A Systematic Quantitative and Qualitative Analysis of Participants' Opinions on Peer Assessment in Surveys and Course Forum Discussions of MOOCs

Thomas Staubitz, Christoph Meinel
Hasso Plattner Institute, University of Potsdam, Potsdam, Germany
{thomas.staubitz, christoph.meinel}@hpi.de

***Abstract*—Peer assessment has become a regular feature of many MOOC[1] platforms and also has potential for other contexts where learning and teaching are required to scale because of growing numbers of students. Where manual grading is not possible due to the large number of submissions and the tasks to be assessed are to complex or open-ended to be assessed by machines, peer assessment offers a valuable alternative. However, particularly in the context of MOOCS, courses featuring peer assessments often have lower completion rates. Furthermore, participants with negative expectations and opinions about this form of assessment are generally 'louder' in their communication with the teaching teams than their counterparts who respond more positively. We have, therefore, set out to establish a broader understanding how the participants perceive the appropriateness and effectiveness of peer assessed tasks, and the quality of the received reviews on the *X1*, *X2*, and *X3*[2] MOOC platforms. For this purpose, we have conducted post-course surveys in a large number of courses that included peer assessments. Additionally, we analyzed the discussions in the forums of these courses as the post-course surveys often are biased due to the low proportion of unsuccessful participants that are still around at the end of a course.**

Keywords*—*MOOC, Peer Assessment, Project-based Learning, Active Learning.

I. Introduction and Motivation

Generally, we speak about peer assessment (PA) when one course participant assesses the work of one or more other course participants. PA can be used for summative or formative assessment or in a combination of both. PA has been introduced to MOOCs in 2013 in Scott Klemmer's course on human-centered interaction design on the *coursera* platform [7]. PA is particularly relevant in MOOCs as it is currently the only option to assess complex, open-ended, or creative tasks at scale. Nevertheless, it also comes with some challenges. Particularly, a certain mistrust of some participants in the judgmental capabilities of their peers and comparably low completion rates in courses that heavily rely on peer assessed tasks to measure the success of a participant [5]. On our MOOC platforms, PA was introduced in 2014 and by now has been employed in more than 60 courses. When PA was employed in the first courses, the participants' opinions have been strongly diverging. As often, the negative voices have been way louder than the positive ones.

[1]Massive Open Online Course
[2]The platforms' names have been obscured for double-blind review

The acceptance of a certain form of assessment strongly relies on the participants' perception whether the grading is trustworthy and how effective the type of exercise is for the learning outcome. We have, therefore, added questions to determine the participants' perception of PA to the post-course surveys of the courses that included peer assessments. As the post-course surveys often are biased due to the low proportion of unsuccessful participants that are still around at the end of a course, we additionally set out to do a thorough and systematic analysis of the conversations in the course discussion forums of these courses.

II. Related Work

Individualized feedback is an integral part of education. This feedback is particularly important in more complex exercises and assignments. But, exactly for these exercises it cannot be delivered in an automated form [10] and neither can it be delivered manually by the instructors of a MOOC as the number of participants is too high. Peer assessment is employed in today's MOOCs as an attempt to address these issues. Benefits of PA include improvement of higher-order thinking skills, consolidation of topical knowledge, and individualized feedback for each participant [3], [6]. PA as a form of educational assessment is very flexible and can be used to serve summative and formative assessment alike [12]. It is a quite common application of formative PA that students are reviewing each other's work and are giving written feedback [4], [11]. Summative PA of fellow students' work, however, is a more complicated matter and requires careful guidance by a teacher, since grades should be fair, consistent, and comparable for all students [1], [11], [2]. Feedback generally is perceived useful by students. Some studies even suggest that some students take comments from their peers more seriously than teacher comments [2], [8], [9].

III. Peer Assessment on X1 and X2

On our platforms a PA consists of three mandatory and two optional steps (see Figure 1). After accepting the honor code, the participants work on their task and submit the result. Depending on the course and the task, the participants have about two to six weeks to complete the PA. In the second step, they have to review and grade a certain number of peers. How many peers have to be graded is defined by the instructors—the recommended minimum is three, the optimum is five. Grading rubrics are defined by the teaching team individually for each task. Additionally to the summative assessment, the participants

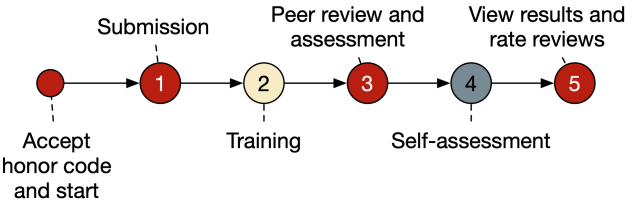

Figure 1: Peer assessment steps on X1 and X2.

are encouraged to provide a formative assessment of their peers' work. In the final step, the received feedback can be rated by the participants. Furthermore, optional training and self-assessment steps can be added.

IV. Data Set and Methodology

We have approached the questions from two different angles. The sources for the first data set are the post-course surveys that have been conducted in several courses including peer assessments on X1 and X2. These surveys will be examined in Section V. Post-course surveys, however, have an inherent bias, as mostly participants who have passed the course will still be around at the end of the course to provide feedback. We have, therefore, additionally conducted a quantitative/qualitative analysis of the discussions in the course forums of the relevant courses. The research was restricted to those courses that already had been finished when this examination started. Since then, further courses including peer assessments have been conducted on our platforms. These have no more been considered as

1) The analysis of this data comes with a high workload. Including more and more course data would have turned out to be an endless effort.
2) Fundamentally different results were not to be expected.

We searched the data for a list of keywords, such as "peer" or "assessment" and verified each match manually. The matching posts—and, if reasonable, the posts adjacent to the matches—have been examined in detail. The results of this study will be discussed in Section VI.

Table I gives a general impression about the number of courses that have been included in this study. It shows the number of started (C-2) and completed (C-3) submissions, the peer assessment completion rate (C-4), the amount of active participants[3] (C-5), and the percentage of active participants that have engaged in the peer-assessed assignment (C-6).

In the following, we list some very basic background information on the examined courses on our platforms. On X2 most of the courses and their assignments were about IT topics. Starting with UML-modeling in several iterations of the *javaeinstieg**[4] course, business process modeling in *bpm2016*, small HTML/CSS/Javascript projects in *homepage2016*[5] and two iterations of *webtech**, to a Java-programming project in several

[3]Active as in: the participant has visited at least one course item.
[4]Object-oriented programming for beginners
[5]A course targeting pupils.

Table I: Peer Assessments on X1, X2, X3
C-1: # of courses containing a peer assessment
C-2: # of started peer assessments (total)
C-3: # of completed peer assessments (total)
C-4: % completion rate
C-5: # of active users at course middle (total)
C-6: % of active course participants engaging in peer assessed assignment
The amount of submissions per course ranges from 4 to 4000.

	C-1	C-2	C-3	C-4	C-5	C-6
X1	33	23393	18408	79%	137906	17%
X2	26	17233	9702	53%	48112	36%
X3	3	371	271	73%	1504	25%

iterations of *javawork**[6]. Additionally, we had some courses with less tech-oriented tasks such as setting up a contract in *it-recht2016*[7], interviewing and observing in two iterations of *insights**[8], or creating a business model in *startup2016*. The courses *javaeinstieg-mint-ec-2018* and *javaeinstieg-schule-2019* are basically just iterations of *javaeinstieg**, but they have been stretched in length to decrease the weekly workload to better fit them in the context of schools.

The examined courses on X1 mostly had a background in the context of innovation. Business model innovation (*bmi*), design (*dfnd**, *dafie1*), design research (*dr**), Design Thinking (*dt**), digital talent management (*dtm1*) Fewer courses and assignments had a more tech-oriented background, such as an introduction to SAP Fiori (*fiori1* and *fiux**). Furthermore, we had courses on copy writing (*cwr**) and Internet of Things (*iot**), and a few others. The course *java1* is basically a translation of the *javaeinstieg** courses on X2 to English.

V. Peer Assessment in Post-Course Surveys

On X1 each course contains a post-course survey asking the participants about their satisfaction with the course. In total, about 30 of these surveys in courses that included a peer assessment have been evaluated. The questions to be answered by examining these data are:

- How appropriate is peer assessment considered to be for the assessment of complex tasks?
- How effective in terms of the learning outcome is peer assessment considered to be, compared to other elements of the platform, such as video and quizzes, etc.?
- How useful is the peer assessment's training phase considered considered to be?
- How is the review quality perceived by the participants?

A. Perceived Appropriateness of Peer Assessment to grade Hands-on Tasks

The most basic question to be asked in this context is if the participants generally accept peer assessment as an appropriate measurement of their performance. To broaden the spectrum,

[6]Follow-ups to the introductory courses
[7]IT-Law
[8]Design Thinking

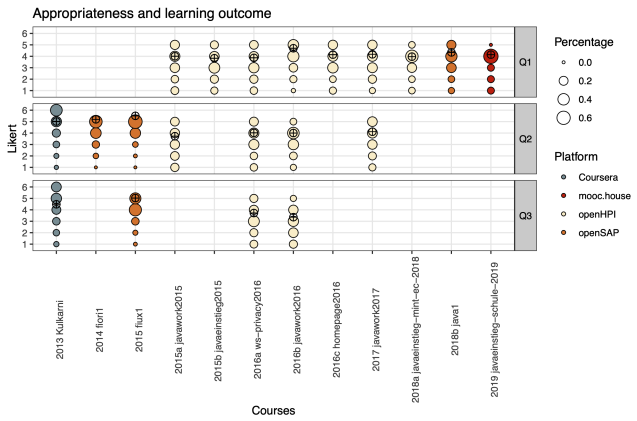

Figure 2: Q1: Is peer assessment an appropriate way to assess such tasks? Q2: Did you learn sth. by reviewing the work of your peers? Q3: Did you learn sth. by reviewing your own work? The results of Kulkarni et al. [7] serve as a reference.

we compared our results to the results of a similar survey by Kulkarni et al. [7] (see Figure 2). Unfortunately, not all surveys contained the same set of questions, a fact that is responsible for the gaps in the data. The crosshair in Figure 2 marks a normalized coefficient that eliminates the differences caused by the different Likert scales of the surveys. The majority of the participants perceives the peer assessment as an appropriate tool to grade hands-on tasks in the given contexts. The same applies for the perceived learning outcome of reviewing the work of peers and to a lesser extent to reviewing their own work. The values are generally a little lower in the more tech-oriented courses on X2 than in the more design/innovation-oriented courses on X1.

The courses *javaeinstieg-mint-ec-2018* and *javaeinstieg-schule-2019* are particularly interesting in this context as they have addressed a very specific target group: high-school pupils (16-19 years old). While the number of participants in the other examined courses are in the hundreds or even thousands, these particular courses were rather small with only a few hundred participants. As Figure 2 shows, there are no differences to the larger courses that are targeting a, generally, more adult audience[9].

The comments in the surveys' free text questions, indicated that the participants often mixed up their opinion on the methodology of peer assessment with their opinion of the given task. In *webtech2017*, the questions have, therefore, been refined. First, the participants have been asked what they, generally, think about hands-on tasks in the MOOC and then have been asked about their opinion on grading such tasks by the means of peer assessment.

Figure 3 shows that the vast majority—close to 90%—approves the possibility to work on hands-on tasks. Another question in the same survey, however, revealed that only ~45% actually also have the time to make use of this opportunity.

[9]The courses *javaeinstieg2017*, *javaeinstieg-mint-ec-2018*, and *javaeinstieg-schule-2019* are more or less identical. The only difference is that the courses that are offered in the school context run over a significantly longer timeframe ($\tilde{3}$ months instead of 4 weeks) with no additional content

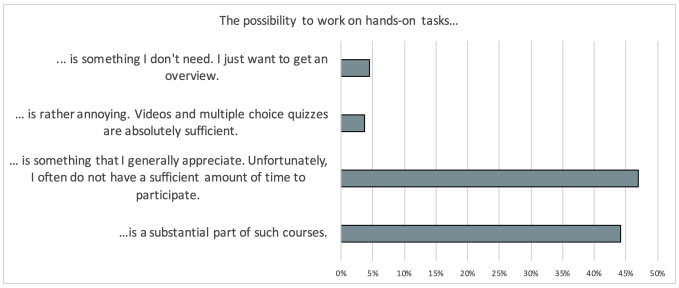

Figure 3: How important is the possibility to work on hands-on tasks in MOOCs. Survey in *webtech2017*, *n*=1016

About 89% of the participants also think that the peer assessment is a proper way to grade such hands-on tasks. 36% of the participants also sees an additional value for their learning in assessing the work of their peers and the reviews that they have received from their peers. Another 14%, considers peer assessment to be a proper tool for that purpose and received appropriate reviews, but didn't see the additional benefit in assessing the work of the others. Of the 11% who see peer assessment as problematic, less than 1% would be willing to **pay** for an alternative assessment by an expert. See also Figure 4.

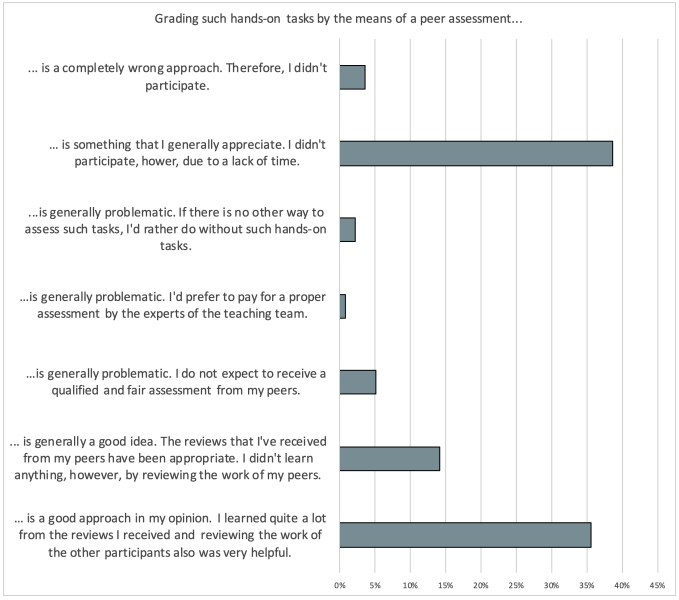

Figure 4: The participants opinion about peer assessment as a means to grade such hands-on tasks. Survey in *webtech2017*, *n*=982

B. Perceived Effectiveness of Peer Assessment to grade Hands-on Tasks

The one particularly interesting question in this context, is about the perceived effectiveness of the different learning elements for the participant's learning. Multiple selections were possible and not all courses provided all answer options.

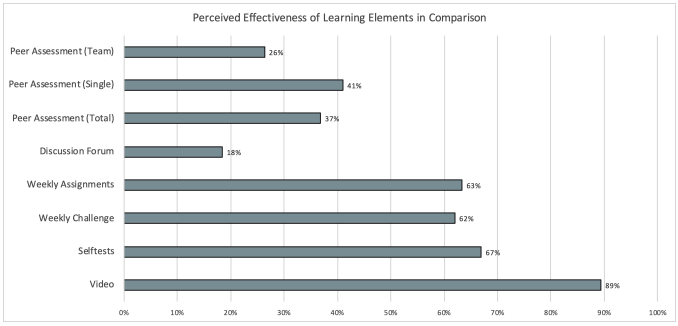

Figure 5: Perceived effectiveness of peer assessment compared to other course elements: Which learning elements did you find effective for your learning in this course? Multiple selections have been possible. Not all answer options have been available in all surveys. All percentages are relative to the respective option's *n*.) Video, Weekly Assignments, Discussion Forum: n=14173, 23 courses. Self-tests: n=10389, 13 courses. Weekly challenge: n=3885, 7 courses. Peer Assessment(total): n=9651, 18 courses (12 single user, 6 team)

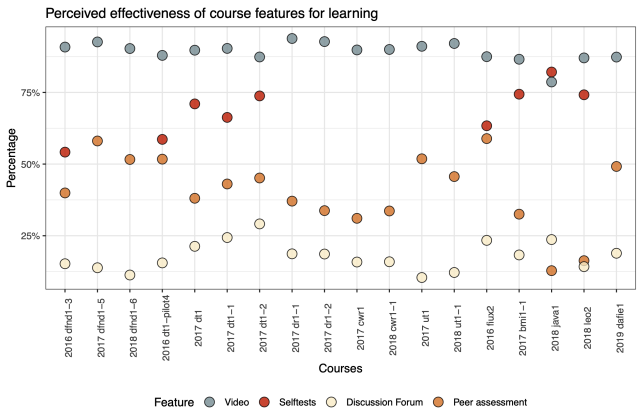

Figure 6: Perceived effectiveness of peer assessment compared to other course elements (same survey question and *n*'s as in Figure 5 but no differentiation between team and single peer assessments). The courses are ordered by course iterations and date. Multiple selections have been possible.

Figure 5 shows the summarized results for this question. Regarding the results of the peer assessment, it has to be considered that, in the examined courses on X1, only about 17% of the course participants in total have engaged in a peer assessed assignment, while a way higher percentage of participants has learned with videos, self-tests, or weekly assignments. This taken into account, it is safe to assume that the perceived effectiveness, for those that actually have used it, is higher to some extent.

The same data is shown course by course in Figure 6 in more detail, but focussing on the most common "traditional" features: videos and self-tests in comparison to the more social features: discussion forum and peer assessment. The data here is ordered by course and course iteration, starting on the left with three iterations of the course "Software Design for Non-Designers", followed by three iterations of "Developing Software Using Design Thinking", two iterations of "Basics of Design Research", two iterations of "Copywriting: Improve User Experience One Word at a Time", two iterations of "Basics of Design Testing", and five courses for which only data for only one iteration was available. There is no clear trend whether the perceived effectiveness is increasing or decreasing towards newer iterations of a course. The most obvious explanation for an increase in perceived effectiveness is that the instructors have improved the design of the task and their communication strategy. One explanation for a decrease in a later iteration is that the teaching team didn't put the same effort in the communication with the participants as some sort of routine was established. Another explanation for a shift in both directions could be a different composition of the courses' learning communities. For pilots, such as *dt1-pilot4* many other rules apply, particularly, their significantly smaller number of participants—often hand-selected from employees, students, or partner organizations—so that they cannot really serve as a reference for the regular courses. A closer look at the *dr1** and the *ut1** courses, where the perceived effectiveness is decreasing doesn't reveal a lot. The iterations are almost identical in content, the teaching teams are also more or less the same, the number of participants doesn't differ too much, the announcements are identical, and the forum participation is similar as well. There is no obvious reason to be found. Once there is a third iteration available, it will become interesting to dive deeper if the trend continues and does not turn out to be just a local minimum. Just to make sure, we have also verified if there are correlations between the features but have not found any. The very low results in *java1* and *leo2* can easily be explained by the low value (in terms of points) that the peer assessed task had in these course contexts and the resulting low participation rate in this activity. Furthermore, these two courses are on tech topics while all other examined courses are more design and innovation oriented. Particularly in *java1*, practical programming exercises have been offered additionally to the more common features, which have been received very well and obviously are an ideal fit for this particular type of courses.

C. Perceived Usefulness of Training Phase

The peer assessment system of the X1 and X2 platforms allows to add an additional training phase. In most cases, the teaching teams, however, have refrained from adding this phase as it faces the instructors with a significant amount of extra work. In the few cases where the training phase had been added, the participants have been asked in a survey about their opinion (see Figure 7).

While in *flux1* the training was perceived particularly good, the results have been less clear in the following courses. Determining the cause for this decrease is not trivial. The two most promising theories are

1) the participants already have been more experienced with peer assessment in the more recent courses and didn't need the training anymore,
2) the quality of the more recent trainings was less good than the quality of the earlier trainings.

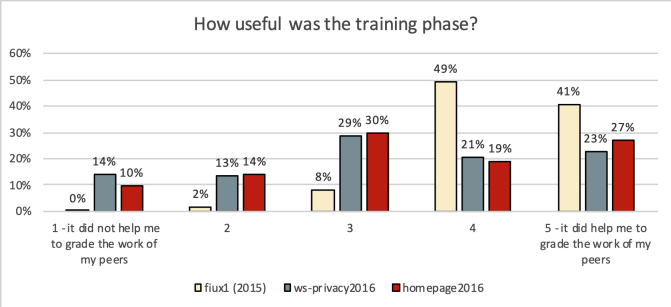

Figure 7: Surveys in *fiux1*, *n*=468, *ws-privacy2016*, *n*=344, and *homepage2016*, *n*=184

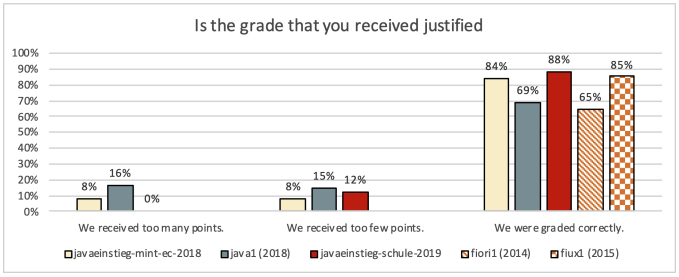

Figure 9: Justness of grade. Surveys in *javaeinstieg-mint-ec-2018*, *n*=26, *java1*, *n*=179, and *javaeinstieg-schule-2019*, *n*=8. In comparison *fiori1*, *n*=54 and *fiux1*, *n*=466

The quality of the training depends on the quality of a set of sample reviews that have to be provided by the instructors for about 10 (real) submissions of the actual task. Obviously the quality of those reviews might differ, but there is no evidence that the quality of the reviews has differed as much as the perceived usefulness of the training. On the other hand, many of the participants on the X1 and X2 platforms are enrolled in a number of courses and many of them also have accounts on both platforms. Particularly cross-platform, it is hard to determine if the participants in the latter courses have worked on previous peer assessments already. However, *homepage2016* featured only 22% first-time participants and *ws-privacy2016* even had only 13% first-time participants, which could be considered as an argument towards the theory that the participants have been more experienced.

D. Perceived Review Quality

The quality of the reviews and their perception/acceptance by the reviewed is the crucial factor in each peer assessment. In the post-course surveys of *ws-privacy2016*, *javawork2016*, and *homepage2016*, the participants have been asked how they perceived the effort that they, respectively their peers had put in writing the reviews. The results align with the results of Kulkarni et al. [7] in a similar survey (see Figure 8).

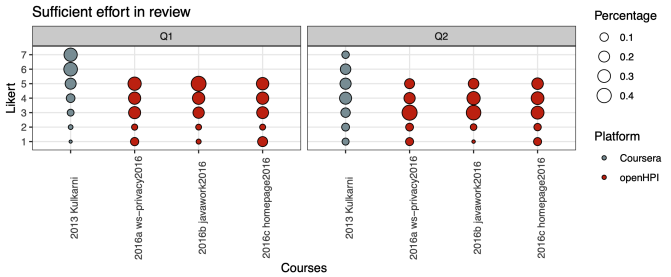

Figure 8: Review effort
Q1: I put sufficient effort in the grading of my peers' work.
Q2: My peers put sufficient effort in the grading of my work.
The results of the surveys on X2 are compared to the results of a similar survey by Kulkarni et al. [7]. Surveys in *ws-privacy2016*, *n*=349, *javawork2016*, *n*=81, and *homepage2016*, *n*=164

As expected, the participants' perception is biased. They tend to think that their own effort in reviewing and grading is higher as everybody else's.

The participants have also been asked if the grade they received is just. In *fiori1* and *fiux1* they just had the choice between "yes" and "no", in the more recent *javaeinstieg** courses "no" was further specified into "I received too many points" and "I did't receive enough points". Figure 9 shows the results. It has to be kept in mind that the *n* for the school courses (*javaeinstieg-mint-ec-2018*, *javaeinstieg-schule-2019*) is very small in comparison.

The majority (65% to 88%) of the participants perceived the received grades from the peer assessment as justified. Except for *fiori1*, more than 80% of the participants in all courses, found that they have received a fair amount of points or even more than expected. Again, it is interesting that the results for the school courses (*javaeinstieg-mint-ec-2018*, *javaeinstieg-schule-2019*) are well aligned with the courses targeting a more adult audience. Finally, the participants have been asked about the mechanism to rate reviews as helpful and providing bonus points for those who have received a good rating. In the post-course surveys of *fiori1* and *fiux1* about 80% of the participants agreed that the bonus points are a good way to motivate better reviews, about 90% agreed that the review rating is a good way to give feedback to the reviewer. Two years later, in 2016, the participants of *javawork2016* and *ws-privacy2016* have been asked a similar question; this time they had the possibility to answer on a five point likert scale. The results are still very good (see Figure 10). The less positive results in *ws-privacy2016* result from a strong bias that many participants had against the given task rather than from a negative attitude towards peer assessment as a form of grading. In another question in the same survey of this course, about 70% of the participants stated that they did not like the given task at all.

VI. Discussion Forum Analysis on Peer Assessment

In early 2018, when a sufficient amount of peer assessments in different contexts had been conducted on our platforms, we started to run a qualitative analysis of the course forum discussions in all courses containing peer assessments. We used the tool MaxQDA for a systematic, qualitative analysis of the forum data. First we defined a basic set of tags that we used to code the forum posts. We use the term *tag* to define a category, while the term *coding* will be used for the text snippets that

Figure 10: Motivation to write reviews.
Q1: Is the possibility to report reviews a sufficient tool to protect your work against inappropriate reviews?
Q2: Are you more motivated to write helpful reviews by the possibility to earn extra points for those reviews?
Surveys in *javawork2016*, *n*=94, *ws-privacy2016*, *n*=370

have been coded with a certain tag category. We started with tags, such as

- Good experience
- Bad experience
- Positive attitude
- Negative attitude

and added further tags whenever a new category became necessary. Each relevant post has been coded manually with these tags. MaxQDA allows a very fine grained coding. Multiple tags can be applied to the same text snippet. Figure 11 shows an example of a coded forum post. The basis for the evaluation have been the discussion forum exports of all courses containing a peer assessment at that point of time on the X2 and X1 platforms[10]. In these exports we searched for the terms "peer assessment", "peer grading", "peer assignment", and prominent terms in the task descriptions of each course's peer assessment. Then we coded the posts in the surrounding threads with the pre-defined tags when applicable; or created new tags whenever necessary. The values in the following figures represent the co-efficient of the found codings per tag category in relation to the total posts of the course. We're aware that the total word count of all forum posts would be the more exact reference value. We decided, however, to work with the count of posts as a close enough approximation.

[10]There have been about ten more courses on the mooc.house platform that also have used peer assessments. However, these courses did not have noteworthy forum discussions on the topic and, therefore, have not been included here.

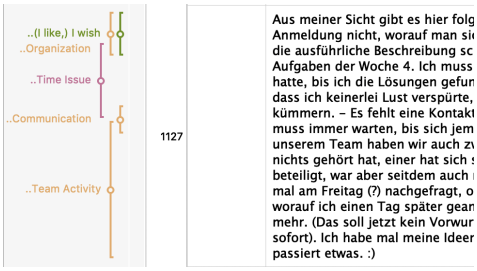

Figure 11: Example of a coded post, including overlapping tags.

When the peer assessment has been introduced on our platforms, few but very "loud" participants, protested strongly. Therefore, we started this endeavour to establish a broad view on the general opinion and mood of the bulk of participants that goes beyond a mere post-course survey. In the examined courses[11], about 2-20% of the participants have been actively posting in the course forums. The average in the examined courses on X2 was about 7% active posters. It was insignificantly lower on X1. About 25-50% of the participants are passively consuming the discussion posts. We examined about 70.000 posts in 30 courses on X1 and X2 and had more than 5.000 hits on the word "peer". The percentage of posts that contain the search term ranges from 1 to 20% within the courses. Often, the whole thread that contained the post with the hit dealt with the peer assessment, but the word itself was not repeated in the other posts. So, while in total, about 8% of the forum posts contain the search term, we can assume that about 10-15% of the posts are related to the peer assessment.

First, we examined the participants experience and attitude. Statements have been coded as "good experience" whenever they deal with situations that the participant really has encountered in the current or in previous peer assessments. Statements have been tagged as "positive attitude" whenever they mention a participant's expectation to what might happen in the current or future peer assessments. The same applies for the negative variant. The courses in Figures 12-16 are ordered first by platform and then alphabetically, to visualize differences between several course iterations. The first twenty courses have been conducted on X1, the rest of the courses have been conducted on X2. In Figure 12, the positive statements are color-coded in green, the negative statements are color-coded in red. In the other figures, the courses on X1 are color-coded in burgundy while the courses on X2 are color-coded in grey. The values represent the number of found codings per tag category in relation to the total posts in the course's forum.

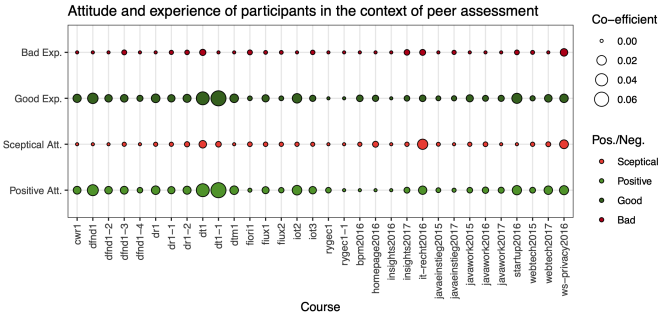

Figure 12: Participants' attitude and quality of experience in the context of peer assessment

In general, the positive statements outweigh the negative statements by far in the vast majority of the courses, both in experience and attitude. For 9 of the 32 examined courses, Figure 12 is hard to read as all of the values are very small. Therefore, we listed these values in Table II for more clarity. The table shows that we have a positive tendency here as

[11]The selected courses all contained a peer assessment. All courses that contained a peer assessment at the time the research was started and featured a sufficiently interesting amount of forum discussions have been selected.

Table II: Details for Figure 12. Negative and positive experience and attitude towards peer assessment. Bold: higher value

	Att.(-)	Att.(+)	Exp.(-)	Exp.(+)
fiori1	**0.0012**	0.0004	**0.0016**	0.0012
fiux2	0.0007	**0.0024**	0.0002	**0.0013**
bpm2016	**0.0006**	0.0000	0.0000	**0.0058**
homepage2016	**0.0038**	0.0000	0.0000	**0.0048**
insights2017	0.0008	**0.0038**	0.0030	0.0030
it-recht2016	**0.0249**	0.0134	0.0038	**0.0057**
javaeinstieg2015	0.0003	**0.0014**	0.0000	**0.0015**
javaeinstieg2017	0.0000	**0.0005**	0.0006	**0.0020**
webtech2015	**0.0007**	0.0034	0.0000	**0.0097**

well. It also shows that even in courses where the attitude is rather skeptical (*fiori1*, *bpm2016*, *homepage2016*, *it-recht2016*, *webtech2015*), the experience is predominantly positive.

Figure 13 shows which technical issues have been strongly discussed in the course forum and the context of peer assessment. As the technical issues in each course are well understood, we used this examination as a sanity check for the whole endeavour. The question is if the forum discussions—and the approach to code these discussions with tags—can serve as a barometer for the issues in a course. Based on the comparison

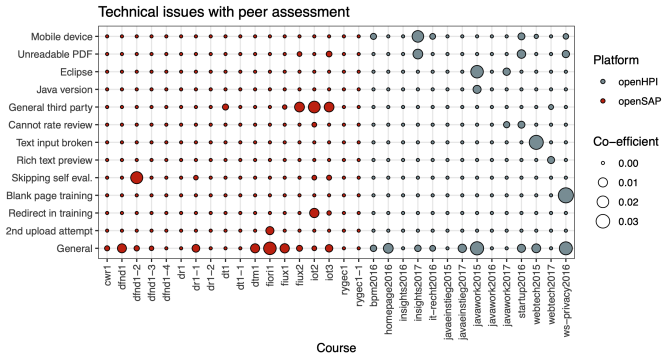

Figure 13: Sanity check of the examined data against known technical issues.

of the codings in the forums and the well-known technical issues, the question can be answered with a definitive *yes*. For example, *fiori1* was the first course on any of our platforms that included a peer assessment. The feature itself was still in its infancy and many teething troubles needed to be fixed. This is reflected clearly in a high amount of general problems being discussed in the forum. Another example is *dfnd1-2*. During the peer assessment in this course, a bug emerged that led to a blank page whenever a participant tried to skip the optional self-assessment. Again this is clearly reflected in the forum discussions. In *webtech2015* the task for the participants has been to implement a given webpage design in HTML. Due to a miscommunication by the teaching team, many participants did not upload their HTML code as a text file, but pasted it into the peer assessment tool's text input field. Unfortunately, there, the HTML was not escaped due to the wrong indentation of a line of code of the platform's source code. The issues listed so far, affected **all** participants of the peer assessment; the following issues affected only a (smaller or larger) **subgroup** of the peer assessment participants.

In *fiux2*, *iot2*, and *iot3* X1 employed third party tools for various purposes, in *javawork2015* the participants have been introduced to Eclipse—a Java programming IDE[12]—for the first time. In comparison, in *javawork2016*, which addressed a more experienced audience[13], the participants have not shown any problems with Eclipse. All the described issues are exactly reflected in the codings of the course discussions[14].

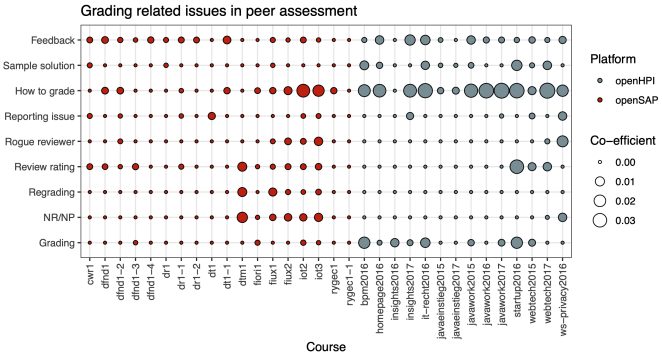

Figure 14: Forum discussions about grading issues.

We can, therefore, assume that a similar accuracy can be found in the discussions coded with grading issues or basic misunderstanding of the peer assessment concept. Figure 14 shows the discussions that have been coded with several issues in the context of grading.

- Feedback–Remarks about the received reviews.
- Sample solution–Discussions about the correctness of the given sample solution or training examples. Also requests for a sample solution when none has been provided (in this case often in combination with a complaint about the received grade).
- How to grade–Discussions about how certain aspects of a peer's solution should be graded or have been graded. It was often not quite easy to make a clear distinction between, sample solution issues, how to grade issues, grading issues, or even task related issues. How-to-grade issues are more about understanding what the teaching team meant with the task. Grading issues are mostly mere complaints about the perceived injustice of the received grades.
- Reporting issue–Questions about reported submissions or reviews. The peer assessment system allows to report submissions as well as reviews for a variety of reasons (e.g. plagiarism, offensive language, etc.).
- Rogue reviewer–The reviewer did not play to the rules. Example: s/he refused a review due to the submitted format, although the format fit the requirements (e.g. pdf was requested, pdf was submitted but the reviewer

[12]Integrated Development Environment

[13]Except for the name and the facilitators, the courses had few in common, while javawork2015 addressed beginners and introduced them to the Eclipse IDE, javawork2016 addressed more advanced users and introduced them to test-driven development with JUnit.

[14]Most of the listed bugs, generally, have been fixed before the respective deadline of the peer assessments. In a few cases, a workaround has been developed as a quick solution while the actual bug has been fixed later on. In none of the cases, did the bugs have an influence on the participants grades.

has security concerns). The reviewer gave substantially less points than other reviewers.

- Review rating–Complaints about missing review ratings, or the review rating process in general. The peer assessment systems asks the peers to rate the received reviews. The reviewers are awarded additional bonus points for reviews with a good rating. Review rating is optional and participants often just forget to rate the received reviews. Another common reason that reviews are not rated is that the participant who submitted the work that was reviewed has dropped out in the meantime.
- Regrading–Misconceptions about re-grading. Under certain conditions, the peer assessment system allows the participant to ask for a re-grading. The most important of these conditions is that there are significant differences in the amount of points that a submission has received from each reviewer. Requested re-gradings are re-assessed by the teaching team.
- NR/NP–The system follows the simple rule "No reviews, no points". No matter how good the work is that has been submitted by the participant, if s/he has not submitted the required amount of reviews, s/he will receive zero points. The codings here are mostly complaints about this practice.
- Grading issues–The grading was unfair. I deserve a better grade.

"Sample solution" discussions and "how-to-grade" issues are more common in courses on X2 than in the X1 courses. In *webtech2017* one of the tasks for the javascript exercise was to write a function that uses Regular Expressions (RegEx) to check if a passed email address is valid. As this can get quite complex, we simplified the task, so that only a subset of email addresses had to be recognized. Although this has been clearly explained in the requirements, many participants did just not believe it and reasoned along the lines of: the course is offered by an elite institute. In elite institutes, questions are not that easy. Therefore, the description in the task requirements has to be wrong. Several straight-forward answers by members of the teaching team have not convinced the discussants, that even in an elite institute some questions might be simple when they are addressing an inexperienced audience. Another possible reason could be that some of the peer assessments contained a training phase whereas others did not. Therefore, the point-biserial correlation between the how-to-grade codings and the existence/absence of a training phase has been calculated. A correlation does not exist here (n=32, r=0.045, p=0.805). Having a closer look at the courses with many *how-to-grade* codings, the most probable explanation, therefore, seems to be the clarity of grading rubrics and instructions.

This assumption is not unrealistic as on the X1 courses a team of well-trained and experienced instructional designers, copywriters and native speakers conducts several levels of quality control, the approach of the X2 teaching teams is generally more ad-hoc and learning by doing. There is, however, a significant correlation between the how-to-grade coefficient and the complaints about the received grade (Pearson-correlation: n=32, r=0.44, p=0.01).

Rogue reviewers, generally, seem to be less of a problem than we expected when the system was developed. It seems that we created a self-defeating prophecy, by taking care of rogue reviewers through mechanisms such as review rating and review reporting.

In the context of review rating, one of our assumptions was that the participants simply forget to rate the reviews they've received. Therefore, we started to send course announcements to remind the participants not to forget the rating. In all three courses on X2 where we have particularly many complaints about missing reviews, no particular reminder has been sent to the participants. A double-check on a handful of courses with particularly low complaints about missing ratings revealed that in these courses such announcements have been sent. So we can assume that reminding the participants on this task is a sufficiently successful measure that should definitely be taken to ensure the most optimal outcome.

Now we're looking at organizational issues, issues with the user interface, time issues, and the participants' misconceptions. Figure 15 shows the analysis of the posts in these contexts. Discussions on organizational issues around the peer assessments most often addressed the teaching teams' communication policy.

- It was announced too late that course success (in terms of certificates) will heavily rely on peer assessment.
- Due to the peer assessment, the actual course duration was longer than the communicated course duration.
- Deadlines have been missed due to missing announcements.

Discussions that have been coded as "time issues", include complaints about the task's time-frame that was designed too tight by the teaching team, statements about the lack of time on the participant's side due to new work assignments or a change in the family situation, illnesses, etc. In *javawork2015* the main time issue resulted from 1. postponing the course from Spring to Autumn, and 2. a lack of communication that the peer assessment required more time than the actual workshop runtime of two weeks. *ws-privacy2016* on the other hand featured two peer assessments in a two week workshop. Other than in *javawork2015* the actual workload was rather low, the organizational overhead that comes with a peer assessment, however, was too much for this short time frame. In addition, the participants heavily disliked the given task.

"User interface (UI) issues" and "misconceptions" have

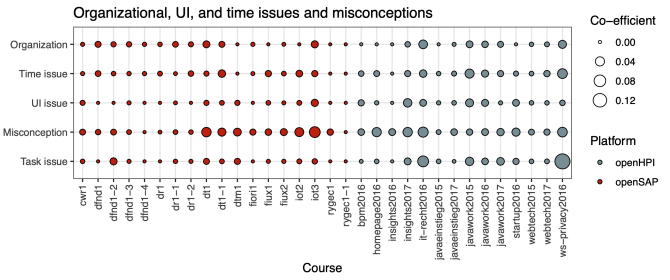

Figure 15: Forum discussions about general misconceptions of peer assessment and MOOCs or about particular misunderstandings on organizational issues of the current peer assessment.

often similar symptoms and are not always easy to distinguish. Statements have been coded as "UI issues" when they directly referred to a UI element and, ideally, included a suggestion for improvement. They have been coded as misconceptions when the problem can either be solved by more detailed, more frequent, or more precise information, or when they've been obviously the result of a certain resistance to learning on the participant's side. Frequent misconceptions covered the amount of reviews that have to be written, how to find the peer assessment or the grading rubrics, what happens when a review is not finished in time. The most basic misconception, however, is that some participants do not understand the underlying concept of peer assessment and/or the necessity for peer assessment as a scalable grading mechanism that allows to provide open ended, complex tasks in such courses. Additionally, there is a lack in understanding of the economics of such courses, which at least on our platforms still are offered completely for free.

Finally, Figure 15 shows where the participants had issues with the task itself. *it-recht2016* and *ws-privacy2016* caused the biggest trouble here for different reasons. The course *it-recht2016* covered topics such as contracts, patents, copyrights, privacy, etc. and addressed software engineers and IT entrepreneurs. The task has been to set up a simple contract. The grading rubrics only checked for the absence or presence of certain terms in the text that was submitted by the participants. As a result, some participants who more or less copied the text from the course material, received full points, while others who approached the task in different way and rather tried to apply what they've learned, received less points. The strongest criticism was, that, due to the given rubrics, the task could have been implemented as a multiple choice test with much less effort for the participants. Another criticism was that the target group, even having completed a course like this, hardly would be able to set up legally waterproof contracts on their own, which rendered the given task inept in their eyes. In a later iteration of this course, which is not part of this evaluation, the task of writing a contract has been replaced by finding errors in a given contract. This approach worked much better. The course *ws-privacy2016*, on the other hand, covered topics such as privacy issues in social media and how to protect yourself and your accounts. The given task was to write an essay describing a privacy violation scenario that results from uploading an image to a social network. The task was divided in two parts 1. doing research and collecting information, 2. based on this research, write a short story about a real or fictitious privacy violation that followed the requested pattern. In this case, the participants had three major concerns/issues with the task.

1) Many did not understand that in part 1, they only had to do the preparation for the actual task in part 2. They handed in their final result right away (and then complained about a lack of time in task 1 and that they did not know what to do in task 2.)
2) Many complained about the form of the task itself, that they had to write an essay in English language. The course itself was also offered in English language. One of the grading rubrics asked for a basic correctness in grammar and orthography. Although, this rubric only provided a marginal amount of points, it caused a substantial amount of forum activity. In the participants' opinion, the task was completely inapt for a "tech" course. The approach was too academic for the bulk of the audience.
3) Two peer assessments during the relatively short timeframe of the course was too much. Furthermore, due to the restricted time frame, they did not have enough opportunity, to use the reviews of the first task as an input to improve the second task.

Finally, Figure 16 shows the teaching teams' communication activities. It has been differentiated between "Announcements" and "Answers". Announcements refer to pro-active posts by the teaching team in the discussion forum. Answers refer to re-active posts by the teaching team.

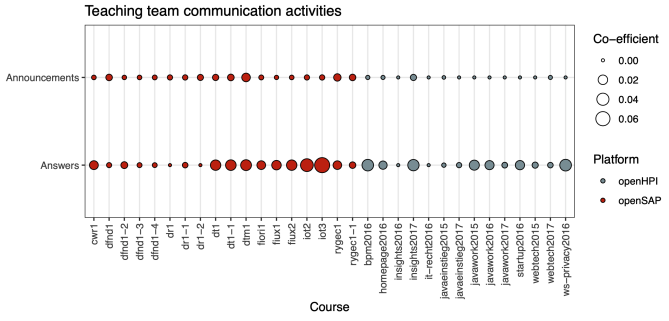

Figure 16: Teaching team activities in the discussion forums.

To round off the evaluation, the correlation between the teaching teams' communication activities and some of the other codings have been calculated. Particularly, we wanted to know if the number of announcements by the teaching team has reduced the amount of answers and if there is a correlation between the teaching team's forum activity and a positive experience of the participant. However, we have not found any significant correlation (see Table III for details). Comparing Figures 15 and 16 shows that there seems to be a correlation between misconceptions of the participants and the amount of teaching team answers. This indicates that the learners receive proper support from the teaching teams.

VII. Future Work

Since we evaluated this set of forum discussions, a new generation of MOOCs containing peer assessments has been launched and conducted. Although we stated in this paper's first section that we do not expect many new insights from running the same analysis on this data, at least we would be interested

Table III: Correlation between teaching team forum activities and positive experience, sum of all issues. (No statistical significant correlation.)

n = 32	**Announcements**	**TT answers**
Announcements	—	r=0.17, p=0.34
TT answers	r=0.16, p=0.34	—
Pos. Exp.	r=0.29, p=0.10	r=0.20, p=0.26
Sum of issues	r=0.30, p=0.08	r=0.24, p=0.18

to confirm this assumption. Starting the manual analysis over again, however, would cost a lot of time and resources. A more interesting approach would be to use the coded posts to train a machine learning algorithm, so that this could be done with less additional effort whenever another course is finished.

VIII. Conclusion

The results of our analysis show that peer assessments, as a tool to enable complex tasks within MOOCs or other forms of large scale educational systems, are perceived very well among the participants. Furthermore, a large number of participants considers such tasks to be essential for such courses. On the other hand, many participants stated that, although they consider such tasks important, they do not have enough time to work on them. The range of topics and target groups in the examined courses was wide enough to consider these findings to be general. The results were similar in courses that addressed school children, developers, and professionals up to the management level. The results are essential, as currently, there is no alternative to grade such assignments at scale. It is important to distinguish between the assignment itself and the grading of the assignment. Courses that rely on peer assessed tasks to determine the course outcome, often have lower completion rates. We have shown that this is more due to the complexity of the task and the required time effort than to the grading by peer assessment. The decision if a peer assessed task is to be added to a course is, therefore, strongly depending on the instructors' objectives for the course. Completion rates, however, cannot be the **only** ultima ratio for the design of a MOOC. We will have to accept the fact, that the completion rates in courses with a simpler form of examination is higher than in courses that request the participants to get out of their comfort zone.

References

[1] J. R. Baird and J. R. Northfield. *Learning from the PEEL experience*. School of Graduate Studies, Faculty of Education, Monash University, 1995.

[2] J. Fermelis, R. Tucker, and S. Palmer. Online self and peer assessment in large, multi-campus, multi-cohort contexts. In *Providing choices for learners and learning Proceedings ASCILITE Singapore 2007*, pages 271–281, 2007.

[3] E. Gehringer. Strategies and mechanisms for electronic peer review. In *Frontiers in Education Conference, 2000. FIE 2000. 30th Annual*, volume 1, pages F1B/2–F1B/7 vol.1, 2000.

[4] J. Hamer, K. T. Ma, and H. H. Kwong. A method of automatic grade calibration in peer assessment. In *of Conferences in Research and Practice in Information Technology, Australian Computer Society*, pages 67–72, 2005.

[5] K. Jordan. "mooc completion rates: The data," the katy jordan mooc project. http://www.irrodl.org/index.php/irrodl/article/view/2112/3340. [Online; accessed 26-04-2019].

[6] L. Knight and T. Steinbach. The pedagogical foundations of massive open online courses. *Journal of Information Technology Education*, 10:81–100, 2011.

[7] C. Kulkarni, K. P. Wei, H. Le, D. Chia, K. Papadopoulos, J. Cheng, D. Koller, and S. R. Klemmer. Peer and self assessment in massive online classes. *ACM Trans. Comput.-Hum. Interact.*, 20(6):33:1–33:31, Dec. 2013.

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

[9] K. Reily, P. L. Finnerty, and L. Terveen. 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, pages 115–124, New York, NY, USA, 2009. ACM.

[10] H. Suen. Peer assessment for massive open online courses (moocs). *The International Review of Research in Open and Distributed Learning*, 15(3), 2014.

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

[12] S. Trahasch. From peer assessment towards collaborative learning. In *Frontiers in Education, 2004. FIE 2004. 34th Annual*, pages F3F–16–20 Vol. 2, Oct 2004.