member of your group (For example: If you have four members in your group you need to complete three evaluation forms). **Check only one box for each question!**

Evaluator Name:_____

Team Member Name:_____

Attendance at Strategic Analysis Meetings (6 points)

☐ Habitually Absent

☐ Missed Close to 50% of our meetings

☐ Missed About 20-30% of our meetings

☐ Missed About 10-20% of our meetings

☐ Very dependable, missed less than 10% of our meetings

☐ Always Present

2. Promptness at Strategic Analysis Meetings (6 points)

☐ Habitually Late

☐ Late to about 50% of our meetings

☐ Late to about 20-30% of our meetings

☐ Late to about 10-20% of our meetings

☐ Late to less than 10% of our meetings

☐ Never kept team members waiting

3. Caliber of Preparation for Strategic Analysis Meetings (familiar with case and did outside research) (6 points)

☐ Always behind rest of the team

☐ Marginal; usually had to catch up during meeting

☐ Adequate; about as well prepared as others

☐ Good; somewhat better prepared than others

☐ Excellent; usually well prepared

☐ Exceptional; generally best prepared of all team members

4. Understanding Company Operations (skills in interpreting and analyzing financial reports) (12 points)

☐ Quite weak

☐ Marginal; sub-par

☐ Adequate

☐ Good

☐ Excellent; very impressive

☐ Exceptional; strongest of all team members

5. Skills in Diagnosing the Company's Problems, Issues, and Competitiveness (12 points)

☐ Quite weak

☐ Marginal; sub-par

☐ Adequate

☐ Good

☐ Excellent; very impressive

☐ Exceptional; strongest of all team members

6. Skills in Proposing "What to do" and Strategic Approaches to Take (12 points)

☐ Quite weak

☐ Marginal; sub-par

☐ Adequate

☐ Good

☐ Excellent; very impressive

☐ Exceptional; strongest of all team members

7. Caliber of Contribution of Team Performance (12 points)

☐ Quite weak; had almost no impact (or took actions which hurt performance)

☐ Had little positive impact (or even a negative impact) in shaping team performance

☐ Adequate; played a supporting role in shaping team performance

☐ Good; played an important role in shaping team performance

☐ Excellent; played a major and positive role in shaping team performance

☐ Exceptional; highest positive impact of all team members

8. Enthusiasm and Commitment (6 points)

☐ Almost none

☐ Inadequate

☐ Adequate; acceptable

☐ Good enthusiasm and commitment

☐ Very enthusiastic and committed

☐ Exceptional; strongest of all team members

9. Teamwork and Cooperativeness (6 points)

☐ Quite weak; gave team many problems

☐ Marginal; prone to make decisions without telling anyone

☐ Adequate

☐ Good

☐ Excellent; very impressive

☐ Exceptional; strongest of all team members

10. Exercise of Leadership (6 points)

☐ Had little to say and little to offer

☐ Ineffective; had a hard time winning support for ideas

☐ Adequate ability to present views and make a case for proposed actions

☐ Good ability to present views and make a case for proposed actions

☐ Effective and persuasive in convincing others to go along with proposed actions

☐ Exceptional; the clear leader of our management team

11. Carried a Fair Share of Overall Workload (6 points)

☐ Far less than a fair share

☐ Slightly below a fair share

☐ Roughly a fair share

☐ Slightly above a fair share

☐ Well above a fair share

☐ Far beyond what other team members did

12. Overall Evaluation (10 points)

☐ Below 50   I would like to have fired this person as a team member

☐ 50-59   Very weak (I would definitely not want to be teamed with this person again)

☐ 60-64   Marginal; sub-par

☐ 65-69   Slightly below average

☐ 70-74   Average

☐ 75-79   Slightly above average

☐ 80-84   Good

☐ 85-89   Very good

☐ 90-95   Excellent; very impressive

☐ 96-100   Exceptional; strongest of all team members

Additional Comments

_____

_____

_____

_____

_____

_____

_____

_____

Rank

_____   Percent   of Work _____

_____   Percent   of Work _____

_____   Percent   of Work _____

_____   Percent   of Work _____

_____   Percent   of Work _____

Total Percent   =   100%