Peer assessments in Engineering: A pilot project

Karl Thurnhofer-Hemsi, Miguel A. Molina-Cabello, Esteban J. Palomo, Ezequiel López-Rubio, and Enrique Domínguez

[1] Department of Computer Languages and Computer Science, University of Malaga
Bulevar Louis Pasteur, 35. 29071 Málaga. Spain
[2] Instituto de Investigación Biomédica de Málaga – IBIMA, Spain
{karlkhader,miguelangel,ejpalomo,ezeqlr,enriqued}@lcc.uma.es

Abstract. The evaluation methods employed in a course are the most important point for the students, above any other learning aspect. For teachers, this task is arduous when the number of students is high. Traditional evaluation requires the teacher to grade all the assignments and exams, while peer assessments have become a valuable tool to involve students effectively in the correction of exercises. This paper applied and analyzes the evaluation through peer review in a course of Computer Sciences Engineering. A total of six assignments and a mid-term exam were evaluated by both teachers (individually) and students (cooperatively), and the differences were discussed to extract conclusions about the viability of this evaluation model.

Keywords: Peer assessment, cooperative assessment, peer review evaluation, web-based education, higher education

1 Introduction

More and more modern methodologies are replacing traditional approaches to develop professional and personal skills in students producing highly knowledgeable individuals, stressing problem-solving in real-life contexts [10].

Assessment-based methodologies are useful for giving students valuable feedback and information on their marks. However, the number of students in higher education is higher than in primary and secondary education, which makes increases the teacher's workload.

One strategy to minimize that heavy workload is to implement peer assessment (PA), where students evaluate the work of their classmates from which an aggregate grade is obtained [17, 1, 6]. Such a paradigm is not only ideal for reducing the teacher's workload, but it also allows students to learn from alternative solutions to the same problems proposed by their peers [15, 9]. Thus, both student interests and skills are positively affected by using this kind of methodology [4, 5, 7]. In fact, PA is considered by researchers as an effective pedagogical strategy involving peer evaluations with scoring methods and feedback to improve learning performance [16]. Peer-assisted evaluation can be seen as a part of peer-assisted learning [2, 3, 18].

This strategy is also known as peer review, which can be defined as a reciprocal process whereby students evaluate and make judgments about the work of their peers and construct a written feedback commentary [9]. Therefore, students both provide feedback reviews on others' work and receive feedback reviews on their own work. This represents an important alternative to teacher feedback without increasing teacher workload. However, some drawbacks are also presented in peer reviews, such as the students' ability to produce meaningful feedback, fairness, and biases in reviewing, collusion and plagiarism, etc.

Nevertheless, most of these drawbacks can be faced through well-designed peer review tasks [13]. One of the tools to facilitate peer review is the use of a set of evaluation rules or rubrics so that students have limitations when it comes to evaluating the work of their peers, as well as they are forced to produce a numerical score for each evaluated activity [15, 12]. Scoring rubrics provide a reliable and valid assessment in a structured format that makes students more confident in this task [11].

In this paper, a study of peer assessment in a course taught in four different Computer Science degrees is presented, where the marks obtained from 900 students and the teacher during 3 academic years in different assignments and mid-term exams are analyzed.

The paper is structured as follows. The academic context of this work is given in 2, whereas the evaluation methodologies are shown in 3. Section 4 provides the project results. Finally, some discussions and the conclusions are given in Sections 5 and 6.

2 Academic Context

The Spanish higher education system is regulated by the Universities Organic Law (UOL). Moreover, Spain is also a member of the World Conference on Higher Education, which is sponsored by the United Nations Educational, Scientific and Cultural Organization (UNESCO) and the Bologna process.

On the one hand, the World Conference on Higher Education has the main goal to address the most important challenges in higher education with their suggested indications. Thus, different conferences have been developed in order to prepare the work of the four commissions of the world conference: relevance, quality, management and financing, and international cooperation. Also, several considerations have been performed such as the take of actions to execute projects and support international cooperation based on solidarity and the construction of an equal society thanks to research, community projects, and specialist training.

On the other hand, the Bologna process involves many European countries in order to achieve different objectives, such as the support of student mobility, the proposition of comparable teaching methodologies and evaluations, and the motivation of institutional cooperation. This process also establishes a comprehensible and comparable education system. This way, several considerations can be yielded easily such as the search for jobs. The proposed degree system is

based on two levels: undergraduate (B.Sc.), and graduate (M.Sc.). In order to organize those degrees, the European Credit Transfer System (ECTS) is used, so that student interchange is achieved among other considerations.

Spanish universities can design their own degrees, but always under the umbrella of the current regulations. At this moment, three different levels compose the Spanish higher education system: undergraduate (B.Sc.), graduate (M.Sc.), and doctorate (Ph.D.). Therefore, the Spanish national common degree catalog from the previous system has been removed.

Currently, undergraduate degrees have between 180 and 240 ECTS credits (there are some exceptions with 300 or 360 ECTS credits), while most graduate programs have between 60 and 120 ECTS credits. The common undergraduate program has 240 ECTS credits and the usual graduate program has 60 ECTS credits, where the academic year has 60 ECTS credits for both undergraduate and graduate programs.

The organization of the courses has changed from the previous contents system to the competencies system with the Bologna process. Therefore, students are now given their degrees if they acquire the corresponding competencies, which have to be assessable.

The current educational innovation project has been applied to students from the Intelligent Systems course during 3 academic years. This course is taught during the fifth semester (third academic year) in 4 different degrees at the School of Computer Science at the University of Malaga (Spain), namely, the Computer Science, Software Engineering, Computer Engineering, and Health Engineering degrees. The results obtained from the peer assessment in this course are shown in the next section.

3 Evaluation Methodologies

There are several tools that support peer assessment in an online way. In this work Moodle has been the selected software due to our university uses it as learning management system [14]. This software facilities many resources. In particular, the workshop module, which may be considered as one of the most powerful peer assessment tool [8]. This way, the activities that students do are carried out with that workshop tool.

Intelligent Systems is the first course of the degree related to Artificial Intelligence (AI); therefore, this course deals with the fundamentals concepts of AI by introducing the diverse and wide area of AI. This is a practical course in which most of the lessons implies lab activities where students apply the theoretical concepts by managing and solving different problems. These lab sessions are complemented with some lecturing where theoretical fundamentals are presented. Lab sessions deal with simple applications of the algorithms presented in the theoretical lessons. Apart from the lab sessions (10% of the final grade), there are three assignments (each one of 30% of the final grade) about the theoretical/practical concepts: Data mining, Advanced models, and Search and Logic.

These assignments comprise the correct application and explanation of different algorithms in order to solve the addressed problems.

The procedure for a given activity or task is as follows. In the first step students do their work and submit it. Then, the teacher assigns a fixed number of submissions to each student and provide a rubric. After that, students use the rubric to evaluate their assigned submissions from other students. In order to achieve an impartial assessment, submissions are anonymous. And finally, the instructor gives a mark for each work by selecting its own evaluation and those provided by the students.

Instructor mark may be computed by using a simple grading or a more complex one. In this work, a numerical evaluation between 0 and 100 has been used as a grade. This kind of evaluation allows a clearer interpretation of the grading process.

4 Project Results

The innovative education project, as commented above, involves 4 courses with a total of more than 900 students during 3 academic years. All courses share a common program with the same evaluation process, which is composed by 6 mandatory assignments, a mid-term exam, and a final exam. The evaluation of the works are carried out by rubrics, which are specifically designed for each assignment. The mandatory assignments and the mid-term exam are peer reviewed by the students and professors. Due to the time limitation in the period exam, the last final exam is not peer reviewed by students, but we are studying other choices in order to include this activity in the peer review process by changing the dates and the evaluation process in the following years.

The deviation of students grade compared with teachers grade for each assignment is shown in Fig. 1. The assignments are evaluated over 100 points, and the mark differences are calculated by subtracting the teacher grade of students grade. In this sense, a negative mark difference means peer reviewers have been more strict than teacher evaluating the work. In contrast, a positive mark difference means the student has been over graded, where reviewers have shown a less rigorous evaluation than teachers.

According to Fig. 1, most of evaluation done by peer reviewers were over graded around 10% respect to the teacher grade. Interestingly, this over rating was greater in mid-term exams, as Fig. 2 shows. Furthermore, most of the outliers are located in the negative part (i.e., reviews are more rigorous) for the assignments. However, these outliers are located in the positive part for the mid-term exam. Regarding to the medians of marks, the behavior is similar: the median of the marks differences are close to 20 (over 100) for the mid-term exam, while the medians are close to zero in the rest of assignments. That is, except for the mid-term exam, the reviews were very fair on average comparing with the teacher grade.

These imbalances in the evaluation process are clearly observable in Fig. 3 where the histograms of marks differences are shown for both mid-term exam

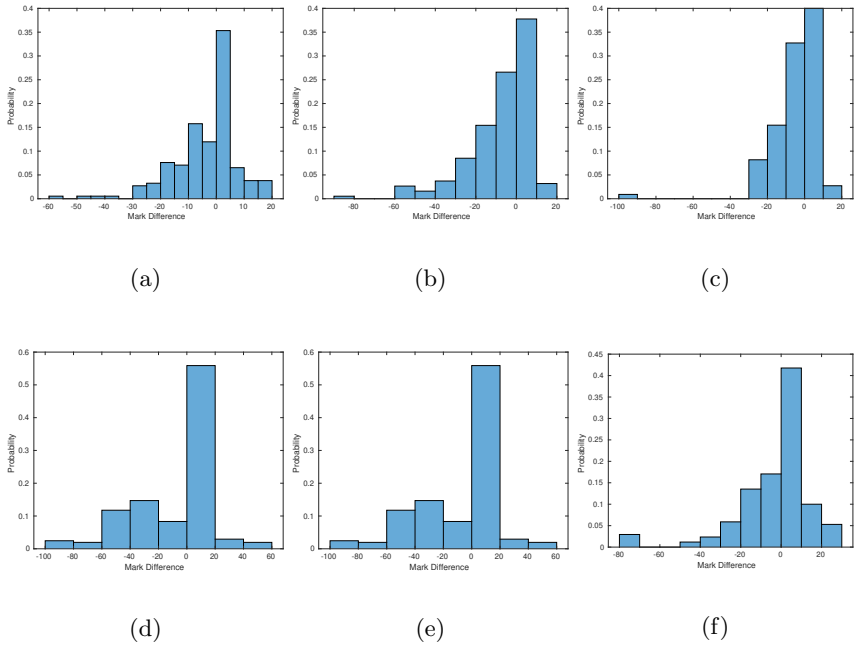


Fig. 1: Histogram of mark differences for several assignments.

and the rest of assignments. Possible causes of this variation are mainly related to the weight of the activity in the final grade of the course, since the contribution of the mid-term exam grade to the final grade of the course is greater than the rest of assignments due to the current evaluation procedure.

The comparison of Fig. 4 analyzes the relationship between the reviewer grades obtained by the students and the difference from the instructor's mark. The first appreciation is that the best reviewers have the lowest difference with respect to the teacher correction. Moreover, successively, these ranges are more extent for the students that obtained lower reviewer grades. It is remarkable that there are good reviewers evaluating more strictly, 40% lower than the teacher (the outlying points), which indicates they are very critical with peers. Nevertheless, the majority of the students are optimistic in their assessments.

Finally, we have compared the evolution of the reviews in peer assessment during the last three academic courses in Fig. 5. The median shows that the results are very similar for all years, despite the fact that the academic contexts are different for each course. In 2018/2019, there is exceptional stability of the mark differences around zero. However, in the following course, 2019/2020, lots of students tended to rate very strictly, having many negative differences. However, during the last year, in which the lessons were mainly given online, the evaluation of the reviewers was positive, as the median reflects. Therefore, the background of the students and of the teaching modality might affect a bit the performance of peer assessment tasks, but with no many deviations.

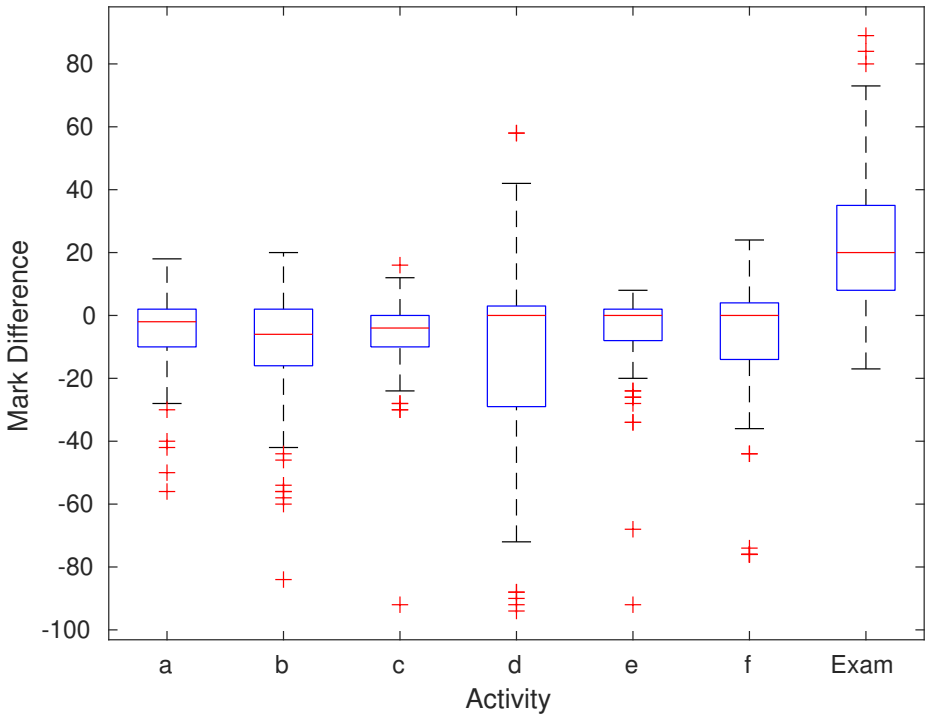
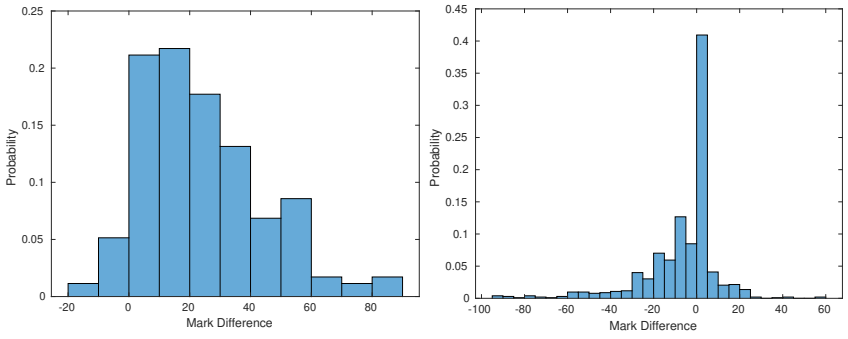


Fig. 2: Comparison of mark differences between the mid-term exam and the rest of assignments.



(a) Mid-term exam.

(b) Rest of activities.

Fig. 3: Comparing histograms between exams and the rest of assignments.

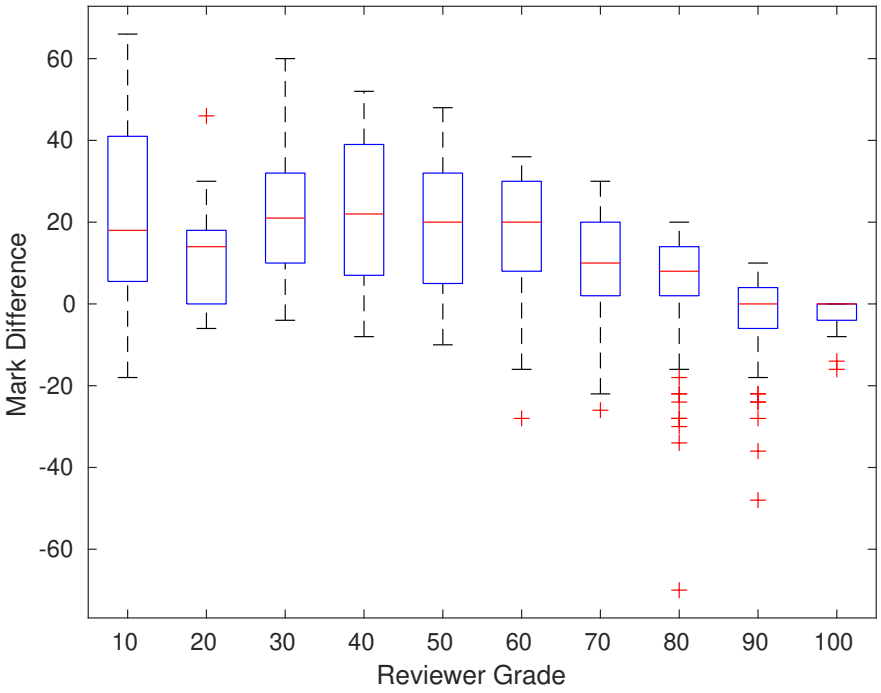


Fig. 4: Mark differences according the reviewer grades.

5 Discussion

An important point that needs to be discussed is the presence of outliers in the mark differences represented in Figs. 2, 4 and 5. As we can see in the first outcome, there are many negative outliers in the mid-course activities, while in the exam, the students tend to be optimistic. The reason might be the level of importance of the activity; that is, students consider that they can be more strict during the year with small activities since they are not determinant for the final mark. So a balance mechanism could be included by weighting the review's marks strongly in those mid-term activities.

Related to that, it is important to consider the type of students that provoke this effect. Fig. 4 reflects that the best students make more restrictive assessments. A way to avoid this problem is by providing a template of a generalized solution combined with a detailed rubric to extend the range of possible valid solutions for these students.

On the other hand, when the course 2018-2019 finished, the mark difference analysis showed that there were outliers on both sides (optimistic and pessimistic), so it was difficult to determine a prevention mechanism for the next years. Actually, the following academic courses showed disparate tendencies,

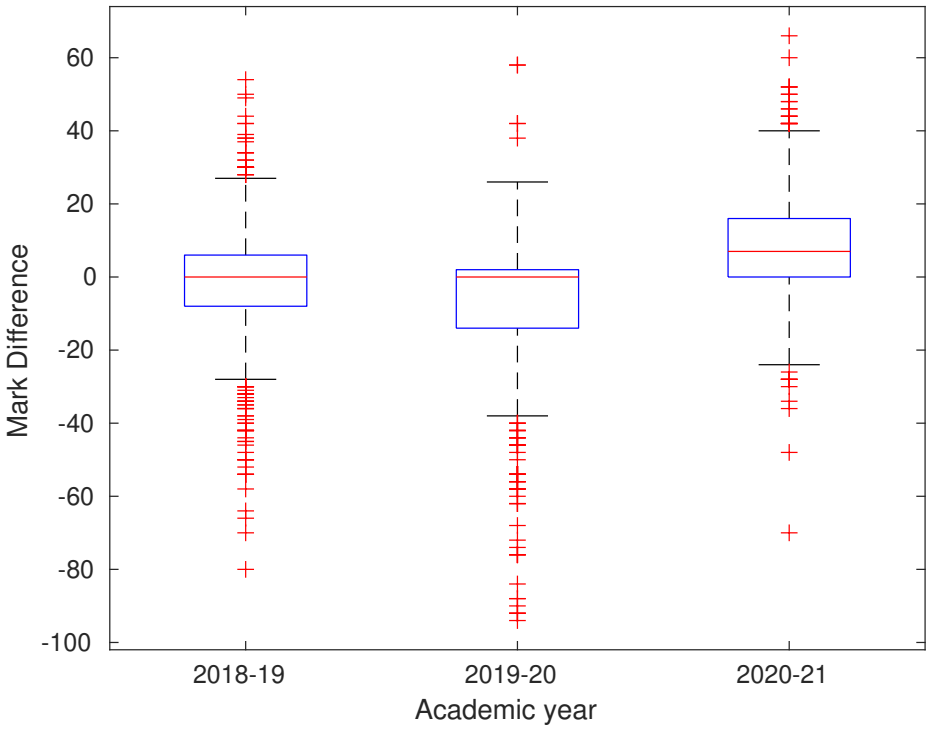


Fig. 5: Comparison of mark differences along academic years.

which implicate that the student context (type of students) strongly influences the peer assessment.

On the other hand, these peer assessment tasks were applied to a specific subject, which comprises a set of assignments. The type of exercise is also relevant in the results presented in this work. These assignments consisted of the application of the algorithms taught in the lessons, giving the results and an explanation of those. Thus, the evaluation by peer assessment can be completely objective since the results are fixed, and the reasoning is the same for all students. However, if the teaching objectives are different and the required activities are subjective, peer assessment could not be adequate. For example, if answers are long texts that need to be interpreted, one student may think differently than others, and the assessments would be quite different.

Therefore, this evaluation method may be only suitable for any objective evaluation. In addition, the workload is not that much when the evaluation criteria are clear. Only those sets of task assessments whose marks are scattered need to be analyzed by the teacher. If the evaluation marks are similar among the student, the teacher does not have to re-evaluate the task.

6 Conclusions

Cooperative evaluation allows teachers to reduce their workload and have additional points of view of the quality of the exercises done by students. The case study analyzed in this work showed a good effectiveness of the evaluation. Students typically have no more than 20 % error on instructor grades, and even they tend to be very strict with their mates. Nevertheless, the importance of an exam in the overall evaluation may affect the student's grades award, since most of them provided optimist marks.

Although students reported a positive feedback about being involved in the evaluation procedure, this task still have a dependency on the teacher, that is, it must not be used without supervision. Students with the highest reviewer grades are generally more precise in the evaluation, and the worst have a lower grade, which indicates peer assessment is coherent and suitable to be included as an evaluation method.

The tendency over the last three years is stable, but the impact of remote lessons might affect the corrections of the students, which have to be taken carefully in future studies. The application of this peer review process to other courses and degrees is being considered, so more data would be available to extract more relevant conclusions.

Acknowledgments

This work is partially supported by the University of Málaga (Spain) under the grants PIE19-161, project name Cooperative Evaluation Procedures in Active Methodologies, and PIE19-219, project name Critical review of evaluation systems. The authors thankfully acknowledge the grant of the Universidad de Málaga, I Plan Propio Integral de Docencia.

References

1. Brooks, C.M., Ammons, J.L.: Free riding in group projects and the effects of timing, frequency, and specificity of criteria in peer assessments. Journal of Education for Business 78(5), 268–272 (2003)
2. Glynn, L.G., MacFarlane, A., Kelly, M., Cantillon, P., Murphy, A.W.: Helping each other to learn – a process evaluation of peer assisted learning. BMC Medical Education 6(1), 18 (2006)
3. Hodgson, Y., Benson, R., Brack, C.: Student conceptions of peer-assisted learning. Journal of Further and Higher Education 39(4), 579–597 (2015)
4. Hodson, R.J., Chan, W.M., Chin, K.N., Nagami, M., Suthiwan, T.: Effects of online cooperative evaluation and group rewards on high school students' argumentative writing attitudes and interaction behaviors, pp. 383–412. De Gruyter Mouton, Boston (2011)
5. Kim, C.J.: The effects of peer assessment and peer feedback in writing education for premedical students. Ewha Medical Journal 40(1), 41–49 (2017)

6. McLarnon, M.J., O'Neill, T.A., Taras, V., Law, D., Donia, M.B., Steel, P.: Global Virtual Team Communication, Coordination, and Performance Across Three Peer Feedback Strategies. Canadian Journal of Behavioural Science (2019)
7. Montanero, M., Lucero, M., Fernández, M.J.: Iterative co-evaluation with a rubric of narrative texts in primary education. Journal for the Study of Education and Development 37(1), 184–220 (2014)
8. Nash, S.S.: Moodle 3.x Teaching Techniques. Packt Publishing Limited (2016)
9. Nicol, D., Thomson, A., Breslin, C.: Rethinking feedback practices in higher education: a peer review perspective. Assessment and Evaluation in Higher Education 39(1), 102–122 (2014)
10. Palomba, Catherine A.; Banta, T.W.: Assessment Essentials: Planning, Implementing, and Improving Assessment in Higher Education. Higher and Adult Education Series. Jossey-Bass (1999)
11. Panadero, E., Jonsson, A.: The use of scoring rubrics for formative assessment purposes revisited: A review (2013)
12. Panadero, E., Romero, M., Strijbos, J.W.: The impact of a rubric and friendship on peer assessment: Effects on construct validity, performance, and perceptions of fairness and comfort. Studies in Educational Evaluation 39(4), 195–203 (2013)
13. Pearce, J., Mulder, R.A., Baik, C.: Involving Students in Peer Review: Case Studies and Practical Strategies for University Teaching. Centre for the Study of Higher Education pp. 1–29 (2009)
14. Rice, W.: Moodle E-Learning Course Development. Packt Publishing Limited (2015)
15. Rico-Juan, J.R., Gallego, A.J., Calvo-Zaragoza, J.: Automatic detection of inconsistencies between numerical scores and textual feedback in peer-assessment processes with machine learning. Computers & Education 140, 103609 (2019)
16. Rico-Juan, J.R., Gallego, A.J., Valero-Mas, J.J., Calvo-Zaragoza, J.: Statistical semi-supervised system for grading multiple peer-reviewed open-ended works. Computers and Education 126, 264–282 (2018)
17. Román-Calderón, J.P., Robledo-Ardila, C., Velez-Calle, A.: Global virtual teams in education: Do peer assessments motivate student effort? Studies in Educational Evaluation 70, 101021 (2021)
18. Ross, M.T., Cameron, H.S.: Peer assisted learning: a planning and implementation framework: AMEE Guide no. 30. Medical Teacher 29(6), 527–545 (2007)