TÉCNICO LISBOA

GradeMe – Studying peer-grading in in-person learning

Madalena Mendes da Fonseca de Cardoso Assembleia

Thesis to obtain the Master of Science Degree in

Information Systems and Computer Engineering

Supervisors: Prof. Daniel Jorge Viegas Gonçalves
Prof. Hugo Miguel Aleixo Albuquerque Nicolau

Examination Committee

Chairperson: Prof. Alexandre Paulo Lourenço Francisco
Supervisor: Prof. Daniel Jorge Viegas Gonçalves
Member of the Committee: Prof. Sandra Pereira Gama

October 2019

Acknowledgments

Firstly, I would like to thank my parents and grandparents for their caring, motivation and unconditional support throughout all these years. Without them always believing in me, this project would have not been possible.

I would also like to express my highest gratitude and appreciation to my dissertation Professor Daniel Gonçalves for all the insight and support given as well as for providing a platform to develop and test my study. To professor Hugo Nicolau, thank you for all insight and sharing of knowledge that helped made this thesis a reality.

To my friends and colleagues from Técnico Lisboa, thank you for all these years of support and patience throughout the good and bad times. There was nothing random about your acts of kindness. Thank you for all you have done.

This work was partly supported by national funds through Fundação para Ciência e a Tecnologia (FCT) grants GameCourse PTDC/CCI-CIF/30754/2017 and UID/CEC/50021/2019.

Abstract

The use of peer grading in educational environments has been studied for several years and the results showed it is possible to obtain similar grades between students and professors, making the assessment reliable and also giving the possibility to increase the scale of courses, as this assessment method alleviates some of the grading burden from professors, allowing for more courses that use blended learning with gamification capacities.

Current research, however, has not investigated if social relationships between students affect grading. Our work focuses on whether social relationships between higher education students and their sociometric status affects the accuracy and quality of peer grading in a gamified blended learning environment. We used an improved version of an existing plug-in, meant to bring peer grading capacities to the Moodle platform. This new version includes new features that allow the collection of information regarding the student's relationships with their peers.

This plug-in was then used in the Multimedia Content Production course and, at the end of the semester, the data was retrieved and analyzed. We did not find any indication of significant bias in the peer grading assignments from students which had shown preference or displeasure towards a given colleague. Students from different sociometric status groups showed similar matching grades with the professors overall, though certain groups showed more inclination to provide deflated grades and other groups to inflate. In short, we conclude that, in this context, peergrading can be used as a legitimate, unbiased form of grading.

Keywords

Peer grading, Blended Learning, Virtual Learning Environments, Gamification, Moodle.

Resumo

O uso da classificação por pares em ambientes educacionais tem sido estudado há vários anos e os resultados mostraram que é possível obter notas semelhantes entre alunos e professores, tornando a avaliação confiável e também aumentando a escala de determinados cursos, pois este método de avaliação alivia a carga de trabalho dos professores, permitindo mais cursos que usam aprendizagem mista combinado com capacidades de gamificação.

A pesquisa atual, no entanto, não investigou se as relações sociais entre os alunos afetam as classificações dadas. Este trabalho foca-se nas relações sociais entre os estudantes do ensino superior, no seu estatuto sociométrico e como este pode afetar a precisão e a qualidade da classificação por pares num ambiente de aprendizagem mista gamificado. Usou-se uma versão aprimorada de um plug-in existente, destinado a trazer capacidades de classificação por alunos para a plataforma Moodle. Esta nova versão inclui novos recursos que permitem a recolha de informações sobre os relacionamentos de alunos com os seus colegas.

No final do semestre, os dados foram recuperados e analisados. Não encontramos nenhuma indicação de bias significativo nas tarefas de avaliação por pares de alunos que demonstraram uma preferência ou descontentamento em relação a um determinado colega. Alunos de diferentes grupos de estatuto sociométrico apresentaram notas semelhantes com os professores em geral, embora certos grupos tenham demonstrado mais inclinação para fornecer notas deflacionadas e outros grupos para inflacionar. Em resumo, concluímos que, nesse contexto, a classificação por pares pode ser usada como uma forma legítima e imparcial de classificação.

Palavras Chave

Avaliação de Pares, Aprendizagem Mista, Ambientes de Aprendizagem Virtuais, Gamificação, Moodle.

Contents

List of Figures

List of Tables

Acronyms

AJAX	Asynchronous Javascript and XML
CDC	Coie,Dodge & Coppotelli
ICC	Intarclass Correlation Coeficient
MCP	Multimedia Content Production
MOOC	Massive Online Open Course
Moodle	Modular Object Oriented Development Learning Environment
MSc	Master of Science
UI	User Interface
VLE	Virtual Learning Environment
XP	Experience Points

1

Introduction

Contents

Blended Learning has shown to be a successful procedure in higher education courses, displaying improvements in students' learning and an overall stronger sense of community (Bonk & Graham, 2012). Blended Learning consists of the combination of the strengths of classroom face-to-face instruction with online learning experiences and has the potential to change how higher education is taught and how students interact with it (Garrison & Vaughan, 2008). It maintains the traditional class lectures and provides a way for students to interact directly between themselves and their professors, adding Virtual Learning Environments (VLEs) as a classroom expansion to provide to a large number of students an educational experience that can still lead to engaging relations between all members involved (Dillenbourg, Schneider & Synteta, 2002). Furthermore, professors have been including gamification features in blended learning environments. The adoption of game features such as badges and leaderboards has grown substantially in the last few years, resulting in increasing students' engagement and motivation to participate in class activities throughout the assessment period (Johnson et al., 2014).

However, one of the main challenges in this approach is student evaluation. The large number of students enrolled in this type of classes makes it unfeasible and possibly costly for professors to grade all assignments delivered by students in an acceptable amount of time. Thus, it is essential to come up with an alternative solution to traditional student evaluation, so that the scaling of these practices is not an issue.

In the case of activities where the assignments are in the form of multiple-choice or direct answer, is it possible to use a web platform which grades the assignments automatically; when dealing with more open-ended exercises, manual assessment is necessary, requiring time to provide a fair grade and concise feedback for students to improve their future works (Topping, 2009). In most cases, this kind of exercise is the most important for student learning and outcome of the course. Peer grading – evaluation of student work done by other student(s) – is a promising solution to help the problem of evaluation in large scales as automatic grading of this kind of assessment is difficult. While a more computerized solution is not found, peer grading or peer assessment could be the solution.

Previous work has shown that peer grading can be a valid and reliable assessment method, from primary schools to higher education (Topping, 2009; Tseng and Tsai, 2007; Taylor, Ryan & Pearce, 2015). It offers the chance to scale the grading of more complex assignments, especially in courses with hundreds of students such as Massive Online Open Courses (MOOCs). However, in MOOCs students usually do not know who their peers are, whereas in blended learning students already know each other and have created social relationships. This leads to the possibility of bias from certain students towards others, an issue that has been raised before (Falchikov, 2001), though there are no studies about how social relationships between students influence peer grading.

1.1 Research Problem

New learning environments such as Blended Learning, can include online learning besides the traditional face-to-face learning and have been used more frequently in recent years as a means to engage and motivate students to participate regularly. This type of learning usually resorts to Virtual Learning Environments, such as Moodle, to provide a platform that lets students not only publish their submissions to be assessed by the professor after, but also allows for easy and instant communications between all participants of the course, especially in courses with a large number of students, known as MOOCs.

Gamification is also another method used to try to make students engaged throughout all semester, as its features such as Experience Points, Levels, Achievements, etc., lead to an assessment that is continuous and enjoyable to watch elapse. The more participation and work submissions increase due to these systems, the harder is to scale all work with the current traditional assessment procedures and the number of professors are very small in contrast.

To try to solve this problem, the integration of peer assessment seems to be the best solution. Peer Grading can lead to the decrease of professors' workload, sharing of information between students that is mutually beneficial and furthers involves students in the course as it is a continuous procedure. Still, peer grading may require that students may have to grade peers which they know and have an established social relationship with, that can influence, either positively or negatively the result of their peer assessment and induce bias in the final grade of students' work. Therefore, the aim of this thesis is to investigate if peer grading can be a valid and unbiased assessment method in a gamified blended learning environment.

To answer this question, we will be using a previously developed peer grading plugin to be used in a higher education course that is taught using blended learning through the Moodle platform. While the plugin already provides the ability for students to peer grade their colleagues, some features will be included to measure social relationships between students.

1.2 Contributions

State of the art. To develop this study, we started by gathering and reviewing several research documents related to the topic in question. The type of environment in which this study will occur, a gamified blended learning, as well as peer grading itself, will be described and the current state of art will be discussed. The current offerings and problems of peer grading in various learning environments are also mentioned.

Development of Peer Grading Plugin. The previous version of the plugin developed to allow peer grading in a gamified learning environment had some critical issues, which were fixed. To tackle one of the biggest challenges of peer grading, the possibility of inaccurate grading due to lack of grading skills,

a training system was implemented where students were informed of the requisites each submission had, plus some examples of previous submissions and their respective grading. A peer nomination and peer ranking system were also implemented, together with some enhancements and new features to make the platform more usable and polished.

Study and analysis of results. To understand to which extent sociometric status affects peer grading in courses that have gamification and blended learning as educational approaches implemented, the peer grading plugin was instantiated it in a higher gamified blended learning course. The user activity was retrieved and analyzed in order to assess the applicability of the developed peer grading solution. By compiling and analyzing the peer grades given together with the results of the peer nominations and peer ranking, it was possible to verify that professor and students grades were often in agreement and no significant bias was present - good indicators that peer grading can be used as an assessment procedure in gamified blended learning environments and that the degree of relationships between students did not affect students capacity to provide accurate grades.

1.3 Thesis Outline

This thesis is organized as follows: **Chapter 2** contains an explanation of this thesis main concepts, as well as an extended analysis and discussion of similar previous work to provide better context of the situation. **Chapter 3** consists in the investigation of available sociometric analysis cases and methods and a description of the system that will be implemented in the peer grading plugin, and how it will provide a method to classify students into different sociometric status groups. **Chapter 4** includes a detailed description of all corrections made to the initial version of the plugin and new features that were implement, based on the conclusions of previous research, providing students with a way to give information regrading social preferences. In **Chapter 5**, the data gathered from the plugin will be studied in order to answer this research problem. Finally, **Chapter 6** provides final conclusions and some suggestions for further research and implementation.

2

Related Work

Contents

In this chapter we present the results from searching literature around peer grading and its actual state. First, the concept of peer grading is presented in more detail by giving light to several works which used various types of educational approaches to apply peer grading as an assessment method. Blended Learning and how it has been applied through virtual learning environments, such as Moodle will be then explained. Some of its features are also presented. Learning environments that use this platform can also include gamification elements, which can work as a mechanism to enhance the learning experience and create a community environment where students feel more motivated to engage and can learn from each other.

2.1 Peer Grading

Peer grading or peer assessment is defined as "an arrangement for students to considerate and specify the level, value or quantity of a product or performance of other equivalent students" (Topping, 2009). Peer grading has been implemented in several subjects such as Science, Mathematics, English, Computer Science, Finances and Biomedical Science and it can be applied in all types of work, from review of literature articles and independent studies to design project revisions for another students' work, oral presentations and grading of mock tests/exams. It allows students to learn through its peers' work and their problems, as well as provide feedback in bigger quantities and immediacy than feedback from professors, which could be later used to generate future problems or serve as exemplification (Gehringer, 2001).

In order to use peer grading as a reliable evaluation method, it would be ideal if the grade awarded by the student was the same as the professor grade, which can be easy to achieve when the questions are in the form of multiple choice or short direct answers.

2.1.1 Peer Grading in traditional learning environments

A study conducted by Scott Freeman and John W. Parks (2010), meant to analyze peer grading efficiency by comparing the scores given by students and their professor in an Introductory Biology class.

To obtain answers, a large-scale randomized trial with approximately 340 students was used, without anyone that was participating knowing. Students were asked to evaluate its peers' practice exams and grade them, which was later compared with the grade given by the professor.

Most of the question were written, using short answers. In some cases, there was the need to complement the answer with schemes and/or graphics. The questions were previously ranked according to its type, using Bloom's taxonomy (Crowe et al., 2008). The grades were submitted giving each question a ranking between 0 (no credit), 1 (partial credit), or 2 (full credit) points. Students had also

access to a sample of a correct answer written by the instructor along with the criteria for no, partial or full credit.

It was found that peer grading was good enough to use as the overall quality of the grading reported showed to be typical and with a high correlation value of 0.61. However, it couldn't be concluded whether the accuracy of student grading would improve if given more training. Another observation mentioned that higher-order cognitive questions would increase the differencing in grades between student and professor.

Topping (2009) describes a case of peer assessment in action in a secondary school, where the professor wanted to use an engaging solution so that her English students would participate and learn during the writing analysis. The students were divided in groups of three, and each student had to review the other two peers' assignment. Before any assessment was done, she worked with her students to establish the criteria needed, which ended up looking like the one she would have used herself and provided pieces from last year and explain how she evaluated them as well as offered some other pieces as training.

During the assessment each student had access to the rubric and a short list of tips. Later the professor compared the two results of each paper and if the results showed some discrepancy, she would look at the paper herself and talk with the students involved in the assessment. Overall, the results appear to tend to average with little sign of bias.

The most prevailing goal of peer assessment is providing feedback to the learners. This feedback can be confirmatory, suggestive or educative. If it is done thoughtfully and in a constructive way, it can help reduce errors and enhance the learning experience. By doing this, both sides gain from doing this type of activity in its time, in practicing, achieve a sense of accountability and can identify and analyses better their errors or misconceptions, which can lead the professor to identify knowledge gaps and work on how to closure them.

Feedback from students is available in bigger quantities and immediacy than feedback from professors, which could compensate for the lesser experience and knowledge of assessment from students.

Peer assessment has also showed to be more reliable when done with training, exemplification, professor assistance and monitoring and leads to an overall increase in reliability and validity of the grading. The lack of quality can be complemented with the availability of peer feedback in higher quantities and immediacy. It has also been shown that a student with lesser knowledge than its professor, but more time to evaluate can produce a similar evaluation (Topping, 2009).

The use of peer grading also makes the assessment process more understandable for students, since instead of relying solely on professor judgement or objective test scores, becomes the reaching of consensus between individuals (Taylor, Ryan & Pearce, 2015). In higher education environments, it can serve to help students in adjusting to university, as it provides a way for students from all kinds of

cultures and educational backgrounds to communicate and have some type of assistance.

Peer assessment was introduced in the second year of Financial Accounting in an Australian university to solve students' biggest complaints on the course: exams and other practices not being aligned with their professional needs. The insertion of peer grading occurred in four phases, each one, enhancing the assessment experience and trying to solve the problems found in each previous iteration.

Phase One occurred during the first semester in AYB200 - a Financial Accounting class - with traditional peer assessment. The main goal was to emphasize the social constructivism view of the assessment by combining explicit and implicit criteria.

Students had a task assigned during a three-week, continuous cycle. The first was used as a tutorial, with both sides defining the solution and criteria to avoid discrepancies. The following weeks students had to reviews their colleagues' tasks.

Three sets of anonymous and voluntary surveys were distributed, two pre-exam and one after the exam. In the final survey, answered by 31 students, 85% considers the feedback received valuable, 80% agrees with the marking and the valiance of the solution, 75% believes the tree-week cycle was important and 70% says it incentivized them to work weekly. Finally, 65% responded that allowed to gain knowledge and that the work allocated to peer assessment was reflected in their final grade. It is also worth mentioning that the approval rate went up to 91%, from the average of 76% in the previous four years.

However, it required commitment from full time staff to implement, collect and record all the marked tasks, resolve issues found and finally return all tasks to their owners, which leads to a considerable increase in workload stripping peer grading from one of its main advantages, efficiency in the attribution of problems, at least in short to medium term, a problem also mentioned in other literature (Falchikov, 2001).

For the 2nd phase, the peer assessment was done with the Web-based PRAZE framework, which was explicitly prepared to deal with the peer activities. This time, the students were from an International Accounting class, during the 2nd semester of 2011. Students had to draft an individual project. The framework included the rubric and a step-by-step pro-forma of review question.

92 students participated in the peer grading and had to answer a survey in the end in which 87 say that it lead to them starting the assignment earlier, 72 said peer assessing made them understand better what was expected of them, 74 mentioned the quality of comments being useful and 85 concluded that the marks awarded were fair and the exercise should be retained and allow for multiple peer reviews.

The authors verified that the staff load was lesser and the use of PRAZE allowed them not only to supervise and monitor, but also maintain a record of all activity on the site. However, a survey was also offered to the students who did not participate that alleged that they felt under skilled for the task.

2.1.2 Peer Grading in fully online learning environments (MOOCs)

Though most studies look for examples and cases of quantitative peer review, qualitative peer review has also been investigated (Meek, Blakemore & Marks, 2017) and offers its own advantages such as the creation of question with quality without waiving the attribution of a grade.

In a study accomplished with the help of the FutureLearn platform, 314 peer reviews of Biomedical Science were analyzed by 79 students (of various types of ages and previous education), who undertook the task. The assignment consisted in a written paper of approximately 300 words rated between 0 - 3. Each student could review up to seven papers.

The two staff reviewers awarded a grade from A – E and had a different criterion. Each work grade was discussed by the two to decide a final grade.

The results showed an overall high quality of the peer reviews written. Students who had done better in previous assignments were more likely to review multiple submissions and gave higher quality peer reviews to those who did not, most likely due to their higher knowledge of the subject and therefore were able to make a more concise and critical assessment.

When asked to discuss the experience, the comments were mixed, with most comments that identified specific positive aspects of having work reviewed by colleagues were, for instance, being useful and encouraging and confirmed the author's understanding of the topic.

From the reviewer part, the most positive aspects were how it added an additional and valued challenge, could get an insight on other's perspectives, how other's dealt with specific aspects and their good writing and could confirm their knowledge on the topic.

Students who did not enjoy the experience, half received a bad grade and considered that their peers were not qualified /experienced to assess them. One fact worth mentioning is that students with higher education levels were more prone to participate in peer grading. The authors believe that this type of assessment could be more suited to more capable/experienced students and that others may need more guidance and training to enjoy the experience, which goes according to the findings in other literature pieces (Topping, 2009).

Still, most studies showed a high correlation between student and professor grades, especially when the professor defines along with the students a rating criterion, with defined levels of expected answers and their corresponding value. One question that is still left to answer is of is whether more training would help achieve more accurate ratings from students.

Peer assessment provides a way for students to plan their own learning, identify what are their strengths and weakness, understand in which parts need to be revised, develop metacognitive and professional transferable skills (Gehringer, 2001) and enhance their reflective thinking and solving abilities during the learning experience and finally increase students' interpersonal relationships in the classroom (Tsai and Tsung, 2007).

To summarize, the use of peer grading, either in a more traditional ambient or using a Web platform, brings many advantages, both to students and professors, for instance:

- Allows for a more efficient attribution of problems, as it a very time-consuming activity for just one person;
- Forces students to write correctly for their peers, as their grade is dependent on it;
- Can be used to generate future problems and/or resources for future students;
- Professors can come up and offer more problems;
- It is a more scalable option for large courses;
- The use of a defined criteria means that students already know which grade they will be receiving;
- Encourages students to start their assessments early;
- Helps students gain more knowledge and confirm their expertise in the matter.

2.1.3 Peer Grading in Blended Learning

Peer grading still can have some drawbacks. Students are used to the idea of having a professional be responsible for their evaluation and can dislike or even reject the idea, as they see evaluation as a responsibility of the professor and not their own (Topping, 2009). The task of having to evaluate another student, which also happens to be a friend, can lead to lesser exact grade, as it provokes a conflict of loyalty. Factors such as reward difference between professor and student (Gehringer, 1999), the idea from some students that they lack the necessary skills, problems related to validity and reliability, bias and fairness can lead to inferior seriousness and motivation from the student (Taylor, Ryan & Pearce, 2015). Finally, it is worth mentioning that some for these studies found a certain bias in the grades of lower level students, which tend to under grade their higher-level peers.

In terms of concerns from the staff, the difficulty around how to prepare the curriculum so it includes peer review that leads not only to student learning and professional development but also to interest in participation (Taylor, Ryan & Pearce, 2015). Professor workload may also increase greatly, especially in large class, if the peer assessment is done inefficiently. In an effort to solve some of these drawbacks, several approaches were taken, and Web-based platforms started to include special features.

Returning to Taylor, Ryan & Pearce (2015) investigation, although the results from Phase One to Phase Two were considerably better, there was still an issue to be resolved, lack of confidence (and actual skill) from students to feel capable of proving quality peer assessment. To undertake this problem, 3rd phase introduced the 4Rs framework of reflection, which consisted in giving peer review and feedback that was Reporting, Relating, Reasoning and Reconstructing.

To allow for full engagement, students were taught how to 1) write peer reviews, by having explicit support and scaffolding, examples and training with samples and 2) how to address the feedback received, knowing how to weigh up on it and establishing a plan of action. The week before the students had access to a workshop which accounted 10/120 of the mark.

The results this time, showed by the 58 students who wrote the peer review and the 61 who dealt with feedback considered it useful to maximize the benefits of peer review and the final grade of those who participated was at least one grade higher than those who did not, which accounted for 15% of enrolled students.

In the final phase, during the 2nd semester of 2013 and 1st semester of 2014, some enhancements to the workshop were made. The 10 values were now given for students who attended, but also participated by submitting a small draft to be peer reviewed in the session.

The attendance was up to 97% and later non-participation in the peer assessment lower to 11%. The new survey showed that 89% of the 37 students felt the workshop helped their writing highly and 93% supported the use of peer feedback.

The use of four different phases evidences the importance of scaffolding support, in order to obtain quality reflective reviews which will lead to better practice through self-management and web-based technology as a peer grading environment, as it allows for higher efficiency management by staff, as well as providing a familiar interface for the current generation of students, which can lead to an increase in motivation and engagement.

To respond to the lack of motivation by students and make sure that the awarded grades were fair, some platforms allow to the authors of the reviewed documents to give a grade to the peer grader (Patchan, Schunn & Clark, 2017). However, even if the reviewer accuracy is considered good, it doesn't necessary relate to the students' ability to apply reviewing criteria. In that case, the use of some sort of calibration or training component can help solve this problem (Black and Gill, 1995), though it may mean an increase of workload for the student.

To look for improvements of the quality of peer assignment, a study has analyzed the effects of accountability to peer review and their consistency (Patchan, Schunn & Clark, 2017). The participants consisted in a group of undergraduate students enrolled in an Introduction to Laboratory Physics, from a public research university.

With the use of the SWoRD platform, 287 students received each week four randomized reports to be reviewed according to a pre-defined rubric, consisting of general guidelines in addition to a specific reviewing criterion divided into seven dimensions. For each dimension, students had several questions meant to point them to relevant aspects of the reports. The quality of their assessments would count 3% for their final grade.

After the initial feedback received, the authors could try to improve their grade based on the received

feedback and after submitting the final draft, they could rate the helpfulness of the reviewer feedback, in a 7-point scale too. The final draft was then graded by the professor with the same rubric.

The idea was for students to help their peers write a second better draft. To check whether accountability had an effect in the consistency of the grading, a questionnaire was provided to the students to check what they perceived as the part which they were going to be accountable, later dividing students into three groups – those who perceived rating and feedback, feedback helpfulness only and rating quality only.

The results showed significant differences in the effect of accountability on ratings given by students who perceived rating and feedback and feedback only from rating only, as the formers earned higher rating scores. When looking at the effects of accountability on the comment's students provided, even though the number of comments was similar, students that perceived accountability from rating and feedback, provided longer comments, assumed as better as they included more detailed responses. In all cases however, the comments were considered helpful by the authors. These results suggest that accountability on feedback improves the quality of the commentary provided by the reviewers and leads to a higher consistency on ratings.

The level of accuracy in peer assessment could be considered the relation of correspondence between the professional and the learner assessments. Most studies have found the validity and accuracy adequate (Sadler and Good, 2006; Falchikov, 2001). While most of the studies show a tendency of high marks given, some suggest an inclination to give more average grading (Gehringer, 1999). This could be due the type of subject, the level of cognitive skill or even the education level at which peer grading occurs.

Tseng and Tsai (2007) also explored the effects and validity of peer assessment as well as the effects from peer feedback with the help of a web-based platform. Peer grading with the help of the web is suggested to bring additional advantages such as the possibility for anonymity, which can make students more motivated to give a more critical assessment, lets the professor to monitor more efficiently, especially if its dealing with larger classes but also allows for savings in terms of paper costs, as students don't have to print a copy of their work for each student that will assess them.

In Taiwan, 184 10th grade students were assigned to design an itinerary project for their Computer Science class. They also had to grade ten other students' projects in three different rounds. Each student then had a chance of improving their work according with the feedback received by their peers.

Students' performance was evaluated in both qualitative and quantitative terms. The evaluation had three different dimensions: Creativity, Relevance and Feasibility, which were graded with a score from 1 to 7. The authors considered the smaller scale easier for students to score without giving arbitrary values. To determine the final grade in each dimension, the average was calculated. The inter-reliability of the peer assessment was measured and showed substantial high levels of correlation, from 0.71 to

0.80.

To measure whether peer assessment is a valid form of assessment, two professors were assigned to evaluate the projects too, with the same rubric. The correlation between the two was high (0,65; p ¡ 0,001). The authors also decided to include Chi (1996) framework to brand learning feedback in: Corrective, Reinforcing, Suggestive and Didactic.

The results showed that the average scores on all dimensions had increased on every round, both for students' grades as for professor grades', showing that students were progressing with the aid of peer grading. The correlation between students' grades and the professors were also found to be high from 0,49 to 0,79. Finally, the authors checked the relationship between the type of feedback received by students and their performance, with the results showing reinforcing feedback useful in all dimensions throughout the three rounds, suggestive feedback useful in the first round however not so much in the rest of the rounds, corrective feedback useful only in the creativity dimension and finally didactic feedback showing as having a negative impact on the students' scores.

These results reveal difficulties from students in giving didactic feedback which requires higher cognitive skills and knowledge, as seen previously, yet the overall experience allowed for a noteworthy improvement of students' projects quality due to the formative feedback received along with the opportunity to not only learn from other students but also by evaluating them through the exchange of knowledge provided from all social interactions.

The results of the presented research allow to confirm how the use of Web platforms can enhance the peer assessment experience for various reasons:

- Possibility of anonymity, which removes any social constraints from the students;
- Easy to submit work in different formats (Gerhinger, 2001);
- Familiar interface for students;
- Experience of grading peers a social event with multiple benefits through its exchange of information;
- Learning of skills related to Web creation which are essential in today's learning as well as business (Gerhinger, 2001);
- Removal of the costs of printing and photocopying the all work to be reviewed, especially when there are multiple reviews of the same work;
- Providing a way to distribute randomly all work and still making it easier for staff to monitor all events related to peer grading.

Overall, peer grading looks a promising alternative to professor-only evaluation in education, especially in higher education settings. In blended learning environments, peer grading gathers all advantages of peer grading achieved in traditional environments and fully online environments and helps removing the biggest challenges for professors and students when using this approach.

2.2 Blended Learning

Blended Learning is defined as a combination between face-to-face instruction with computer-mediated instructions. It represents a transformation of the way we approach learning, by redesigning and rethinking of how the learning experience should be, by openly evaluate and integrate the strong points of learning face-to-face and online, to achieve educational goals that are worthwhile (Garrison & Vaughan, 2008). It challenges professors to teach differently and offers flexible design possibilities.

Blended learning is already a common approach in Higher Education according to Arabasz and Baker (2003), stating that 80% of the institutions already offer blended learning courses.

However, there are some considerations to have in account. Badrul Khan defined Khan's Octagonal Framework, where each dimension corresponds to a problem that needs to be addressed beforehand. Its various problems are: Pedagogical, Institutional, Ethical, Resource Support, Technological, Interface Design, Evaluation and Management.

Building an effective learning environment with all these problems in mind, can lead to an overall better experience and better outcomes. There are several studies that show the benefits of using blended learning (Singh, 2003), such as:

- Increased range – a virtual classroom removes possible time and space constraints as it allows remote access;
- Development optimization – an entirely web-based system can be very expensive, but by combining both systems, the cost and time attribution is more efficient and optimized;
- Learners engagement – provides a succession of interactive experiences that lead to solving a given problem with the help of other students;
- Blended Learning proved that it works – studies made by the University of Stanford and University of Tennessee offer discretion about mechanisms in which blended learning is better than the traditional methods or web-only methods.

Both in blended learning and traditional learning, the professor is the one responsible for the grading of its students, though this assessment method could be extended with the use of peer assessment.

Some alternative ways to implement peer grading include distance learning, in particular, Massive Open Online Courses (MOOCs).

Blended learning offers a way to transcend the conventional paradigm of classroom. It maintains the possibility of direct communication between professor and students and removes the limitations of traditional learning, by adding features available in fully online courses, and combines it in the most complete and engaging learning experience for students, especially in a time where online presence is very strong. As technology advances, in the future, blended learning systems will distinguish themselves according to how they allow/create blending (Garrison & Vaughan, 2008), especially with the growth of virtual learning environments.

2.2.1 Virtual Learning Environments

A Virtual Learning Environments (VLE), is a web environment based on computers which are relatively open systems, allowing the integration of face-to-face learning interactions with online learning activities and meetings with other participants that provide access to a large variety of resources (Piccoli, Ahmad & Ives, 2001). The Joint Information Systems Committee Managed Learning Environment Steering Group (JISC, 2001) has defined VLEs as "components in which learners and tutors participate in 'online' interactions of various kinds, including online learning".

These systems are best known for their sharing and communications capacities between students and professors (forums, chats) as well as providing various ways evaluate and manage students content (quizzes, grade of assignments, etc.) (Piccoli, Ahmad & Ives, 2001; Chou & Liu, 2005).

To achieve an efficient learning environment, it should be taken in consideration three pre-requisites: quality of course design, selection of appropriate tools and definition of context in which the learning occurs (Stiles,2000; Dillenbourg, Schneider & Synteta, 2002).

With a VLE, it is possible to take full advantage of the WWW flexibility and power, as all technologies associated with it, use virtual (instant) communications and information technology (Stiles, 2000), access of materials by students independently and remotely, achieve high levels of student control, restructure the experience of learning and give immediate, or at least, faster feedback (O'Leary & Ramsden, 2002; Piccoli, Ahmad & Ives, 2001).

Most VLEs consist of MOOCs, where peer grading serves as an essential tool to alleviate grading of complex open-ended assignments to courses that go from hundreds of students to the thousands (Piech et al., 2013).

Ideally, a peer grading in a MOOC should be able to provide assessment with high levels of reliability, guarantee that the workload assigned to the students and the staff is reasonable and limited, scale to the different sizes of the classes and allow for a wide collection of problems to be applicable.

VLEs appear as a remote learning experience that offers an organized way of distributing learning

materials in an electronic and more efficient manner, without insolating its users from each other or its professors (Chou & Liu, 2005). Moodle is an example of a VLE commonly used in fully online and blended courses, due to the enormous capacities it offers to educational settings.

2.2.2 Moodle

The Modular Object Oriented Dynamic Learning Environment, known as Moodle, is an open source course management system used by education systems to provide the addition of Web technology their courses (Cole & Foster, 2007). The platform can be used exclusively as an online course or as an enhancement to traditional learning. It is one of the most used VLE that allows for the creation of a course website, providing access only to enrolled students (Costa, Alvelos & Teixeira, 2012). Meant to be flexible, compatible and easily modifiable, Moodle uses simple technologies such as shared libraries, abstraction and cascading stylesheets. The platform is characterized by a set of functionalities, grouped in 2 classes:

1. Resources – Represent institutional materials, created in a digital format and later submitted. Some examples are quizzes and class presentations;
2. Modules – Components created in Moodle to provide interaction between students and professors towards the creation/transformation of content. This includes upload and revision of work, news, forums and chats, etc.

All these characteristics allow for an easy exchange of information between students dispersed geographically and can be accessed anytime. It is also easy to reconfigure so that professors can evaluate students in a wide range of ways, provide lots of complementary tools to support the learning and teaching process. Lastly, Moodle includes tools for controllable management of tasks according to the timetable of certain courses.

The use of a gamification methodology in Moodle is easily added through some tools which are mostly included in Moodle already such as avatar/profile picture, visualization of student progress, display quizzes results, levels, feedback, badges, etc.

Overall, Moodle shows itself as one of the most capable and robust VLE systems, especially useful in courses with a large number of students. It is also an example of a VLE where gamification features can be included to increase student engagement and motivation. It is already used in the Multimedia Content Production (MCP) course and offers the opportunity to get more concise feedback, ensuing in a better learning experience. Peer grading has been included in this course before (Ribeiro, 2016), through the use of a plugin, in which several students were given assignments and graded their colleagues. This activity included a gamification component. For their participation, they would be awarded with a badge and extra Experience Points (XP), which were added as a bonus to students' final grade.

2.2.3 Gamification

Gamification is the adoption of game experiences and mechanics in non-game activities, leading to the simulation of experiences and practicing the acquisition of skills (Arnold, 2014). This idea is already largely visible seen in mobile apps with the inclusion of challenges, leaderboards and badges and it is becoming a common feature in businesses and education.

Some features that may be included in the experience (Kiryakova, Angelova and Yordanova, 2014) include considering all users as participants, having challenges or tasks which users complete and progress towards a certain objective, the possibility of leveling up after reaching a defined amount of points, being awarded with badges after completing a set of actions or seeing a ranking system that considers users' achievements.

The increasing use of gamification in educational courses recognizes that games stimulate the productivity and creative inquiry between students (Johnson et al., 2014). It allows for entertainment that is awarded with knowledge and new skills.

The involvement with technology nowadays starts right at the beginning of our lives, and the use of this methodology provides a familiar ambient where students feel more involved. It also provides a great way of tracking user progress, which is equally important in education and in games.

Gamification in companies has shown to have several benefits, some examples are (Arnold, 2014):

- Increase of participation in the various surveys regarding the level of satisfaction;
- Higher motivation to complete mandatory and optional training;
- Encouraging to adopt positively changes in management projects.

The use of gamification in education has showed good results in several contexts. A gamified environment used by the IT course at Corvinus University of Budapest, during the years of 2015 and 2016, was created within Moodle and included features such as rewards, alternate learning paths and several feedback and social communication options (Barna and Fodor, 2017).

Gamification is an effective approach that can bring positives changes in the student behavior and attitude towards their learning, by providing better and more familiar conditions and higher comprehension. Peer grading could be included in this context in various ways, for instance, using badges as a reward to their peers' work (O'Connor & McQuigge, 2013), offering a way for students to reward each other as well as enabling better communication and cooperation.

2.3 Discussion

This chapter gives an overview of all existing research literature regarding this work. It started by defining peer grading and reviewing studies where this assessment method was used and what advantages and

shortcomings were found. After this, blended learning was presented and how its use can improve students' learning by joining the strengths of face-to-face instruction with a web-based ambient. To achieve this educational approach, VLEs such as Moodle, offer a set of modules that provide a way for students to create work as well as to submit and exchange with other students easily, especially in the case of large classes, such as MOOCs, which may include students from different geographical locations. The inclusion of gamification in these environments makes the learning and participation in course activities much more appealing.

The inclusion of peer grading in a gamified blended learning ambient could provide as an alternative solution to an evaluation previously only done by the professor's judgement. It started by being implemented in traditional environments (Topping, 2009; Freeman & Parks, 2010) and as blended learning environments later appeared (Gerhinger, 2001; Tseng and Tsai, 2007; Patchan, Schunn & Clark, 2017), it was brought along and was able to profit from computerized features such as anonymity and automated attribution. Fully online environments, such as MOOCs also have used it as an assessment procedure (Meek, Blakemore & Marks, 2017), where scalability takes an important form.

Overall, peer grading shows it can be an efficient and valid learning strategy in higher education courses. In most cases increases the student's motivation and makes students do their work in time (Taylor, Ryan & Pearce, 2015), throughout the semester instead of leaving it to the last weeks of the semester. It raises students' levels of critical thinking (Meek, Blakemore & Marks, 2017) and can lead to the improvement of their own grades through time (Tseng and Tsai, 2007). Lastly, most cases showed students can make fair and high-quality reviews (Freeman & Parks, 2010; Tseng and Tsai, 2007; Meek, Blakemore & Marks, 2017) and that grades were awarded accordingly, especially if the students were scaffolded or had any training/calibration given beforehand (Patchan, Schunn & Clark, 2017; Taylor, Ryan & Pearce, 2015). Features such as anonymity offer a way for students to feel more comfortable and to give honest grades without any fears of repercussions (Tseng and Tsai, 2007). When students feel accountable for their actions and their grading will have an impact on the final, the evaluation given appears to be even more accurate (Patchan, Schunn & Clark, 2017). Still, in cases where anonymity is not possible, such as in blended learning environments, there is no research that shows whether social relationships affect peer grading.

3

Sociometric Analysis

Contents

This chapter presents all information gathered related to the measurement and understanding of social relationships through sociometry. The concept of sociometry is firstly explained and then studies made in order to define and compare the different methods to establish sociometric status groups are presented. Finally, one method to be later used in this study is chosen and explained in more detail.

3.1 Sociometry

Sociometry is a measurement based on the degree of affiliation amongst a group of people. The sociometric technique was developed by Jacob Levy Moreno in 1951, who defined it as "the inquiry into the evolution and organization of groups and the position of individuals within them". Several sociometric studies have been performed ever since in different environments, from business corporations to schools. It is a powerful tool for assessing the hidden dynamics that create the structure of certain groups (Rostampoor-Vajari, 2012). Given that factors such as trust reputation and cooperativeness between students can have impact in the assessments provided, sociometry could provide answers to understand if previous and current relatedness between people are reflected in the way students assess each other works.

Previous research conducted in terms of sociometric analysis showed that occupying a central position in the network of informal peer groups, having friends and having a certain sociometric status, that is, the degree to which someone is liked or disliked by their peers in a group, relates to unique behavioral reputations (Gest, Graham-Bermann & Hartup, 2001), which can lead to bias, whether it is intentional or not, in the assessment of another peers' work.

Classifying people into social status groups has been done to examine children status in peer groups and possible implications of their popularity or rejection. This classification (Terry & Coie, 1991), can be done either through peer ratings – the assessment of relationships achieved through peers rating each of their peers in terms of much they like them or through peer nominations – each peer classifies a small number of peers in a group of who they like the most and who they like the least. The status a person belongs to can be determined simply from the totals of positive and negative nominations, or the two totals can be combined and create a value of social impact and social presence. It is also possible to combine peer ratings with peer nominations by replacing the negative nominations and instead use 1 a rating in a 5-point rating scale.

In the work of Terry & Coie (1991), the four most common procedures for classifying children into social status groups were compared with each other in terms of discriminant validity and temporal stability. These included two positive and negative peer nominations systems, one where students could be left unclassified and another where all students would have to fit in some group; a positive only peer nomination system combined with peer ratings and a rating-scale system only. The results revealed that

in situations in which the population consisted of people that may not all know each other, peer nomination will more likely provide a more accurate picture of the social status. Peer nominations methods also are more capable of captivating peer connections from long standing relationships, due to the part of the population that knows each other, to have connections started years ago that can be expressed when nominating peers. Furthermore, if the goal to the social status grouping is related to find possible behavioral comparisons between groups, having some students left as uncategorized may be more useful.

These results from the work mentioned lead to believe that the most adequate system to use in this research is one that uses peer nominations, both positive and negative ones, and that did not force all students into some category, i.e., the Coie et al. (1982) system.

The Coie et al.(1982), hereafter described as the CDC system, consists in a bi-dimensional system which uses peer nominations to group a given population into five main sociometric status: popular, rejected, controversial, neglected, average. In the original version of this system, it is possible for a group of the population to be left unclassified, as the main objective of this classification system is to group people according to possible behavioral differences and thereafter grouped as other. To calculate to which group each person from a given population belongs to, each absolute frequency of positive and negative nominations is standardized, creating an liked most (LM) and liked least (LL) value. With these values, it is possible to calculate the social preference (SP) and social impact (SI) of each person, as LM-LL and LM+LL respectively.

These values are also standardized, and each person is then attributed to a given group according to the following formulas: Popular, SP score greater than 1.0, an LM score greater than 0, and a LL score less than 0; Rejected, SP score less than — 1.0, a LL score greater than 0, and a LM score less than 0; Neglected, children with an SI score less than — 1.0 and an absolute frequency of positive nominations of 0; Controversial, children with an SI score greater than 1.0 and both LM and LL scores greater than 0; Average, children who receive both an SP score and an SI score between -0.5 and 0.5.

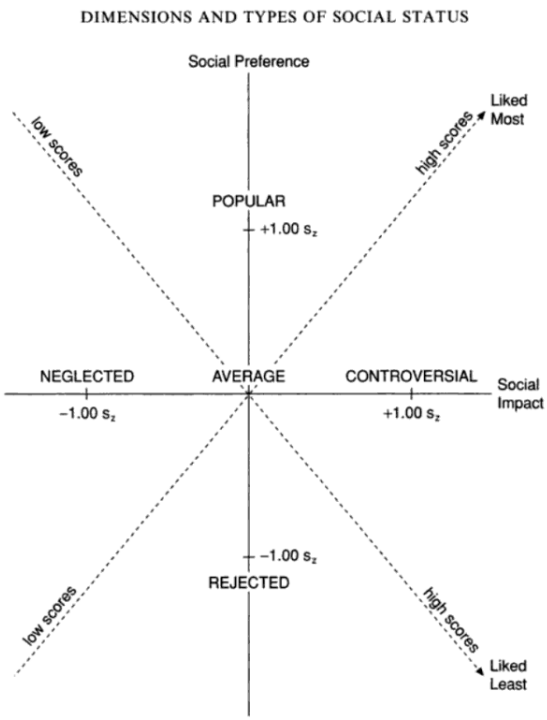


Figure 3.1: The different social status placed according to their characteristics

To reach the objective of this work, it was required a system that could get the most in the most truthful form possible, the sociometric status of students, even if meant leaving some of them unclassified. In addition to this, the system used should be one that does not need for all students to know each other as the course is constituted by a large number of students. While it is infeasible to use a typical rating-scale method, and could also lead to higher error levels, since the core of the population in these course are students from a relatively stable classroom, peer rankings could also provide useful extra information, though, peer nomination seems the most relevant and feasible method to verify whether relationships between students have an impact in the grading provided by students to their peers. The CDC system, consequently, seems the best option for this study.

Students were asked directly how they feel towards certain students. The new version of the plugin includes mandatory peer nominations previous to peer grading and peer ranking will be prompted as an optional inquiry available to do throughout the whole semester. The details in how these metrics will be presented to students are explained in more detail in the next chapter.

3.2 Discussion

Sociometry research has showed that classifying students into different sociometric groups can help discover and even predict behavioral standings. Since the objective of this work is to investigate the

effect of social relations on students' peer grading, given that trust and reputation between students can have impact in the assessment provided, the use of sociometric measures is crucial to understand not only how reliable students can be when peer grading, but also if their placement in a certain sociometric group can lead to bias from other students.

The MCP course has around 130 students each year and in higher education settings, such as this case, most students only have some sort of friendship or close relationship with a portion of their peers. The number of enrolled students also makes almost impossible for students to know all of their peers. In addition, this particular course, is not mandatory and includes a considerable portion of students which are new to this educational environment. However, since this course is part of the MSc Information Systems and Computer Engineering at Instituto Superior Técnico, University of Lisbon, the existing relationships and connections between students that do know each are likely to be long lasting and should be observed as a factor of possible influence when peer grading.

4

PeerForum Plugin

Contents

Peer grading has exhibited to be a valid assessment method in various learning environments. The objective of this work is to verify the validity and reliability of peer grading in a gamified blended learning environment and investigate the effect of social relationships on students' peer grading. To do this, we used the Moodle platform, more precisely, the forums section and included a peer assessment option, through a plugin.

Although the plugin had already been developed and provided peer assessment capabilities to the platform, it contained some issues regarding performance and usability, making it not completely reliable in its current form. Moreover, it did not include any way to gather and extract additional information regarding students' relationships and feelings towards each other.

This study required for two versions of the plugin to be developed – 2.6 and 3.5 versions. Both versions have the same features, and while the 3.5 version corresponds to the most recent version of plugin and compatible with the latest version of Moodle, the 2.6 version was the one used during the study for compatibility reasons with the current version of Moodle used in the higher education course.

In the following sections, a brief description of the current state of the PeerForum plugin is given, followed by a detailed description of the work done to fix the plugin known problems and development of new capabilities to the already existing plugin.

4.1 Peerforum Initial State

The PeerForum plugin offers peer grading capacities to Moodle (Ribeiro, 2016). It was created for students to evaluate their peers' submissions in Moodle, through the activity module PeerForum, and was responsible to hand out work equally throughout the students. It also includes the PeerBlock module, responsible for the management of all activities. Both students and professors have access to this module, with different interfaces and actions available for each group. The students use the block to easily peer grade all posts given to them and can also check previous posts assigned to them, whether they had submitted a peer grade or let the post expire.

The professor can visualize all information related to the peer grading activity, from which students were allocated to which posts, see all peer grades given and have a general vision of the statistics of the peer grading participation/engagement. The professor was also responsible for creating new assignments, provide instructions and the respective peer grade criteria. After this, every enrolled student had the option to submit their work related to the assignment. Each submission was then given to a group of students automatically by the system, yet, the professor could manually select students to grade a certain submission, if the need arose. The students assigned to each submission were listed and all peer grading activity of each post could be managed by the professor, including the option for professors to either remove or block students manually.

Students which are assigned to peer grade have a defined period to provide a grade according to the rating scale and metric defined by the professor, together with some feedback. Both the grade and feedback provided could be edited up to 30 minutes after the grade was posted. The system was also able to detect outliers and the professor could decide to block the grade. The professor could either grade the submissions independently or also grade these submissions as a peer grader. After all assigned students (and possibly professor) provided a grade, the final grade was presented to the evaluated student. The weight of the professor peer grade/rate and the weight of the students' peer grade were configurable by the professor and can go from 0 to 100 percent.

However, the plugin had some known issues that while it did not impede its use, made the system performance and response uncertain at times. Although it was designed to hand out assessments throughout all students and in an equal manner, in practice, only some students received work to assess, with some students receiving an enormous amount of work to grade and others very little to none, which created an unfair workload. Another problem that prevented this plugin utilization was related to the high number of server requests sent each time a page in the forum was loaded, which would send the whole platform down in periods of high traffic. An initial solution to this problem, consisted in having each topic in the forum to have pagination according to a certain number of posts made. However, this implementation brought up another problem. Not only the counter of posts did not include responses to posts, which made the pagination incoherent, but also the notifications provided by the system via email did not have the page included in the link to the new post, which lead to professors not being able to find the post or response to the post made, having to look through every single post and response to find which one was new.

All these described issues were addressed and fixed for this newer version of the plugin. Every change made to the plugin is described in detail in the next section.

4.2 Peerforum Adjustments

In this section all corrections made to the previous version of the plugin are addressed. Some of these issues were left reported and others were found while testing the system functionality and during its usage period.

4.2.1 Unbalanced Student Attribution to Peer Grade Posts

In order to solve the unbalanced attribution of peer grades, the algorithm now prioritizes students which have participated less to be selected when a new post is submitted in the PeerForum. Each time a new post is made, all students who are considered eligible to peer grade are retrieved and from this set of students, instead of choosing one randomly, students are now sorted according to the sum of

posts they have peer graded plus those they have yet to peer grade, from the lowest value to the higher. The student with the lowest resulting sum is then assigned to peer grade the post. This way students have the opportunity to make up for peer grades that they let expire and all have a similar number of peer grades done. Students who do not participate can be blocked by the professor as to avoid being continuously chosen.

4.2.2 Server Requests Efficiency

To achieve an increase of the plugin performance, some elements of the page that display in all replies to a new discussion topic are not loaded when the page is opened initially and only when requested. Some elements from each post such, as the peer grades given by each student and student assignment details, are now loaded when clicking on the collapsed link displayed at the bottom of each post, instead of loading in the beginning with the post contents. This is done with AJAX and allows for less server requests each time a discussion page is opened.

4.2.3 UI Changes

Several user interface changes were made to improve system usability in both modules.

In the Peerforum module, the pages which display all replies to a discussion topic suffered some changes. Every post included extra white space which only made the posts look larger than they actually needed to be. That extra space is now removed, allowing to display more information in the window. Whenever a peer grading activity is considered finished, the message “The peer grading activity is now finished.” is displayed in the post footer. This message only appeared to students initially and now is displayed to everyone.

In the PeerBlock, there were changes in the persistent block that is displayed on the top corner of the page. Students’ view now also informs students of peer rankings available to fill. The professors’ view only served to access the block contents, now it also displays some general information related to the peer grading activities taking place, i.e., the number of posts were peer grading is elapsing and the number of posts were peer grading is expiring.

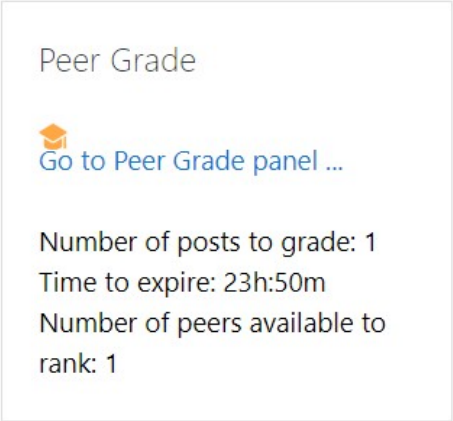


Figure 4.1: PeerBlock view from the student's perspective

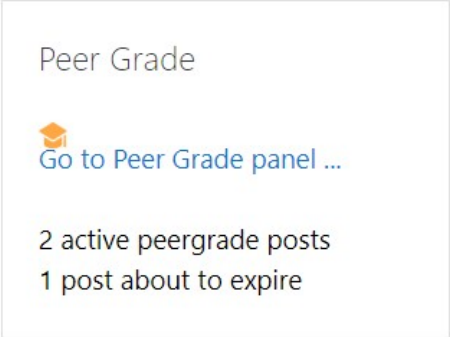


Figure 4.2: PeerBlock view from the professor's perspective

When accessing the PeerBlock, several contents can be observed. To avoid excess of information being displayed and improve usability, the PeerBlock contains several tabs with different information available. However, the existing tabs had long names that made the tabs overlap and many of the tabs repeated parts of their name, therefore all names were shortened, to make the interface look cleaner.

The visualization of graded posts and ungraded posts by each student were in different tabs, which could be confusing sometimes, and since in both cases the displayed information consisted in a single table that shared most rows, the two sections were merged. This table also included one row "Feedback" which displayed the written comments given by students. Since commentary can go from a couple of words to dozens of lines, the result would be a very unbalanced and unusable table. To avoid this, while at the same time not losing the information, each feedback was limited to a certain space and if it is too long, the remaining feedback would not be presented and replaced by "(...)". The feedback visible is now also a link and clicking on it makes it possible to see it in full in another browser tab.

The graders' statistics tab was the only one were the information was displayed in a static table and also included pagination that divided information in many pages, which made it hard to observe and get an overview of peer grading participation. Pagination is now removed from this tab and it's possible to

filter students by name or surname and/or order students according to the different rows displayed.

Lastly, the Manage Posts tab was changed to display the information about posts that have not expired only when it's loaded initially, instead of displaying all posts and the filtering options in this section were corrected so that when the current page displayed changes, the filtering options are not reset.

4.3 Peerforum New Features

This section presents all new features implemented in the plugin to allow the collection of all information needed to later built the sociometric status groups and also to further improve the plugin and provide more capabilities that can help peer grading achieve better results.

4.3.1 Sociometric Status Data Collection and Visualization

To solve the research problem, the PeerBlock can now be configured to include a form where students nominate their favorite peers and least favorite peers. The number of students nominated can be configured to be a fixed number or a minimum number, giving students the option to give more answers if they want.

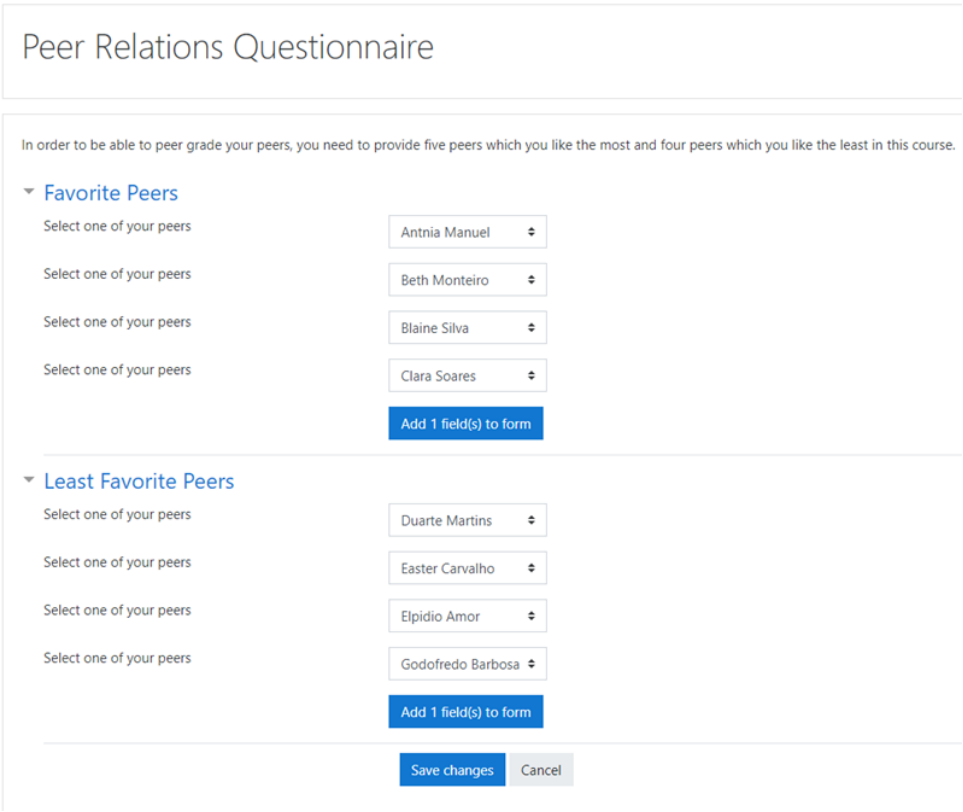


Figure 4.3: Peer nomination section for students

Besides peer nomination, peer ranking can also be done. If this option is activated, the student PeerBlock version will include a new section called “Peers to Rank”, in which students can rank peers which they have peer graded.

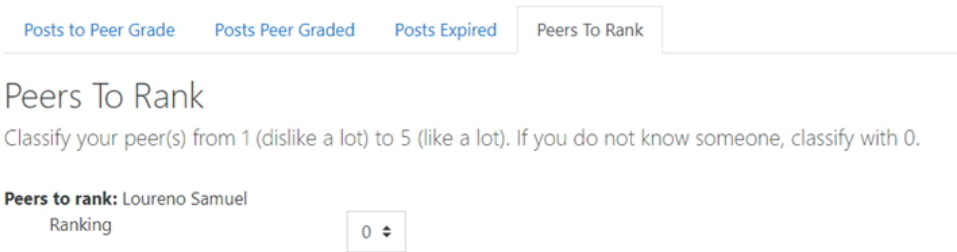


Figure 4.4: Peer ranking section for students

The answers submitted by the students in either peer classification systems can be seen, and posteriorly edited if needed, in the professor’s version of the PeerBlock, in the tab “Students Relationships”.

Posts | Graders posts | Peergrades | Graders statistics | Conflicts | Training | Student Relationships | Threaded Grading

See Peer Nominations

Student	Favorite Students	Edit	Least Favorite Students	Edit
Antnia Manuel	Blaine Silva	Remove	Torrance Machado	Remove
	Beth Monteiro	Remove	Orpheus Cruz	Remove
	Easter Carvalho	Remove	Xanthe Santos	Remove
	Godofredo Barbosa	Remove	Loureno Samuel	Remove
	-	Choose a student... Add	-	Choose a student... Add
Clara Soares	Antnia Manuel	Remove	Xanthe Santos	Remove
	Beth Monteiro	Remove	Waylon Azevedo	Remove
	Blaine Silva	Remove	Torrance Machado	Remove
	Carlota Neves	Remove		Remove
	-	Choose a student... Add	-	Choose a student... Add
Micaela Matos	Easter Carvalho	Remove	Clara Soares	Remove
	Ricardo Nunes	Remove	Antnia Manuel	Remove
	Nella Blair	Remove	Nuno Cardoso	Remove
	Duarte Martins	Remove	Torrance Machado	Remove
	-	Choose a student... Add	-	Choose a student... Add

Figure 4.5: Example of visualization of peer nominations results

Posts | Graders posts | Peergrades | Graders statistics | Conflicts | Training | Student Relationships | Threaded Grading

See Peer Rankings

Student	#Students Not Ranked	Students Ranked	Rankings	Edit
Antnia Manuel	0	-	-	Add
Clara Soares	1	-	-	Add
Micaela Matos	0	-	-	Add
Nuno Cardoso	0	Filipe Barros (id:4)	3	Remove

Figure 4.6: Example of visualization of peer ranking results

4.3.2 Training Pages

Previous work on peer grading in Chapter 2 stated how the inclusion of training leads to higher levels of student participation in peer grading activities, as it helps students to become more confident that they possess the capacities required for performing the task, and consequently to better quality evaluations. Therefore, it was included the possibility for students to consult training pages related to the subject of the post they had been given to peer grade. In the PeerBlock section where all posts to be peer graded are displayed, above each post, a link would appear that would open the training page related to the subject of the post. These peer grade training pages explained briefly what was intended to be submitted, a list of requirements for the submission to be accepted and examples of previous submissions, the rate given and some explanation of why it received the following grade. In the professor version of PeerBlock, a “Training” tab is available in which all training pages can be accessed and configured.

Links to the training pages are also available in the general Peerforum page, each skill link is next to the discussion name, as to be easily accessible anytime and not just when a student has to peer grade.

4.3.3 Hidden Professors' Reply Posts

To avoid that students could be influenced by the professor output on the posts, the professor rate could be configured to be temporarily hidden, according to criteria defined in the PeerForum settings. However, the professor feedback, which was in a reply post, was still visible to everyone and students could still be influenced in their peer grading and hence the professor would usually wait until peer grading was finished before posting feedback. It is now included an option to also allow hiding of professor responses to students submissions. This way, the professor can rate and provide feedback to posts whenever it decides to so and avoids accumulating submissions to grade because it had to wait for students to peer grade first.

Instead of displaying the feedback given, a message saying "This reply is not available yet." is shown. In the cases of peer grades given, both the resulting peer grade and each individual peer grade are also now possible to hide to both students and professor until the peer grade activity of a post is considered finished.

Figure 4.7: View of professor post reply from the professor perspective

Figure 4.8: View of professor reply from the student perspective

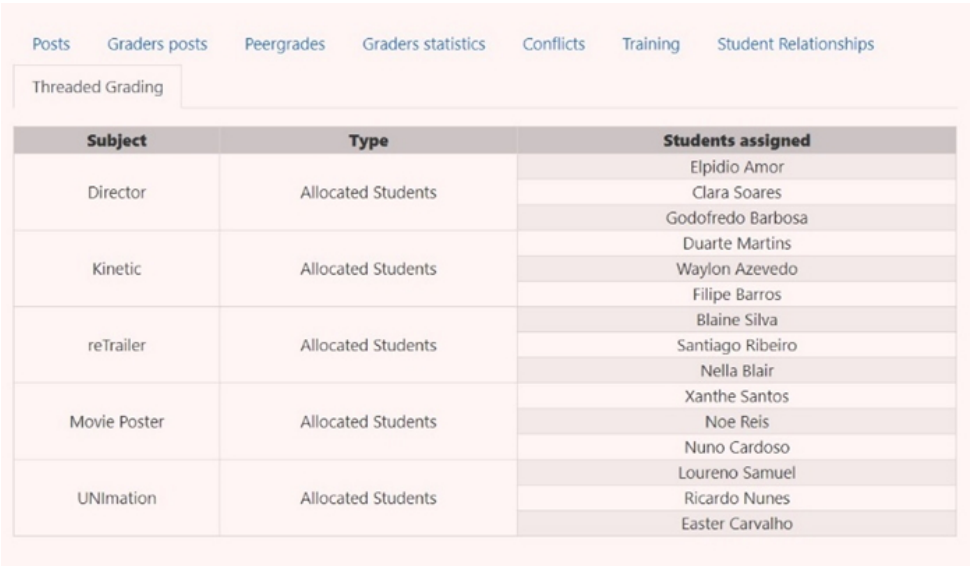
4.3.4 Automatic Assignment of Improvement Posts

This new version of the plugin also allows for students replies to their initial reply to have automatically assigned the same set of peer graders from the first reply, as these cases are usually related to submissions that were modified and/or improved given the initial feedback given by both the students and professor. Previously, it was only possible for the professor to manually assign the same set of students, but these required for professors to go to every one of these posts, which could become bothersome in cases where the subjects have dozens or hundreds of submissions.

Currently, whenever a post is submitted, before trying to assign any number of students, it is verified if the post is not a reply to the main post of this discussion and if the user submitting the post is a student. In the situations where both these conditions are true, the students from the post parent, which has students assigned, will be assigned to this new post.

4.3.5 Advanced Topic Attribution Algorithm

In addition to the changes already made to the peer grading attribution algorithm, it was also introduced the possibility to divide students throughout the different subjects available. This distribution can be done either by randomly divide all students as equally as possible through all subjects or in a more detailed manner. In this case, for each available subject, the professor could decide whether that subject would be considered a “Specified” topic, i.e., one that would require to designate a given number of students and only they can peer grade this topic submissions or a “Randomized” topic, one that all students left could be potential peer graders. By default, all created discussion topics are considered “Randomized” topics and any students can have assigned all posts.



Subject	Type	Students assigned
Director	Allocated Students	Elpidio Amor
		Clara Soares
		Godofredo Barbosa
Kinetic	Allocated Students	Duarte Martins
		Waylon Azevedo
		Filipe Barros
reTrailer	Allocated Students	Blaine Silva
		Santiago Ribeiro
		Nella Blair
Movie Poster	Allocated Students	Xanthe Santos
		Noe Reis
		Nuno Cardoso
UNImation	Allocated Students	Loureno Samuel
		Ricardo Nunes
		Easter Carvalho

Figure 4.9: Case in which all students were randomly assigned to the available topics

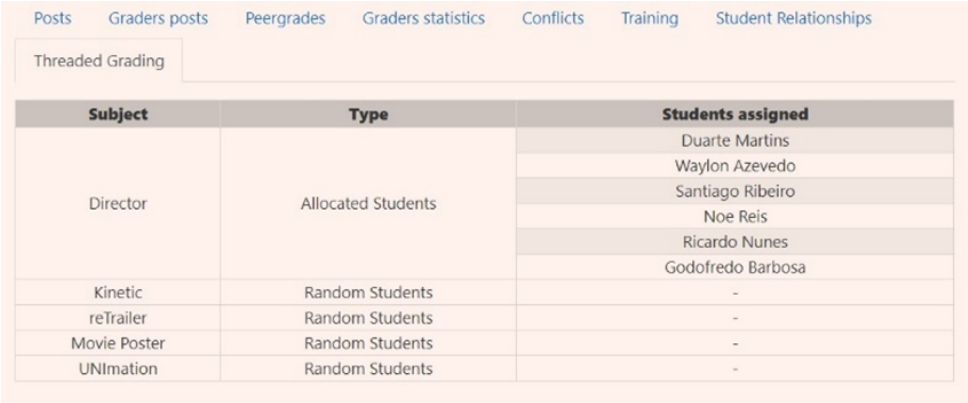


Figure 4.10: Case in which the professor has one topic selected as ”Specified” and remaining topics stay in their default form

4.3.6 Email and Internal Notification of Peer Grading Assignments

For a student to be aware that they had peer grades assigned, it was required for them to visit Moodle regularly, since that information was only available in the PeerBlock, which could affect the participation rates. In the most recent version of the PeerForum, every time that a new post is submitted, all students assigned to peer grade the post now receive a message in Moodle from the professor, with the message “You have a new post to peer grade.”. By default, when a student receives a message in Moodle, an email will be sent with the content of the message, informing them that they have work to peer grade. This implementation avoids the need to create a new type of notification.

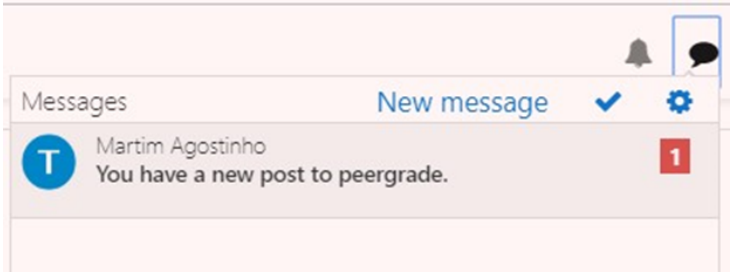


Figure 4.11: Message received when a student has a post assigned seen through the Moodle platform

4.3.7 Data Exportation

Lastly, in order to facilitate the analysis and study of both the peer relations and the peer grading activities, it was included in the PeerBlock sections of “Manage Posts”, “Graders Statistics” and “Peer Relationships”, links to export the data related to posts made by the students, overall student participa-

tion and answers to the peer nominations and peer rankings respectively. When clicking on these links available at the top right corner of the different sections of the PeerBlock, the user would be prompted to save a generated .csv file containg the information related to the PeerBlock tab it was displaying.

5

Peer Grading Analysis

Contents

The newer version of the plugin was developed with the purpose of gathering information that would let us understand if social relationships between students influence students' ability to provide peer grading that is unbiased and reliable. To solve this research problem, the plugin was included in a gamified blended learning environment, specifically the Multimedia Content Production (MCP) course, where students could assess each other's submissions along with the course professors and share their social preferences.

5.1 MCP Course Overview

This study was conducted in the Multimedia Content Production course for the MSc in Information Systems and Computer Engineering at Instituto Superior Técnico, University of Lisbon. This course takes a blended learning approach with face-to-face components, consisting of two theoretical and one laboratory class every week, and an online component, supported by the Moodle platform, where students can obtain all necessary content to work successfully in the face-to-face lectures, discuss any topics related to the course curriculum and submit assignments to be assessed by the professors. The Moodle platform includes a system that is able to provide gamification features to the MCP course. By including features such as levels, leaderboard, badges, etc., it was intended to make the course more appealing and motivating for students to engage actively throughout the whole semester.

While completing the course activities, instead of receiving a traditional grade, students gain experience points (XP). The former 20-point grade system is replaced by 20 levels, each separated by 1000XP, going from 0 to 20000XP. For every 1000XP, students level up. In the leaderboard, students can see how much XP they have accumulated compared to their colleagues. There are several ways for students to gain XP such as, completion of laboratory exercises, weekly quizzes performed on the Moodle platform at the end of theoretical classes, receiving badges and submitting multimedia related work on the forum.

The section where the students submit work to be later evaluated is called the Skill Tree. This component is done through the forums and consists in producing various types of multimedia content throughout the semester. The skill tree is divided into four levels. Each level has several nodes, each corresponding to a different skill that can be achieved. By completing a certain node, it is granted a fixed amount of XP, depending of the level it belongs and other nodes in the upper level of tree may become available to the student. Each student can choose which nodes it wants to complete and follow the path in the skill tree it wants according to its interests.

The evaluation process is usually done by the professor, who has the responsibility of going through all new submissions to rate them according to the defined criteria as well as leave some feedback, whether is just some congratulatory message or constructive criticism pointing out the flaws found. Students can resubmit their work to improve the initial grade received.

5.2 Experimental Protocol

The first version of the new plugin was used during the final nine weeks (out of 13) of the second semester of the academic year of 2018/2019, in the Multimedia Content Production course. As an incentive to participation, students were given an extra XP if they participated in the peer grading activity. Since this iteration was meant to evaluate if peer grading could be used as a reliable source of evaluation, all grades provided by the students have been not accounted for the students' final grades in their posts.

5.2.1 Peerforum Plugin

The peer grading plugin was introduced at the end of March. At this time, students had already been submitting their work assignments in the traditional MCP course forum for their level one and two skills, therefore the PeerForum created only included subjects corresponding to level three and four of the skill tree, as to not create confusion to students on the right place their submissions.

Students were given five days to grade each peer grade they received, and all peer grades given were anonymous. During this time professors could also rate and give their feedback on submissions. To avoid students (and professor) being influenced by the other responses submitted already, all rates and feedback given were temporary hidden. They were only made visible to everyone once the peer grading activity to a given post was considered ended, i.e., there was a minimum of three peer grades submitted or the time to peer grade had expired for all students. Students who had to peer grade a certain post could not see neither their peers nor professor responses until they peer graded the submission, even if there was already a minimum of three peer grades. After peer grading it, they would also be affected by the restrictions mentioned above.

5.2.2 Research Questions

Students who had nominated peers and during the course performed peer grading to the same people were checked for biased grading. Peer rankings were also observed. These results were used to answer the following research questions:

1. Did the peers who nominated or ranked other peers favorably showed some sign of bias when evaluating their works?

2. Did the peers who nominated or ranked unfavorably showed some sign of bias when evaluating their works?

3. Is peer grading a valid and reliable evaluation method in a higher education gamified blended learning environment?

5.2.3 Data Overview

From the 112 students enrolled in this course, 97 of these participated at least once in the peer grading activity (86.6%). To be able to peer grade their colleagues, students had to first answer to a questionnaire where they would nominate the peers they liked most and peers which they liked least. These results will later be used to form the sociometric status groups of these course students through the CDC system.

Throughout the seven discussion topics created on the PeerForum, a total of 180 posts were submitted, of which 96 were graded (53.3%). From these 96 graded posts, resulted 235 peer grades. Students who participated were assigned five posts on average and graded 68.1% of their assigned posts. It is worth mentioning that the participation levels decreased severely on the final days of the semester, possibly due to higher participation from students on the other components of the Moodle platform as to make remaining submissions before the deadline and higher load of work to do in other courses that they were attending during this semester. Also due to the fact that students can only be evaluated for their submissions and possible improvements made until the end of the semester, the time to which students had to peer grade decreased from five days to only one day.

As previously mentioned, students were asked to peer nominate their colleagues before being able to participate in the peer grading activities and to peer rank the peers they had graded during the semester. The peer nomination consisted of a questionnaire in which students nominated at least four peers they liked most and four peers which they liked least. Optionally, students could provide more answers. This questionnaire had a total of 101 answers and provided a total of 846 nominations.

After grading an assigned post, students had the option to peer rank their colleague(s) from one (dislike a lot) to five (like a lot). Since this course takes place in a higher education learning environment, with hundreds of students enrolled, an extra option, zero, was provided in case students did not know who the peer they graded was. In the peer ranking activity, 80 students ranked at least one student, providing a total of 180 rankings to 50 different students, with some of these rankings coinciding to students they had previously nominated too.

With this information, it was possible to apply the CDC system (Terry and Coie, 1991), which only considers peer nominations and organize all students into the different sociometric status groups described in Chapter 3. Students were split into different five sociometric groups: popular, rejected, neglected, controversial and average. Students who did not fit any of these five groups' criteria were left unclassified - other. The results obtained are displayed below:

Popular	Controversial	Rejected	Neglected	Average	Other
12	12	3	19	25	43

Table 5.1: Sociometric status results of all nominated students(N=114)

Surprisingly, there were two students who were attributed to both Controversial and Average categories. While the Coie,Dodge & Coppotelli (CDC) formula makes it possible, there is no information in how it should be perceived.

5.3 Results

All gathered data regarding peer grading will be studied, firstly observing how students performed compared to each other, followed then by the observation of student peer grades correlation and agreement to the ideal case, the professor. For each of the following sections the data will be studied in terms of scores and feedback given, final grade obtained by the students, access to the training pages, participation in the skill tree, more specifically in the levels three and four. In the last section, the previous section results will be associated with the sociometric status groups computed as to find any answers to the research problem.

5.3.1 Peer Grading Between Students

To measure the agreement among student graders assigned to grade the same submission, the concordance, also known as inter-reliability, was calculated. Since the posts were assigned to five students chosen at random from a pool of students enrolled, the statistical model chosen to find the concordance agreement between students was the case 1 Intraclass Correlation Coefficient [ICC(1)].

The inter-rater agreement between the students randomly selected to rate the same post is defined as Single Measure. Its result of 0.567 suggests a significant agreement among students when peer grading the same post. The Average Measures results from the concordance of each student peer grade to a given post and the final peer grade score. The result obtained of 0.691 shows that there is a strong agreement between the peer grades given and the average final result, meaning the average final result does not originate from very differentiated peer grades.

For this study, the minimum number of students required to peer grade a post in order to consider its peer grading activity sufficiently, and consequently, display the results of both the peer grading activity and rating from the professor (except to students who could still peer grade the post) was of three students.

To understand how the number of students that peer graded the posts affects the mean score, and what should be the minimum number of students peer grading posts in order to achieve a sufficient level of agreement between them, the ICC was also used to compare peer grade scores from posts with different number of peer grades submitted. Once again the Single Measure represents the agreement among students grading those posts whereas the Average Measure represents the agreement of students' peer grading the post compared to the final average of the peer grade.

	5 graders	4 graders	3 graders	2 graders
Single Measure	0.444	0.391	0.106	-0.191
Average Measure	0.564	0.497	0.299	0.257

Table 5.2: Intraclass Correlation Coefficients (Case 1) for different number of student graders (N=68)

The results in Table 5.2. show that the level of agreement between students becomes reasonable when there is four peer graders scores given to a post. When looking at the Average Measure, that measures the concordance between students peer grades and the final mean peer grade, the coefficient resulting is higher and becomes moderate starting from the three peer graders. Looking at this data in more detail, it is possible to observe that in most cases, students agreed whether a submission should or not be acceptable, and the differences found in the scores given were of one point, mainly between score three or four and in less instances between the score four or five. In the few cases, in which students disagreed between the work presented being acceptable or not, having more than three peer graders helped the final grade become more decided.

In terms of the feedback provided by the students, the data obtained was analyzed with objective and subjective terms. Overall students had an average of 216 characters per peer grade. While it seems a reasonably decent feedback length value, it does not mean much by itself, thus to further observe and understand if the feedback was unique and diverse, and more likely to provide more insight to help students who submit defective works, a similarity algorithm was chosen to compare feedback given - Levenshtein distance similarity, an edit distance-based algorithm which takes two strings at a time and computes how many times does one string need to be changed in order to transform into other string. A smaller result indicates that two strings are very similar whereas a higher result means that it requires a lot of changes made in order to make one string to another. This algorithm was used in two situations: to check similarity in feedback overall and to find how unique feedbacks given by each student are.

Looking at each student feedback, as to understand how much the feedback given to each post differs, the Levenshtein algorithm was run to compare strings from the same student, meaning students who only did one peer grade were not considered for this segment. After computing the results for each set of scores, all values were averaged and grouped between uniform ranges. In most cases (17.7%), a student scores a similarity value in the range [108;158], followed by [58;108], in 15.6% of the time and [158;208] at 10.4% of the cases, meaning that there is some similarity between feedback made by the same student. When looking into more detail, these results can be explained due to students, especially in cases where they award higher grades, using the same or very similar expressions in their feedback, such as, ”Good/Great/Nice job!”, ”I really liked/had fun watching your submission”, etc. It was also retrieved the average of each student’s rates as to understand if there is some correlation between students’ grade and the length of the feedback given.

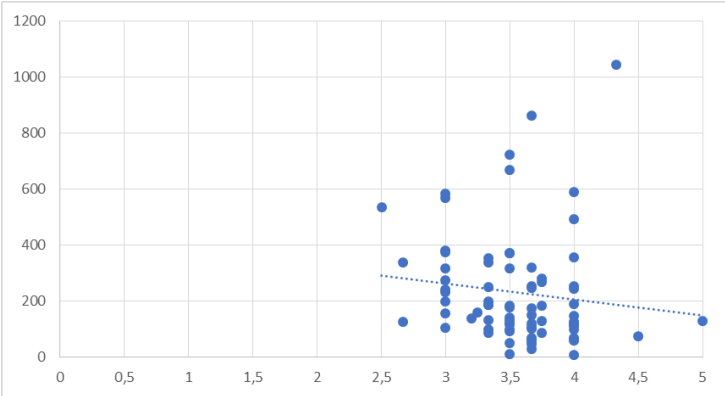


Figure 5.1: Relation between students' feedback length and average rate given

Figure 5.1 shows a tendency for students to give longer feedback when they rate students works lower, which could indicate that feedback in lower quality submissions may include useful criticism and recommendations to the graded students.

Running the algorithm between all existing pair of feedbacks available (expect with itself), provided the results displayed in figure 5.2. Most of the feedbacks given had a similarity score between the ranges [70;140] followed by [140;210] and [210;280]. Considering the results found previously that showed lower similarity scores when comparing for each student all its feedback entries, it looks like feedback given by the different students is considerably distinct, making it more useful to graded students if they decide to make future improvements to their work based on the initial reception of their work.

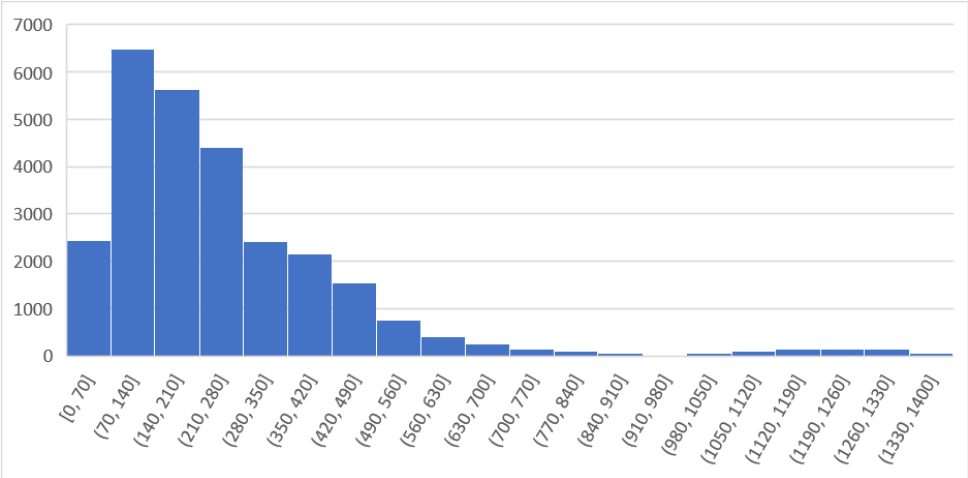


Figure 5.2: Overall student feedback similarity

To understand if the content of the feedback given had any value to the students graded, using more direct observation, each peer grade was evaluated according to three criteria: (a) peer grader feed back matched the rate given, (b) pointed out clearly flaws found in the submission, if there were any,

and (c) included suggestions on how to improve the submission, if needed. In 92.31% of the time the feedback matched the grade given. In the 18 cases which the grade did not match the feedback, in 14 situations the grade given by the students was higher than the professor's, and while in almost all cases the difference was of only one point, there were some situations in which the difference meant passing or not in the skill. In the remaining four situations, in two occasions the grade ended up matching the professor's and the other two remaining cases the grade was lower than the professor's grade, once again by one point of difference only and all cases did not affect the students' passing of the skill.

Looking more in depth at these cases respectively, it was noticeable that either students would point out major flaws, but not realize what is more important for a given submission to achieve in order to be acceptable, or would provide simple and mainly congratulatory feedback for a submission that they rated relatively low.

Regarding the other two criteria evaluated, all cases where the grade was four or five were excluded, as in these cases is not necessary or sometimes does not make sense to provide suggestions or point out any problems (as there should not be any or at least they should be minimal). Therefore, for the 129 cases considered, 108 contained feedback that pointed problems in the submission in a clear manner and 49 included suggestions to possible changes that could be made. Only in 5 cases the feedback contained neither of these. Nevertheless, it is noticeable that students can produce quality feedback, even if relatively short and that students produce feedback substantially different from each other's, which can benefit the graded student.

To help students providing better peer grades, the plugin included training pages which included a description of the type of work that is expected, which is also mentioned in the skill node of each skill, the respective requirements each submission needed to have in order to be accepted and some examples of various graded submissions, together with an explanation of the grade given. For a given post that was assigned, the respective training page was made accessible through the link "Check out some examples" on the top of the post to be peer graded. 31 students checked these pages and nine students also looked through the pages for topics that they did not peer grade but submitted assignments. This indicates that students who first checked the training pages for peer grading purposes have found them useful beyond just help in peer grading.

When looking more in depth to the 39 peer grades given by students to posts in which the topic it belonged to match the pages they checked, there was an agreement between professors and students in 51% of time. In the remaining 49%, in most of the times the disagreement of grade between student and professor came from the fact that students did not verify if the submission contained all of the additional required contents (script of the work, detailed explanation of the process, etc.), as the grades ended up matching the professor after these submissions corrections were made. When calculating the ICC case 1 to understand how in agreement the professor and students were, the resulting value of

0.067 indicates that there is almost no concordance between the two, though the observation made to the scores previously shows that this is not the case. However, in 75% of the cases the grade given by these students was in consensus with the final average peer grade and its respective ICC value of 0.659 confirms the concordance between the student who peer graded with previous training and the final score.

This goes in according with previous findings that students show a tendency to have better agreement with each other than the professor. Finally, there were seven instances in which there was two peer graders which had checked the training page corresponding to the topic which they had to peer grade in the same post. In these cases, in less than half the cases (three), there was consensus between the two. While these results are not particularly impressive, it is worth mentioning that it is not possible to know whether these students actually read carefully these pages or just opened those and skimmed through them.

To understand if checking the training pages may have had any impact in the student's submission, the rates received between students who did check the training pages before submitting work in the skill tree with the students who did not were compared. The results are displayed in Table 5.3. and indicate that students who checked training display a slight increase in the overall average rates obtained from the professor compared to the students who did not checked the training pages. This seems to indicate that students who checked training have better results in their submissions. This could be due to having access to previous submissions, and not only have a better understanding of what is required from them, but also having concrete examples of what they should do and not do.

Comparisons	Mean	Std. Deviation	Error Std. Deviation
Students not Trained	3.591	0.634	0.054
Students Trained	3.657	0.630	0.106

Table 5.3: Comparison between students who did check training and students who did not check training

5.3.2 Peer Grading Between Students and Professor

Through inter-rater agreement, the results showed that from the 96 posts graded by both professor and student(s), 55 peer grades matched the professor rates (57%), thus there was a medium agreement between professor and students. On the remaining 41 posts, 15 had a lower grade than the professor and 16 had a higher grade than the professor, in which there were six exceptional cases where the difference was of two points, with the professor giving a rate of two and the students a rate of four. In all these cases, the students give a wrong scoring to the posts as they did not verify if all the requirements for a submission to be approved were met and only graded the submission according to the work presented by the student.

When comparing each grade given by a student to the professor grade, the percentage of peer

grades that match the professor decreases to 53%, a good indicator that having more than one student peer grade the post can lead to a more accurate peer grade. In most cases, the grades given by each student to the assignments tend to convergence with the professor the more they participate; this could be due to the student gaining more experience from peer grading.

Comparing the overall mean scores given by professors and students, there seems to be a similar value between the two, students seem to be more conservative in their results, awarding mostly scores of three and four, and give slightly higher scores than the professor.

Comparisons	Mean	Std. Deviation	Error Std. Deviation
Peer Grades	3.595	0.469	0.048
professor Grades	3.448	0.720	0.073

Table 5.4: Comparison between professor rates and peer grades (N=96)

For this study the ICC was calculated again to understand how the student peer grades and professor rate correlates. When comparing each peer grade given with the corresponding rate for the same post, the coefficient value obtained was of 0.339, while the coefficient value between the professor rate and the average of students peer grades was of 0.204, which indicates that there are some significant differences between the professors and the students and that having multiple students evaluating actually seems to accentuate the differences. The results together with the previous findings seem to indicate that students are much more in agreement with each other than with the professor.

Finally, to understand the validity of peer grading, that is, assuming the professor grade is the right grade, it was studied the similarity between the professor rate and peer grading final scores. This study used the convergent validity through the Pearson product-moment correlation coefficient (r). When applying the statistical model, the result of 0.237 was achieved. The small value confirms that the students and professor tend to grade considerably different scores, though this result could be affected by the fact that the scale used for peer grading is considerably small (1-5), making the disagreement of one value of difference impact highly the validity results. Most of the disagreement cases between students and professor were of one point difference or less, and both sides tend to agree on whether the submission is acceptable to pass the skill or not. The following box plot shows the discrepancy between students' peer grades and professor score according to the 5-point scale, displaying the differences in ascending order of the scale. The cases in which the professor rated the submission three and four were the ones which had high consensus between the two parts as expected. The box for the first series is the largest mainly due to the cases in which the professor rated the submission two as not all requirements were present, but students did not have this detail in consideration when peer grading.

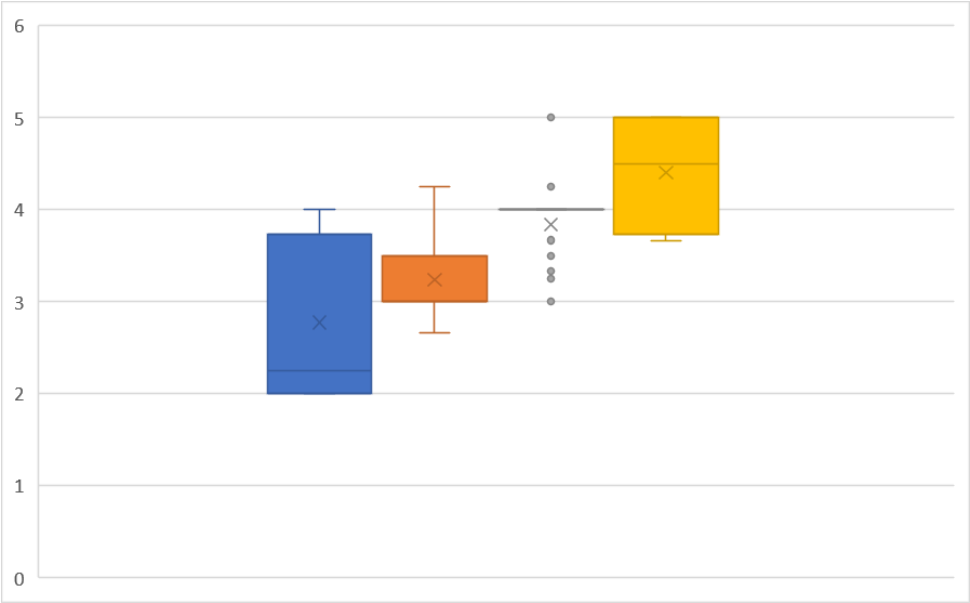


Figure 5.3: Differences in peer grading across the 5-point scale

As to understand better the similarity results obtained between the professor rates and students final peer grade, a more detailed analysis, covering different aspects of the course, was done in the following paragraphs.

Looking at students' overall participation in the skill tree activity, it was found that students with the highest number of skills achieved in the skill tree usually give the same grade as the professor, which could be due to having accumulate more experience in how these type of submissions should be graded. Students who did a relatively low number of submissions have more inconsistent peer grades. The correlation value obtained between students' participation in the skill tree and matching of peer grades given was of 0.515, corresponding to moderate strength, indicating that participating in the skill tree and being familiar with the grading process helps achieving better peer grades.

When looking at students' participation in the levels three and four of the skill tree only, the more skills students achieved, the higher occasions occur in which the student peer graded the same score as the professor. The correlation between students peer grades and professor rates was of 0.781, confirming that higher participation in the skill tree helps peer grade validity.

After grouping students according to their leaderboard level the results found were similar to the ones already found when grouping students by participation in the skill tree activity, which makes sense as students with more skills achieved have earned more XP from this segment of the course. Students which finished the course with the higher grades (19 and 20) had their grades match with the professors the most, in 61.95% of the cases and students which had the lowest match of grades with the professors were students with a final grade in the course either in the eleventh or twelfth level. The correlation

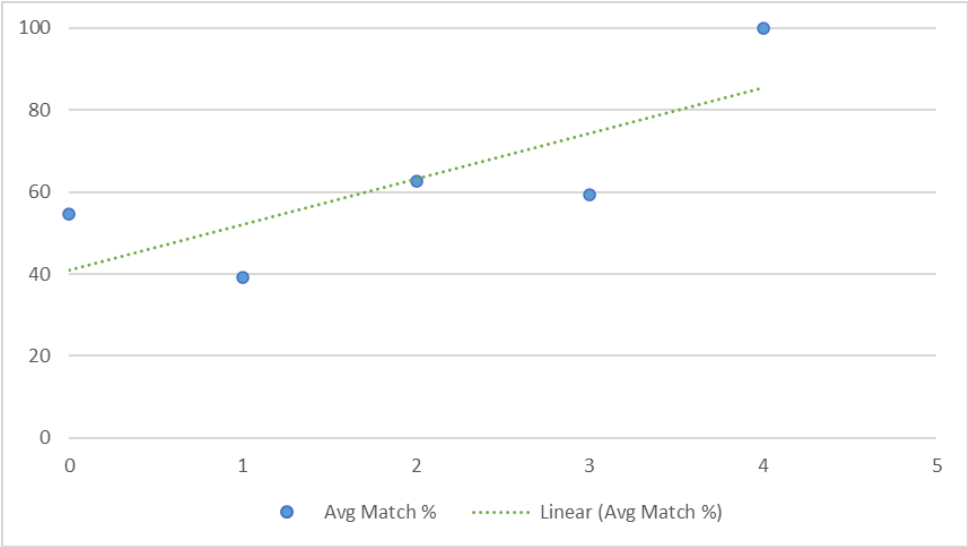


Figure 5.4: Relation between the number of skills of level 3 and 4 achieved and percentage of agreement between student and professors in grading

value obtained between final grades obtained and the matching of peer grading from students was of 0.377. This result is rather low, compared to the value previously obtained through the skill tree participation since the skill tree only accounts for one fourth of the grade, meaning students can still attain a reasonably high grade without too much participation or knowledge of the skill tree.

In the interval of level 13-14 and level 10 or less, there was an substantial increase in matching between student and professors rate in relation to the respective upper levels. Looking into more detail to the data, while number of students in the levels 10 or lower where were only three, and therefore the average can be easily swayed, in the level 13-14, the number of students was similar to other levels ranges though there is not enough data to find any particular reason as to why the level of agreement between student-professor has increased.

5.3.3 Peer grading and Sociometric Status Groups

In order to answer to the research problem, that is, understand if social relationships have an impact in student's ability to provide valid assessments, the sociometric status results previously obtained in 5.2.3 can be used to observe how each resulting group performed overall in the course, in the peer grading activities as peer graders and if students positioned in any of the groups have biased grades. It is also studied if students who had nominated peers which they later peer graded, display any bias towards them. Rankings are also used to find any inconsistencies.

Looking at the data in more detail in terms of peer nominations, it was found that 20 peer grades were given to students by peer graders which also had been nominated by them. These 20 peer graded students had previously received 10 positive nominations from their grader and other 10 had received

a negative nomination. In 13 of those cases, the grades given by the students matched the professor grade. Of seven remaining cases, all of them were different by a point difference. Only in two situations, the discrepancy in professor and student scores came from instances in which a negative nomination was given beforehand.

The other five cases came from students which had given a positive nomination to their colleague and in three of them, the students gave a higher grade than the professor. This indicates that it does not look like social relationships have an impact in student's capacity to peer grade a colleague fairly, at least in situations where their identity is hidden.

Using the ICC method again, it was calculated the coefficient of concordance between professors and students who peer graded posts of students which they peer nominated previously. The results are displayed in the following table.

Overall Nominations	Positive Nominations	Negative Nominations
0.617	-0.053	0.802

Table 5.5: Comparison of ICC coefficient between different nominations previously received by students

Observing the results, it is noticeable that there is a moderate agreement between students and professors in the submissions score given, especially when students gave a negative nomination to the student that they were peer grading, which were mostly identical, but the opposite case happened in cases where students had given a positive nomination beforehand. While this results may not convey it, there was actually some agreement between professor and student when looking at positive nominations, with almost half of the cases observing actually having matching scores and none of the cases in which the grade was different from the professor, affecting the acceptance or not of the submission.

In terms of peer rankings given, from the 180 answers provided by the students, only 86 (47.8%) were rankings form one to five, the 94 remaining being zero rankings and therefore not many conclusions can be gathered from them. In the rankings from one to five, observing the results, we found almost identical scores between professor and students, and in the few cases where there was not, the student usually gave a higher score than the professor, even if after ranked the student lowly.

Using the groupings of students by sociometric status and averaging their final grades in the course, it was found that the Controversial students are definitely the better students, with an average final grade of 17.579, followed by the Popular students who had an average of 15.833. Average and unclassified students had a similar average final grade of 15.364 and 15.057 respectively. The worst performers in the course were the Rejected and Neglected students, with average grading of 13 and 12.5.

	Popular	Rejected	Controversial	Neglected	Average	Other
Mean	15.364	17.579	12.5	15.057	15.833	13
Std. Deviation	2.789	2.347	1.5	2.828	1.951	2.345
Error Std. Deviation	0.595	0.538	1.061	0.478	0.563	0.829

Table 5.6: Comparison between sociometric status groups average final grade (N=114)

Even though controversial students are the best students in the course, they are the sociometric group which the peer grades agrees the least with the professor rate. More than half of the students either deflated or inflated at least one grade, with most cases being a deflated grade, and 32% of them did both.

Rejected students only have cases of grade inflation compared to the professor and most students had multiple cases of inflation when peer grading. These results go in according to the research work done, mentioned in Chapter 2, in which it had been noticed that students with higher grades have a tendency to deflate their peers' submissions whereas worst performing students, such as the case of rejected students, do have a tendency to inflate grades. Neglected students were the best group, however, only two students were classified as such, therefore, its group population is too small to make any reliable conclusions.

To understand better these results, a more in depth observation of data is made, firstly by observing how each group of student acted when peer grading their colleagues and then observing if there any sociometric group that tends to receive biased peer grades.

To examine if students belonging to a certain sociometric status were displaying some bias towards students from a certain sociometric group when grading (whether these group of students was its own or another) or to students in general, a new sociometric status table was formed, including only the 97 peer graders. The results are displayed bellow.

Popular	Rejected	Controversial	Neglected	Average	Other
12	8	2	19	22	35

Table 5.7: Sociometric status of students who peer graded

The popular and controversial groups were the only ones that maintained the same number of students in the respective groups. This is possibly due to students from these groups being mainly students which ended in the upper levels of the leaderboard and have the highest participation levels throughout the various sections of the course.

Using the sociometric table previously mentioned, each group and their respective students were analyzed in terms of peer grades provided. All groups, with the exception of the neglected group, which also only had two students, had similar matching performances between professor and students, ranging from 49.05% (controversial students) to the highest, 54,31% (popular students). Using the ICC coefficient to compare how agreement between professor and students from the different groups changed, the results found are displayed in Table 5.8.

Popular	Rejected	Controversial	Neglected	Average	Other
0.32	0.246	0.311	1	0.287	0.342

Table 5.8: Comparison of ICC coefficient between professor rates and students peer grades

The results display low concordance with the professor, as expected from the previous results in

3.5.2, though is possible to see that the difference between the groups is very little meaning that the sociometric status group in which students are insert into does not seem not affect their reliability of peer grading. Neglected students were in complete agreement with the professor, however they are a group consisting of two students each only peer graded once. The sociometric status from the 55 evaluated students, resulted in the following division:

Popular	Rejected	Controversial	Neglected	Average	Other
8	2	14	0	12	21

Table 5.9: Sociometric status of peer graded students

In this case, no sociometric group maintained the same number of members to the original number. To see if the placement in these groups could have an influence in the grading that the students received, it was again observed how the students in each groups had received peer grades that matched the professor and when it did not, if any of these groups had a tendency to have more inflated or deflated grades.

When looking at each group, the students' scores received show no indication of bias towards students of one certain group, with each group receiving a similar proportion of peer grades from two to four. The sociometric group which had the best peer grades, including several scores of five were the Rejected and the groups of Popular, Controversial and Average received identical amount of peer grades scored two and similar number of the remaining scores.

5.4 Discussion

The results presented after studying and analyzing the data regarding peer grading reveal that students can produce moderately consistent peer grades in a gamified blended environment, especially if the number of students is of four or more. It was found some cases in which the grade given by the student was different from the professor, however looking into these cases in more detail it was possible to see that in most cases, the difference was of one point only, and there was a general agreement that the submission was acceptable or not. In the remaining cases in which there was a one or two point difference, and that differentiated a student passing or not the skill, it was understandable by reading the feedback provided that that these situations happened due to wrong assessment from the student, for not verifying if all submission required materials were presented, and had no connection with students' capabilities of peer grading or social connections. Students also have a significant tendency to agree more with each other than with the professor. The students who checked training seemed to agree more with the professor, a good indicator that more training, and including all students, can lead to better peer grading results.

In the situations were a student peer graded a colleague which had been nominated in advance, it

was found that there was some disagreement between professor and students and in most cases, it was due to the professor awarding the student peer graded with a higher grade, once again usually by one point of difference only and in situations that did not mean disagreement between the peer passing or not passing the skill, therefore no relevant bias was found. In all cases analyzed, when the nomination given before was a negative one, there was almost complete agreement between students and professor regarding the scores given, so it is possible to conclude that there is not any sign of significant bias in these situations.

When looking at the different sociometric status groups formed through the CDC system, it was possible to notice that students belonging to the Controversial group were the better students overall, however, they also had the highest and most frequent situations of disagreement with the professor, in most cases assigning a lower grade than the professor. Rejected students displayed the exact opposite behavior, by having the worst performance in the course and showing a tendency to inflate grades, which goes in accord with previous findings that the better students deflate grades, as they are more demanding, whereas they the worse students tend to inflate the grade, possibly due to not having as much knowledge in the course materials as they should in order to better perform as student and grader. However, when submitting work, the Rejected students were the ones with the highest peer grades received and the remaining groups received a similar number of peer grade scores from across the scale.

6

Conclusion

Contents

Peer grading in blended learning environments has shown in previous research that can be a reliable form of grading and help solve scalability problems that may arise from this type of learning environment. In spite of the promising results, it has not been studied if the use of students which have developed social connections between each other may result in biased peer grades, which could lead to unfairness in students' evaluations. To understand this, a plugin was developed for Moodle which would bring peer grading capacities to the platform as well as several methods for retrieving information regarding students' connections. The plugin was deployed to the Multimedia Content Production course and the data collected revealed that students tend to have high agreement rates with each other, especially when three or more students peer grade a post, however, the results of agreement between teacher and student indicated that there was some disagreement between professor and student, though mostly of one-point difference in a five-point scale, and in most cases, both professor and student agreed that the work presented was good enough to be accepted. Students with training, appear to be more in accordance with the professor.

Looking at the results of peer nominations, no significant bias was found when observing the cases in which the students peer graded peers each they had nominated in advance, neither in cases where they afterwards peer ranked them between 1-5. When looking at peer grade results by dividing students in different sociometric status groups, it was also found that the agreement levels of peer grades and professor rates was similar throughout all groups, meaning no group seems to provide more biased grades. Finally, when observing the peer grades received by students from the different sociometric groups it was also not found any group suffering from biased peer grading, therefore concluding that the sociometric status of students do not lead to biased grading and students can be considered a reliable source of grading, especially if more practice to the task is done beforehand.

6.1 Future Work

Although the Peerforum plugin is now in a stable form, there are some aspects that can be improved regarding plugin performance as the system is still considerably slow. In terms of performance, the PeerForum can still be further improved to reduce load times by trying to include less elements in first loading of the discussion pages, which already contain numerous different content to be loaded. The results showed that training does help students becomes better peer graders and achieve better scores in their submissions, therefore training pages could become mandatory in the beginning of the course and include some exercises to make sure that there is some effort made when students consult those pages.

In terms of the study itself, having the PeerForum plugin working since the begining of the semester can lead to more students participating and more chances for students to improve their peer grading

capacities and increase their level of concordance with the professor, someone who has extensive training in the same type of work. The use of a increased scale can also help to decrease the high level of apparent difference in reuslts of peer grading versus the professor rating.

Bibliography

[1] L. Alves, D. M. V. Barros, and A. Okada, “Moodle: estratégias pedagógicas e estudos de caso,” 2009.

[2] P. Arabasz and M. B. Baker, “Evolving campus support models for e-learning courses,” *Educause Center for Applied Research Bulletin*, vol. 1, no. 9, p. 9, 2003.

[3] B. J. Arnold, “Gamification in education,” *ASBBS Proceedings*, vol. 21, no. 1, p. 32, 2014.

[4] B. Barna and S. Fodor, “An empirical study on the use of gamification on it courses at higher education,” in *International Conference on Interactive Collaborative Learning*. Springer, 2017, pp. 684–692.

[5] C. J. Bonk and C. R. Graham, *The handbook of blended learning: Global perspectives, local designs*. John Wiley & Sons, 2012.

[6] M. T. Chi, “Constructing self-explanations and scaffolded explanations in tutoring,” *Applied Cognitive Psychology*, vol. 10, no. 7, pp. 33–49, 1996.

[7] S.-W. Chou and C.-H. Liu, “Learning effectiveness in a web-based virtual learning environment: a learner control perspective,” *Journal of computer assisted learning*, vol. 21, no. 1, pp. 65–76, 2005.

[8] J. D. Coie, K. A. Dodge, and H. Coppotelli, “Dimensions and types of social status: A cross-age perspective.” *Developmental psychology*, vol. 18, no. 4, p. 557, 1982.

[9] J. Cole and H. Foster, *Using Moodle: Teaching with the popular open source course management system*. ” O’Reilly Media, Inc.”, 2007.

[10] C. Costa, H. Alvelos, and L. Teixeira, “The use of moodle e-learning platform: a study in a portuguese university,” *Procedia Technology*, vol. 5, pp. 334–343, 2012.

[11] C. Piech, J. Huang, Z. Chen, C. Do, A. Ng, and D. Koller, “Tuned models of peer assessment in moocs,” *arXiv preprint arXiv:1307.2579*, 2013.

[12] P. Black, B. Fairbrother, and P. Gill, *Teachers assessing pupils: Lessons from science classrooms*. Association for Science Education, 1995.

[13] N. Falchikov, *Learning together: Peer tutoring in higher education*. Psychology Press, 2001.

[14] S. Freeman and J. W. Parks, "How accurate is peer grading?" *CBE—Life Sciences Education*, vol. 9, no. 4, pp. 482–488, 2010.

[15] D. R. Garrison and N. D. Vaughan, *Blended learning in higher education: Framework, principles, and guidelines*. John Wiley & Sons, 2008.

[16] E. F. Gehringer, "Electronic peer review and peer grading in computer-science courses," *ACM SIGCSE Bulletin*, vol. 33, no. 1, pp. 139–143, 2001.

[17] ——, "Peer grading over the web: enhancing education in design courses," *age*, vol. 4, p. 1, 1999.

[18] S. D. Gest, S. A. Graham-Bermann, and W. W. Hartup, "Peer experience: Common and unique features of number of friendships, social network centrality, and sociometric status," *Social development*, vol. 10, no. 1, pp. 23–40, 2001.

[19] J. Golbeck and J. Hendler, "Accuracy of metrics for inferring trust and reputation in semantic web-based social networks," in *International conference on knowledge engineering and knowledge management*. Springer, 2004, pp. 116–131.

[20] G. Kiryakova, N. Angelova, and L. Yordanova, "Gamification in education." Proceedings of 9th International Balkan Education and Science Conference, 2014.

[21] S. E. Meek, L. Blakemore, and L. Marks, "Is peer review an appropriate form of assessment in a mooc? student participation and performance in formative peer review," *Assessment & Evaluation in Higher Education*, vol. 42, no. 6, pp. 1000–1013, 2017.

[22] E. A. O'Connor and A. McQuigge, "Exploring badging for peer review, extended learning and evaluation, and reflective/critical feedback within an online graduate course," *Journal of Educational Technology Systems*, vol. 42, no. 2, pp. 87–105, 2013.

[23] R. O'Leary and A. Ramsden, "Virtual learning environments," *Learning and Teaching Support Network Generic Centre/ALT Guides, LTSN. Retrieved July*, vol. 12, p. 2005, 2002.

[24] M. M. Patchan, C. D. Schunn, and R. J. Clark, "Accountability in peer assessment: examining the effects of reviewing grades on peer ratings and peer feedback," *Studies in Higher Education*, vol. 43, no. 12, pp. 2263–2278, 2018.

[25] G. Piccoli, R. Ahmad, and B. Ives, “Web-based virtual learning environments: A research framework and a preliminary assessment of effectiveness in basic it skills training,” *MIS quarterly*, pp. 401–426, 2001.

[26] P. M. Sadler and E. Good, “The impact of self-and peer-grading on student learning,” *Educational assessment*, vol. 11, no. 1, pp. 1–31, 2006.

[27] M. Stiles, “Effective learning and the virtual learning environment,” in *Proceedings: EUNIS 2000–Towards Virtual Universities, Instytut Informatyki Politechniki Poznanskiej*. Citeseer, 2000.

[28] S. Taylor, M. Ryan, and J. Pearce, “Enhanced student learning in accounting utilising web-based technology, peer-review feedback and reflective practices: a learning community approach to assessment,” *Higher Education Research & Development*, vol. 34, no. 6, pp. 1251–1269, 2015.

[29] R. Terry and J. D. Coie, “A comparison of methods for defining sociometric status among children.” *Developmental Psychology*, vol. 27, no. 5, p. 867, 1991.

[30] K. J. Topping, “Peer assessment,” *Theory into practice*, vol. 48, no. 1, pp. 20–27, 2009.

[31] S.-C. Tseng and C.-C. Tsai, “On-line peer assessment and the role of the peer feedback: A study of high school computer course,” *Computers & Education*, vol. 49, no. 4, pp. 1161–1174, 2007.

[32] M. Rostampoor-Vajari, “What is sociometry and how we can apply it in our life,” *Advances in Asian Social Science*, vol. 2, no. 4, pp. 570–573, 2012.

[33] J. L. Moreno, *Sociometry, experimental method and the science of society*. Lulu. com, 1951.

[34] J. Ribeiro, “Peer grading in gamified blended learning environments,” 2016.

[35] L. Johnson, S. Adams Becker, M. Cummins, A. Freeman, D. Ifenthaler, and N. Vardaxis, *Technology Outlook for Australian Tertiary Education 2013-2018: An NMC Horizon Project Regional Analysis.* ERIC, 2013.

A

PeerForum Plugin Database Tables

The tables created and updated in the database of the PeerForum plug-in are displayed bellow. For the tables that were updated, only the new content and its structure is represented.

Column	Type	Null	Default
id	longint(10)	Yes	NULL
iduser	bigint(10)	No	-
courseid	bigint(10)	Yes	NULL
peersfav	longtext	No	-
peersunfav	longtext	No	-
studentsranked	longtext	Yes	NULL

Table A.1: Structure of table mdl_peerforum_relationships

Column	Type	Null	Default
topics	lonttext	Yes	NULL
peergradetype	smallint(5)	No	2
gradesum	bigint(10)	No	0

Table A.2: New fields of table mdl_peerforum_users

Column	Type	Null	Default
threaded_grading	bigint(10)	No	0
autoassignreplies	bigint(10)	No	1
hidereplies	bigint(10)	No	1
peernominations	bigint(10)	No	1
peerrankings	bigint(10)	No	1
peernominationsfields	bigint(10)	No	4
peernominationsaddfields	bigint(10)	No	1
training	bigint(10)	No	1
random_distribution	bigint(10)	No	0

Table A.3: New fields of table mdl_peerforum

Column	Type	Null	Default
type	bigint(10)	No	2
idlink	bigint(10)	Yes	NULL

Table A.4: New fields of table mdl_peerforum_discussions

B

PeerForum Plguin Settings Description

The new local settings for the PeerForum plugin are displayed in Figure B.1,B.2 and B.3. Its decription is included bellow.

B.1 Peer Grade Configurations

- **Auto assign replies:** Allows to automatically assign the same peer graders to a post that is a reply from a previous post of the same student

- **Hide replies to student posts:** Allows for the professor replys to be hidden while peer grading has not ended.

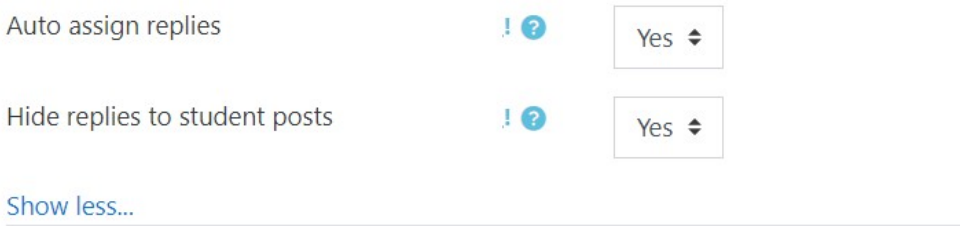


Figure B.1: PeerForum plugin local interface - Peer Grade configurations

B.2 Topic Attribution Configurations

- **Enable advanced topic attribution:** Select to use more complex options of assigning student to posts
- **Random Distribution:** Select if the distribution of students should be uniform throughout all available submission topics
- **Discussion topics:** Select topic(s) to later assign them a type of peer grading. The available options are:

  - Specified Topic: If selected, this topic will only be available for peer grading for a certain group of students;

  - Random Topic: All students not allocated to a certain specific topic;
- **Type of peer grading:** Select the type of peer grading the topic(s) previously selected will have
- **Students assigned to discussion topic:** If Specified Topic was selected, decide how many students to be allocated to each topic previously selected. The scale id goes from the number of students assigned to grade a post plus one to 30.

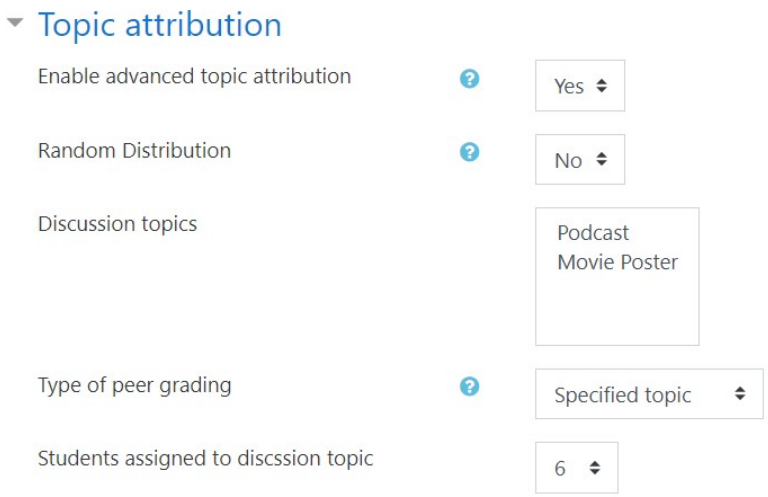


Figure B.2: PeerForum plugin local interface - Topic Attribution configurations

B.3 Miscellaneous Configurations

- **Enable peer rankings:** Select this option to activate peer rankings
- **Enable peer nominations:** Select this option to activate peer nominations
- **Number of peer nominations:** If 'Enable Peer Nominations' is selected, choose how many students should be nominated positively and negatively.
- **Allow students to give more nominations:** If 'Enable Peer Nominations' is selected, allow for students to add more nominations fields
- **Enable peer grading training:** If selected, grants access to training pages in the PeerBlock and page of discussion topics.

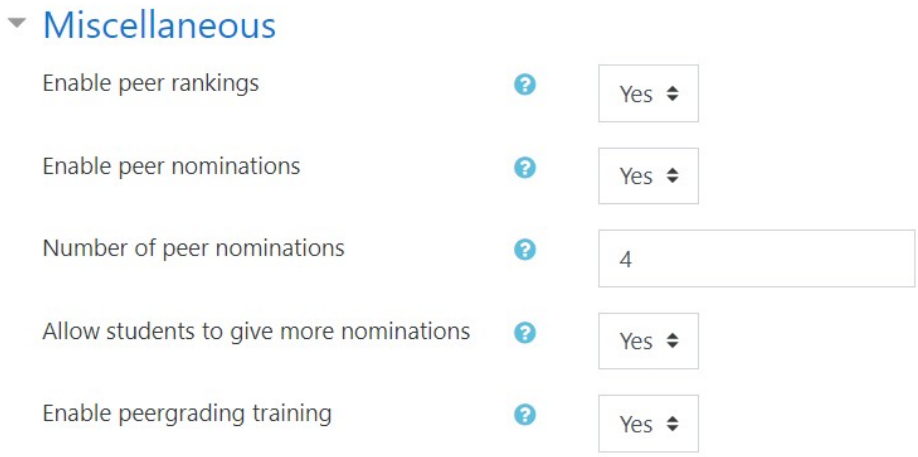


Figure B.3: PeerForum plugin local interface - Miscellaneous configurations