AC 2010-1842: STUDENTS' PEER EVALUATION CALIBRATION THROUGH THE ADMINISTRATION OF VIGNETTES

Junqiu Wang, Purdue University

P.K. Imbrie, Purdue University

Students' Peer Evaluation Calibration Through the Administration of Vignettes

Abstract

Peer evaluation has been widely used for measuring student performance in collaborative team work. However, students tend to be biased when rating their peers. Halo effect, central tendency effect and leniency effect are very common bias in peer evaluation. One technique to reduce the possible bias is to calibrate student peer evaluation with vignettes. A Vignette describes a hypothetic team member with specific attributes demonstrating specific characteristics meant to be assessed with peer evaluation. Through evaluating this hypothetic person and comparing students' evaluation results with results from trained experts, we expect to measure students' biases and provide a training opportunity to improve student rating skills and reduce rating bias.

The theoretical framework in our study operationally defines team effectiveness as interdependency, goal setting and potency. A vignette designed to illustrate attributes of interdependency, goal setting and potency was administered at different periods of the semester. Participants in the study were enrolled in the first year engineering course and assigned to work on real engineering related projects in authentic teams of 3 or 4. The authentic means that the students are put into a team working towards course related projects.

Student ratings were compared with expert ratings considering the expert's ratings as unbiased. The differences between the students' rating and expert's rating were defined as students' bias. The biases of students' rating performance were analyzed at the item-level and the construct level. From both the item and construct level, our data show that students did not perform better with repeated vignette administrations. However, after taking the students' bias calibration into consideration, students' peer evaluation performance move closer to the expert's score.

Introduction

The Accreditation Board for Engineering and Technology (ABET)[1] Engineering Criteria 2000 requires engineering students should "be able to function effectively in a multidisciplinary team". The question is: how can students' team skills be taught and assessed[2,3,4]? In a previous study, we defined student's team skills through a three-construct theoretical model: interdependency, goal setting and potency. This model entails possible application in both pedagogy and assessment. Peer evaluation has been used as an effective instrumentation tool to assess students' team skills and performance[5,6,7,8,9]. We developed a 9-item peer evaluation questionnaire to measure student's individual perceptions on their teammates along our three-constructs theoretical model[10,11].

When conducting peer evaluation, students tend to create their own social situations, leading to different rater biases. Three biases are most common in the peer evaluation process: Halo effect, central tendency and leniency[12,13,14]. Halo effect occurs when students (rater) does not differentiate differences between subscales. When the raters do not make use of the full range of the rating scale, central tendency effect bias occurs. A rater might consistently give higher or lower scores than appropriate; this bias is defined as leniency effect. One possible solution to reduce the rater's bias is to calibrate raters' bias with vignettes[15].

A vignette is defined as "short descriptions of a person or a social situation which contain precise references to what are thought to be the most important factors in the decision-making or judgment-making process of respondents "[15]. A vignette provides a near real-life judgment situation, thus reducing the possibility of the respondents creating their own situations. In the administration of a vignette, students' rater bias can be captured by comparing their rating scores with those of the experts. Therefore, through repeated vignette administration, it is theorized that students might become a better peer evaluator [15].

Theoretical model

In our peer evaluation model, we developed a 9-item questionnaire along three constructs, interdependency, goal setting and potency, to measure a student's individual perception on their teammates' effectiveness. The detailed description of the 9-item questionnaire is listed in table 1. The first letter item ID column represents the corresponding construct: I= Interdependency; G= Goal Setting and P= Potency.

Table 1 9-item Peer evaluation questionnaire

Item ID	Item Description
I1	Collaborates well with my team on all in-class and out of the class assignments.
I2	Contributes to my team's effectiveness by having a clearly defined role(s).
I3	Is a reliable team member.
G1	Often helps my team think of what we were/were not achieving.
G2	Articulates individual goals that can be achieved with the help of my team.
G3	Actively helps my team establish goals.
P1	Helps my team to build a shared confidence in its ability to successfully work together on course assignments.
P2	Often encourages each team member to believe in my team's ability to succeed no matter what the task.
P3	Often makes my team feel confident in its ability to resolve disagreements.

The vignette was designed based on our three-construct model peer evaluation questionnaire. The detailed description of the vignette is as follows:

> "You are a member of an engineering team that is responsible for completing both in-class and out-of-class assignments. Successful completion of the course assignments requires your team to meet regularly as well as have equal contributions from all members. For this evaluation you should reflect on the performance of hypothetical member of your team, Kris.
>
> During the team meetings, you noticed that Kris never came prepared. Furthermore, Kris did not work well with the other team members to complete assignments. Finally, while working on the course assignments, Kris's role on the team was never clear, which did not help ensure a synergistic effort by everyone else on the team.

> During the team meetings, you noticed that Kris never came up with ideas that gave the team a clear long term direction. Kris failed to actively participate in developing a timeline to accomplish assignment goals. Kris never communicated personal goals for assignments (and the course) that could have been achieved by involving the rest of the team.
>
> While working on the course assignments, you noticed that Kris collaborated well with others and that you made you feel confident in the team's ability to successfully complete assignments. Furthermore, Kris worked with the team through interpersonal conflicts which made you believe in the team's ability to overcome adversity. Finally, Kris did not prejudge teammates on their individual ability to perform but rather supported the team's collective abilities to succeed. ”

Method

Participants. The participants in this study are from a large Midwestern University. Participants were enrolled a freshmen engineering problem solving course. These students were provided a simple questionnaire on their math and computer background knowledge. Students were divided into teams of 3 or 4 at the beginning of the semester based on their background variety in math and computer skill. Students worked with the same team on team projects throughout the semester. For this specific study, we are looking at students enrolled at fall semester 2008.

Immediately after reading the vignette, students were asked to peer evaluate this hypothetic team member using questionnaire in table 1. The same vignette was administrated four times during the semester, each after the completion of a team project.

Experimental design. The goal of this study is to assess students' peer evaluation bias, thus we begin with the vignette calibration process. In this study, we are using within subject design, meaning that each participant receives all conditions. In our study, all conditions mean that each student receives each of four instances of peer calibration administration. The detailed collection time and the corresponding collection number and date are listed in table 2. The parenthesis after the date period gives the sequences of the nine questions presented to the students after reading the vignette. Students use the same vignette and same question sequences throughout the semester.

Table 2 Vignette administration time and number

collection 1: NO. 1115	09/22/2008—12/04/2008 (I1,P2, G1, I2, P3, G2, I3, G3, P1)
collection 2: NO. 1123	10/15/2008—10/28/2008 (I1,P2, G1, I2, P3, G2, I3, G3, P1)
collection 3: NO. 1147	11/07/2008—12/11/2008 (I1,P2, G1, I2, P3, G2, I3, G3, P1)
collection 4: NO. 1151	12/03/2008—12/13/2008 (I1,P2, G1, I2, P3, G2, I3, G3, P1)

Data analysis and discuss of results

In this study, the 9-item data was presented as 9-dimensional vector, (I1, I2, I3, G1, G2, G3, P1, P2, P3). We designed the expert's evaluation on the hypothetical member as unbiased. The expert's evaluation on this hypothetic team is (0, 0, 0, 0, 0, 0, 100, 100, 100). The differences between the expert's and a student's evaluation score is defined as the student's bias value. When student's evaluation scores of the hypothetic team member approaches the expert's value, we define this as improved peer evaluation skills for an individual student.

In this study, we compared the four collections on individual item level and we hypothesized that student peer evaluation scores on the hypothetic team member will tend to move towards expert's score. We are looking for two results: a) the four collections should be significantly different at the item level; b) data collections later in the semester should approach the expert score. Based on these two expectations, we proposed the hypothesis:
H_0 : The four peer calibration collections are not significantly different as measured by the average calibration scores of the students: $\mu 1 = \mu 2= \mu 3 = \mu 4$
H_a : The mean calibration values of the 9-items are significantly different for the four collections

Table 3 ANOVA one factor data analysis result of peer calibration data, fall, 2008

	coll1	coll2	coll3	coll4	expert	F-Critical	F-Value	P-Value
I1	44.72	48.50	50.24	48.39	0.00	2.61	2.57	0.053
I2	25.27	27.97	31.93	33.62	0.00	2.61	13.52	0
I3	38.02	37.60	39.46	39.92	0.00	2.61	2.08	0.10
G1	32.60	36.02	40.24	39.65	0.00	2.61	5.32	0.001
G2	31.04	32.10	37.80	36.70	0.00	2.61	7.44	0
G3	32.73	35.24	38.27	37.31	0.00	2.61	3.84	0.009
P1	60.04	66.54	61.95	60.45	100.00	2.61	3.73	0.011
P2	58.38	63.65	63.05	60.60	100.00	2.61	2.06	0.10
P3	65.01	66.98	64.90	62.57	100.00	2.61	1.63	0.18

Table 3 lists the ANOVA one factor comparison of means results for all the four different collections of the 392 participants. The difference between expert score and the students' score is significant, which means students did have large biases based on our definition of bias. F-test results show that, for items I2, G1, G1, G3, and P1, the F-values is greater than their corresponding F-critical values, so we reject the null hypothesis and conclude that these items are significantly different for the four different collections. The rest of the items, I1, I3, P2 and P3 have no statistically significant differences When we take a closer look at the items with significant differences, for example, I2 of the first collection, coll1 has the lowest score This indicate that students actually did better in coll1 based on our definition of good peer evaluation performance. This directly contradicts our expectation. Figure 1 provided a better representation of the result in table 3. In figure, we can see that the gap between expert's score and students' score did not reduce with repeated vignette administration. We can conclude that there is no evidence to show improved peer evaluation skills as measured by comparing individual item with unbiased expert's score; in other words, we can conclude that repeated vignette administration calibration did not guarantee better peer evaluation skills.

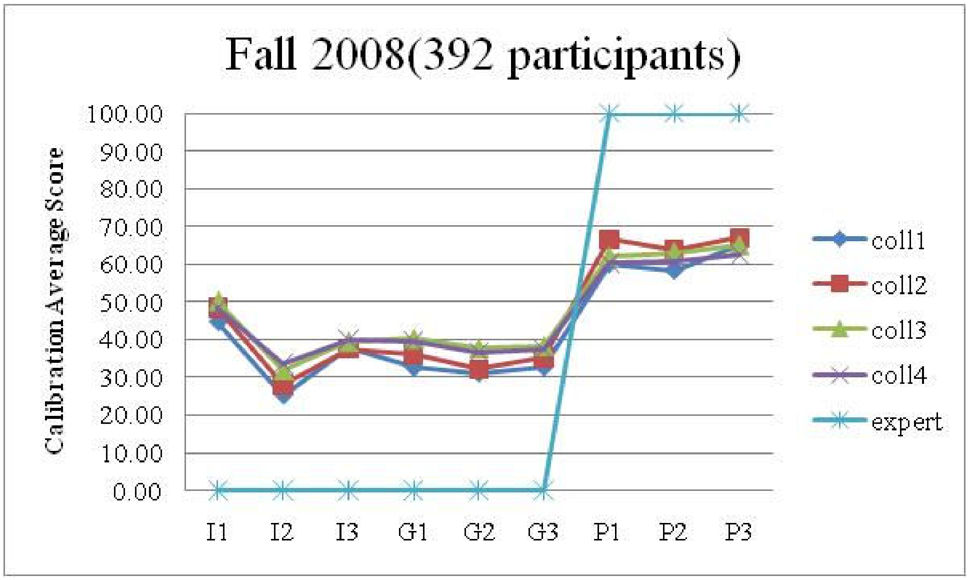

Figure 1 Students Peer Calibration Average at different collection time

Three dimensional data analysis

We will now look at the data from multivariate perspective. We reduced the 9-item (I1, I2, I3, G1, G2, G3, P1, P2, P3) into three-dimensional vector (I, G, P) by averaging the three items for each construct, i.e. I=(I1+I2+I3)/3, and the same with G and P. Thus the expert has the score vector of (0, 0, 100).

In this section, we are going to take bias into consideration. We use the data collection 1 to get the bias of every student. The algorithm of including bias is:

1) Find the i-th student's scores for collection 1, (I_{1i}, G_{1i}, P_{1i}), calculation the bias for this student with expert reference score, the bias is (I_{1i}, G_{1i}, P_{1i}-100)
2) Add this bias for the specific student into the second, third and fourth collection. For example, if the i-th student evaluation score at collection 2 is (I_{2i}, G_{2i}, P_{2i}), the calibrated score is (I_{2i}- I_{1i}, G_{2i}-G_{1i}, P_{2i}-P_{1i}+100)=(I_{2i}', G_{2i}', P_{2i}');
3) If I_{2i}' and G_{2i}' are less than 0, constrain it to zero, if P_{2i} is larger than 100, make it 100 since we cannot have score value more than 100 or less than 0.

After the bias was taken into account, we calculated the centroids of the new data sets of collection 2, 3 and 4 and put them into the same plot showing the centroid coordinates of the original data, marked as (I, P, G), as shown in figure 2. The new centroid is marked as (I-Calib, G-Calib, P-Calib). From the figure, we can see that the new centroids are moving towards to the expert point (0, 0, 100). Also from the plot we can find that repeated data collection using the same vignette does not result in better students' peer evaluation performance. The P value tend to move away from 100 score line with repeated vignette administration; while I and G values are moving away from 0 score, which means students' performance is getting worse.

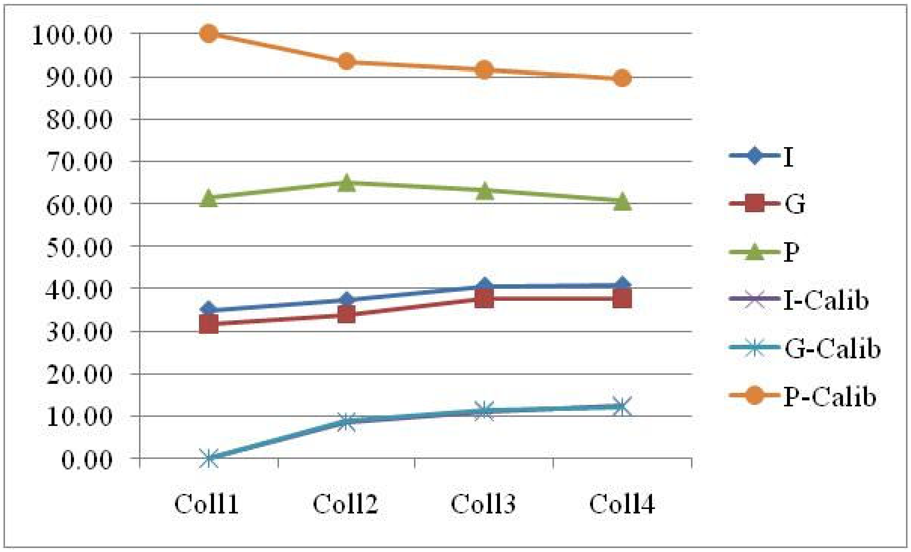

Figure 2 The original centroids coordinate and the calibrated centroids coordinate

In order to understand how the inclusion of bias changed the data set, we viewed our data sets in three-dimensional plots, using collection as example. Both the original data and the calibrated data points for collection 2 are plotted in figure 3. The calibrated data collection 2 is more densely distributed around the expert value, as shown in the figure 3.

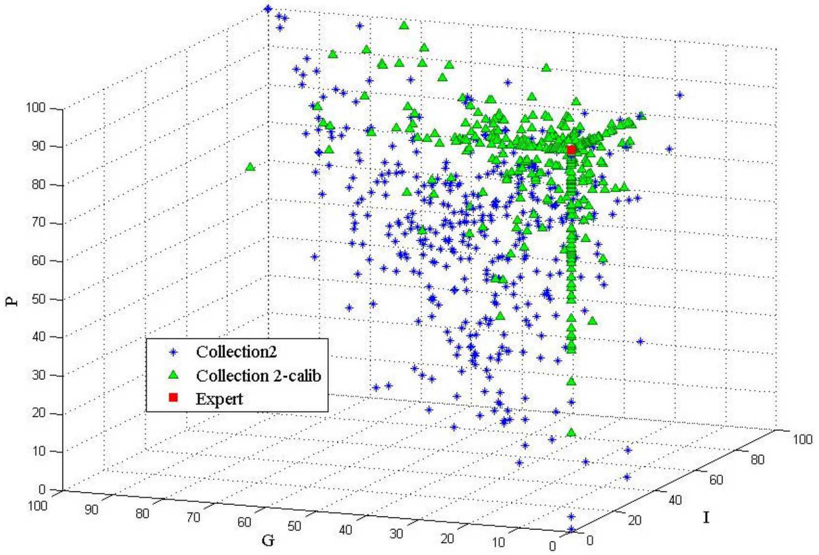

Figure 3 Three-dimensional plot of collection2 and calibrated collection 2.

While variance can be used to calculate the spread of data sets, we can also use the 'variance' concept to quantify the spread of the data sets around the expert reference score. For example, if a student gives the hypothetic team member in vignette a score point (I_i, G_i, P_i) and the expert of the whole data set is (0, 0, 100), the spread factor can be calculated by the average distance of the points in the data sets to the expert reference point.:

$d= [\sum(0-I_i)^2+(0-G_i)^2+(100-P_i)^2)^{1/2}]/n$

The results for both the original data and the calibrated data for collection 2, 3 and 4, are shown in figure 4. There is a significant decrease in the spread factor d for the calibrated data. Also as shown in figure 4, there is a tendency of increase in the spread factor with repeated data collection, which means that the average distance from the data points to the reference increased slightly. This further supports the conclusion that student does not appear to become a better peer evaluator with repeated administration of the same vignette.

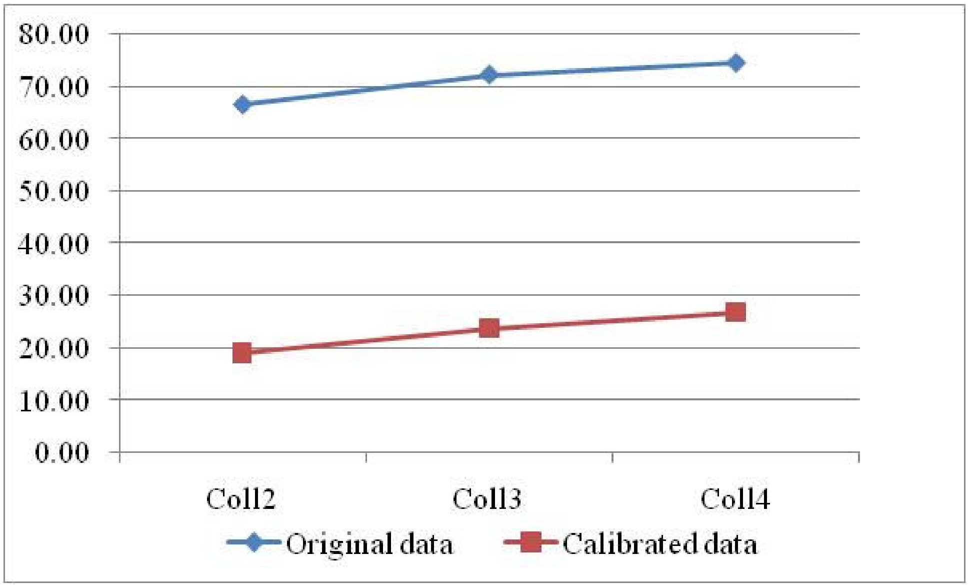

Figure 4: the spread factor of the calibrated data and the original data, fall 2008

Conclusion and further study

Based on our current study, we can conclude that students do not appear to become a better peer evaluator (i.e., approach the peer evaluation results of an 'expert' evaluator) with repeated administration of the same vignette. Research studies show that feedback is a necessary step to enforce students' behavior change[6]. In the future study, we will employ a feedback system and study how the implication of a feedback device will change the students peer evaluation performance.

Bibliography

1 ABET (2002). Engineering Criteria 2002-2003. Accreditation Board for Engineering and Technology, http://www.abet.org/criteria.html, accessed 2/1/2009.

2 J. McGourty, P. Dominick, and R.R. Reilly. Incorporating student peer review and feedback into the assessment process. presented at theBest Assessment Processes in Engineering Education: A Working Symposium Sponsored by National Science Foundation, ABET, andRose-Hulman Institute of Technology, Terre Haute, IN, Apr. 1997.

3 M. Besterfield-Sacre, L.J. Shuman, H. Wolfe, C.J. Atman, J. McGourty, R.L. Miller, B.M. Olds, and G.M. Rogers. Defining the outcomes: A framework for EC-2000. *IEEE Transactions on Education*, 43(2):100–110, 2000.

4 L.J. Shuman, M. Besterfield-Sacre, and J. McGourty. The ABET " professional skills"- can they be taught? Can they be assessed? *Journal of Engineering Education*, 94(1):41–55, 2005.

5 M.W. Ohland and R.A. Layton. Comparing the reliability of two peer evaluation instruments. In *Proceedings. ASEE Annual Conference & Exposition, St. Louis, MO*, 2000.

6 D.F. Baker. Peer assessment in small groups: A comparison of methods. *Journal of Management Education*, 32(2):183, 2008.

7 P.G. Dominick, R.R. Reilly, and J.W. McGourty. The effects of peer feedback on team member behavior. *Group & Organization Management*, 22(4):508, 1997.

8 N. Falchikov and J. Goldfinch. Student peer assessment in higher education: A meta-analysis comparing peer and teacher marks. *Review of Educational Research*, 70(3):287, 2000.

9 L.E. Gueldenzoph and G.L. May. Collaborative peer evaluation: Best practices for group member assessments. *Business Communication Quarterly*, 65(1):9, 2002.

10 Imbrie, P.K., Maller, S.J., and Immekus, J.C. (2005) Assessing Team Effectiveness, *Proceedings of the 2005 American Society for Engineering Education Annual Conference & Exposition*, Portland, Oregon

11 Wang, J., Imbrie, P.K., "Assessing Team Effectiveness: Comparing Peer-evaluations to a Team Effectiveness Instrument" Proceedings of the 2009 American Society for Engineering Education Annual Conference & Exposition, Austin, TX

12 R.L. Holzbach. Rater bias in performance ratings: Superior, self-, and peer ratings. *Journal of Applied Psychology*, 63(5):579–588, 1978.

13 K.G. Love. Comparison of peer assessment methods: Reliability, validity, friendship bias, and user reaction. *Journal of Applied Psychology*, 66(4):451–457, 1981.

14 R.J. Klimoski and M. London. Role of the rater in performance appraisal. *Journal of Applied Psychology*, 59(4):445–451, 1974.

15 C.S. Alexander and H.J. Becker. The use of vignettes in survey research. *Public opinion quarterly*, 42(1):93, 1978.