Self-Assessment to Improve Learning and Evaluation

Dr. Edward F. Gehringer, North Carolina State University

Dr. Gehringer is an associate professor in the Departments of Computer Science, and Electrical & Computer Engineering. His research interests include computerized assessment systems, and the use of natural-language processing to improve the quality of reviewing. He teaches courses in the area of programming, computer architecture, object-oriented design, and ethics in computing.

Self-Assessment to Improve Learning and Evaluation

Abstract

Self-assessment has many advantages for student learning. By inducing students to think about their own learning, it encourages metacognitive practices that deepen learning. It helps them to gain perspective, by thinking about how assignments fit into the context of their education. By itself, self-assessment is a useful formative exercise. While self-assessment is not reliable when students do not understand the material well, it is possible to combine it with peer assessment or instructor assessment to derive valid grades. There are several approaches to including a self-assessment component in a student's grade, on the basis that accurate self-assessment itself demonstrates learning gains. The contribution of this paper is to summarize research on self-assessment over time, including where it has and has not proved successful, as well as to survey several approaches and software applications for incorporating self-assessment into a course.

Keywords: self-assessment, peer assessment, evaluation rubric

1. Introduction

Self-assessment is a powerful mechanism for enhancing learning. It encourages students to reflect on how their own work meets the goals set for learning concepts and skills. It promotes metacognition about what is being learned, and effective practices for learning. It encourages students to think about how a particular assignment or course fits into the context of their education. It imparts reflective skills that will be useful on the job or in academic research.

Most other kinds of assessment place the student in a passive role. The student simply receives feedback from the instructor or TA. Self-assessment, by contrast, forces students to become autonomous learners, to think about how what they *should* be learning. Having learned self-assessment skills, students can continue to apply them in their career and in other contexts throughout life.

Self-assessment is a useful life skill. In school, students are told what they need to learn, but in the work world, they usually need to figure it out for themselves. Most students do not come into higher education with this skill well developed. But if they are to emerge as graduates who can take responsibility for their own learning, they must understand what they have learned so far and why they need to study the next topic at hand. To some extent, this skill is discipline specific, expressed quite differently in, say, history than in calculus. Students who are able to self-assess are more likely to continue their learning and increase their competence after graduation [1].

2. How Is Self-Assessment Used?

Most uses of self-assessment are purely formative [1], i.e., undertaken to help students improve their work, rather than to assign a final grade. Students may learn self-assessment in first-year courses devoted to teaching them how to study and learn. They can use self-assessment to monitor their own learning, either to keep them on track to reach their goals, or to fulfill requirements of their course. If they are in difficulty, their advisor may help them think about what might be the most effective ways to improve their learning.

Self-assessment is often used in medical schools as a way of improving students' clinical skills. Nursing education, like teacher education, emphasizes becoming a reflective practitioner. Thinking about how current material relates to material learned in other courses can benefit anyone in a degree program. Capstone courses serve to encourage self-assessment by requiring students to apply previously-learned skills to new contexts.

The e-portfolio movement is another practice that encourages self-assessment. Students put together work that they have produced in various contexts. The very act of assembling it promotes self-assessment, even if it is not identified as such.

Self-assessment has summative uses as well. The most direct summative use is for the staff to grade students on the quality of their self-assessment, perhaps as one component of the grade for an assignment. There are also several approaches to combining self-assessment grades with peer grades or staff grades, as we shall see.

3. How Do Self-Assessment Scores Compare?

The Dunning-Kruger effect [2] tells us that people often overestimate their knowledge in fields where they have limited competence. This tells us that even if students didn't have a self-serving interest in grading themselves highly, we should be skeptical about using self-assessments summatively (e.g., as homework grades). A perusal of the evidence, though, indicates that the situation is not entirely bleak. In many cases, self-assessments do approximate disinterested assessments, especially in formative situations. Let us consider several sub-questions.

3.1. Do students tend to underrate or overrate themselves? Usually, staff-assigned grades are taken as the "gold standard" for student ratings. This is not to say they are always accurate; staff may not have the time to completely assess each student's work, and they may grade the first paper differently from the last, after they have been exposed to the common errors students make. But in most instances (Piech et al. [3] being a notable exception), staff grades are used to judge the accuracy of student assessments.

Many studies have addressed this question. In their 1995 survey [4], Boud and Falchikov tally 16 studies that report overrating by students, against 11 that report underrating. More recently, Gonzales and Murthy [5] studied self-assessment of problem sets in a Signals and Systems course. They found that self-assessment scores were "slightly" higher than instructor assessments. Liang et al. [6] looked at self-assessment vs. peer and instructor assessment in a 3rd-year course using Problem-Based Learning (PBL) at Fudan University's medical school. They found, over 117 observations, that self-assessed grades (mean of 95.5%) exceeded peer-assessed grades (94.7%) and tutor assessments (mean of 90.9%). However, Papinczak et al. [7], at the University of Queensland medical school, noted the opposite: on each assignment, students scored themselves significantly lower than the staff did ($p < 0.001$ for a *t*-test).

Kulkarni et al. [8] reported on two iterations of a MOOC on computer-human interfaces. In the first iteration, self-grades exceeded peer grades by a median value of 6%; in the second running, the self-grades were higher by a median of 7.5%. Not all submissions were graded by staff, but among those that were, the staff grades were higher the first time (48% of the staff grades exceeded peer grades, but only 40% of peer grades exceeded staff grades). The second time around, the situation was reversed, with 46% of the peer grades exceeding staff grades, and only 36% falling short of staff grades.

Contrast these results with those reported by Wilkowski et al. [9]. In a Google MOOC on advanced power searching, 2708 of 3853 students (70.3%) gave themselves full credit. Of these, 9.9% were blank or nonsense. Further, 8.5% of these full-credit submissions were duplicates of work submitted by others. Three of these artifacts were each submitted by more than 40 students. However, in the other MOOC they studied, on mapping with Google, results were much more encouraging. There were 5160 submissions, out of which course staff graded a random sample of 384 artifacts that were also self-assessed. In 52.3% of the cases, self- and TA scores were within 1 point of each other (on a 27-point scale); 71.6% were within 2 points of each other. Only 0.3% of full-credit submissions were found to be duplicates of other artifacts.

Given the lack of consistency in these results, it seems that other factors may play a dominating role in determining whether students underrate or overrate themselves. These may include the level of detail of the rubric, which may call students' attention to areas where their work could be improved; the amount of training they have received in self-assessment; how much opportunity they have had to compare self-grades with instructor grades; and whether they think that scoring themselves highly will help or hurt their course grade (it could do either, as explained in Section 4).

3.2. Do students at different ends of the performance spectrum underrate or overrate themselves? Here the result is clearer. High-performing students tend to underrate themselves, and low-performing students overrate themselves. Boud and Falchikov [4]

report results from eleven studies that looked at this issue between 1932 and 1988 (all but one of which were after 1962). Low performers overrated themselves in all of these studies. In nine of the eleven studies, high performers underrated their performance (the remaining two papers said merely that high performers were more accurate raters).

3.3. Do students underrate or overrate themselves when self-assessment is summative? Boud and Falchikov [4] found 7 studies that addressed this point between 1969 and 1988 (this group was distinct from the studies covered in 3.2). Five of the seven said that students overrated themselves in summative situations. In the remaining two studies, it depended on level (undergraduates overrated themselves while graduate students underrated themselves) or discipline (civil engineers underrated themselves).

3.4. Do students improve their ability to self-assess over time? Here Boud and Falchikov [4] found that students tended to improve over time (4 studies) than stay the same (1 study) or get worse (1 study). In one of the cases, the improvement occurred after "hearing detailed criticism of [the students'] work." However, the results were not clear, especially because the papers did not track the amount of training students received in self-assessment. But studies [10, 11] report that training improves students' peer-assessment performance, so it seems likely that training benefits self-assessment performance too.

3.5. Are there gender differences in self-rating? Falchikov and Boud [4] enumerate six studies that addressed the issue between 1969 and 1988. Three of them found no differences. One found that women's self-ratings were closer to instructor ratings. One found that men underrated themselves less than women. One study on teacher training found that male teacher trainees in elementary classrooms overrated themselves, but in secondary classrooms, women trainees overrated themselves.

The problem of getting accurate ratings from self-assessment has much in common with the problem of engaging accurate peer assessments. Twenty years ago, Calibrated Peer Review™ [12] pioneered the *calibration* approach, where students are asked to rate three works that had previously been rated by the instructor. Students who assign scores that closely track the instructor's are assigned a higher Reviewer Competency Index, and their work is accorded more weight in computing a peer grade. Since then, other approaches have been developed, such as *reputation systems* [13–15] that compare student-assigned scores on the same work to infer which ratings are more trustworthy. Natural-language processing has been used [16] to provide feedback to reviewers (and to the instructor) on the quality of their reviews, and give automated advice on how to improve them. Work has been done on combining these techniques [17] to predict when a peer assessment is accurate.

4. How Can Self-Assessment Scores Be Combined with Other Scores?

As Section 2 described, the simplest approach to including self-review in a student's grade is for the staff to assign the self-review a grade, which counts for part of the assignment. While this is straightforward to do, it does require extra time on the part of the instructor. Several other strategies include the results of self-assessment without requiring it to be graded separately.

Most well known, perhaps, is the approach of Calibrated Peer Review (CPR) [12, 18]. In CPR, students grade three artifacts submitted by their peers, but do not see the peer assessments that they have received before they complete a self-review. If the instructor sets it up that way, the difference between a self-review and the weighted average of peer reviews can be one of the factors in computing the student's grade for the assignment. The Coursera MOOC [8] employs a similar strategy. When the student's self-assessed score was within 5% of the median peer-assessed score, a student received the higher of the two scores as the grade on the assignment.

The Fair Contribution Scoring (FCS) approach [19] is designed for semester-long team projects. Students are given a fixed number of points (20 per team member) to divide among all members of the team, including themselves. Each week they fill out an assessment comparing their contribution with peers' contributions in a number of dimensions. The point totals awarded to each student are added, with self-ratings being included in the total. The total is then divided by the number of students in the group to derive a weighting factor. This weighting factor is multiplied by the grade awarded to the student's work to derive mid-term and final grades. Teams whose weighting factors are nearly equal (smallest standard deviation) are deemed to have better cooperation.

Van den Bogaard et al. [20] identify three way that self-evaluations can be combined with peer evaluations.

1. Ranking. Students can be asked to rank the contributions of their partners and themselves from 1 (best) to *n* (worst), where *n* is the number of students in the group.
2. Rating. Students are given a rubric based on learning objectives, and asked to rate themselves and their partners on the rubric.
3. Division of a certain amount of assets, as in FCS, where the initial asset pool is 20 × the number of students in the group.

The authors designed a rating-based system because it made it easiest to give students feedback on different aspects of their performance. It is named PeEv, for "peer evaluation." Students rate their partners on the rubric criteria, and the system calculates the average rating of each student, and the composite average rating of all students. They do this once in mid-term and once at the end of the term. Tutors also rate the students, but tutor data is not included in the final score. The final grade is computed as

$$G_i = G_g + (AV_i - AV_g)/2$$

where G_i is the individual grade, G_g is the grade for the group, and AV_i and AV_g are the respective average grades.

SPARK^{PLUS} [21] is a self- and peer-assessment system developed at the University of Technology, Sydney. It is also a rating-based system,where students fill out a rubric assessing their and their teammates' contributions. It produces two assessment factors. To compute the first factor, it multiplies the project score by a student-specific contribution ratio (called the "Self- and Peer Assessment, or SPA) to derive a final grade (this is similar to the FCS approach). The second factor is known as the Self-Assessment to Peer Assessment factor (SAPA). A SAPA of > 1 means that the student rated his/her contribution higher than the other team members did; a SAPA of < 1 means that the teammates rated the team member's performance higher than the team member did. This factor is not used in calculating the final grade, but the authors state [22], "The potential embarrassment of receiving a SAPA factor much higher than 1.0 appeared to be a significant motivator in promoting honest assessment."

Other online systems, including CATME [23], iPeer [24], PEAR [25], and Teammates [26] collect peer and self-assessments of students' contributions to teams. The instructor, of course, can use the output of these systems to adjust students' project grades. But since they don't build in any way to adjust grades based on self-assessment, they are not considered further in this section.

Summarizing these four approaches (CPR, FCS, PeEv, and SPARK), we can see three different ways in which self-evaluation scores can contribute to a composite score for an assignment: (1) the *agreement* between self- and peer-assessment scores can be one component of the overall grade; (2) self-assessment scores can be combined with peer-assessment scores to produce a weighting factor by which a team grade can be multiplied to produce a final grade; and (3) the amount by which a student exceeded his/her teammates scores can be added to the group grade to produce a final grade. Note that all of these approaches assume that peer assement is also performed. In principle, staff assessment could be substituted for peer assessment, but (1) this would consume much more staff time, and (2) students would miss out on the metacognitive benefits of evaluating others' work. It is true, however, that efficiently processing peer assessments requires significant IT support (see Babik et al. [27] for a discussion of the options).

Table 1 shows how the four approaches compare. CPR (and the similar training program used by Coursera) contrasts with the other three approaches because (i) it is used to assess artifacts (writing, reporting, etc.) rather than student contributions to a team, and because it employs calibration to train the students in effective reviewing. In using self-assessed scores, PeEv *adds* (or subtracts) an amount to/from the student's grade based on how individual grades compare with group grades; the other approaches determine

individual grades by *multiplying* the overall grade by a factor computed from self- and peer reviews. CPR also uses weighting, but it computes a final grade by weighting the scores given by peer reviewers on the basis of calibration results, with students who did better in calibration having their grades count more heavily in determining the grades of the projects they review. The Coursera approach is the only method that can assign a self-grade as the student's final grade; it does this when the self-grade exceeds the median peer grade by less than 5%. Two approaches (PeEv and SPARK) collect information that is not used in the final grade. In PeEv, it is the staff assessment; in SPARK, it's the SAPA factor.

Table 1. How Different Approaches Use Self-Assessments

	CPR	**FCS**	**PeEv**	**SPARK**
What is assessed?	Artifact	Contribution	Contribution	Contribution
Calibration?	Yes	No	No	No
Rating, ranking, division?	Rating	Division	Rating	Rating
Peer assessment performed?	Yes	Yes	Yes	Yes
Staff assessment performed?	No	No	(Yes)	No
Collected but not used in grading	—	—	Staff assessment	SAPA
Weighting	Based on calibration	Multiplicative	Additive	Multiplicative
How does self-assessment contribute?	Bonus if close to peer grade	Weighting	Weighting	Weighting

5. Summary

Self-assessment is an important metacognitive strategy, and like other metacognitive approaches, it can produce large learning gains. Research on the validity of student ratings can be best summarized by saying that results in some settings are better than others, but it is generally true that better students are better self-assessors than weaker students, and that summative use of self-assessment is associated with over-rating by students. For that reason, self-assessments are rarely used directly as components of grades, but "accurate" self-assessments (those that closely match peer or instructor assessments) may be rewarded with higher grades.

References

[1] Boud, D., *Enhancing Learning Through Self-Assessment*, London: Kogan-Page, 1995.

[2] Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of personality and social psychology, 77*(6), 1121.

[3] Piech, C.; Huang, J; Chen, Z.; Do, C., Ng, A., Koller, D. "Tuned models of peer assessment in MOOCs," *Proc. 6th International Conf. on Educational Data Mining*, Memphis, TN, July 2013.

[4] Boud, D. and Falchikov, N., "What does research tell us about self-assessment?" Chapter 12 of Boud [1].

[5] González, J.C. and Murthy, A., "Including peer and self-assessment in a continuous assessment scheme in electrical and electronics engineering courses," *Frontiers in Education 2014*, Madrid, Spain, Oct. 2014.

[6] Liang, Y.; Wang, Q.; Lu, Y.; Qian, R.; and Yiqing, Y., "Using a web-based system to explore peer, self, and tutor assessment in problem-based learning tutorials," 2011 *International Symposium on IT in Medicine and Education* (ITME), Guangzhou, China, Dec. 2011

[7] Papinczak T, Young L, Groves M, Haynes M. An analysis of peer, self, and tutor assessment in problem-based learning tutorials. *Medical Teacher* [serial online]. June 2007;29(5):122-132. Available from: Professional Development Collection, Ipswich, MA.

[8] Kulkarni, C., Wei, K.P., Le, H., Chia, D., Papadopoulos, K., Cheng, J., Koller, D. and Klemmer, S.R., 2015. Peer and self assessment in massive online classes. In *Design thinking research* (pp. 131-168). Springer International Publishing.

[9] Wilkowski, Julia, Daniel M. Russell, and Amit Deutsch. "Self-evaluation in advanced power searching and mapping with google MOOCs." In *Proceedings of the first ACM conference on Learning@ scale conference*, pp. 109-116. ACM, 2014.

[10] Sluijsmans, Dominique; Zundert, Marjo van; Merriënboer, J.J.G. van, "Effective peer assessment processes: Research findings and future directions," *Learning and Instruction*, Vol. 20, No. 4, p.270-279. ISSN 0959-4752.

[11] Song, Y., Hu, Z., Morris, J., Kidd, J., Gehringer, E., "Toward better training in peer assessment: does calibration help?" Proc. Computer-Supported Peer Review in Education, 2016, workshop at *9th International Conference on Educational Data Mining*, Raleigh, NC, June 2016.

[12] Robinson, Ralph. "Calibrated Peer Review™: an application to increase student reading & writing skills." *The American Biology Teacher* 63.7 (2001): 474-480.

[13] Lauw, H. W., Lim, E.-P., and Wang, K. (2007). Summarizing review scores of "unequal" reviewers, in *2007 SIAM International Conference on Data Mining*, Minneapolis, April 26–28, 2007, pp. 539–544.

[14] Hamer, J., Ma, K. T., and Kwong, H. H. (2005). A method of automatic grade calibration in peer assessment. In *Proceedings of the 7th Australasian Conference on Computing Education– Vol. 42* (Newcastle, New South Wales, Australia). A. Young and D. Tolhurst, Eds. ACM International Conference Proceeding Series, vol. 106. Australian Computer Society, Darlinghurst, Australia, 67-72.

[15] Cho, K., Schunn, C.D., and Wilson, R. W. (2006). Validity and reliability of scaffolded peer assessment of writing from instructor and student perspectives, *J. Educ. Psych.* 98:4, pp. 891–901.

[16] Ramachandran, L., Gehringer, E., Yadav, R. Automated assessment of the quality of peer reviews using natural language processing techniques, *International Journal of Artificial Intelligence in Education*, January 2017.

[17] Lee, D., Pramudianto, F. and Gehringer, E. Prediction of grades for reviewing with automated peer-review and reputation metrics. Second Workshop on Computer-Supported Peer Review in Education, associated with Educational Data Mining 2016, Raleigh, NC, June 29, 2016.

[18] Carlson, Patricia A., and Frederick C. Berry. "Calibrated Peer Review and assessing learning outcomes." *Frontiers in education conference*. Vol. 2. STIPES, 2003.

[19] Cvetkovic, Dean. "Evaluation of FCS self and peer-assessment approach based on Cooperative and Engineering Design learning." Engineering in Medicine and Biology Society (EMBC), 2013 35th Annual International Conference of the IEEE. IEEE, 2013.

[20] Van Den Bogaard, Maartje ED, and Gillian N. Saunders-Smits. "Peer & self evaluations as means to improve the assessment of project based learning." 37th Frontiers In Education Conference, Milwaukee, WI, October 2007.

[21] Willey, K., and A. P. Gardner. "Investigating the capacity of self and peer assessment to engage students and increase their desire to learn." Attracting Young People to Engineering, Proceedings of the SEFI 37th Annual Conference, Rotterdam. Vol. 14. 2009.

[22] Willey, Keith, and Mark Freeman. "Completing the learning cycle: The role of formative feedback when using self and peer assessment to improve teamwork and engagement." Proceedings of the 17th Annual Conference of the Australasian Association for Engineering Education: Creativity, Challenge, Change; Partnerships in Engineering Education. Australasian Association for Engineering Education, 2006.

[23] Ohland, Matthew W., Misty L. Loughry, David J. Woehr, Lisa G. Bullard, Richard M. Felder, Cynthia J. Finelli, Richard A. Layton, Hal R. Pomeranz, and Douglas G. Schmucker. "The comprehensive assessment of team member effectiveness: Development of a behaviorally anchored rating scale for self- and peer evaluation." *Academy of Management Learning & Education* 11, no. 4 (2012): 609-630.

[24] Spiridonoff, Sophie. "iPeer Software: Online Rubric-Based Peer Evaluation." In *8th Annual WebCT User Conference*, pp. 10-14.

[25] Magluilo, Steven, Abdullah Konak, Sadan Kulturel-Konak, Ivan Esparragoza, and G. Okudan Kremer. "PEAR: Peer Evaluation & Assessment Resource." In *Proceedings of the Spring 2015 Mid-Atlantic ASEE Conference, Villanova University, PA*, pp. 1-13.

[26] Goh, G., Lai, X., & Rajapakse, D. C. (2011, May). Teammates: A cloud-based peer evaluation tool for student team projects. In *Software Engineering Education and Training (CSEE&T), 2011 24th IEEE-CS Conference on* (pp. 555-555). IEEE.

[27] Babik, Dmytro, Gehringer, E., Kidd, J., Pramudianto, F. and Tinapple, D. "Probing the landscape: toward a systematic taxonomy of online peer assessment systems in education," Second Workshop on Computer-Supported Peer Review in Education, associated with Educational Data Mining 2016, Raleigh, NC, June 29, 2016.