Analyzing the Validity of the Peer Assessment Process in a Project-Based Learning Scenario: Preliminary Results

Gabriel Badea
Computers and Information Technology Department
University of Craiova
Craiova, Romania
gabriel.badea@edu.ucv.ro

Elvira Popescu
Computers and Information Technology Department
University of Craiova
Craiova, Romania
elvira.popescu@edu.ucv.ro

***Abstract*—An important challenge in effectively implementing the peer assessment process is represented by the validity of the grades assigned by the students to their peers. Validity is defined as the level of agreement between the grades given by the students and the reference ones, given by the teacher. The literature offers conflicting perspectives on the validity of the peer assessment process, and very few works investigate the validity of peer grading in project-based learning (PBL) settings. Hence in this paper we address this less explored direction, by applying peer assessment in conjunction with PBL in a Human-Computer Interaction course; a dedicated platform called LearnEval is used to support the peer assessment process and 27 students participate in the study. Two main research questions are investigated: (1) How does the validity of the peer grades evolve throughout the semester, over several peer assessment sessions? (2) How does the grading mechanism provided by LearnEval compare to a baseline approach which relies on the mean of the peer grades? The preliminary findings are encouraging but the study also reveals some limitations and areas for improvement.**

Keywords*— *peer assessment; project-based learning; peer grading validity; grade computation mechanism

I. INTRODUCTION

Evaluation plays a central role in allowing learners to achieve higher levels of knowledge [5]. The continuous shift from traditional learning to blended learning modes and the emergent technologies have changed the way evaluation is performed. Novel assessment methods and techniques are necessary in order to appraise the constructivist approaches, such as active learning, that are becoming more widely applied and accepted [12].

Thus, in recent years, peer assessment (also known as peer evaluation or peer review), has started to be adopted in a wide range of learning settings, as well as course designs, such as MOOCs [14]. Peer review denotes the approach in which learners evaluate the amount, worth, value, quality or success of the work results of peers with similar status [13]. Various peer evaluation systems have been developed, offering more and more tailored experiences for the instructors in terms of assessment [1, 9].

An important challenge in implementing the peer assessment process represents the validity of the evaluations offered by the students [7]. The notion of validity refers to the level of agreement between the grades assigned by the learners and the ones granted by the teacher [4]; for the peer assessment process to be applied effectively, it is essential to attain high levels of validity [12]. Furthermore, the process can be affected by students' skepticism regarding peers' reviewing abilities and issues such as "friendship marking, fear of 'tit-for-tat' scoring, or lack of honesty" [8]. The literature offers conflicting perspectives on the validity of the peer assessment process: in some studies the students are regarded as competent assessors [11], whereas in others the evaluations performed by the students are very different from the teacher's assessments [10].

When it comes to the use of peer assessment in project-based learning (PBL) settings, evidence regarding validity is quite limited; there are few studies which analyze the correlation between the evaluations performed by the peers and the ones performed by the instructor. Therefore, in this paper we conduct an analysis of the validity of the peer assessment approach in the context of a PBL scenario. The platform employed for supporting the peer assessment process in our study is called LearnEval and was introduced in [2, 3]. Apart from providing effective support for the submission and reviewing phases, the platform allows the teacher to tailor the peer assessment workflow based on the requirements of the course; it also offers features such as: several automatic allocation mechanisms of the submissions to reviewers, configurable grade computation mechanism, as well as detailed tracking of the learners' activity by means of scores and statistics [3].

The current paper adds to the literature by offering an analysis of the validity of the peer assessment process in a PBL context, involving 27 students. Furthermore, the effectiveness of the grade computation mechanism employed by LearnEval is examined by comparing it with a reference mechanism that is commonly used in other peer assessment systems, based on simply averaging the peer grades. More specifically, the following research questions are addressed in the paper: (1) How does the validity of the peer grades evolve throughout the semester, over several peer assessment sessions? (2) How does the grading mechanism provided by LearnEval compare to a baseline approach which relies on the mean of the peer grades?

In this context, we start with several related works that report on the validity of the peer assessment process in various settings (section II), followed by an overview of the PBL context of our study and the methodology applied (section III). The results of the analysis are subsequently reported (section IV) and discussed (section V). We conclude by summarizing the findings, mentioning the limitations of the current study and outlining future improvements (section VI).

II. Related Work

The application of peer review in education has been broadly investigated over the last decades [6]. Furthermore, the validity of the peer assessment process has also been studied for several decades [4], but the literature reports on contrasting results. In the following we will offer an overview of several works that present the findings in terms of correlation between the peer evaluations and teacher assessments in order to gain a better understanding on the validity of the peer assessment process.

A method based on statistical analysis for detecting biased peer reviews in works that require open-ended answers was proposed in [11]. The instructor is required to intervene only when such biased evaluations are automatically found by the tool. The activities from two different assignments were randomly allocated to learners and each submission was reviewed on average by five students. Even from the start, in order to mitigate the risk for providing biased reviews, the learners were informed that the peer assessment activity represented 30% of the final grade. The peer evaluations were performed using rubrics, in the form of closed Likert scales. The agreement and consistency between the evaluations were analyzed using Pearson Correlation and Intraclass Correlation Coefficient. The findings show that the resulting grades after applying the tool were very similar with the ones offered by manual teacher assessment of the activities ($r = 0.98$ for the first activity, $r = 0.95$ for the second activity). In the study few outliers were detected by the tool, however, the correlation between the two data sets of assessments still increased by removing them. The tool is based on the assumption that outliers are scarcely found and the peers rarely provide reviews that are very different from the teacher assessments. Furthermore, better results were obtained by the tool when comparing it with the results offered by applying the Moodle peer assessment algorithm.

The effect of an assessment training module on peer review results is studied in [7]. The sample consisted in 78 undergraduate students from a mid-western US university following a technology application course. The peer assessment was performed using a forum created on Blackboard course management system. Prior to performing peer assessment, the students were asked to complete a training exercise involving tasks such as: review learning concepts, discuss assessment criteria, evaluate sample projects, or compare their reviews with teacher evaluations. Paired t-tests and Pearson Correlation were applied to examine whether there are significant differences before and after training between peer and teacher assessments. The findings revealed that the training resulted in a lower disparity between peer reviews and teacher assessments. Furthermore, the disparity was a predictor of the feedback quality offered by the students and of the quality of the subsequent revisions submitted by learners. Lower disparity was correlated with providing higher quality feedback and better revisions.

The validity of the peer assessment process is also examined in [12] by comparing the degree of similarity between peer and teacher assessments. The sample consisted in 48 students from a Turkish University following a Specific Teaching Methods I course during 2006-2007 academic year. The students reviewed the term projects of their peers using an evaluation form at the end of the semester and delivered it through e-mail to the instructor. The review form comprised two sections totaling 30 assessment criteria needed to be rated on a scale from 0 to 3 (ranging from very bad to very good). After normalizing the ratings on a scale from 1 to 100, Pearson Correlation Coefficient was applied to study the relationship between peer and teacher ratings. The coefficient value showed a very high and significant correlation between the two sets of ratings (0.991). Additional metrics were provided to study the similarity between the peer and teacher evaluations: mean, which was slightly lower for the peer scores (71.22 vs 74.43); mode, which was slightly higher for the peer scores (66 vs 62); standard deviation, which was slightly lower for the peer scores (14.08 vs 16.35); and range, which was slightly higher for the peer scores (57 vs 50).

The correlation between peer and teacher assessments after applying the Authentic Assessment for Sustainable Learning (AASL) model is analyzed in [6]. The sample consisted in around 280 first-year undergraduate students from University of Notre Dame, Sydney, following a Bachelor of Education degree and studying a unit in Mathematics. The students attended a pilot marking session prior to developing the real assignments. The grading was on a scale up to 100, and in the beginning of the process the learners were informed that a score which deviated with 15 points from the teacher's rating will be considered an outlier and discarded when computing the average solution score, in order to mitigate the risk of inflated grades. However, only five outliers were found (less than 2% of the evaluations). The final grade computed for each solution was based on the self-assessment, peer reviews and teacher's evaluation. The lecturer's assessment represented 40% of the final grade, thus lowering the risk of bias from self and peer assessments. On the other hand, peer assessments represented 30% of the final grade; two students collaboratively reviewed another peer's anonymous solution, each evaluation representing 15% of the final grade. The findings showed that learners, even the ones without prior experience in peer assessment, were capable to fairly evaluate the work of their colleagues. Noteworthy, 45% of the students reviewed their peers within a 5% margin variance of the teacher's evaluation and the mean final grade was only 0.01 variance apart from the instructor's grade.

The peer and teacher assessment of performance in a problem-based learning scenario is studied in [8]. The sample consisted in 125 first-year medical students from University of Queensland following a Bachelor of Medicine and Bachelor of Surgery Program and attending a course on metacognition over a period of half a year. The peer review involved the evaluation of a student presentation and the level of fulfillment of his/her responsibilities. There were two assessment sessions along the semester and each session entailed one teacher evaluation, one self-assessment and nine peer reviews. Direct comparison of the means and paired t-tests indicated that many students were offering higher grades to their peers than the teacher. The peer evaluations moderately correlated with the teacher assessments in the beginning ($r = 0.4$), but the correlation increased over time ($r = 0.6$). From frequency histograms it was found that the peer grades were not normally distributed, with many students offering maximum grades. The qualitative data revealed that students intentionally assigned maximum grades to their peers, especially the learners doubting the peer assessment process. An algorithm was applied to discard the highly

skewed evaluations, resulting in the removal of 4.6% of the grades. However, the remaining evaluations still reflected the bias of the students.

To sum up, the validity of the peer assessment process has been investigated in various contexts over the time, ranging from computer science to mathematics and medical education, from studies with dozens of students to studies with hundreds of participants. The current paper adds to the literature by exploring the validity of the peer assessment process applied in conjunction with PBL in a Human-Computer Interaction course. Furthermore, the results from this analysis could be considered a stepping-stone for further studies on the joint application of peer assessment and project-based learning.

III. Study Settings and Methodology

A. Context of Study

We analyzed the peer assessment data gathered from employing LearnEval platform in a PBL context at the University of Craiova, Romania. The study took place during 10 weeks of the second semester of 2018-2019 academic year and involved 27 undergraduate students following a Computer Science program and studying a Human-Computer Interaction (HCI) course. The theme of the project was individually chosen by each student and consisted in the requirements analysis, design, implementation and evaluation of the user interface (UI) and user experience (UX) for a web application. Throughout the semester the learners had to prepare four incremental assignments (deliverables) related to their project. The assignments were evenly distributed in time starting with the third week of the semester and occurring at every two weeks. As the peer review process was a novel activity for the learners, the first two weeks were dedicated for performing briefing sessions regarding the advantages for both reviewers and reviewees of engaging in such an activity.

The first assignment asked students to gather the requirements for the UI and perform user modeling by generating roles, stereotypes, use cases and scenarios, thus the deliverable was a requirements document. For the second assignment the students designed wireframes, mockups, as well as low/high fidelity prototypes for the UI. The students had the alternative to use a wide range of tools for developing the wireframes and mockups (e.g., Balsamiq, Moqups etc.), thus no specific instructions were asserted regarding the deliverable format. The third assignment required students to effectively implement the UI using the programming languages and technologies at their own choice. HTML5, CSS3 and native JavaScript were extensively used, however, in some cases the students applied frameworks such as Vue.js or AngularJS; the deliverable for this assignment was represented by the web application interface. An experimental study followed by the application of evaluation techniques (such as heuristic evaluation) for testing the usability and accessibility of the UI had to be performed in the last assignment. The deliverable was a document comprising the plan for the experiment as well as the results of the evaluation.

The development of the project was complemented by a peer assessment scenario supported by LearnEval. The four assignments were utilized to perform four peer assessment sessions. At the beginning of the semester the students were informed that the reviewing activity represented a part of the final course grade. Following the in-class presentation of the project deliverable, the student had a timeframe of one week to upload it into LeanEval (in the form of an URL where it can be downloaded), followed by a timeframe of one week for providing double-blind reviews for the peers. Once the submission deadline for a session was reached, the deliverable was automatically assigned to three reviewers with varied assessment skills. An evaluation consisted in assigning a grade on a scale from 1 to 10 and the provision of feedback for each of the review criteria defined by the instructor. The review criteria were different for each assignment depending on the specific requirements. A grade and a confidence factor were automatically assigned to each deliverable by the platform depending on the peer evaluations received once the review deadline was reached. Afterwards, various scores and statistics were made available for both students and instructor.

The involvement in the peer assessment activity was satisfactory during the semester. The entire cohort of 27 students enrolled in the HCI course created an account in LearnEval, however, there were two students who decided not to participate in the activity; thus 25 out of 27 students (or 93%[1]) attended at least one of the peer assessment sessions. Table I summarizes the number of project deliverables and reviews submitted for each milestone. Relatively equal levels of involvement can be noticed across the sessions, with the exception of the second assignment where students provided less deliverables. Considering the fact that all the students presented the second assignment on time (during face-to-face class sessions), the decision to not upload it might have been influenced by the fact that some students did not know clearly how to export the wireframes and mockups from the various tools they have employed in a format their peers can assess.

TABLE I. Numbers and Percentages of Deliverables and Reviews Submitted for each Assignment (A1 - A4)

Artifact	A1	A2	A3	A4	Total
Deliverables	24 (89%)	15 (56%)	23 (85%)	19 (70%)	81
Reviews	47 (65%)	30 (67%)	50 (72%)	50 (88%)	177

B. Grade Computation Mechanism

In LearnEval the reviewers are ordered descendant based on their assessment skills and split into three categories: students with high reviewing skills (HRS) – the first third, students with medium reviewing skills (MRS) – the second third, and students with low reviewing skills (LRS) – the last third. Once the review deadline for an assignment is reached, each deliverable is automatically assigned a grade based on the peer evaluations received. In addition, each evaluation is weighted based on the assessment skills of the reviewer. The platform allows the instructor to configure these weights. In the current study, the grades provided by the students associated to HRS, MRS and LRS categories represented 50%, 33.33%, and 16.67% respectively, of the final grade assigned to a deliverable.

One of the goals of our study is to investigate whether this grade computation mechanism implemented in

[1] Percentages are rounded to the nearest integer throughout the paper

LearnEval provides more valid grades compared to a baseline approach which is commonly used in the literature and is based on the simple mean of the individual peer grades. This is the second research question of our study and it will be addressed in section IV.C.

C. Review Assignment Procedure

The submitted project deliverables are allocated to three reviewers with various assessment capabilities (one HRS, one MRS, and one LRS) once the submission deadline for an assignment is reached. When distributing the deliverables to reviewers, one of the goals is to assign for each evaluator a similar number of submissions to assess throughout the semester; hence the review assignment procedure dynamically picks from a review category the student with the lowest cumulated number of deliverables assigned to review in the previous sessions.

IV. Results

A. Central Tendency Measures

The data gathered in the study was analyzed to compare the peer evaluations with the teacher assessments.

Descriptive statistics regarding central tendency measures and central distribution of the grades assigned for the four peer assessment sessions are available in Table II. The mean of the grades was lower for the peer reviews in most of the sessions. The same results were achieved when examining the median. On the other hand, the standard deviation was higher for peer grades in the first two assignments, thus students offered more scattered grades than the instructor. Furthermore, the range was larger for peer grades in three of the sessions, emphasizing that students did not assign only a narrow interval of grades to their peers. The negative skewness points out that more grades were concentrated to the right side of the axis, thus both the students and the instructor offered high grades; this is especially true for the peer grades in the third assignment. These findings are similar with the results reported in other studies such as [8]. At a higher granularity level, the *Overall* column displays the figures obtained when considering aggregated data for all the four sessions. The values in this column disclose there were no major differences between the two datasets of grades. The mean of the peer grades was slightly lower than the teacher grades thus the "friendship marking" effect was not seen. Furthermore, the standard deviation and range were slightly larger for the peer assessments signaling that students provided a wider range of grades compared to the instructor.

B. Validity of Peer Grades – Evolution Throughout the Semester

The validity of peer grades was assessed by analyzing the correlation between the grades computed by LearnEval and the teacher grades for each of the four peer assessment sessions as students incrementally developed their projects.

We decided to use Pearson Correlation as it has been successfully applied in several peer assessment data analysis studies [7, 8]. The results are included in Table III. A moderate positive correlation was obtained in the first session. An even stronger positive correlation was achieved in the last session. By contrast, weaker levels of positive correlation were obtained in the third, and especially in the second session, which were not statistically significant. The *Overall* column displays the figures obtained when considering aggregated peer grades for all the four sessions. We can notice a weak to moderate positive correlation, which is lower than similar figures reported in the literature (i.e., r = 0.69 in [4]).

TABLE III. Correlation Between Peer Grades (Using LearnEval Grade Assignment Mechanism) And Teacher Grades for each Assignment (A1-A4)

	A1	**A2**	**A3**	**A4**	**Overall**
Pearson Correlation Coefficient	0.54	0.21	0.29	0.84	0.39
p value	< 0.01	0.45	0.18	< 0.01	< 0.01

C. Validity of the Grade Assignment Mechanism Provided by LearnEval

The validity of the grade assignment mechanism provided by LearnEval was assessed by comparing it to a baseline approach that is commonly used in the literature and is based on the simple mean of the peer grades [6].

Again we applied Pearson Correlation, this time between the baseline grades (i.e., average of the peer grades, without using LearnEval formula, so without weighting these grades according to the reviewing skills of the evaluators) and the teacher grades; the results are included in Table IV. The *Overall* column displays the figures obtained when considering aggregated peer grades for all the four sessions. Lower correlation values were obtained for all peer assessment sessions compared to the figures in Table III. Therefore, we can conclude that the LearnEval grade computing mechanism provides better results compared with the baseline approach.

TABLE IV. Correlation Between Peer Grades (Using Baseline Mechanism Based on Mean Value) And Teacher Grades for each Assignment (A1-A4)

	A1	**A2**	**A3**	**A4**	**Overall**
Pearson Correlation Coefficient	0.52	0.15	0.13	0.78	0.32
p value	0.01	0.59	0.54	< 0.01	< 0.01

TABLE II. Statistics on Central Tendency Measures and Central Distribution of the Grades Assigned to the Submitted Project Deliverables

Statistic	**Assignment I**		**Assignment II**		**Assignment III**		**Assignment IV**		**Overall**	
	Peer	Teacher	Peer	Teacher	Peer	Teacher	Peer	Teacher	Peer	Teacher
Mean	7.66	8.46	7.98	9.2	9.07	8.39	8.71	8.79	8.36	8.71
Median	8.86	9	8.11	10	9.25	9	9	9	8.9	9
Std. dev.	2.11	1.32	1.25	0.98	1	1.86	1.27	1.32	1.62	1.48
Range	6.76	5	5	3	3.56	6	4.4	4	6.76	6
Skewness	-0.75	-0.78	-0.96	-0.83	-1.39	-0.86	-0.89	-0.84	-1.28	-0.97

V. Discussion

A. Validity of Peer Grades – Evolution Throughout the Semester

As shown in section IV.B, the validity of the grades assigned to the project deliverables did not follow an ascending trend throughout the semester, as the highest levels of correlation were attained in the first and fourth session, with lower levels of correlation achieved in the second and the third session. Nevertheless, the correlation follows an ascending trend starting with the second assignment. Despite this, in the first session the correlation was higher than in the subsequent two sessions. Two potential explanations are behind this inconsistency: firstly, in the second assignment the number of project deliverables was relatively low and there is a likelihood for the sample to be too small to offer significant results; secondly, several deliverables were not correctly uploaded into the LearnEval platform. The second assignment required students to design and implement wireframes and mockups using various tools at their own preference. This caused an issue, as several students were not able to export the wireframes and mockups in a format that was known to their peers. Therefore, the students either decided to not upload some of the files or they uploaded them in a format their colleagues could not open. Given this circumstance, several assessments were not valid and the outcome was a high discrepancy between the peer grades and the teacher grades in some cases. To mitigate this effect, several measures could be taken in the future: integration of an outlier detection mechanism and adding the option for a reviewer to flag a submission as not valid for evaluation.

As far as the third session is concerned, the students faced a different issue. The third assignment required students to provide the actual implementation of the UI in a programming language of their own choice. In several cases, the students decided to implement the UI in a programming language that required complex set up, such as the configuration of a server. Therefore, some students were not able to correctly configure their local environment to run the peers' project deliverables and subsequently review them; thus they decided to assign high grades to their peers, even if the actual quality of the deliverables was not known. In the fourth assignment this issue was eliminated as the deliverable consisted in a plain Word document, hence the students could easily access and evaluate their peers' work.

B. Validity of the Grade Assignment Mechanism Provided by LearnEval

As shown in section IV.C, the grade assignment mechanism provided by LearnEval offers higher validity levels compared to the baseline approach. In the first assignment session this improvement was not very clear, as the system did not have any knowledge regarding the students' assessment skills and the distribution of the peers to one of the three reviewing categories was done in a random manner. However, given the high number of submissions in the first session and the increased knowledge of the system before the second assignment, a clear separation of the students in the reviewing categories was available and a higher correlation could be noticed compared with the previous session. The difference was especially visible in the third session, with a correlation coefficient of 0.29 compared to 0.13. An important factor influencing this increase was again the higher number of submissions cumulated from the first two sessions combined with more knowledge of the system regarding the students' assessment capabilities and subsequently a more clear separation in the three categories. However, the difference was less marked in the fourth session, when the correlation value was relatively high even when applying the baseline approach.

VI. Conclusions

The purpose of the current study was to analyze the validity of the peer grades throughout the semester in the context of a PBL scenario and to examine the validity of the grade assignment mechanism provided by LearnEval platform.

The findings show that considering the reviewing skills of the students offers more valid grades, and that relatively high levels of correlation can be achieved between the peer and teacher grades provided that project deliverables are uploaded in a format which is easily accessible by the evaluators. In this respect, our study had an important limitation: in sessions 2 and 3 several project deliverables were not uploaded or were uploaded in a format that reviewers could not access, which led to an inconsistent grading. A potential solution is to allow reviewers to flag a submission as invalid or incorrectly uploaded; in addition, we aim to integrate a mechanism for automatically detecting outliers and rogue reviews.

A further limitation of our study is the relatively small sample size, consisting in only 27 students. This did not allow for more in-depth analyses, at finer granularity levels, such as investigating the validity of the grades for each of the three categories of reviewing skills. Hence this study provides only some preliminary findings; additional experiments are currently under development in order to extend the analysis of the peer assessment process.

Acknowledgment

This work was supported by the grant POCU380/6/13/123990, co-financed by the European Social Fund within the Sectorial Operational Program Human Capital 2014 – 2020.

References

[1] L. D. Alfaro and M. Shavlovsky, "CrowdGrader: Crowdsourcing the Evaluation of Homework Assignments", Proceedings of the ACM Technical Symposium on Computer Science Education (SIGCSE 2014), pp. 415–420, 2014.

[2] G. Badea and E. Popescu, "A Web-Based Platform for Peer Assessment in Technology Enhanced Learning: Student Module Prototype", Proceedings of the 19th International Conference on Advanced Learning Technologies (ICALT 2019), IEEE Computer Society Press, pp. 372-374, 2019.

[3] G. Badea and E. Popescu, "Instructor Support Module in a Web-Based Peer Assessment Platform", Proceedings ICSTCC 2019, IEEE, pp. 691-696, 2019.

[4] N. Falchikov and J. Goldfinch, "Student Peer Assessment in Higher Education: A Meta-Analysis Comparing Peer and Teacher Marks", Review of Educational Research, vol. 70(3), pp. 287-322, 2000.

[5] G. Joughin, "The Hidden Curriculum Revisited: A Critical Review of Research into the Influence of Summative Assessment on Learning",

Assessment & Evaluation in Higher Education, vol. 35(3), pp. 335-345, 2010.

[6] S. Kearney, T. Perkins, and S. Kennedy-Clark, "Using Self- and Peer-Assessments for Summative Purposes: Analysing the Relative Validity of the AASL (Authentic Assessment for Sustainable Learning) Model", Assessment & Evaluation in Higher Education, vol. 41(6), pp. 840-853, 2016.

[7] X. Liu and L. Li, "Assessment Training Effects on Student Assessment Skills and Task Performance in a Technology-Facilitated Peer Assessment", Assessment & Evaluation in Higher Education, vol. 39(3), pp. 275-292, 2014.

[8] T. Papinczak, L. Young, M. Groves, and M. Haynes, "An Analysis of Peer, Self, and Tutor Assessment in Problem-based Learning Tutorials", Medical Teacher, vol. 29(5), pp. e122–e132, 2007.

[9] J. G. Politz, D. Patterson, S. Krishnamurthi, and K. Fisler, "CaptainTeach: Multi-Stage, In-Flow Peer Review for Programming Assignments", Proceedings of the 2014 Conference on Innovation & Technology in Computer Science Education (ITICSE'14), pp. 267-272, 2014.

[10] H. I. Reiter, K. W. Eva, R. M. Hatala, and G. R. Norman, "Self and Peer Assessment in Tutorials: Application of a Relative-Ranking Model", Academic Medicine, vol. 77(11), pp. 1134-1139, 2002.

[11] J. R. Rico-Juan, A. J. Gallego, J. J. Valero-Mas, and J. Calvo-Zaragoza, "Statistical Semi-Supervised System for Grading Multiple Peer-Reviewed Open-Ended Works", Computers & Education, vol. 126, pp. 264-282, 2018.

[12] S. Şahin, "An Application of Peer Assessment in Higher Education", The Turkish Online Journal of Educational Technology, vol. 7(2), pp. 5-10, 2008.

[13] K. J. Topping, E. F. Smith, I. Swanson, and A. Elliott, "Formative Peer Assessment of Academic Writing Between Postgraduate Students", Assessment & Evaluation in Higher Education, vol. 25(2), pp. 149–169, 2000.

[14] A. Vista, E. Care, and P. Griffin, "A New Approach towards Marking Large-Scale Complex Assessments: Developing a Distributed Marking System that Uses an Automatically Scaffolding and Rubric-Targeted Interface for Guided Peer-Review", Assessing Writing, vol. 24, pp. 1–15, 2015.