New Orleans JAZZED about Engineering Education

Paper ID #16485

Exploring the Value of Peer Assessment

Mrs. Sally Sue Richmond, Pennsylvania State University, Great Valley

Sally Sue Richmond is a Lecturer in Information Science at the School of Graduate Professional Studies, Penn State Great Valley. Richmond has a B.A. in Art and an M.S. in Information Science from The Pennsylvania State University. She has 25+ years experience in industry as a software developer, network analyst, trainer, and Help Desk supervisor. She teaches courses in Human-Computer Interaction, Computer Organization and Design, Computer Forensics, Microprocessors and Embedded Systems, Networking, and IS Architecture. She has published articles in conference proceedings and journals in the areas of concept mapping, cognitive style, and engineering education.

Dr. Kailasam Satyamurthy, Penn State University

Dr. Kailasam Satyamurthy is an Assistant Professor in Engineering at Penn State University. He earned his Ph.D. in Engineering Mechanics from Clemson University and an MBA from Penn State. Before joining Penn State, he was a senior manager at Vanguard for 8 years and head of the engineering department at GenCorp for 20 years. He teaches Decision and Risk Analysis, Business Statistics, Finance and Economics for Engineers, Quantitative Methods in Finance and Quality and Continuous Improvement courses at Penn State. At GenCorp, he did extensive research in the mathematical modeling and developed methodologies and algorithms for the nonlinear finite element analysis of mechanical systems under mechanical and thermal loadings. He is also a six sigma master blackbelt and trained numerous professionals in manufacturing, transactional and healthcare industries.

Dr. Joanna F. DeFranco, Pennsylvania State University, Great Valley

Joanna F. DeFranco is Assistant Professor of Software Engineering in the School of Graduate Professional Studies, Penn State Great Valley. Dr. DeFranco holds a B.S. in Electrical Engineering from The Pennsylvania State University, a M.S. in Computer Engineering from Villanova University and a Ph.D. in Computer and Information Science from New Jersey Institute of Technology. She teaches in both the resident and online software engineering, systems engineering, and engineering management graduate degrees. She has published a number of articles in journals and conference proceedings in the area of technical teams and engineering education.

Exploring the Value of Peer Assessment

Sally S. Richmond, Kailasam Satyamurthy, and Joanna F. DeFranco
The Pennsylvania State University

We have collected peer-assessment (PA) and self-assessment (SA) data from two resident sections of a software construction course. This course is a core requirement in a graduate program in software engineering at a large research university. While the body of research gives strong evidence that there are many benefits to implementing peer and self-assessment, concerns remain. Two concerns are that students will inflate their evaluation of themselves and that they may collude to give each other high ratings ("cronyism"). These concerns motivated this exploratory study of student bias in peer and self-assessment in a graduate engineering program. Our results confirm previous research that students tend to rate themselves higher than their peers, but we found no evidence of cronyism.

I. Introduction

Student assessment can be complex task for an engineering instructor. Assessment is a vast topic with many options. There are generally two types of knowledge to assess: declarative and procedural. Instructors vary the assessment method depending on that category of knowledge. For example, instructors can easily measure a student's declarative knowledge with a written exam and procedural knowledge with a project. However, in a graduate engineering course assessing procedural knowledge is more prevalent, since the student must apply the declarative knowledge appropriately in order to succeed in our competitive workforce. In addition to an instructor's method to evaluate a project, both peer and self-evaluation are often used as a complement to an instructor's evaluation. Assessing software engineering projects poses many challenges. One challenge stems from the fact that a software engineering project consists of multiple requirements, and there can be multiple effective solutions for each. For example, in a software construction course where students develop a software system, one requirement may be to capture user input. Students may capture this requirement with varying degrees of complexity and sophistication based on their development experience. Since there is no single correct answer, instructors may utilize additional modes of assessment for this project, such as peer assessment. This also mirrors common practice in industry, since software engineers are usually expected to evaluate their colleagues work and provide feedback. So peer assessment in a software construction course provides two benefits to students, they both receive suggestions for improving their own work and gain practice in giving feedback.

Peer and self-evaluation are not universally accepted by engineering faculty. Some instructors question its reliability, and some question whether it improves learning. In other words, the concern is that the student assessments (whether peer or self) may be biased. Some instructors may hesitate from using student peer evaluation to assess projects due to concerns about its reliability. One concern is whether students can assess themselves and classmates as accurately as the instructor, who has greater knowledge and understanding of the material. Other concerns are that students may rate themselves higher than their classmates, or engage in cronyism where students come to an informal agreement to give each other the highest assessment. Despite these concerns, research has shown that self-assessment and peer assessment are more effective in improving learning than instructor formative assessment alone (De Sande & Godino-

Llorente, 2014). Numerous studies conducted over decades have explored a variety of aspects about peer and self-assessment, including reliability. The results regarding reliability when taken overall are inconclusive, and these studies usually compared peer and self-assessment with the instructor's assessment. Fewer studies looked at student peer and self-assessments with each other. In addition, the majority of these studies include only undergraduate students.

In this paper we are exploring the effectiveness of a peer evaluation. Graduate software engineering students enrolled in a software construction course individually designed and developed a software system. This paper describes exploratory study that looked primarily at bias in peer and self-assessment of the software systems developed in that class.

II. Background

The literature on PA and SA fall into three broad categories – research to determine whether use of PA and SA affects learning, use of PA and SA to foster collaboration and assess contribution of individual members within teams, and use of PA and SA as assessment tools.

The significant amount of research on how PA and SA affect learning indicates that it generally improves learning. These studies span a range of disciplines, such as computer science (Lin, Liu, & Yuan, 2001), history (Van den Berg, Admiraal, & Pilot, 2006), and psychology (Sung, Lin, Lee, & Chang, 2003. Studies conducted with undergraduate engineering students concluded that SA and PA improved oral presentation skills (De Grez, & Valcke, 2013), writing skills (McConlogue, Mueller, & Shelton, 2010), was more effective than instructor-only assessment in a required Signals and Systems course (De Sande & Godino-Llorente, 2014), and improved understanding of dimensioning techniques (Study, 2015).

Other research focuses on the use of PA and SA within teams. Assigning course work to student teams is a common practice in higher education. The benefits of student teams are many, including improved learning and enhanced teamwork skills (Elliott & Higgins, 2005; Willey & Freeman, 2006). Team skills are critical for engineering students, since they will likely work in teams throughout their careers. As with individual student work, it can be used to improve learning, it can also help teams form a common vision. Two motivations are given for using PA and SA in teams. The first is to encourage all team members to contribute equitably, since typically all team members receive the same grade regardless of individual contribution (the so-called "free loader" effect) unless peer-assessment is included in grades. While instructors and students alike express the most concern regarding the reliability of using PA and SA to determine individual contribution to the work produced by the team, multiple studies conclude that using peer evaluation to assess individual performance within a team discourages free-loading and increases students' perception of grading fairness (Elliott & Higgins, 2005).

The third area of research looks at the use of PA and SA as an assessment tool, where the assessment is included in calculating the student's grade. Instructors express concerns that PA and SA are inherently biased because students rate themselves higher than their classmates in SA, may rate their peers lower in PA, or may participate in cronyism (where two or more students mutually agree to give each other the highest possible assessment). Multiple studies report the reliability of PA when compared with instructor assessment (IA) is acceptable (De Sande & Godino-Llorente, 2014; Falchikov & Boud, 1989; Marin-Garcia & Miralles, 2008; and Ward, Gruppen, & Regehr, 2001). There is less evidence to support the reliability of SA, rather, multiple studies have concluded that it is unreliable (Falchikov & Boud, 1989, Mishra, 2015; Gopinath, 1999; Ryan, Marshall, Porter & Jia, 2007; and Burchfield & Sappington, 1999). To the best of our knowledge, cronyism has not been studied with regards to peer assessment.

While SA appears to be unreliable and therefore should not be used in assessment, research suggests PA exhibits sufficient reliability to include it as part of student assessment (grades). One reason to use PA is that students, particularly engineering students, are expected to give peer assessments in the workplace. Marin-Garcia and Miralles (2008), Wellington, Thomas, Powell, and Clarke (2002)., and Vidic (2010) all note that peer assessment is an important professional skill that engineers need to succeed in their careers, along with critical thinking ability, effective team and communication skills, and ability to engage in lifelong learning. Nonetheless, students may take part in cronyism or be reluctant to give their classmates poor peer assessments even when doing so is appropriate because they do not want to negatively impact the classmate's final grade. This is particularly relevant when students are working adults and rely on employer tuition reimbursement to fund their education, as employers often pro-rate the amount of reimbursement based on the student's course grade.

Despite the large body of work in this area, most of the studies were conducted with undergraduate students. This study investigates use of PA and SA as an assessment tool for a graduate-level software engineering course.

III. Course Description

The course used in this study is a graduate level software engineering course in software construction taught at a large public research university in the Mid-Atlantic. The course is intended to help students learn and apply software engineering principles through developing a large-scale application that uses distributed objects and web technologies. Students are given relevant reading to complete prior to each class. Class time is spent on lecture, followed by hands-on activities and classroom discussion to reinforce and apply concepts covered by the reading and lecture.

Students are assessed on class participation and activities, a timed final exam, the development project, and a peer evaluation of their project presentation. The project consists of five deliverables that students complete individually throughout the semester. For the final project deliverable, each student demonstrates the working application and explains how they applied software engineering principles to design and develop the software and any challenges they experienced during the process. The instructor assesses all components other than the peer evaluation. Weights for each component are shown in Table 1.

Assessment	**% of Final Grade**
Project – Deliverable 1	5%
Project – Deliverable 2	10%
Project – Deliverable 3	15%
Project – Deliverable 4	15%
Project Presentation and Software Demonstration	10%
Class Participation and Homework	20%
Peer Reviews	5%
Final exam	20%
Total	**100%**

Table 1 Grade point allocation of all assignments toward final grade

IV. Peer Assessment

This section first discusses the use of assessment in education and software engineering practice in general, then focuses more specifically on the use of assessment to evaluate performance, and finishes by describing how this general view was used to develop the peer assessment form used for the courses used in this analysis.

Assessment is common in both education and in the practice of software engineering in commercial organizations. As noted above, peer assessment may be done in higher education to facilitate learning, foster collaboration, and assess performance. Peer assessment is done in software engineering practice to find errors, verify conformance to requirements and standards, suggest improvements, and to assess performance. A variety of assessment modes are employed in education, this is also true of software engineering practice. The Software Engineering Body of Knowledge (SWEBOKv3) includes sections on assessing the product (the artifacts created throughout the process of developing, deploying, and maintaining an application), the process (the activities and work done to create the product), and software engineering management ("planning, coordinating, measuring, monitoring, controlling, and reporting activities to ensure that software products and software engineering services are delivered efficiently, effectively, and to the benefit of stakeholders", SWEBOKv3, 7-1). Performance assessment falls within the area of software engineering management (SWEBOKv3, 7-9). To relate the idea that software engineering includes assessment of the product, process, and performance to education, we consider the learning objectives for a course to be the product, the teaching techniques employed as the process, and the individual student's learning to be performance.

In both disciplines a number of measures, techniques, and modes can be employed to assessment product, process, and performance. However, those used in education are developed with greater concern for validity and reliability than typically seen in software engineering practice, particularly when assessing performance. In general, the measures, measurement techniques, and analysis of the gathered metrics used by software engineering practitioners are less formal than those used in education research and are not statistically validated. This is particularly true when assessing performance. SWEBOK identifies multiple techniques to assess the correctness of the software engineering product and process that are quite rigorous. However, little guidance is given in the SWEBOK regarding how to evaluate performance, other than it should be done and the process and measures used to do so should also be evaluated periodically (SWEBOKv3, 7-0).

Given the relative informality of assessment in software engineering practice and the absence of any specific performance evaluation methods in SWEBOK, we chose to use a peer evaluation form where students assess the work at a high-level, and defined performance as the amount of effort expended on the development project. The peer assessment evaluates two aspects, the quality of the project and the quality of the presentation. Students are given a form instructing them to rate each aspect on a scale of 1-5, with a one representing the poorest quality and a five representing the highest. Therefore the assessment rating could range from 1-10. Project quality is defined as addressing the software requirements as well as the user interface configuration, quality and detail of output, software structure and extensibility, and number of errors detected. The form includes a rubric that provides qualitative descriptions for each aspect.

This study looked at the data from two sections of the software construction course. There were 15 and 17 graduate software engineering students respectively. In the first section students rated the presentations of each of their classmates but did not rate themselves. In the second section students rated themselves along with their classmates.

IV. Results

The most notable observation is that students do not use the entire scale range; no student was assessed at the low end of the possible range (3, 2, or 1). This may be due to a tendency for students to inflate how they evaluate their peers, since the students know their peer evaluations account for 5% of their final grade. Regardless of the reason, the effect of this reluctance to use the full range is that the data are essentially ordinal data and not interval or ratio. For example, if an evaluator gave student S1 a rating of nine and student S2 a rating of ten, all we can conclude is that the evaluator thought student S2 performed better than S1. We cannot conclude that the evaluator thought student S2 performed 10% better than student S1 (which would be the case with interval data, since the scale ranges from 1-10 and therefore each scale value reflects a 10% change). The data are also not ratio data, since zero is not a possible rating (a student would only receive a zero if they did not present their work, in which case there would be no peer evaluation). This restricts the types of analysis that can be applied to the data, and reduces the sensitivity of any conclusions (Trochim, 2006). Nonetheless, interesting conclusions regarding bias and potential cronyism can be drawn when we organized the ratings into ranges and looked at the frequency of ratings in each range, as discussed below.

Given the small amount of data and that reliability has been studied extensively, we chose to investigate to what extent students used the entire range of assessment values when evaluating themselves and their peers. We also looked for any student who had a low mean peer evaluation but one or two high peer ratings, as this could suggest cronyism. We grouped the range of possible ratings into four groups, ratings from 1 to 3, ratings from 4 to 6, ratings from 7 to 9, and a final group for 10, the highest possible rating. For each student we summed the number of peer assessment ratings in each of the groups, as shown in Table 2 for the first section. The second section also did self-assessment, and this is shown in Table 3. We gave each student's data an arbitrary identifier to maintain privacy. Note that the total number of ratings is one less than the total number of students in section 1 since those students did not assess themselves.

	Number of peer assessments in range			
Student	**10**	**9, 8, or 7**	**6, 5, or 4**	**3, 2, or 1**
S1	14	0	0	0
S2	13	1	0	0
S3	7	7	0	0
S4	7	6	1	0
S5	6	8	0	0
S6	5	8	1	0
S7	4	10	0	0
S8	4	10	0	0
S9	4	9	1	0
S10	3	10	1	0
S11	3	10	1	0
S12	3	9	2	0
S13	3	9	1	0
S14	0	10	4	0
S15	0	9	5	0

Table 2 Frequency of peer assessment within given ranges (n=15, students did not assess themselves)

We observed that the ratings for eight of the fifteen students (53%) span only one or two adjacent frequency groups (students S1, S2, S3, S5, S7, S8, S14, and S15). Of the remaining seven students, six had only one rating in a third frequency group (students S4, S6, S9, S10, S11, and S13), and the other (student S12) had two in a third frequency group. This suggests that cronyism is not evident in this section. Cronyism would be suspected if a student had one rating in a non-adjacent frequency group, e.g., one in frequency group 10 (the highest rating) and the remainder in lowest two frequency groups (6, 5, 4 group and 3, 2, 1 group), or a pattern of one rating in frequency group 10, just a few in the adjacent frequency group, and the majority of ratings in the next adjacent group (e.g., one in frequency group 10, one or two in frequency group 9, 8, 7, and the remainder in frequency group 6, 5, 4).

Students in the second section completed self-assessment along with peer assessment. The results are shown in Table 3.

	Number of peer assessments in range				
Student	**10**	**9, 8, or 7**	**6, 5, or 4**	**3, 2, or 1**	**Self-assessment**
S1	14	3	0	0	8
S2	13	4	0	0	10
S3	10	7	0	0	10
S4	9	8	0	0	9
S5	8	9	0	0	10
S6	8	9	0	0	10
S7	7	10	0	0	10
S8	7	10	0	0	10
S9	7	10	0	0	10
S10	7	9	1	0	10
S11	6	11	0	0	10
S12	6	11	0	0	8
S13	6	11	0	0	10
S14	6	9	2	0	8
S15	5	12	0	0	9
S16	5	12	0	0	9
S17	3	7	7	0	10

Table 3 Frequency of peer assessment within given ranges (n=17, students assessed themselves)

As with section 1, students again gave more ratings in the two highest frequency groups. They also tended to rate themselves higher than their peers did. Almost 53% of the students rated themselves higher that the most frequent rating given by their peers (students S5, S6, S7, S8, S9, S10, S11, S13, and S17), whereas only 12% rated themselves lower than the most frequent peer rating (students S1 and S4). The remaining 35% rated themselves the same as the most frequent peer rating. We conclude that students do tend to rate themselves higher than their peers.

Since fourteen of the seventeen students (82%) had ratings spanning only two adjacent frequency groups, there is less evidence of widespread cronyism. It is noteworthy that student S17 had three ratings in

frequency group 10 and the remainder of ratings split evenly in the next two frequency group ratings. This could suggest possible cronyism, but if so, student S17 would have negotiated with two other students (S17 gave a self-assessment of 10, meaning that two other students would have had to agree to give student S17 the highest rating). This implies that the ratings for two other students would also exhibit a similar pattern, which is not the case.

The self-assessment rating suggest that students tend to rate themselves higher than their peers, which is consistent with Liow (2008). Nine of seventeen students rated themselves higher than the most frequent peer rating. Only two students (S1 and S4) rated themselves lower than the most frequent peer rating. is the lone exception. Fourteen of seventeen students (82%) rated S1 with a 10, whereas S1's self-assessment is only an 8. This could be due to gender, as one study found female students tended to rate themselves lower than their male counterparts (Lind, 2002), however, this is not conclusive since Gopinath (1999) found gender was not a factor in self-assessment. It could also be due to the tendency of successful students to rate themselves lower than their peers and unsuccessful students to rate themselves higher (Mishra, Ostrovska, & Hacaloglu 2015; Falchikov & Boud, 1989; and Gopinath, 1999). Since the instructor did not keep any personally identifiable information we are unable to explore this further here, but plan to do so in future work.

V. Discussion

The research clearly shows that PA provides value that is distinct from the value of SA in improving learning, fostering collaboration and weighting individual contribution to work done by teams, and as an assessment tool. It provides additional feedback beyond instructor assessment alone and engineering students are expected to regularly conduct peer assessments throughout their careers. Hence it is prudent for engineering educators to continue its use. Best practices suggest that PA is most reliable when students are given operational definitions for the assessment criteria along with training in conducting peer assessments.

Consistent with the literature, the authors found that students tend to give themselves higher self-assessments than how their peers assess them. On a more positive note, there was no evidence of cronyism. However, this needs to be explored further, as does the question of whether the tendency to give peer assessments at the higher end of the scale is due to realistic assessment of the quality of work done or student reluctance to mark their classmates negatively.

This study was exploratory in nature. We have held the assumption that peer and self-evaluation can be valuable for learning and assessment. Since it is common in software engineering practice and our students are working as software engineers or developers, we adopted the use of peer assessment in a manner similar to that done in commercial organizations, where the measures and methods are developed far more informally than when used for education research. However, the concerns regarding peer assessment expressed in both previous research and collected anecdotally from colleagues, and the smaller body of research done with graduate students, made it apparent that a more robust study of peer assessment could be useful.

This study has several limitations. The samples are small and the only attempt to validate the assessment instrument was done through subjective review by our colleagues. The study made it clear that multiple questions remain. For example, we would like to understand more fully why students do not use the full scale range, determine whether the results would be different if students received training on the use of the assessment tool or if students were graded on the quality of their peer assessment of others (perhaps

by awarding grade points for using a wider range of the scale)? Would results differ if peer assessments were used simply to provide feedback to students and not used as part of the final grade?

The authors intend to continue the use of PA as an assessment tool, incorporating more operational assessment criteria and training. In addition, future research will include data collection in a more rigorous manner as well as designing a larger research plan to address some of these issues.

VI. References

De Grez, L., and Valcke, M. (2013). Student response system and how to make engineering students learn oral presentation skills. *International Journal of Engineering Education, 29*(4), 940-947.

De Sande, J.C.G., and Godino-Llorente, J.I. (2014). Peer assessment and self-assessment: effective learning tools in higher education. *International Journal of Engineering Education, 30*(3), 711-721.

Elliott, N., and Higgins, A. (2005). Self and peer assessment – does it make a difference to student group work? *Nurse Education in Practice, 5*(1), 40-48.

Falchikov, N., and Boud, D. (1989). Student self-assessment in higher education: A Meta-Analysis. *Review of Educational Research, 59*(4), 395-430.

Gopinath, C. (1999). Alternatives to instructor assessment of class assessing team work in engineering projects participation. *Journal of Education for Business, **75***(1), 10–14.

Lin, S. S. J., Liu, E. Z. F., and Yuan, S. M. (2001). Web-based peer assessment: feedback for students with various thinking-styles. *Journal of Computer Assisted Learning*, 17, 420-432.

Lind, D. S. , Rekkas, S. , Bui, V. , Lam, T., Beierle, E., and Copeland, III, E. M. (2002). Competency-based student self-assessment on a surgery rotation, *Journal of Surgical Research, 105*(1), 31–34.

Liow, J. (2008). Peer assessment in thesis oral presentation. *European Journal of Engineering Education,* 33(5–6), 525–537.

Marin-Garcia, J. A., and Miralles, C. (2008). Oral presentations and assessment skills for engineering education. *International Journal of Engineering Education, 24*(5), 926-935.

McConlogue, T., Mueller, J., and Shelton, J. (2010). Challenges of developing engineering students' writing through peer assessment. *The Higher Education Academy Engineering Subject Centre, EE2010*.

Mishra, D., Ostrovska, S., and Hacaloglu, T. (2015). Assessing team work in engineering project. *International Journal of Engineering Education, 31,*(2), 627-634.

Study, N., E. (2015, June14 -17). *Using peer review in a freshmen engineering graphics course to enhance understanding of basic dimensioning techniques.* Paper presented at the 2015 ASEE Annual Conference and Exposition. doi: 10.18260/p.25010.

Sung, Y. T., Lin, C. S., Lee, C. L., and Chang, K. E. (2003). Evaluating proposals for experiments: An application of web-based self-assessment and peer-assessment. *Teaching of Psychology*, 30, 331-334.

Trochim, W. M. (2006). *The Research Methods Knowledge Base* (2nd ed.). Levels of Measurement: http://www.socialresearchmethods.net/kb/. Retrieved 16 March 2016.

Van den Berg, I., Admiraal, W., and Pilot, A. (2006). Design principles and outcomes of peer assessment in higher education. *Studies in Higher Education*, 31, 341-356.

Wellington, P., Thomas, I., Powell, I., and Clarke, B. (2002). Authentic assessment applied to engineering and business undergraduate consulting teams. *International Journal of Engineering Education, 18*(2), 168-179.