Studying how social relationships affect peer assessment in an E-learning environment

Francisco Sousa[1] · Tomás Alves[1,2] · Sandra Gama[1,2] · Joaquim Jorge[1,2] · Daniel Gonçalves[1,2]

Abstract

Peer assessment has been widely studied as a replacement for traditional evaluation, not only by reducing the professor's workload but mainly by benefiting students' engagement and learning. Although several works successfully validate its accuracy and fairness, more research must be done on how students' pre-existing social relationships affect the grades they give their peers in an e-learning course. We developed a Moodle plugin to provide the platform with peer assessment capabilities in forums and used it on an MSc course. The plugin curated the reviewer set for a post based on the author's relationships and included rubrics to counter the possible interpersonal effects of peer assessment. Results confirm that peer assessment is reliable and accurate for works with at least three peer assessments, although students' grades are slightly higher. The impact of social relationships is noticeable when students who do not like another colleague peer grade their work consistently lower than students with whom they have a positive connection. However, this has little influence on the final aggregate peer grade. Our findings show that peer assessment can replace traditional evaluation in an e-learning environment where students are familiar with each other.

✉ Francisco Sousa
francisco.sousa.98@tecnico.ulisboa.pt

Tomás Alves
tomas.alves@tecnico.ulisboa.pt

Sandra Gama
sandra.gama@tecnico.ulisboa.pt

Joaquim Jorge
jorgej@tecnico.ulisboa.pt

Daniel Gonçalves
daniel.j.goncalves@tecnico.ulisboa.pt

[1] Instituto Superior Técnico, University of Lisbon, Av. Rovisco Pais 1, Lisboa 1049-001, Portugal

[2] INESC-ID, R. Alves Redol 9, Lisboa 1000-029, Portugal

Introduction

Peer assessment has gained followers as an alternative model to traditional student evaluation. Peer assessment consists of students with similar backgrounds judging each other's work (Topping, 1998; Na and Liu, 2019), promoting reflection and discovery of new understandings by finding the difference between others and themselves (Chang et al., 2020). Recent meta-analyses showcase the significance of such mechanisms in teaching and learning nowadays (Zheng et al., 2020; Li et al., 2020; Yan et al., 2022) by improving students' academic performance (Black and Wiliam, 2009; Yan et al., 2022) through pedagogical activities that facilitate learning (Chie Adachi and Dawson, 2018; Double et al., 2020). Previous work has shown that peer assessment can obtain the same accuracy and fairness of grades one would get from the professor (Topping, 2009). Due to their scalability, e-learning courses popularized peer grading, especially in Massive Open Online Courses (MOOC). There, the traditional assessment approach, which consists of manual grading, would be unfeasible and costly for professor's to complete within a reasonable period, particularly in settings with many students and assignments. However, recent literature reviews point toward a large research gap in the field. Few past studies focused on understanding how social factors may explain the variance observed in peer assessment environments (Dilrukshi Gamage and Whiting, 2021; Panadero et al., 2023).

In most e-learning courses, students usually do not know each other. However, virtualizing traditional learning environments due to the COVID-19 pandemic (Dhawan, 2020) brought peer assessment to e-learning environments where students interact outside of classes and are familiar with each other from previous years. Researchers must consider a peer assessment bias related to student relationships in this student familiarity scenario. Furthermore, peer assessment is a social activity requiring a mutual trust relationship (Panadero, 2016). To minimize the effect of social relationships between students, researchers argued that anonymizing uni- or bidirectionally the parties could improve the peer assessment (Li, 2017). A recent review shows mixed results of using anonymity to minimize interpersonal effects in peer assessment environments (Panadero and Alqassab, 2019). For instance, Lin (2018) shows that anonymity positively impacts providing more critical peer feedback and fosters different types of peer feedback compared to students in non-anonymous conditions. Research also shows that anonymity provides more comfort and less peer pressure to the students (Raes et al., 2015; Vanderhoven et al., 2015; Seifert and Feliks, 2019). In contrast, the meta-analysis by Panadero and Alqassab (2019) and Li et al. (2016) suggests that non-anonymity is better for increasing students' peer grading accuracy when compared to teachers' assessment. These mixed results highlight the complexity of factors that are present in peer assessment.

The current state of the art led us to reconsider whether concealing the identity of peers is always beneficial to control interpersonal effects in peer assessment. In particular, we believe that providing peer assessment aids such as criteria, rubrics, and training may help alleviate the interpersonal effects. Other researchers have already considered social relationships through the nomination of intimate friends (Azarnoosh, 2013) or prediction of relationships based on student interactions (He et al., 2015) but did not find solid and reliable results. However, to the best of our knowledge, there is insufficient empirical evidence on whether using peer assessment aids helps minimize the interpersonal effects of peer

assessment. In this work, we aim to understand how the social relationships between students affect the quality and students' perceptions of the peer assessment environment in an e-learning course. We collected multiple self-reported relationships through peer nominations, peer ratings throughout the semester, and students' perceptions of the peer assessment process. Then, we analyzed whether it is possible to overcome the relationship bias in the peer assessment environment, considering that faculty provided rubrics to train the students' grading and where most students were previously familiar with each other.

The remainder of this paper is organized as follows: First, we discuss the related work on peer assessment. Then, we present our research methodology, including the peer-grading environment, data collection, and processing. Next, we discuss how we analyzed the data, the results obtained, and the implications for designing future experiences. Finally, we present our conclusions and pointers for future work.

Related work

According to Topping (1998), peer assessment—or peer grading—consists of students with similar backgrounds judging each other's work, including the number, level, value, practicality, quality, success, and the result of their daily study, with substantial evidence that it can improve the effectiveness and quality of learning (Li et al., 2020). The most common method to determine the quality of the peer assessment is by measuring the differences between grades provided by students and faculty members (AlFallay, 2004; Kulkarni et al., 2013; Yan et al., 2022). For instance, we can compare the grade given to exercises by professor's with that given by peers to infer if peer assessment is *accurate*—high correlation between student- and professor-assigned scores—in awarding the same score, and *fair*—consistency of scores given by multiple student graders—to all the students involved. Most literature agrees that peer assessment is sufficiently similar to a professor's evaluation and, therefore, accurate to use with students, concluding there is no significant difference in who is grading (Azarnoosh, 2013; Luo et al., 2014; Usher and Barak, 2018; He et al., 2015). By comparing the influence of different numbers of graders in peer assessment fairness, Luo et al. (2014) found that more is best, recommending between three and five graders. Cho et al. (2006) suggest four to six graders. Other researchers decided to tackle the subjectiveness bias by applying other strategies such as anonymity (Bostock, 2000; Lin, 2018; Hoang Phuoc et al., 2022).

Anonymity in peer assessments

Although primarily used in e-learning courses, like MOOC, peer assessment is progressively growing in traditional classes and blended learning courses, the integration of face-to-face with both online instruction (Graham, 2013) and online learning experiences (Garrison and Kanuka, 2004). These are composed of students who know each other, and due to peers being peers, inevitably form relationships, resulting in bias when awarding grades.

There are several approaches to account for relationship biases in peer assessment. For instance, researchers argue that anonymizing through single- or double-anonymized methods improves peer assessment (e.g., Li (2017)). Past research (e.g., Raes et al. (2015); Vanderhoven et al. (2015)) found that anonymity allows students to overcome inhibitions and improve their evaluation skills. Initial work by Howard et al. (2010) reported that anonymizing the peer assessment led students to be five times more likely to create

critical feedback and four times more likely to provide justifications for the improvements they suggested than those in a non-anonymous group. Later, Güler (2016) found that an anonymous group provided peer ratings more correlated with the instructor ratings than a non-anonymized group, and Gamage et al. (2017) showed that non-anonymized reviewers resulted in improved feedback and interaction in the peer assessment process. More recent work by van den Bos and Tan (2019) showed that an anonymous group of reviewers processed more directive higher-order feedback (e.g., feedback on ideas, organization, and argumentation) and obtained higher scores on their revised essays than a non-anonymous group.

Regarding students' perceptions, Lin (2018) investigated the role of anonymity in an online peer assessment within a Facebook-based learning application. The authors leveraged two groups of students. One group had the assessors' identities hidden throughout the peer-assessment process, and the other showed the full real names of the graders to their reviewers. Results show that anonymity increased cognitive comments but reduced affective comments. Lin (2018) also observed that the anonymous group had a more positive attitude toward the single-anonymized system, particularly by reporting a higher level of perceived learning. However, the authors reported that the anonymous group had a lower perception that peer assessment was fair. Further work by Seifert and Feliks (2019) compared self- and anonymous peer-assessment to understand if these strategies encourage students to take more responsibility for the learning process. Results showed that anonymity also provides students with more comfort and less peer pressure. Recent results by Kumar et al. (2019) also show that students had a favorable perception and attitude toward the online peer assessment method when the assessor and assessee are kept anonymous. More recently, Su (2023) found that students preferred an anonymous peer assessment since it provided more comfort than a non-anonymous approach.

Nevertheless, recent meta-analyses show mixed results of using anonymity to minimize interpersonal effects in peer assessment environments. In particular, Panadero and Alqassab (2019) and Li et al. (2016) suggest that non-anonymity is better for increasing students' peer grading accuracy when compared to teachers' assessment. These mixed results highlight the complexity of factors present in peer assessment and, in particular, led researchers to opt for considering relationship biases in the peer assessment rather than trying to remove them.

Relationship bias in peer assessments

Social relationships can be looked at more closely to find a friendship bias in peer assessment. Peer ratings—rating friends on a Likert scale—and peer nominations—classify a small number of peers in a group by who they like the most and the least—are sociometry tools to study and infer student relationships (Coie et al., 1982). Azarnoosh (2013) asked each student to nominate their three most intimate friends in class and compared friends' awarded grades with acquaintances. It found no significant difference but justified the result with the small class of 26 students who all know each other.

He et al. (2015) went further with more significant class sizes and, instead of only associating student relationships with peer assessments, tried to reduce the possible bias. He et al. (2015) tried to take advantage of interactions in the online discussion board to connect students and reassign graders to distant acquaintances since the course was primarily remote. A control group was left untouched for comparison. Although students reported they did not grade friends' exercises, the result comparison with the ones who

did is inconclusive. Inferring relationships from online data is valid but does not accurately reflect students' lives offline.

State-of-the-art research points toward an effect of friendship as a social factor on peer assessment. In particular, researchers reported this bias affecting peer assessment scores or students' perceptions of the peer assessment (e.g., Harris and Brown (2013); Domínguez et al. (2016); Kilickaya (2017); Ersöz and Şad (2018)). The primary concern is over-scoring based on friendship biases (Panadero et al., 2013) since students believe that they may lose friendships if they provide poor grades (Kilickaya, 2017). Panadero et al. (2013) considered using rubrics to counter this bias, but this approach only reduced biases of low and moderate-level friendships. In particular, high relationship levels produced significantly more over-scoring than low relationship levels between students. Nevertheless, the state-of-the-art provides limited knowledge regarding the effect of relationship biases in peer assessment and, more precisely, how to assess these relationships between students. Therefore, leveraging student relationships with peer nominations and peer ratings from a peer assessment environment may provide more robust and broader insights to help us understand the effect of the assessment relationship bias in an e-learning course.

Research methodology

This section describes the research questions tackled in the experiment, the remote learning course, the peer assessment environment, and the data collection and analysis.

Research questions

Our objective is to understand whether student relationships affect the peer assessment environment in an e-learning course. However, we need to tackle the quality of the peer assessment considering the tailoring of assessors and whether students perceived a bias based on the relationships. We decided to run an experiment in an e-learning course for a semester. At the beginning of the course, students anonymously self-reported their relationships with each peer by stating which colleagues were their most and least favorite. Along the course, we also tracked how relationships between peers who had no previous relationship before the course emerged and considered them in the peer assessment environment dynamics. During the course, students peer-assessed their colleagues' posts in a single-anonymized approach, i.e., only the reviewer knew the identity of the post's author. To understand the role of social relationships in this process, we specifically choose the reviewers of a post based on the relationship between the post's author and their peers. The faculty provided rubrics to train each student in the peer assessment to assist assessors in judging the quality of student performance.

In this light, a significant component of the peer assessment environment was tailoring reviewers for a specific post. As we mentioned, this curation can affect the quality of the peer assessment process and how students perceive it. Therefore, we defined two main research questions:

RQ1: Do self-reported social relationships affect the peer assessment quality in an e-learning course?

RQ2: Do students perceive social relationships bias the peer assessment process in an e-learning course?

Fig. 1 Visual depiction of the Skill Tree used in MCP

We analyzed the quality of the peer assessment environment through its accuracy and fairness. We also employed questionnaires to collect user perceptions. All collected data originate from a university course that supports e-learning and peer assessment, as presented in the next section.

E-learning course

We leverage a course named Multimedia Content Production (MCP), from the MSc in Information Systems and Computer Engineering at Instituto Superior Técnico - University of Lisbon, which uses Moodle[1] as its virtual learning environment. The Moodle platform enables students to obtain all the necessary content to work successfully in lectures, discuss any topics related to the course curriculum, and submit assignments to be assessed by the professor's. Researchers have already used the MCP course for several research studies in the past (e.g., Barata et al., 2017; Nabizadeh et al., 2021; Alves et al., 2024).

Students typically have two weekly lectures and one laboratory class. Although MCP is traditionally a blended learning course, the faculty ran it exceptionally in a remote setting due to the COVID-19 pandemic. In this case, students attended theoretical lectures and practical laboratories through a video-conference platform. The *skill tree* is among the different grading components. This element focuses on producing several types of multimedia content throughout the semester (Fig. 1). In addition to class exercises, students must deliver and have more than average grades in the *skill tree*. This strategy allows students
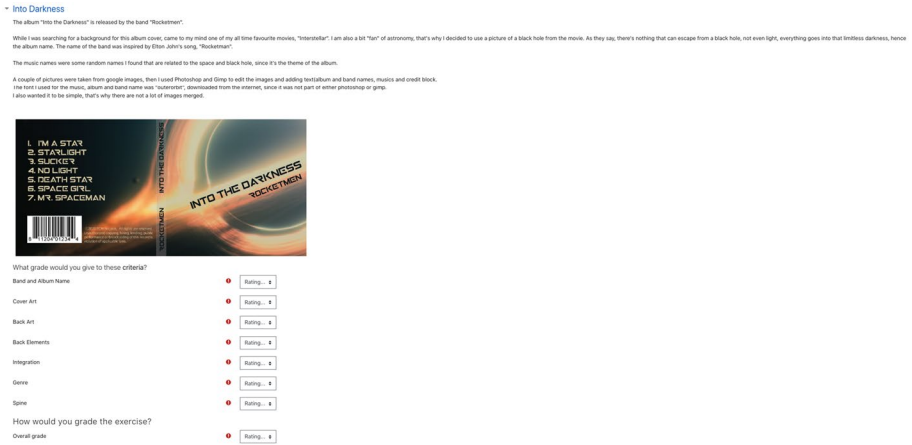

Fig. 2 PeerForum screen of an example of a training post

to obtain a final high mark. Students submit *skill tree* exercises (*skills*) to Moodle Forums for reviewing and assessment by professor's. There, professor's review and assess the skills on a 6-point scale and some written feedback. Afterward, the student can reply with an improved version based on the critique if the minimal acceptance requirements were unmet or if they wish to improve their exercise's quality (and rating). We took advantage of these Moodle Forums and submissions to have peers assess and give feedback to each other's *skill tree* submissions using a Moodle plugin.

Peer assessment environment

We created PeerForum, a Moodle plugin to enable peer assessment of students' *skill tree* submissions. As part of the functionality provided by the plugin, at the beginning of the semester, students had to answer a two-item survey where they nominated the peers they liked the most and the peers who they liked the least (Coie et al., 1982). At least four peers per category (liked and disliked) were required to be nominated by each student. After selecting peers, all enrolled students could reply to the professor's post with their solution for the *skill*. The plugin assigned each post to five students for assessment. We chose this amount according to Luo et al. (2014) and Cho et al. (2006), who recommended using between three and five graders and four to six graders, respectively. This assignment was not wholly random, consisting of a weighted sum of peers who nominated the post's author (if there were enough) and peers who did not. This strategy allowed for more data to be analyzed. Each of these five students received a Moodle notification and had 48 h to complete the peer assessment until it expired.

professor's also created a training page for each *skill* for students to practice the peer assessment and get familiar with examples and exercises. First, each training page contained an example of the multimedia piece expected to complete the skill and which criteria the faculty used to grade the students' submissions. Then, the page contained at least two sample submissions for the students to grade: skill tree submissions from previous years and curated by the faculty. The page required students to grade between 0 (*the post does not meet the requirements*) and 5 (*excellent*), each sample following the criteria previously presented (see Fig. 2). We did not train or
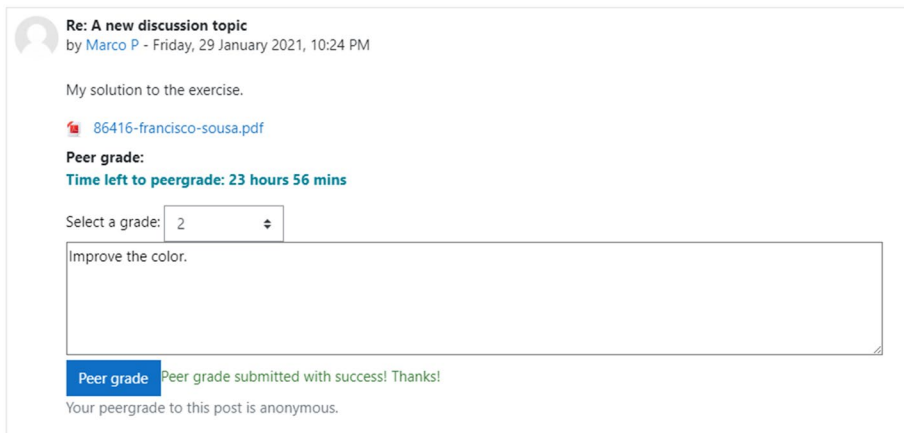
Fig. 3 PeerForum screen of a submission being assessed

process the written feedback that the students provided to the post. However, the faculty trained students to fill out these rubrics in the theoretical classes. We considered that the student completed their training only when the grades the student provided matched exactly the grades the faculty provided for each sample. Furthermore, students could only submit their work for peer assessment after completing the training page. This approach enforced that students had the same baseline on how to grade and acknowledged which criteria were more relevant for each skill, thus preparing them for upcoming peer assessment assignments and helping them produce submissions to the skill better aligned with the course's objectives.

Upon receiving a peer assessment assignment, we required students to complete the training exercises from the corresponding *skill*, and only then would they be allowed to perform the peer assessment: evaluate the exercise by submitting a grade (peer grade) and some written feedback (Fig. 3). Besides the five student graders, the professor also graded (rating) and gave feedback on each post by making a post reply. Both grades were on a 6-point scale (0 to 5), with clearly defined criteria for each point. A well-defined and extensive scale allows us to more effectively identify subtle differences in grading (Barata et al., 2017).

During the peer assessment process of each post, all the submitted peer grades, ratings, feedback, and professor replies were hidden from students until the post reached a minimum number of three peer assessments, a number found reasonable in the related work. At that point, all students could see the professor's rating and reply for each post and the average of all peer grades (final peer grade), except those who could still assess the post or until the 48 h expired. Individual peer grades and feedback were only visible to the post's author, but each grader's identity remained anonymous. The student author could improve their work and resubmit it for appraisal based on the feedback. The original peer graders, already familiar with the work and able to better understand those improvements, would, in that case, be assigned to the resubmission post. If the algorithm selected a student to assess a post of a colleague not present in their nominations list, we later asked them to rate that colleague on a 6-point scale (0 to 5, where 0 signifies not being acquainted with that colleague).

Data collection and processing

The PeerForum plugin replaced the current Moodle Forum in the MCP course 2021. In this edition, 69 students completed the course. By the end of the semester, we also presented them with a final questionnaire we developed on their opinion on peer assessment. The students were asked to rate, on a 5-point Likert scale, their thoughts on peer grading usefulness, fairness, and difficulty before and after the course; the quality of the feedback provided and the attention given to it; the perceived amount of effect the students' relationships had on the peer grades given and received; the usefulness of the training pages in peer grading and completing the skills; and the usefulness of peer grading in achieving the skills. A text box was also open to general feedback and suggestions for course improvement. Although the literature has shown that peer assessment works best when mandatory (He et al., 2015), we could not implement that condition in MCP due to school policies, so we awarded extra grade points for participation. The scale had a high level of internal consistency, as determined by a Cronbach's alpha of 0.958.

There were 26 *skills* with submissions, counting 863 posts with at least one peer assessment and 625 posts peer-assessed at least three times. The 77 students who nominated peers summed up to a total of 699 nominations and performed 1144 peer ratings. Only 56 students answered the final questionnaire. Regarding faculty members, eight professor's oversaw grading the skill tree. Each was solely responsible for a subset of skills, thus ensuring consistency of criteria in evaluating skills.

We removed the students who never assessed and ended up with a final set of 64 students (45 men, 17 women, and two identifying with another gender) aged 22 ± 1.61 years old. Moving forward, we will only analyze peer assessments from these students. Then, we removed from the dataset all peer assessments that contained no feedback related to the post, leaving 2681. The final post's peer grade is the rounded average of all peer grades for each post. For each peer assessment, the relationship between the post's author and the peer grader depends on whether the first was peer-nominated or peer-rated by the second.

Data analysis and results

This section describes the results of the experiment. It starts by analyzing the peer assessment fairness and accuracy and then the impact of the students' relationships on the metrics. We also discuss our results and state the limitations of our study.

Fairness

Fairness measures the consistency of scores given by multiple student graders, reflecting the general agreement among the students assigned to assess the same post—inter-rater reliability. Since the algorithm selected peer graders from a larger pool of enrolled students and a different set of students assessed each post, we measured their agreement with form 1 of the Intraclass Correlation Coefficient [ICC(1)].

The guidelines suggested in Koo and Li (2016) will be used to interpret the ICC results: Values lower than.50 are indicative of poor reliability, values between.50 and.75 indicate

Table 1 ICC estimates and their 95% CI by number of peer assessments

No. of Peer Assessments		Intraclass correlation	95% CI	
			Lower bound	Upper bound
2 (N = 124)	Single measures	.188	.013	.352
	Average measures	.316	.026	.520
3 (N = 199)	Single measures	.435	.350	.519
	Average measures	.698	.617	.764
4 (N = 203)	Single measures	.238	.168	.315
	Average measures	.556	.447	.648
5 (N = 186)	Single measures	.222	.159	.293
	Average measures	.588	.486	.675

moderate reliability, between.75 and.90 indicate good reliability and values greater than.90 mean excellent reliability.

The ICC(1) test was applied to the data obtained throughout our study. This test can only be made to data sets with the same number of item ratings. Hence, we grouped posts by their number of completed peer assessments. ICC estimates and their 95% confidence intervals (CI) were calculated using SPSS statistical package version 26 (SPSS Inc, Chicago, IL) based on single and average ratings, absolute agreement, one-way random-effects model (see Table 1).

The ICC Single Measures estimates the reliability of each of the randomly selected student graders when grading the same assignment. For posts with five peer assessments, the coefficient value of.222, with a 95% CI of [.159, .293], is considered poor in strength by the guidelines above, suggesting peer grades vary significantly among individual student graders and a single student's grade is not very reliable. The ICC Average Measures of.588, with 95% CI in [.486, .675], shows poor to mostly moderate reliability. In our use case, we plan to use the average peer grade of the five students as the assessment basis instead of one. Hence, we should use the ICC Average Measures as the measurement index. According to the results, there is barely any difference between having four and five peer assessments for one post. However, both ICC Single and Average Measures stand out when considering posts with only three peer assessments, reflecting consistently moderate reliability, almost good, overall higher than the four and five values. On the contrary, the ICC results drop to their lowest value when checking posts with only two peer assessments.

This result confirms what has also been observed in Nabizadeh et al. (2021): Three peer assessments are usually enough for students to reach an agreement on the post's final peer grade, being the minimum number of assessments to consider the peer assessment activity sufficient and display the results of both it and the professor rating.

Accuracy

In this study, we used the professor's grade as the ground truth on whether the final peer grade of a post is accurate: the professor rating. The means and standard deviations of the posts' metrics are in Table 2. Accuracy measures the similarity between the final peer grade and the professor's grade in the same post by assuming professor's award fair and accurate scores—a convergent validity.

Table 2 Mean and standard deviations of post's metrics: rating, average of all peer grades and its rounded value (final peer grade). N = 588

	Professor rating	Final peer grade	Average of peer grades
Mean	3.16	3.62	3.58
Standard deviation	.92	.61	.57

(a) final peer grade.

(b) average of peer grades.
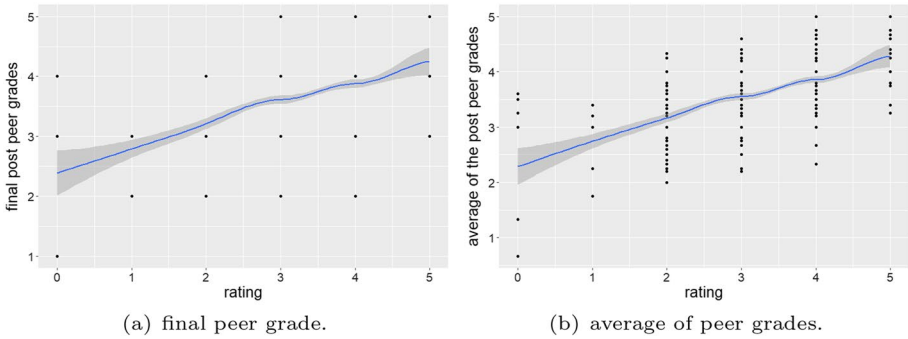
Fig. 4 Correlation between the posts' ratings and the peer grades

Table 3 Spearman's Correlation Coefficient and significance value between the professor rating and the average post peer grades by a minimum number of peer assessments

	Minimum number of peer assessments in post				
	5 (N = 186)	4 (N = 389)	3 (N = 588)	2 (N = 712)	1 (N = 806)
r_s	.596	.549	.558	.545	.512
p	<.001	<.001	<.001	<.001	<.001

Since submission grades are ordinal, we assessed the relationship between each post's professor grade and final peer grade with a Spearman's rank-order correlation. Five hundred eighty-eight posts were considered, each with at least three peer assessments. Preliminary analysis showed the relationship to be monotonic, as assessed by visual inspection of a scatterplot. There was a statistically significant moderate positive correlation between each post's professor rating and the final peer grade, $r_s(586) = .49$, $p < .001$ (Fig. 4(a)). When considering the average of the peer grades (without rounding), the correlation becomes stronger, with $r_s(586) = .56$, $p < .001$ (Fig. 4(b)).

The minimum number of peer assessments required to be considered for a valid final peer grade was studied by looking at its impact on accuracy. Table 3 shows Spearman's Correlation Coefficient r_s and its significance p-value between the professor rating and average of peer grades of posts with different minimum peer assessments. The correlation accuracy pattern seems consistent with the one observed in the reliability ICC (Table 1): there is a strong and significant correlation in all cases; having at least three peer assessments is recommended to obtain the most accuracy when comparing with the professor grade; and have five peer assessments is the ideal. These results confirm the literature findings that peer assessment is accurate: Students can overall give a grade like the professor's, even if all are slightly higher on average (Table 2).

We also compared the difference between professor ratings and students' peer grades grouped by the post's rating to check which latter was more (and less) in line with the student's grades. The results in Table 4 show that students tend to be more positive and moderate than professor's when giving their assessments: in posts with negative grades (below 3), students award more points, rarely giving the complete fail. In posts with 3, students give them 4. Students only agree (strongly) with professor's when giving 4's. As for the maximum grade, students usually choose it fewer times than expected.

Relationships

The main focus of this study is to infer whether the personal relationships between students impact their assessments. Hence, we collected some existing relationships (like or dislike) and compared them to a control group of neutral acquaintances. We did not analyze the peer assessments in which the relationship between the post's author and the student grader is unknown.

From 2681 peer assessments, we know the relationships between post author and peer grader of 2561 (95.5%), but only 2238 (83.5%) are in posts with at least three peer assessments. Of these, 684 (30.6%) are between students who like each other, 152 (6.8%) are between students who do not like each other, and the rest, 1402 (62.6%) report they have neutral feelings toward the author of the post.

We conducted two one-way Welch ANOVAs to study the effect of student relationships on the fairness and accuracy of peer assessment, determining if the peer grade difference from the final post grade and professor rating was distinct for different relationship types (like, dislike, neutral).

Both the final peer grade and rating differences were statistically significantly different between the three relationship types, Welch's $F(2, 412.373) = 7.472$, $p = .001$, Welch's $F(2, 411.337) = 3.561$, $p = .029$, respectively. The peer grade difference from the final post's peer grade increased from the dislike group (M = − .12, SD = .597) to the neutral (M = − .07, SD = .639) and then to the like group (M = .04, SD = .69). The same for the peer grade difference from professor rating: increased from the dislike group (M = .36, SD = .881) to the neutral (M = .37, SD = .931) and then to the like group (M = .49, SD = .987). Games-Howell post hoc analysis revealed that the mean increase from neutral to like was statistically significant in both the final peer grade difference (.111, 95% CI [.04, .19], $p = .001$) and the rating difference (.118, 95% CI [.01, .22], $p = .025$), as well as the increase from increase from dislike to like (.156, 95% CI [.03, .20], $p = .014$) in final peer grade difference.

Although students who like their peers tend to give them higher grades, the mean difference is meager, reaching much less than half a point (Fig. 5). This analysis shows that although there is a trend in students grading peers they like with higher grades, the difference and probability are not strong enough to impact the post's final grade for our scale. The same conclusions would not be applicable in a high-stakes assessment with a larger scale, as this discrepancy could be more accentuated and meaningful for the final grade.

Student perceptions

At the end of the semester, we asked students to reply to a questionnaire about the course, with some questions regarding peer assessment. Of the 56 students who responded, only 53 had peer-graded any post and were part of the 64 students mentioned

Table 4 Final peer grade difference from rating depending on the post rating

		Final peer grade difference from rating								
		− 4	− 3	− 2	− 1	0	1	2	3	4
Post Rating	0	–	–	–	–	0 (0,0%)	**3 (42,9%)**	0 (0,0%)	2 (28,6%)	2 (28,6%)
	1	–	–	–	0 (0,0%)	0 (0,0%)	2 (40,0%)	**3 (60,0%)**	0 (0,0%)	0 (0,0%)
	2	–	–	0 (0,0%)	0 (0,0%)	10 (7,8%)	**82 (63,6%)**	37 (28,7%)	0 (0,0%)	–
	3	–	0 (0,0%)	0 (0,0%)	2 (0,9%)	80 (37,6%)	**130 (61,0%)**	1 (0,5%)	–	–
	4	0 (0,0%)	0 (0,0%)	1 (0,5%)	31 (14,5%)	**175 (81,8%)**	7 (3,3%)	–	–	–
	5	0 (0,0%)	0 (0,0%)	1 (10,0%)	**11 (55,0%)**	7 (35,0%)	–	–	–	–
	Total	0 (0,0%)	0 (0,0%)	3 (0,5%)	44 (7,5%)	**272 (46,3%)**	224 (38,1%)	41 (7,0%)	2 (0,3%)	2 (0,3%)

The most common difference for each post rating is highlighted in bold

(a) final peer grade

(b) professor rating
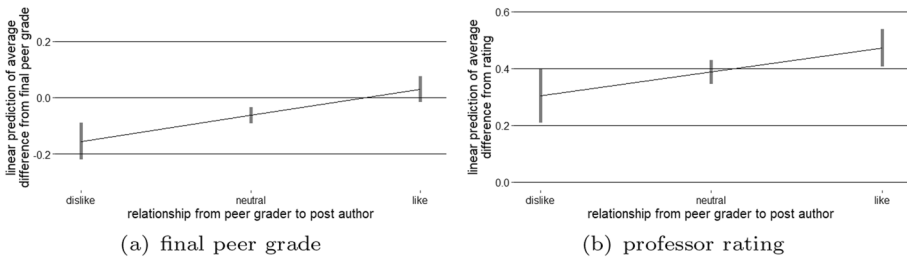
Fig. 5 Linear prediction of the differences from grades by the relationship between students

in 3.4. The following analysis only takes into consideration the replies from those 53 students. From that final set of students, 27 (50.9%) had already heard of peer assessment before the course, and 21 (39.6%) had already peer-graded someone. When asked if they thought peer assessment was helpful before the course, 23 agreed (43.4%), and 13 (24.5%) disagreed; after, 29 (54.7%) agreed, with 10 changing their opinion for the better. When asked about fairness, 13 students changed their opinion for the better compared to what they thought before the course, now with 28 (52.8%) agreeing and 11 (20.8%) disagreeing that peer assessment is fair. At the end of the course, 16 students (30.2%) agreed, and 25 (47.2%) disagreed that peer assessment is difficult, as 11 lowered their agreement after.

Regarding the impact and quality of the given feedback, although a great portion of students (45.3%) agree/completely agree their feedback was heard, many (37.7%) are indecisive. Similarly, many students agree (47.2%) that they provided quality feedback, although only 15.1% completely agree with the statement (26.1% are indecisive). Finally, around 11.3% completely disagree, and 34% disagree with the statement, "I think peer assessing my colleague's submissions did not help me when completing the *skills*." At the same time, 17% of students completely agree with it.

Regarding their perceptions of relationships, students disagree (75.5%) that their relationship with their colleagues affected the assessment they gave them. Only six students (11.3%) agreed. The reverse is similar: Many students (67.9%) disagree that their relations with them influenced the feedback they received. Students have confidence in their peers' grades, with 49.1% disagreeing/completely disagreeing that their colleagues are bad at evaluating compared to the professor's. 34% are neutral.

Furthermore, in the open text boxes, most mentions to peer grading praised the feature. From the 53 replies, eight students mentioned peer grading and the peer grading extra grade points as the most effective/interesting achievement from the course. It was because they could "give feedback to their colleagues' work," "learn about the skills (some before submitting work to them) and what the teachers expected." It helped students improve their work and made them more engaged in the course by making them "keep up with Moodle" and be attentive to the skills and how their peers were doing in each.

They were motivated to collaborate, engage in the course, and "give good and fast feedback to help colleagues." Students also reported that peer grading was a "useful feature" as it helped them learn and fix things the professor missed, but their colleagues pointed that out. Two other students suggested that the peer grade should have a bigger impact on the student's final grade, meaning they find peer grading fair.

Additionally, five students pointed out peer grading should be mandatory because it is an "important process" and it "helps students improve their work," both by "the possibly

good feedback I might give my peers," as well as because "they would get to see the professor's feedback sooner."

Only one student suggested that the person being graded should always be able to see the professor's reply. The remaining students did not report issues with having to wait for the peer grades before having access to the professor's reply/grade, although one pointed out this was a problem at the beginning of the semester and it improved throughout; others noted that this also worsened in the final days of the semester when students were focusing on delivering last minute work.

Three other students mentioned the quality of the written feedback in some peer grades, suggesting that the faculty should reward thoughtful/helpful feedback. They even proposed a mechanism to rate the feedback. No student said peer grading increased their workload, being classified as "easy enough work for a good chunk of credits," taking "not even 5 min of the day," and "offering some utility to the rest of the students."

Only one student who peer graded said it was their least effective/interesting achievement from the course. They said they were "unsure what to write because the examples given in some of the training pages were insufficient to completely understand the criteria for a skill." The same student added that their feedback consisted of small "nice work" variations to unblock their peers' grades and gain bonus points.

Finally, from the three students who replied to the final questionnaire and whose replies we did not consider in the above analysis, two mentioned peer grading as their most ineffective/uninteresting course achievement because, in their opinion, "it helps nothing to the experience and just slows down the turnaround time to receive a teacher's grade." Their replies were not included in the remaining data because they did not grade any posts.

Discussion

This study aimed to understand how pre-existing personal relationships between students affect peer assessment in an e-learning environment and whether it can accurately and fairly replace traditional evaluation. Regarding the quality of the peer assessment, we observed that students converge with similar grades to a specific post. In particular, our algorithm to pick the reviewers of a post according to their relationship with the post's author requires a minimum of three assessments to agree on the post's final peer grade. Our result aligns with the work of Luo et al. (2014) regarding the minimal number of reviewers for rater convergence. The exception is when students need to rate posts in the extremes of the rating scale, i.e., negative grades and excellent grades. For instance, we found that these posts generated an average standard deviation of all the peer grades larger than one point in posts with a final peer grade of one. Our finding suggests that the rubrics were not precise enough to capture the worst and the best posts' grading features. Similar to Panadero et al. (2013), our results showcase the limitations of not using fine-tuned rubrics.

Next, we consider the professor's rating as the reference point for each post's grade. Although students tend to grade slightly higher than the professor's, it is at most half a point on average. Considering that reviewers grade with discrete values with a granularity of one, we assume that students give peer ratings similar to those of an instructor. In particular, we found a strong correlation between the professor's grade and the final peer grade of each post with at least three peer assessments. Contrary to recent meta-analyses (e.g., Panadero and Alqassab (2019); Li et al. (2016)), our results are in line with previous literature suggesting that anonymous groups tend to provide more accurate grades

(e.g., Azarnoosh (2013); Luo et al. (2014); He et al. (2015); Güler (2016); Usher and Barak (2018)). Moreover, only a few students had been peer-assessed before taking the course. Students reported that the peer assessment process was fair and valuable. They were also confident in their peers' grades compared to the professor's and stated it has helped them complete the *skills*. Similar to Lin (2018), the students received the single-anonymized approach positively. However, our sample believes it was fair compared to Lin (2018). We hypothesize that students believe our approach was fair since they found the peer assessment valuable and the final grades of the posts aligned with the professor's grades. Kaufman and Schunn (2010) reported a similar finding. In Lin (2018), the usefulness of the peer feedback students received did not meet their expectations, hence the lower fairness perception.

With the overall fairness and accuracy of peer assessment confirmed, we continued by observing the effect of students' relationships on the process. Our findings show that the peer grader liking or disliking the author of the post they are assessing impacts accuracy and fairness. On average, students who like each other tend to give higher grades than the rest of the peer graders or the professor, and students who dislike each other tend to give lower grades than the rest of the peer graders or closer to the professor (which already grades lower than the final peer grade). These findings are in line with past results on the over- or under-scoring phenomena triggered by a relationship bias (e.g., Harris and Brown (2013); Domínguez et al. (2016); Kilickaya (2017); Ersöz and Şad (2018)). However, when we consider the reviewers tailoring we applied when picking the set of reviewers for a post based on the relationships, we found that these differences can be ignored, not exceeding half a grade point. We hypothesize that the assignment algorithm, which ensures a post is only assigned to at most one student who likes the author and one who does not, helped balance the fairness of the final peer grade. Without this algorithm, the chances of a post being assigned mostly to students who know the author are low. By ensuring that the post's reviewers cover a heterogeneous range of relationship statuses regarding the reviewer, we could balance out the peer ratings and still converge them to final ratings similar to the professor's. Furthermore, students disagreed that their relationship influenced the feedback both given and received. Our results provide evidence to understand the role of relationship bias in peer assessment and how to control it.

Design implications

Based on our results, we propose a set of implications that can be useful in designing the pipeline of peer assessment systems.

There is no need for peer assessment to be anonymous. In Usher and Barak (2018), the peer assessment was anonymous, and on-campus students still refrained from writing negative comments, contrasting with the unkind MOOC's feedback, showing students' awareness of grading peers with whom they have personal acquaintance. The authors conclude that although Li et al. (2020) observed better results in anonymous peer assessment, a bias is present even without knowledge of the identity of the exercise's author. Therefore, solely relying on anonymity may not improve the peer assessment process. Moreover, recent meta-analyses (e.g., Panadero and Alqassab (2019)) suggest that non-anonymity is better for increasing students' peer grading accuracy when compared to teachers' assessment. Our study consisted of a mix of interactions between students who knew and did not know each other, and we focused on studying the bias of these relationships. Our results do not support the idea that anonymity is also necessary if we consider the students' relationships

and curate the set of reviewers for a post based on the author. The overall peer assessment was accurate, and when analyzing the grades of students who knew each other, the difference was not significantly different from those who did not. Based on these results, we do not recommend anonymizing the peer assessment.

There should be prior training for peer assessment. Students are not experienced graders, nor are they fluent in the subjects being evaluated when assessing peers, which influences the accuracy and fairness of peer assessment. For example, in Luo et al. (2014), the grading criterion with the lowest peer agreement is a criterion more related to the course content, being more affected by different students' levels of prior knowledge and learning outcomes; the prior training of the grader is the most decisive factor that explains the variation of the peer assessment in Li et al. (2020); and Azarnoosh (2013) specifically accredit the high agreement between professor and peer assessments to the training and practice sessions before the actual peer assessment experience, as well as the usage of clear scoring criteria. In our study, we observed that the rubrics were useful to train students to grade a post except if it had poor or excellent quality. Panadero et al. (2013) also leveraged rubrics to counter the relationship bias. All findings point toward the importance of creating fine-tuned rubrics and training students with them to improve the results of peer assessment.

The relationship bias can be controlled. Our results show that the fairness and accuracy of peer assessment are the highest for posts with at least three peer assessments, confirming the literature findings that three should be the minimum of peer assessments required for a fair and accurate final peer grade (Luo et al., 2014). To reach this minimum of three graders, we advise future researchers to use peer nominations and ratings of student relationships to tailor the reviewers' choice for a post based on the author's relationships. Our approach shows that we could balance out the sample of reviewers and still provide grades close to the professor's ratings.

Limitations and future work

Regarding the analysis, we collected a substantial data trove from relevant student and educator interactions for future work. The MCP course is a Master's course that usually follows a blended learning approach, with face-to-face interactions and a balanced amount of students who know each other from previous/simultaneous courses and new students. Due to exceptional reasons, the semester in which we ran the experiment was completely online and at a distance, which may have reduced the number of relationships between students. Additionally, several plausible data combinations were not checked and thoroughly analyzed to maintain focus on the intended objective and for brevity. Hence, more levels and aggregations of the collected data can be explored, including the evolution through the semester and questionnaire comparisons, among others. Another limitation is using a grading scale with few levels (the 6-point scale), not allowing for a subtle distinction between grades, increasing the difference in results of the final peer grade and the professor rating.

Although the data set we gathered is comprehensive, there were some issues with the training variables and not an abundance of student relationships, which would benefit from a rerun of the study in a blended learning environment instead of e-learning. For instance, the training pages were an ongoing process created throughout the semester by the faculty, with several bugs, typos, and unfinished configurations in the pages' exercises. The students eventually spotted and fixed them, but this recurrent process polluted the gathered training data, so purging was impossible. Furthermore, the sample size in this experience does not allow us to draw substantial conclusions. Future studies should consider

increasing the number of participants to generate more data and relationship dynamics, allowing us to investigate the effect of social relationships with increased validity. Finally, our sample mainly comprises Portuguese individuals. Thus, it includes a cultural bias in our results.

Conclusions

Peer assessment has been widely studied as a replacement for traditional evaluation, not only by reducing the professor's workload but mainly by benefiting students' engagement and learning. This study uses a Master's e-learning course to research the influence of pre-existing social relationships between students in peer assessment's accuracy and fairness upon prior training and through the self-collection of relationship data. The results confirm the literature findings that peer assessment is reliable—with students agreeing on each post's final grade—and accurate—students giving the same grade as the professor—for posts with at least three peer assessments. However, the peer grades are slightly higher.

Students' social relationships are noticeable when students who do not like the other grade their work consistently lower than students who have a positive connection. However, this difference has minimal influence on the final peer grade, and students are unaware of it. Through self-reported feedback, they agree that peer assessment is valuable and fair. It allowed them to learn throughout the semester while grading, improving their work and engagement. These results allow us to conclude that peer assessment can replace traditional evaluation in an e-learning environment, as long as it is not a high-stakes assessment, benefiting students independently of their social relationships.

Author Contributions Francisco Sousa is responsible for Conceptualization, Methodology, Software, Validation, Investigation, Writing—Original Draft, and Visualization. Tomás Alves is responsible for Methodology, Validation, Formal analysis, Investigation, Writing—Review & Editing, and Funding acquisition. Sandra Gama is responsible for Methodology, Writing—Review and Editing, Supervision, and Funding acquisition. Joaquim Jorge and Daniel Gonçalves are responsible for Methodology, Writing—Review and Editing, Supervision, Project administration, and Funding acquisition.

Data availability The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

Code Availability The PeerForum Moodle extension code is avaliable in the author's GitHub page at https://github.com/TheMrKiko/peerforum.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Plagiarism or self-plagiarism This paper or a similar version is not currently under review by a journal or conference. This paper is void of plagiarism or self-plagiarism as defined by the Committee on Publication Ethics and Springer Guidelines.

Ethical approval Not applicable.

Consent to participate Not applicable.

Consent for publication Not applicable.

References

AlFallay, I. (2004). The role of some selected psychological and personality traits of the rater in the accuracy of self- and peer-assessment. *System, 32*(3), 407–425. https://doi.org/10.1016/j.system.2004.04.006

Alves, T., Sousa, F., Gama, S., Jorge, J., & Gonçalves, D. (2024). How personality traits affect peer assessment in distance learning. *Tech Know Learn, 29*, 371–396. https://doi.org/10.1007/s10758-023-09694-2

Azarnoosh, M. (2013). Peer assessment in an EFL context: Attitudes and friendship bias. *Language Testing in Asia, 3*(1), 11. https://doi.org/10.1186/2229-0443-3-11

Barata, G., Gama, S., Jorge, J., & Gonçalves, D. (2017). Studying student differentiation in gamified education: A long-term study. *Computers in Human Behavior, 71*, 550–585. https://doi.org/10.1016/j.chb.2016.08.049

Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability (formerly: Journal of Personnel Evaluation in Education), 21*, 5–31. https://doi.org/10.1007/S11092-008-9068-5

Bostock, S. (2000). Student peer assessment. *Learning Technology, 5*(1), 245–249.

Chang, S. C., Hsu, T. C., & Jong, M. S. Y. (2020). Integration of the peer assessment approach with a virtual reality design system for learning earth science. *Computers & Education, 146*(103), 758. https://doi.org/10.1016/j.compedu.2019.103758

Chie Adachi, J. H. M. T., & Dawson, P. (2018). Academics' perceptions of the benefits and challenges of self and peer assessment in higher education. *Assessment & Evaluation in Higher Education, 43*(2), 294–306. https://doi.org/10.1080/02602938.2017.1339775

Cho, K., Schunn, C., & Wilson, R. (2006). Validity and reliability of scaffolded peer assessment of writing from instructor and student perspectives. *Journal of Educational Psychology, 98*, 891–901. https://doi.org/10.1037/0022-0663.98.4.891

Coie, J. D., Dodge, K. A., & Coppotelli, H. (1982). Dimensions and types of social status: A cross-age perspective. *Developmental Psychology, 18*, 557–570.

Dhawan, S. (2020). Online learning: A panacea in the time of covid-19 crisis. *Journal of Educational Technology Systems, 49*(1), 5–22. https://doi.org/10.1177/0047239520934018

Dilrukshi Gamage, T. S., & Whiting, M. (2021). Peer assessment in moocs: Systematic literature review. *Distance Education, 42*(2), 268–289. https://doi.org/10.1080/01587919.2021.1911626

Domínguez, C., Jaime, A., Sánchez, A., et al. (2016). A comparative analysis of the consistency and difference among online self-, peer-, external- and instructor-assessments: Thecompetitive effect. *Computers in Human Behavior, 60*, 112–120. https://doi.org/10.1016/j.chb.2016.02.061

Double, K., McGrane, J., & Hopfenbeck, T. (2020). The impact of peer assessment on academic performance: A meta-analysis of control group studies. *Educational Psychology Review*. https://doi.org/10.1007/s10648-019-09510-3

Ersöz, Y., & Şad, S. N. (2018). Facebook as a peer-assessment platform: A case study in art teacher education context. *International Journal of Assessment Tools in Education, 5*(4), 740–753. https://doi.org/10.21449/IJATE.478277

Gamage, D., Whiting, M. E., Rajapakshe, T., et al. (2017). Improving assessment on moocs through peer identification and aligned incentives. In: Proceedings of the Fourth (2017) ACM Conference on Learning @ Scale. Association for Computing Machinery, New York, NY, USA, L@S '17, pp. 315-318, https://doi.org/10.1145/3051457.3054013

Garrison, D. R., & Kanuka, H. (2004). Blended learning: Uncovering its transformative potential in higher education. *The Internet and Higher Education, 7*(2), 95–105. https://doi.org/10.1016/j.iheduc.2004.02.001

Graham, C. R. (2013). Emerging Practice and Research in Blended Learning. In: Handbook of Distance Education. Routledge, https://doi.org/10.4324/9780203803738.ch21, https://www.taylorfrancis.com/books/9780203803738

Güler, Ç. (2016). Use of whatsapp in higher education: What's up with assessing peers anonymously? *Journal of Educational Computing Research, 55*(2), 272–289. https://doi.org/10.1177/0735633116667359

Harris, L. R., & Brown, G. T. (2013). Opportunities and obstacles to consider when using peer- and self-assessment to improve student learning: Case studies into teachers' implementation. *Teaching and Teacher Education, 36*, 101–111. https://doi.org/10.1016/j.tate.2013.07.008

He, R., Hui, L., Shih, T. K., et al. (2015). Peer grading system based on social relations on MOOC. In *2015 8th International Conference on Ubi-Media Computing (UMEDIA)*. IEEE, Colombo, Sri Lanka, pp. 152–157, https://doi.org/10.1109/UMEDIA.2015.7297446, http://ieeexplore.ieee.org/document/7297446/

Hoang Phuoc, L., Le, H., Tran, H., et al. (2022). Does evaluating peer assessment accuracy and taking it into account in calculating assessor's final score enhance online peer assessment quality? *Education and Information Technologies, 27*, 1–29. https://doi.org/10.1007/s10639-021-10763-1

Howard, C. D., Barrett, A. F., & Frick, T. W. (2010). Anonymity to promote peer feedback: Pre-service teachers' comments in asynchronous computer-mediated communication. *Journal of Educational Computing Research, 43*(1), 89–112. https://doi.org/10.2190/EC.43.1.f

Kaufman, J., & Schunn, C. (2010). Students' perceptions about peer assessment for writing: Their origin and impact on revision work. *Instructional Science, 39*, 387–406. https://doi.org/10.1007/s11251-010-9133-6

Kilickaya, F. (2017). Peer assessment of group members in tertiary contexts. Innovations in Languages for Specific Purposes pp. 329–343.

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine, 15*(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

Kulkarni, C., Wei, K. P., Le, H., et al. (2013). Peer and self assessment in massive online classes. *ACM Transactions on Computer-Human Interaction (TOCHI), 20*(6), 1–31.

Kumar, K., Sharma, B. N., Nusair, S., et al. (2019). Anonymous online peer assessment in an undergraduate course: An analysis of students' perceptions and attitudes in the south pacific. In *2019 IEEE International Conference on Engineering, Technology and Education (TALE)*, pp. 1–8, https://doi.org/10.1109/TALE48000.2019.9226008

Li, H., Xiong, Y., Zang, X., et al. (2016). Peer assessment in the digital age: A meta-analysis comparing peer and teacher ratings. *Assessment & Evaluation in Higher Education, 41*(2), 245–264. https://doi.org/10.1080/02602938.2014.999746

Li, H., Xiong, Y., Hunter, C., et al. (2020). Does peer assessment promote student learning? A meta-analysis. *Assessment & Evaluation in Higher Education, 45*(2), 193–211. https://doi.org/10.1080/02602938.2019.1620679

Li, L. (2017). The role of anonymity in peer assessment. *Assessment & Evaluation in Higher Education, 42*(4), 645–656. https://doi.org/10.1080/02602938.2016.1174766

Lin, G. Y. (2018). Anonymous versus identified peer assessment via a facebook-based learning application: Effects on quality of peer feedback, perceived learning, perceived fairness, and attitude toward the system. *Computers & Education, 116*, 81–92. https://doi.org/10.1016/j.compedu.2017.08.010

Luo, H., Robinson, A., & Park, J. Y. (2014). Peer grading in a mooc: Reliability, validity, and perceived effects. *Online Learning: Official Journal of the Online Learning Consortium, 18*, 1–14. https://doi.org/10.24059/olj.v18i2.429

Na, J., & Liu, Y. (2019). A quantitative revision method to improve usability of self-and peer assessment in moocs. In: Proceedings of the ACM Turing Celebration Conference - China. Association for Computing Machinery, New York, NY, USA, ACM TURC '19, https://doi.org/10.1145/3321408.3322846,

Nash, S. S., & Rice, W. (2018). Moodle 3 E-Learning Course Development: Create highly engaging and interactive e-learning courses with Moodle 3. Packt Publishing Ltd

Nabizadeh, A. H., Jorge, J., Gama, S., & Gonçalves, D. (2021). How do students behave in a gamified course?—A ten-year study. *IEEE Access, 9*, 81008–81031. https://doi.org/10.1109/ACCESS.2021.3083238

Panadero, E. (2016). Is it safe? social, interpersonal, and human effects of peer assessment: A review and future directions. Handbook of human and social conditions in assessment pp. 247–266.

Panadero, E., & Alqassab, M. (2019). An empirical review of anonymity effects in peer assessment, peer feedback, peer review, peer evaluation and peer grading. *Assessment & Evaluation in Higher Education, 44*(8), 1253–1278. https://doi.org/10.1080/02602938.2019.1600186

Panadero, E., Romero, M., & Strijbos, J. W. (2013). The impact of a rubric and friendship on peer assessment: Effects on construct validity, performance, and perceptions of fairness and comfort. *Studies in Educational Evaluation, 39*(4), 195–203. https://doi.org/10.1016/j.stueduc.2013.10.005

Panadero, E., Alqassab, M., Fernández Ruiz, J., et al. (2023). A systematic review on peer assessment: intrapersonal and interpersonal factors. *Assessment & Evaluation in Higher Education, 48*(8), 1053–1075. https://doi.org/10.1080/02602938.2023.2164884

Raes, A., Vanderhoven, E., & Schellens, T. (2015). Increasing anonymity in peer assessment by using classroom response technology within face-to-face higher education. *Studies in Higher Education, 40*(1), 178–193. https://doi.org/10.1080/03075079.2013.823930

Seifert, T., & Feliks, O. (2019). Online self-assessment and peer-assessment as a tool to enhance student-teachers' assessment skills. *Assessment & Evaluation in Higher Education, 44*(2), 169–185. https://doi.org/10.1080/02602938.2018.1487023

Su, W. (2023). Masked ball for all: How anonymity affects students' perceived comfort levels in peer feedback. *Assessment & Evaluation in Higher Education, 48*(4), 502–512. https://doi.org/10.1080/02602938.2022.2089348

Topping, K. J. (1998). Peer assessment between students in colleges and universities. *Review of Educational Research, 68*(3), 249–276.

Topping, K. J. (2009). Peer assessment. *Theory Into Practice, 48*(1), 20–27. https://doi.org/10.1080/00405840802577569

Usher, M., & Barak, M. (2018). Peer assessment in a project-based engineering course: Comparing between on-campus and online learning environments. *Assessment & Evaluation in Higher Education, 43*(5), 745–759. https://doi.org/10.1080/02602938.2017.1405238

van den Bos, A. H., & Tan, E. (2019). Effects of anonymity on online peer review in second-language writing. *Computers & Education, 142*(103), 638. https://doi.org/10.1016/j.compedu.2019.103638

Vanderhoven, E., Raes, A., Montrieux, H., et al. (2015). What if pupils can assess their peers anonymously? A quasi-experimental study. *Computers & Education, 81*, 123–132. https://doi.org/10.1016/j.compedu.2014.10.001

Yan, Z., Lao, H., Panadero, E., et al. (2022). Effects of self-assessment and peer-assessment interventions on academic performance: A meta-analysis. *Educational Research Review, 37*(100), 484. https://doi.org/10.1016/j.edurev.2022.100484

Zheng, L., Zhang, X., & Cui, P. (2020). The role of technology-facilitated peer assessment and supporting strategies: A meta-analysis. *Assessment & Evaluation in Higher Education, 45*(3), 372–386. https://doi.org/10.1080/02602938.2019.1644603