published as: Thomas Staubitz, Dominic Petrick, Matthias Bauer, Jan Renz, Christoph Meinel: Improving the Peer Assessment Experience on MOOC Platforms, In Proceedings of 3rd Annual Learning@Scale Conference (L@S2016), 25-26 April, 2016, Edinburgh

Improving the Peer Assessment Experience on MOOC Platforms

Thomas Staubitz, Dominic Petrick, Matthias Bauer, Jan Renz, Christoph Meinel
Hasso Plattner Institute, Potsdam, Germany
{thomas.staubitz, matthias.bauer, jan.renz, christoph.meinel}@hpi.de
dominic.petrick@student.hpi.de

ABSTRACT

Massive Open Online Courses (MOOCs) have revolutionized higher education by offering university-like courses for a large amount of learners via the Internet. The paper at hand takes a closer look on peer assessment as a tool for delivering individualized feedback and engaging assignments to MOOC participants. Benefits, such as scalability for MOOCs and higher order learning, and challenges, such as grading accuracy and rogue reviewers, are described. Common practices and the state-of-the-art to counteract challenges are highlighted. Based on this research, the paper at hand describes a peer assessment workflow and its implementation on the openHPI and openSAP MOOC platforms. This workflow combines the best practices of existing peer assessment tools and introduces some small but crucial improvements.

Author Keywords

MOOC; Online Learning; Peer Assessment; Assessment.

INTRODUCTION

MOOCs can have tens of thousands of participants, which is why assignments are often offered in machine-gradable formats. However, machine-grading of hardly quantifiable criteria, e.g. elegance, style, and creativity, is difficult or next to impossible. Additionally, individualized feedback is an integral part of education, but can not be delivered by using automated assessments [30].

Peer assessment (PA) is employed in today's MOOCs as an attempt to tackle these issues. This method allows participants to receive personalized feedback and to engage in challenges that go beyond the capabilities of automatic machine-grading by allowing the participants to grade and comment each other's work [30].

While PA seems to be an ideal tool for MOOCs to bring complex assignments and feedback to participants, the various implementations across the large MOOC platforms face some challenges. PA is the subject of active research, which attempts to find new ways of presenting, conducting, and structuring peer assessments to overcome its current limitations. Based on this research, the paper at hand presents a concept that incorporates the state of the art of peer assessment implementations of selected MOOC platforms and explores new ideas, which are added throughout the various steps of the workflow. This concept has been implemented on our own MOOC platforms. Three courses, which have piloted in employing this technology have been evaluated.

The rest of this paper is structured as follows: Section *Foundations* presents a summary on theoretical background, advantages, didactic importance, and common criticisms of peer assessment. Section *State of the Art and Best Practices* provides an overview on current practices of peer assessment on selected MOOC platforms. Section *Our Implementation* presents our concept and its implementation of a peer assessment system. Section *Evaluation* evaluates our system based on feedback from participants in several peer assessment pilots that we have conducted until the time of writing. Finally, the last two sections present the work that remains to be done and sum up our contribution and findings.

FOUNDATIONS

Benefits of peer assessment include improvement of higher order thinking skills, consolidation of topical knowledge, and individualized feedback for each participant [10, 14]. Learning effects can thus not only be seen as welcome byproducts of the process itself, the learning might even be put in the foreground. Educational assessment has the purpose of deepening a student's understanding, measuring student achievement, and evaluating the effectiveness of an educational program [22]. PA as a form of educational assessment is very flexible and can be used to serve either purpose–summative or formative assessment–or even both at the same time by combining several components [33]. Students reviewing each other's work and giving written feedback is a widely accepted application of formative PA [12, 32]. Summative PA of fellow students' work is a more complicated matter and requires careful guidance by a teacher, since grades should be fair, consistent, and comparable for all students [1, 32, 9].

Rubrics

A common way to enable students to peer-grade other students' work is to use teacher-designed grading rubrics. Depending on the task at hand, an assignment can be objective or subjective. Objective assignments have clearly defined what is a right or a wrong answer, which in its simplest form could be a yes-or-no test. Subjective assignments have no clear way of expressing a correct answer, meaning that the student has multiple ways to solve the task at hand [31]. Creative writing is an example of a highly subjective task. Gauging the quality of such a piece of work through PA is difficult and highly subjective. On the other hand in a MOOC environment, such tasks are the ones that most likely will end up to be handled by peer assessment, as the massive nature of these courses forbids a manual grading by the teaching team, and an automatic grading of creative work is only hard to imagine. Rubrics counteract the subjectiveness during the grading by providing students with teacher-designed categories that communicate the quality expectations that a piece of work should fulfill [31]. They seek to guide students through the grading and often elaborate on the criteria, e.g. by giving examples how many points to award for which expertise and completeness displayed in the work of a peer [23].

Knight and Steinbach divide peer assessments into three guidance categories: (i) open-ended, where little guidance, e.g., in form of rubrics, is given, (ii) guided, where general hints and questions to consider are provided, and (iii) directed, which provides reviewers a detailed, checklist-like guidance to grade peers. They argue that directed peer assessment is superior to the other categories, since it also enables less knowledgeable students to assess their peers' work [14].

Self Assessment

A special type of assessment, often mentioned in conjunction with PA, is self-assessment. Having seen and assessed the work of their peers, students evaluate their own work on the same criteria that they used to evaluate the work of their peers [4]. Due to similar didactical and cognitive benefits and emerging synergies they are often used together [25, 3].

Studies examining student's performance improvements through PA participation report varying results. The consensus being that performance improvements largely depend on the specific application and learning environment where PA is employed [26, 27]. If applied responsibly, students have been found to improve in overall course performance if they previously participated in PAs for this particular course [26, 27]. Feedback in general is perceived useful by students. Some studies suggest that in some cases, students take comments from their peers more seriously than teacher comments [9, 25, 26]. Accurate grades have been reported for both peer and self-assessment, with more reviews increasing the accuracy relative to an accepted standard, such as teacher-assigned grades [26, 27].

Responsibly-applied PA should have explicit training or workshop sessions beforehand to increase student eligibility and confidence to assess peers [27]. The biggest concerns voiced regarding PA are grading bias and rogue reviewers, which both are related to the question whether or not students are eligible to assess peers.

Issues

Although large in-person classes also can amount to several hundreds of students, according to Suen [30], PA in this setting mostly has been employed in the context of smaller groups guided by a teacher or teaching assistant and often as a supplement to teacher assessment. In the context of MOOCs, increasingly heterogeneous groups of students participate in PA. Due to the different backgrounds and knowledge of students, student eligibility and grading accuracy is doubted and PA itself, as a valid assessment form, is sometimes being challenged by course participants [14, 30, 11]. Therefore, it depends on the teaching teams to narrow down quality expectations, for example by providing detailed rubrics to ensure the success of PAs and to keep assessments comparable, consistent, and fair [9, 27]. Students' subjectiveness based on their culture, education, and knowledge of the topic at hand will influence the way a student grades to a certain degree [16]. Bias can partially be counteracted with anonymity, multiple reviews per peer (averaging a grade), clear expectations, and trainings [14, 3].

Another factor linked to the problem of student eligibility are rogue reviews [26]. Rogue reviews are insufficient reviews caused by laziness, collusion, dishonesty, retaliation, competition, or malevolence [14, 26]. These were always a problem for PA, but have found to be a bigger problem online due to increased anonymity and a decreased feeling of community affiliation [12, 18].

STATE OF THE ART AND BEST PRACTICES

This section explores best practices and the state of the art of peer assessment implementations on Coursera, edX, and Iversity[1]. We have selected these MOOC platforms for the quality of the peer assessment tools that they feature.

Platform-specific Features

Across all platforms, PA is implemented as a work-flow consisting of several steps that have to be completed one after the other. The essential parts of the PA workflow on all platforms are a submission step, followed by a peer evaluation step, and finally a result step.

Coursera and edX additionally offer a calibration and a self-assessment step. On Coursera, these are optional [6, 5]. Iversity offers a Cloud Teaching Assistant System (CTAS) for customers that have voted to pay for this option [35]. In this model, freelance teaching assistants–hired by Iversity–are assigned to the course and grade the submissions of the participants. Thus providing sort of a missing link between crowd-sourced peer-grading and classic instructor-based grading. Vogelsang and Ruppertz [35] admit that it merely was assumed that these assistants (CTAs) possess the required qualities due to their professional status. In an experiment they conducted, CTAs and peers graded rather different from the course instructor but close to each other. In terms of the examined MOOC platforms, a step unique to edX is Artificial Intelligence (AI) grading, which allows to machine-grade essays based on advanced machine-learning algorithms [6, 7].

[1] Iversity is the largest German MOOC platform. https://iversity.org

Coursera offers a self-assessment step for the students to rethink their submission. As an incentive for doing self-assessment accurately, the maximum of both grades will be taken as the final grade if a participant grades herself within a 5% distance to the grade received from the peers [5].

Submission handling

Missing a deadline has different consequences across the platforms. On Iversity, no points are awarded if any deadline has been missed, whereas Coursera awards no points if participants missed the submission deadline, but only penalizes the final assessment grade by 20% if the evaluation deadline has been missed [5, 13].

edX offers the participants a plain text input field and the possibility to upload a file to submit their work. Iversity and Coursera offer the possibility of multiple subquestions per assignment, each one with an input and file upload. All platforms allow to save drafts as long as the submission deadline has not yet passed. The participants are required to explicitly submit their work at the end of the submission step to reduce the amount of empty submissions to be offered for grading. edX employs a quality filter, which determines, by using metrics, such as word diversity and text complexity, if a submission is accepted [6]. Based on the work of Kulkarni [16], Coursera was the first MOOC platform to use calibrated peer review (CPR) [5, 2], which transfers the idea of training sessions to online environments [2]. Participants learn to use grading rubrics correctly by comparing their grading to a grading sample of the teaching team. This allows them to adjust their grading, which leads to improved grading accuracy [2]. Coursera determines a competency index for participants, which is then used as a weight for the actual grading [2]. edX requires participants to reach an accuracy threshold before they can proceed. For the actual reviews Coursera and edX both advise teaching teams to opt for three to five reviews, whereas Iversity usually requires people to review seven peers [6, 5, 13]. Coursera and edX also offer motivated participants the possibility to grade additional reviews beyond the required amount.

Review distribution

A participant retrieves submissions to review one after another (edX, Iversity) or as a set to grade (Coursera) [7, 5, 13]. Exact details of how reviewers are mapped to reviewees are scarce. Several sources hint that Coursera and Iversity have a fixed mapping, which in turn means that if a reviewer fails to review all peers, reviewees might receive less than the required reviews [5]. edX on the other hand uses a submission pooling principle, where participants pull a submission from a pool and, thereby, lock the submission for everyone else. If it has been reviewed or 30 minutes have passed, the submission automatically returns to the pool [7]. The review process itself is double-blind without exception. Double-blind peer reviewing allows participants to give more critical feedback and freely express their opinions without having to consider interpersonal factors, which in turn results in more honest and, ideally, more useful reviews [18, 28]. To report code of honor violations, submissions can be flagged during the peer evaluation. Iversity has no obvious way to report submissions for misconduct other than plagiarism, whereas edX and Coursera allow further distinction, such as explicit content.

Grading

If participants have finished all reviews in time, they qualify for a grade, which is shown together with individual grades and comments received from their reviewers. For grade computation, Coursera and edX use the median to catch gradings too low or too high, whereas Iversity uses the average of reviewer grades [6, 5, 13]. As shown by Kulkarni [16], the median performs better than the average in this context. Raman [24] suggests to employ ordinal instead of cardinal grading. Each participant orders the submissions she has been assigned to for grading in the form of *a better than b better than c*. On the basis of these "local" orderings, a "global" ordering of all submissions is estimated and from this the cardinal grades are derived.

All platforms offer the possibility to request a re-grading. Coursera and edX suggest posting submissions and reviews in the discussion forums to get further feedback [7, 5]. Iversity offers a built-in re-grading tool [13].

EdX also employs an AI grading step that uses technology commonly referred to as Automated Essay Scoring (AES) [7, 2]. Statistical machine-learning models are employed to predict human-assigned scores. These models are trained based on teacher-assigned grades for a set of submissions and automatically extracted text features, such as number of words, orthographic or grammar errors, word frequency, or sentence complexity. AES dates back to the work of Page [21] in the 1960s and has been been an issue of continuous research since then [29]. Besides the high time investment that is required for training, the results are still problematic, as machines are not able to understand and interpret texts the way humans do [2].

The standalone tool Peerstudio[2], which can be added to any platform that supports the Learning Tools Interoperability (LTI) interface, takes a completely different approach on peer assessment. It focuses on enabling students to give immediate formative feedback on work-in-progress. Kulkarni et al. report that the grades of students that had received fast feedback were significantly higher than the grades of those students who did not receive feedback at all or after a too long time frame. [17]

OUR IMPLEMENTATION

Based on the state of the art PA implementations shown in the previous section, our core design is a workflow model that encapsulates a series of steps.

Basic Workflow

Submit work, *review peers* and *see results* are the core steps of the workflow. A training unit and a self-assessment step can additionally be activated by the teaching team. The additional steps can be mandatory or optional for the user. An unlock date and a deadline can be set for each of the steps. If the deadline of any mandatory step has been missed, the participant has failed the assignment. If the deadline of an optional step is missed, the participant is forwarded to the next

[2] https://www.peerstudio.org

step and can no longer complete the missed step. To start a peer assessment, participants have to acknowledge the code of honor. Furthermore, the system provides the option to create a best-of gallery. The participants, therefore, either have to acknowledge that their work can be used as an example in the gallery or opt-out from being featured in this gallery.

Peer Evaluation–The Core Feature

The system allows the teaching teams to define a summative and a formative part of the PA. The summative part is handled via grading rubrics. Defining the grading rubrics is one of the key tasks for the teaching team during the creation of a PA. A rubric consists of three components: An (ideally) self-explanatory title (e.g., "Writing Style" or "Creativity"), an explanatory text, and rubric options. The explanatory text can be used to give additional hints for the grading. This serves for a better guidance of participants, which helps to understand the rubrics and aims to reduce grading bias. Rubric options represent the different levels of attainment for their respective rubric. The scoring scale of the rubric options is not required to be linear and can weight levels differently, e.g., 1, 2, 4 points, or can be an interval, e.g., 0, 5, 10 points. Rubrics and rubric options are defined by the teaching team according to their needs for each PA.

Participants grade a submission based on the available rubrics. Additionally to this summative feedback, the participants are encouraged to give their peers formative feedback in the form of free text reviews. All reviews are double-blind to protect peers' identities from each other and to prevent call-outs or conflict escalations outside of the peer assessment. Participants who have completed the required reviews in time, advance to the next step. Participants who do not, receive zero points for the assessment. Voluntary additional reviews up to the number of required reviews can be handed in as long as the deadline has not passed.

A large amount of reviews has the advantage of increasing the probability that the average grade of a submission is close to the grade teachers would assign [26, 27]. However, time should be provided for those willing to write thoughtful reviews. With too many reviews to write, participants will rush through the process, which likely lowers the quality of the feedback overall. Therefore, we have designed the system to have a preset of three reviews as the default amount to be completed during the evaluation step. We recommend to increase this amount only if the peer assessment spans multiple weeks, only has a few rubrics to grade, or has a less demanding assignment.

In the following, we will focus on presenting some of those features that we added on top or that we have improved.

1. **A step in the workflow to train the participants in the summative part of the assessment.**
2. **A distribution mechanism that boosts the priority of the work of participants that already did a review.**
3. **Review rating, additional points for writing good reviews.**
4. **Transparent grade computation and re-grading option.**

Training Step

The teaching team has the possibility to add a simple training step to the PA workflow. This training step can be optional or mandatory for the participants, defined by the settings of the teaching team.

The teaching team grades at least 10 samples of the participants' submissions for the current PA to ensure a more diverse training pool. A diverse set of samples benefits the training as it provides the participants with examples for each level of quality. While a sample submission is being reviewed by a teaching team member, it is blocked for all other teaching team members to ensure that it is not accidentally graded twice.

As soon as the training step has been opened for the participants, the submissions that already have been graded by the teaching team are presented to the participant one after the other. Initially, the students only can access the submission, but not the grading of the teaching team. This view is more or less identical to the view of the actual evaluation step. As soon as the participant herself has graded the sample, she is presented an overview, which shows the teaching team's comment and adds indicator arrows to the rubric options that have been selected by the teaching team and those that have been selected by the participant. This way, the participant can easily detect scoring differences.

If the training step is mandatory, the teaching team can define the amount of training reviews that have to be done. If a participant still feels insecure in using the rubrics after grading the required amount of samples, she is allowed to grade as many additional samples as there are left in the pool.

Distributing Submissions to Peers

The distribution mechanism that maps submissions to peers is the core of the evaluation step. Our goal is that each participant receives sufficient feedback. Each participant is required to write a certain amount *(n)* of reviews, which means that each participant ideally receives that amount of reviews in return. The size of *n* is defined by the teaching team.

We are using a dynamic mapping approach that assigns submissions on demand from a pool of available submissions. This allows to rebalance review counts for submissions on the fly as the current submission distribution state can be considered. With additions such as review expiration, the problem of redistributing submissions is solved without requiring heuristics that determine when submissions are considered for redistribution.

Our implementation of the pooling approach introduces a *submission priority principle* to balance submission distribution and the possibility that a submission is reviewed by several reviewers in parallel (see Figure 1).

Particularly, the *submission priority principle* is a novelty that deserves some more detailed explanation. Submissions are assigned a *priority*, which influences their likelihood to receive reviews, and they have a *required reviews counter* to indicate how many reviews are still required for this submission. As soon as a submission is inserted into the pool, both the required

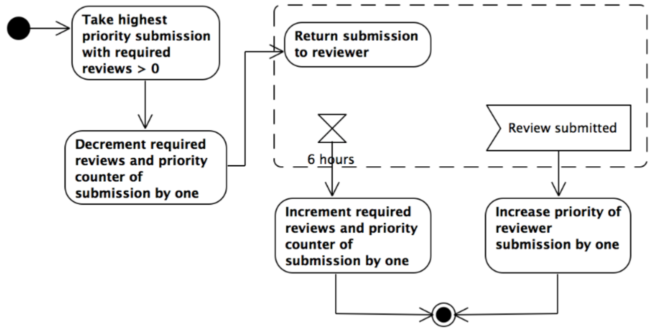
Take highest priority submission with required reviews > 0
Decrement required reviews and priority counter of submission by one
Return submission to reviewer
6 hours
Review submitted
Increment required reviews and priority counter of submission by one
Increase priority of reviewer submission by one

Figure 1. Activity diagram summarizing the distribution mechanism. The dotted line marks the time range from the start to the end of the actual review process. This time range starts when a participant requests a submission to review and ends either when the review is submitted or after 6 hours if the participant fails to return the review in time. A submission is the work that has been submitted to the system by a participant a needs to be reviewed.

reviews counter and the submission priority are initially set to the required amount of reviews of the PA *(n)*.

Submissions with the highest priority that still require reviews are retrieved first. If there are multiple submissions with the same priority, the retrieval of these is randomized to spread reviews more evenly. Retrieving a submission reduces both its counters by one, which means that other submissions are now more likely to be retrieved next, as they now have a higher priority relative to the retrieved one. Reducing both counters before a review is submitted, ensures that a submission cannot accidentally receive too many reviews, since it can be retrieved at most *n* times before other submissions are retrieved next. Otherwise, the balance of the mechanism can be disrupted due to one submission receiving too many reviews, which are then "missing" for other submissions in the worst case.

When a review is submitted, the reviewer's own submission receives a priority boost of one. The philosophy behind this boost is that those who write reviews deserve to get reviews in return, while those who do not should receive less. If a participant does not submit a review within six hours, it will expire. If a review expires, the priority and required reviews count of the corresponding submission are restored by increasing each by one again. Without review expiration, reviewers who forget about their reviews would leave their peers at a disadvantage, because they would no longer be able to receive the required number of reviews.

Review Rating

In the result step, participants can see the reviews they received and–as soon as the deadline for the evaluation step has passed–their final grade for the assessment. Our system introduces the option to report received reviews for misconduct, analog to reporting submissions in the evaluation step. A further improvement is the possibility to rate the received feedback.

Current platforms lack incentives to write thoughtful reviews, as well as a way for peers to give feedback on received reviews [36, 20, 34]. Participants are practically defenseless if the reviews they receive are not worthy of a report, since teaching teams have only limited capacities and cannot investigate minor incidents. Lu et al [19] have conducted an experiment to "grade the graders". They report particularly positive results when the graders receive grades for their grading and also grade the grading of others.

We, therefore, decided to go for a feedback rating approach, which has already been used in online and classroom peer assessments [10, 12]. The basic idea is to allow participants to actually make an impact and "defend" themselves by influencing the grade their reviewer receives, thus motivating reviewers to write more detailed and helpful reviews [14]. We have transferred the idea of feedback rating to MOOC peer assessment by allowing participants to earn up to one bonus point for each review they write. The receiving participants rate reviews on a scale from zero to three stars, based on how useful they perceive the review. Guidance on what a good review looks like is provided during the rating.

A four-point scale has been chosen and translates into no point for zero stars, 0.3 for one star, 0.7 for two stars, and 1.0 for three stars. This bonus can accumulate to a notable amount of points, even considering that participants already can earn a significant amount of base points for their submission in the PA compared to other assessment forms on our platform. This may seem to be a rather high bonus. However, reviews are an integral part of peer assessment and its learning experience. Hence, reviews are small but important work artefacts of a peer assessment process that require a considerable effort to be written in high quality, for which feedback rating can be seen as an appropriate reward. Until the participant has rated the feedback, we only show her the overall grade but not the individual grades that have been assigned by the peers. As soon as the written part of a review is rated, we also show the individual grade given by that peer. This is an attempt to make sure that participants only rate the usefulness and quality of the written review and not the grade that they received from their peer. Furthermore, this method uses the inherent curiosity of the participant as an incentive to rate the reviews.

Grade Computation

We originally opted for an experimental weighted average approach taking multiple factors into account. Some of these were intended to serve as long-term adjustments to reduce the impact of rogue reviewers and reviewers who constantly deliver poor review quality. In the end, however, we skipped that plan as this calculation would have become absolutely intransparent for all stakeholders, such as teaching team, helpdesk, and participants. Kulkarni et al [16] have already shown that simply using the median is good enough in most cases and will be only marginally improved by employing complex weighting mechanisms.

We use the average for those rare cases when a submission receives less than three reviews and the median when it receives three or more reviews.

The final grade consists of multiple components, which are transparently communicated to the participant in the result step:

- Median of review grades (base points)
- Self-assessment bonus points

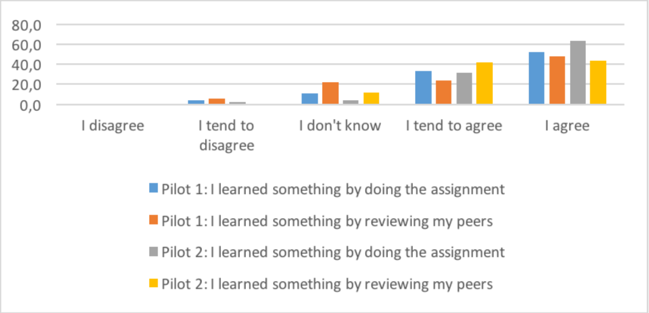

Figure 2. Perceived learning effects for doing the assignment and reviewing peers (OSAP1: n=54, OSAP2: n=463).

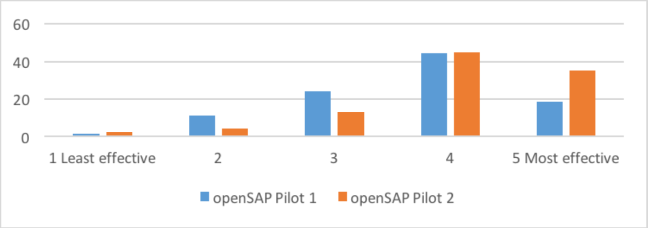

Figure 3. Perceived learning impact of peer grading assignment compared to other assessment types used on our platforms, such as quizzes (OSAP1: n=54, OSAP2: n=467).

- Review feedback rating bonus points
- Grading delta

The median of the review grades is calculated per rubric. We then sum up these values to the final amount of points for the participant[3].

The grading delta can be set by the teaching team members during conflict reconciliation. This delta can be relative or absolute. A relative delta adds a fixed amount of points to the base points, bonus points are left untouched and are added on top. The absolute delta sets the overall grade to a fixed amount of points, ignoring the originally received base and bonus points.

EVALUATION

We will now evaluate the implemented PA design based on surveys we conducted among the participants of our first PA pilots and on an analysis of some key factors of the data that we collected during these pilots. In total we have conducted five PAs in four courses, two on openSAP and two on openHPI, in a variety of settings up to now. The review distribution mechanism has been employed in all of the PAs, the other key features will be listed per PA in the following section.

Key Features of Examined PAs

Introduction to SAP Fiori UX (OSAP1)

- Enrolled participants: ~21,000
- Task: creative design challenge to create an SAP Fiori application
- Extra Steps: none
- Minimum reviews: 3
- Rubrics: 8 (target group definition, creativity, simplicity, delightfulness, etc. (subjective))
- Timeframe: 5 weeks (submission), 1 week (peer assessment)
- 30 bonus points (course total: 360)
- Submissions: 149, qualified for grade: 116, survey: 54 (43%)

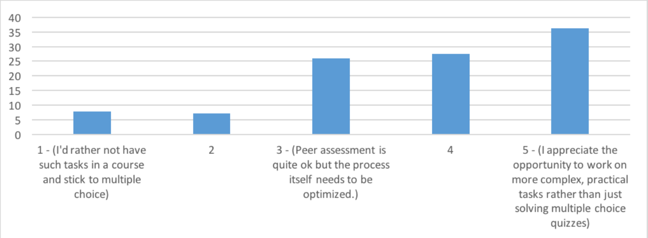

Figure 4. The participants of OHPI2 appreciated the opportunity to work on more complex practical tasks, enabled by the PA system (n=466).

SAP experts reviewed the highest ranking submissions and chose three winners who were rewarded with a tablet computer.

Build Your Own SAP Fiori App in the Cloud (OSAP2)

- Enrolled participants: ~18,000
- Task: creative design challenge to create an SAP Fiori application
- Extra Steps: training, self-assessment, both mandatory
- Minimum reviews: 5
- Rubrics: 8 (Story, Persona, User Experience Journey, User-centricity, Look and Feel, etc. (subjective))
- Timeframe: 2 weeks (submission) + 1 week (training, peer- and self-assessment)
- 150 regular points (Course total: 450)
- Submissions: 1529, qualified for grade: 1332, survey: ~470 (~30%)

SAP experts reviewed the highest ranking submissions and chose three winners who were rewarded with a laptop.

Java for Beginners(OHPI1)

- Enrolled participants: ~11,000
- Task: modeling an object-oriented application, class diagram
- Extra Steps: self-assessment, optional
- Minimum reviews: 3
- Rubrics: 6 (mostly just checking if something useful had been delivered (yes or no) (objective))
- Timeframe: 3 weeks(submission), 1 week(peer- and self-assessment)
- 10 bonus points (Course total 103 + another 8 bonus points)
- Submissions: 337, qualified for grade: 297

Web Technologies(OHPI2 CSS/Pong)

- Enrolled participants: ~10,000
- Task1: create an HTML page including some CSS formatting according to a given design
- Task2: second task was to complete a given Javascript example
- Extra Steps: self-assessment, optional

[3] See: https://open.hpi.de/pages/p_a_grading

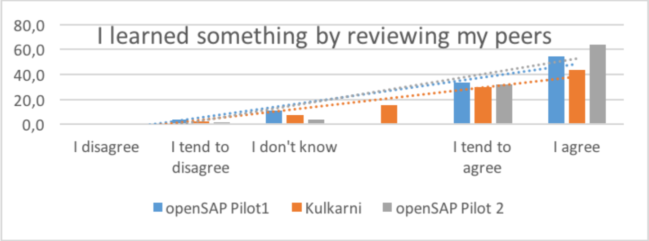

Figure 5. Comparison of the openSAP pilot surveys (OSAP1: n=54, OSAP2: n=464) with Kulkarni's findings on perceived learning by assessing peers.

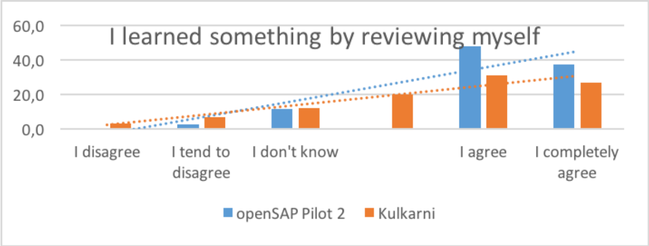

Figure 6. Comparison of the OSAP2 survey (n=469) with Kulkarni's findings on perceived learning by assessing one self.

- Minimum reviews: 3
- Rubrics: 3/6 (mostly simple yes or no questions to check if a certain feature exists or not (objective))
- Timeframe: 3/2 weeks (submission), 1/1 week (peer- and self-assessment)
- 7/9 bonus points, (Course total 180 + another 15 bonus points)
- Participants in assignment: 1371/1010
- Submissions: 761/567, qualified for grade: 476/592

General Feedback

Figure 2 shows how participants perceived the two key steps of the PA tasks in OSAP1 and OSAP2. Both–working on the assignments and reviewing the peers' submissions–were perceived by the participants as having a strong impact on their learning. These results validate PA usage in a creative design challenge context as given in those pilots. This impression is further reinforced by Figure 3, showing that the learning impact of PA is also perceived positively in comparison to other assessment types that are employed on our platforms.

The participants of OHPI1 stated in the course end survey that they appreciate the opportunity to work on more complex practical tasks, enabled by the PA system, instead of doing multiple choice tests (see Figure 4). This finding is also supported by participants' comments, throughout all of the examined courses, which generally expressed appreciation of the activity-driven assessment variation that has been introduced into the courses.

Comparing the survey results of OSAP1 and OSAP2 (in terms of perceived learning impact of assessing peers or one's own) with Kulkarni's results [16] (see Figure 5 and 6) shows at least very similar tendencies.

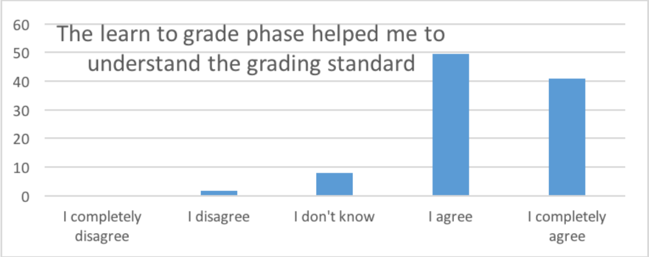

Figure 7. The training step helped me to understand the standards (n=468)

Evaluating the Novelties of our Implementation

Training Step

Almost 90% (n=54) of the participants of the OSAP1 survey stated that they would have preferred to have a training step before the actual peer grading step.

Figure 7 shows that the majority of participants of OSAP2's post-course survey perceived the offered training step as very helpful to understand the expected grading standard. In both pilots almost all participants perceived themselves to have given fair and detailed feedback (OSAP1: 94%, n=54, OSAP2: 96%, n=466). The perception of the received feedback, however, differs significantly between the pilots (OSAP1: 64%, n=54, OSAP2: 85%, n=466). While the difference between the perception of given and received feedback can be accounted to a well-known phenomenon in behavioral psychology (compare e.g. [15], [8]), the improvement in OSAP2 might be a sign for improved grading abilities due to the training step. However, there are too many factors involved to make a definitive statement here.

Distributing Submissions to Peers

Figure 8 shows that, except for OSAP1, the distribution mechanism worked quite well[4]. Starting with OSAP2 we can see a similar tendency throughout all the examined courses, showing that those participants who wrote the minimum amount of reviews, have also received a sufficient amount of reviews, while fewer reviews were "wasted" on those that did not write sufficient reviews. At least for the summative part of the PA the data shows that the distribution works sufficiently well. It has been suggested, however, that strong students are more likely to write reviews than weak students. Therefore, a concern might be that strong students get more reviews that way while weaker students get less. The formative part of the PA thus might increase the "learning gap" between good and bad performers.

Review Rating

Our review rating mechanism has been received very positively as a motivator for review quality. About 80% of the participants of the OSAP1 (n=54) and OSAP2 (n=472) surveys stated that our review rating implementation motivated to write useful feedback and was a good way to review the reviewers. More than 70% of the participants in the OSAP1 survey stated that the feedback they received motivates them to improve their reviews in the future.

[4] During OSAP1, the system was in a very early beta state and a lot of adjustments still needed to be performed

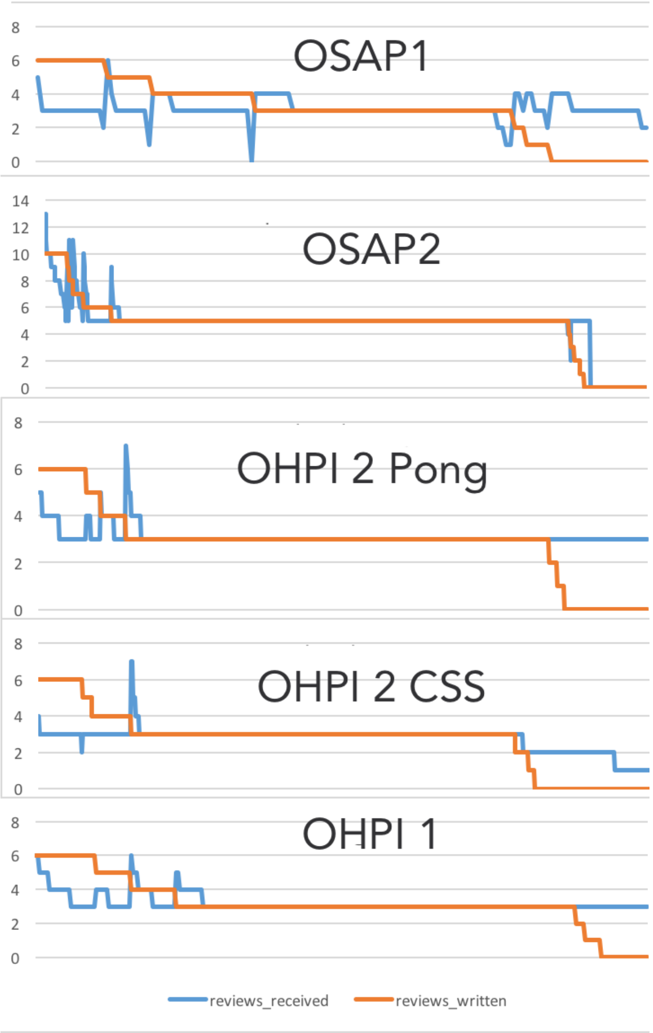

Figure 8. Ratio reviews written to reviews received in all of the examined courses. From top down (required reviews, submissions, sufficient amount of reviews: (3,149,116), (5,1529,1332), (3, 567, 476), (3, 761, 592), (3,337, 297))

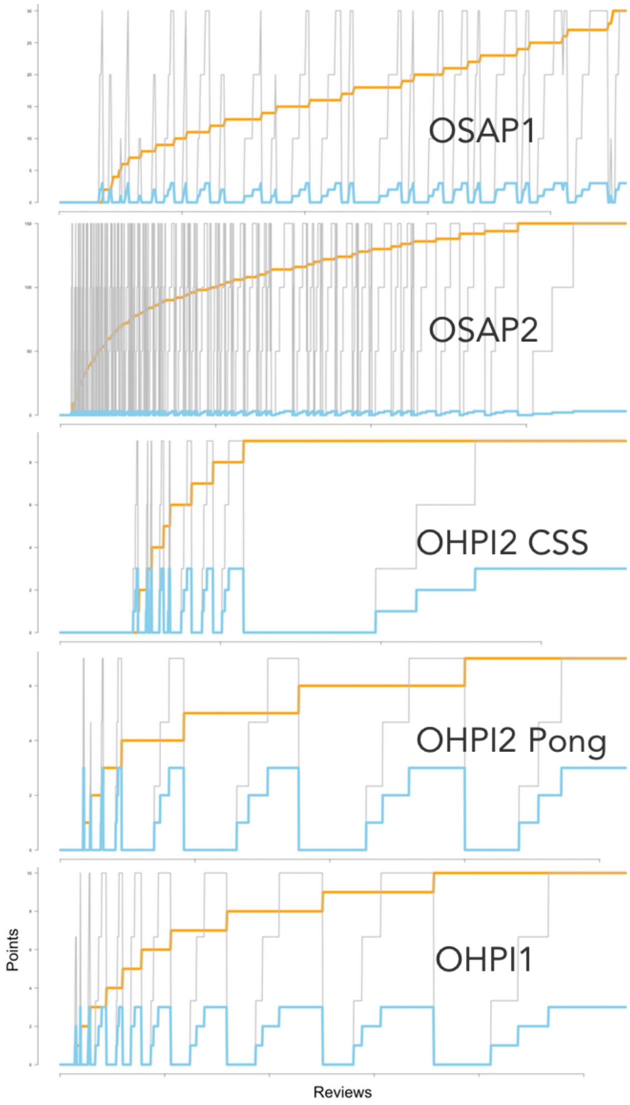

Figure 9. Orange: Points for submission, Blue: Review Rating, Grey: Review rating normalized (blue*maxOrange/maxBlue)

There is an objection that the introduced review rating might lead to a "Tit for Tat" situation where good grades instead of good reviews are rated good. We, therefore, analyzed the data in the the examined PAs to see if we can find evidence for this assumption. First, we compared the points a participant received for the submission to the rating she gave to the review. Figure 9 shows that we can observe a very similar trend in all of the examined courses. The feedback points that a review received, in general, is not related to the amount of points a reviewer gave. We then compared the reviews word count, as a very basic factor for review quality, to the received rating. Figure 10 at least implies that there is a certain relation between the length of a review and the rating it receives.

FUTURE WORK

It has been suggested that the review distribution mechanism might increase the "learning gap" between good and bad performers. A potential tweak in the process might be to explicitly inform the participants about the option to improve their chances for a review by writing reviews themselves. Furthermore, we need to further investigate how to provide a re-grading mechanism that is smart enough to decide who is eligible for a re-grading and who is not. Raman's approach of ordinal grading [24] might be a starting point worth investigating.

CONCLUSION

The paper at hand described the importance of peer assessment for scaling formative assessments and more challenging and engaging assignments in large online learning environments. A peer assessment workflow concept based on research findings, platform implementations of large MOOC platforms, and participant feedback has been presented, which incorporates the state of the art of MOOC peer assessment and adds several technical improvements to the process. Feedback rating has been introduced, which is a concept to let peers rate the reviews

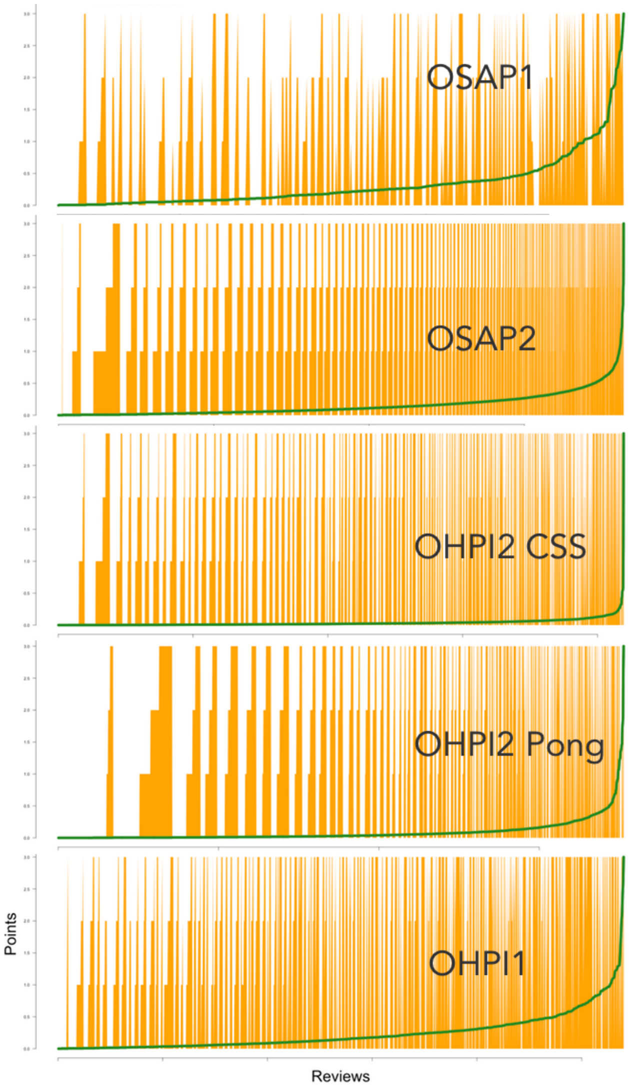

Figure 10. Orange: Review rating, Green: Review word count normalized (green*maxOrange/maxGreen)

they received, with the intention to provide incentives for reviewers to improve their review quality. Rated reviews award reviewers bonus points based on the rating, which is not only meant to motivate reviewers to write better reviews, but also to allow participants to defend themselves against poor review quality. Submission pooling has been developed as a dynamic algorithm that takes the current submission-distribution state into account and balances distribution of submissions on the fly, based on priority and reviews needed per submission. An evaluation based on peer assessments with different settings in four courses on two of our MOOC platforms marked a success of the implemented PA workflow with largely positive feedback from participants. Finally, our implementation is perceived as on-par with other implementations and many participants expressed their appreciation of a more activity-driven and creative challenge, wishing to have more of those in future courses.

ACKNOWLEDGMENTS

We thank openSAP, particularly Katrin Elk, for sharing their insights and the overall great collaboration.

REFERENCES

1. John R. Baird and Jeff R. Northfield. 1995. *Learning from the PEEL experience*. School of Graduate Studies, Faculty of Education, Monash University.
2. Stephen P. Balfour. 2013. Assessing writing in MOOCS: Automated essay scoring and Calibrated Peer Review. *Research & Practice in Assessment* (2013), 40–48.
3. Stephen Bostock. 2000. Student peer assessment. http://www.reading.ac.uk/web/FILES/engageinassessment /Student_peer_assessment_-_Stephen_Bostock.pdf. (2000). Online; accessed 14-September-2015.
4. David Boud. 1995. *Enhancing Learning Through Self-assessment*. Taylor & Francis.
5. Coursera. 2014. Coursera Peer Assessment Student FAQs. http://help.coursera.org/customer/portal/topics/521177-peer-assessments. (2014). Online; accessed 3-November-2014.
6. edX. 2014a. Open Response Assessment Problems. http://edx-partner-course-staff.readthedocs.org/en/latest/exercises_tools/open_response_assessment.html. (2014). Online; accessed 3-November-2014.
7. edX. 2014b. Open Response Assessments for Students. http://edx-partner-course-staff.readthedocs.org/en/latest/students/ora_students.html. (2014). Online; accessed 3-November-2014.
8. Joyce Ehrlinger, Kerri Johnson, Matthew Banner, David Dunning, and Justin Kruger. 2008. Why the Unskilled Are Unaware: Further Explorations of (Absent) Self-Insight Among the Incompetent. *Organizational behavior and human decision processes* 105, 1 (01 2008), 98–121.
9. Jan Fermelis, Richard Tucker, and Stuart Palmer. 2007. Online self and peer assessment in large, multi-campus, multi-cohort contexts. In *Providing choices for learners and learning Proceedings ASCILITE Singapore 2007*. 271–281.
10. Edward.F. Gehringer. 2000. Strategies and mechanisms for electronic peer review. In *Frontiers in Education Conference, 2000. FIE 2000. 30th Annual*, Vol. 1. F1B/2–F1B/7 vol.1.
11. David Glance, Martin Forsey, and Myles Riley. 2013. The pedagogical foundations of massive open online courses. *First Monday* 18, 5 (2013).
12. John Hamer, Kenneth T. Ma, and Hugh H. Kwong. 2005. A Method of Automatic Grade Calibration in Peer Assessment. In *of Conferences in Research and Practice in Information Technology, Australian Computer Society*. 67–72.
13. Iversity. 2014. P2P Grading: Assessing Globally. https://iversity.org/blog/p2p-grading/. (2014). Online; accessed 3-November-2014.

14. Linda Knight and Theresa Steinbach. 2011. The pedagogical foundations of massive open online courses. *Journal of Information Technology Education* 10 (2011), 81–100.

15. Justin Kruger and David Dunning. 1999. Unskilled and Unaware of It: How Difficulties in Recognizing One's Own Incompetence Lead to Inflated Self-Assessments. *Journal of Personality and Social Psychology* 77 (1999), 1121–1134.

16. Chinmay Kulkarni, Koh Pang Wei, Huy Le, Daniel Chia, Kathryn Papadopoulos, Justin Cheng, Daphne Koller, and Scott R. Klemmer. 2013. Peer and Self Assessment in Massive Online Classes. *ACM Trans. Comput.-Hum. Interact.* 20, 6, Article 33 (Dec. 2013), 31 pages.

17. Chinmay E. Kulkarni, Michael S. Bernstein, and Scott R. Klemmer. 2015. PeerStudio: Rapid Peer Feedback Emphasizes Revision and Improves Performance. In *Proceedings of the Second (2015) ACM Conference on Learning @ Scale (L@S '15)*. ACM, New York, NY, USA, 75–84.

18. Ruiling Lu and Linda Bol. 2007. A comparison of anonymous versus identifiable e-peer review on college student writing performance and the extent of critical feedback. *Journal of Interactive Online Learning* (2007), 100–115.

19. Yanxin Lu, Joe Warren, Christopher Jermaine, Swarat Chaudhuri, and Scott Rixner. 2015. Grading the Graders: Motivating Peer Graders in a MOOC. In *Proceedings of the 24th International Conference on World Wide Web (WWW '15)*. 680–690.

20. Debbie Morrison. 2013. Why and When Peer Grading is Effective for Open and Online Learning. http://onlinelearninginsights.wordpress.com/2013/03/09/why-and-when-peer-grading-is-effective-for-open-and-online-learning/. (2013). Online; accessed 2-November-2014.

21. Ellis B. Page. 1967. Statistical and Linguistic Strategies in the Computer Grading of Essays. In *Proceedings of the 1967 Conference on Computational Linguistics (COLING '67)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 1–13.

22. James W. Pellegrino, Naomi Chudowsky, and Robert Glaser (Eds.). 2001. *Knowing What Students Know: The Science and Design of Educational Assessment*. The National Academies Press, Washington, DC.

23. Nancy Pickett and Bernie Dodge. 2007. Rubrics for Web Lessons. http://webquest.org/sdsu/rubrics/weblessons.htm. (2007). Online; accessed 14-September-2015.

24. Karthik Raman and Thorsten Joachims. 2014. Methods for Ordinal Peer Grading. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '14)*. ACM, New York, NY, USA, 1037–1046.

25. E.F. Redish, M. Vicentini, and Società italiana di fisica. 2004. *Research on Physics Education*. Number v. 156 in International School of Physics Enrico Fermi Series. IOS Press.

26. Ken Reily, Pam Ludford Finnerty, and Loren Terveen. 2009. Two Peers Are Better Than One: Aggregating Peer Reviews for Computing Assignments is Surprisingly Accurate. In *Proceedings of the ACM 2009 International Conference on Supporting Group Work (GROUP '09)*. ACM, New York, NY, USA, 115–124.

27. Philip M. Sadler and Eddie Good. 2006. The Impact of Self- and Peer-Grading on Student Learning. *Educational Assessment* 11, 1 (2006), 1–31.

28. Martyn Shuttleworth. 2008. Double Blind Experiment. https://explorable.com/double-blind-experiment. (2008). Online; accessed 8-November-2014.

29. IEEE Intelligent Systems staff. 2000. The Debate on Automated Essay Grading. *IEEE Intelligent Systems* 15, 5 (Sept. 2000), 22–37.

30. Hoi Suen. 2014. Peer assessment for massive open online courses (MOOCs). *The International Review of Research in Open and Distributed Learning* 15, 3 (2014).

31. Germaine L. Taggart, Sandra J. Jones, and Judy A. Nixon. 1999. *Rubrics: A Handbook for Construction and Use*. Rowman & Littlefield Education.

32. Keith Topping. 1998. Peer Assessment between Students in Colleges and Universities. *Review of Educational Research* 68, 3 (1998), pp. 249–276.

33. Stephan Trahasch. 2004. From peer assessment towards collaborative learning. In *Frontiers in Education, 2004. FIE 2004. 34th Annual*. F3F–16–20 Vol. 2.

34. Gregor Ulm. 2012. A Critical View on Coursera's Peer Review Process. http://gregorulm.com/a-critical-view-on-courseras-peer-review-process/. (2012). Online; accessed 3-November-2014.

35. Tim Vogelsang and Lara Ruppertz. 2015. On the Validity of Peer Grading and a Cloud Teaching Assistant System. In *Proceedings of the Fifth International Conference on Learning Analytics And Knowledge (LAK '15)*. ACM, New York, NY, USA, 41–50.

36. Audrey Watters. 2012. The Problems with Peer Grading in Coursera. https://www.insidehighered.com/blogs/hack-higher-education/problems-peer-grading-coursera. (2012). Online; accessed 8-November-2014.