The State of Peer Assessment: Dimensions and Future Challenges

Usman Wahid, Mohamed Amine Chatti, Ulrik Schroeder
Learning Technologies Research Group (Informatik 9), RWTH Aachen University
Aachen, Germany
{Wahid; Schroeder}@cil.rwth-aachen.de; Chatti@informatik.rwth-aachen.de

Abstract— **Modern day education and learning has moved on from brick and mortar institutions to open learning environments. Massive Online Open Courses (MOOCs) are a perfect example of these learning environments. MOOCs provide a cost and time effective choice for learners across the globe. This has led to new challenges for teachers such as providing valuable and quality assessment and feedback on such a large scale. Recent studies have found peer assessment where learners assess the work of their peers to be a viable and cost effective alternative to teacher/staff evaluation. This study systematically analyzes the current research on peer assessment published in the context of MOOCs and the online tools that are being used in MOOCs for peer assessment. 48 peer reviewed papers and 17 peer assessment tools were selected for the comparison in this study and were assessed on three main dimensions, namely, system design, efficiency and effectiveness. Apart from these dimensions, the study highlights the main challenges of peer assessment. In the light of the comparison and discussion of current research in terms of the identified dimensions, we present future visions and research perspectives to improve the peer assessment process in MOOCs.**

Keywords-Open Assessment; Peer Assessment; Open Learning Environments; MOOC; Blended Learning; Peer Reviews; Peer Feedback; Online Assessment.

I. INTRODUCTION

This paper presents an extended and more detailed version of our paper presented at the eighth international conference on mobile, hybrid, and online learning (eLmL 2016), where we reviewed the existing tools and research directions for peer assessment [1]. The field of education has transformed in recent year, with a growing interest in learner-centered, open, and networked learning models. These include Personalized Learning Environments (PLEs), Open Educational Resources (OER) and Massive Open Online Courses (MOOCs).

Massive Online Open Courses (MOOCs) have revolutionized the field of technology-enhanced learning (TEL). MOOCs enable a massive number of learners from all over the world to attend online courses irrespective of their social and academic backgrounds [2]. MOOCs have been classified in different forms by researchers, e.g., Siemens [3] classifies MOOCs into cMOOCs and xMOOCs. In his opinion cMOOCs allow the learners to build their own learning networks by using blogs, wikis, Twitter, Facebook and other social networking tools outside the confines of the learning platform and without any restriction and interference from the teachers [4]. Whereas, xMOOCs follow a more institutional model, having pre-defined learning objectives, e.g., Coursera, edX and Udacity. Apart from these sMOOCs and bMOOCs have also been introduced as variations of the MOOC platform with sMOOCs catering to a relatively smaller number of participants and bMOOCs combining the in-class and online learning activities to form a hybrid learning environment [3].

Irrespective of the classification, MOOCs require their stakeholders to address a number of challenges including and not limited to the role of university/teacher, plagiarism, certification, completion rates, innovating the learning model beyond traditional approaches and last but not the least assessment [5].

Assessment and Feedback are an integral part of the learning process and MOOCs are no different in this regard. Researchers acknowledge that the Teach-Learn-Access cycle in education cannot function in the absence of quality assessment [6]. However, in the case of MOOCs assessment presents a bottleneck issue due to the massiveness of the course participants and requires increased resources (time, money, manpower etc.) on part of the teachers to provide useful feedback to all the learners for a satisfying academic experience. This limitation causes many MOOCs to use automated assessments, e.g., quizzes with closed questions like multiple choice and fill in the blanks. These questions largely focus on the cognitive aspects of learning, and they are unable to capture the semantic meaning of learners' answers; in particular, in open ended questions [7]. Some other methods used in this scenario make use of crowd sourcing techniques to provide assessment and feedback to students. These methods include portfolios and self-assessment, group feedback and last but not the least peer assessment [8].

Peer assessment offers a scalable and cost effective way of providing assessment and feedback to a massive amount of learners where learners can be actively involved in the assessment processes [9]. A significant amount of research is directed towards exploring peer assessment in MOOCs. While this research discusses many issues such as the effective integration of peer assessment in various MOOC platforms and the improvement of the peer assessment process, it does not cover what has been done in this field for the past years from an analysis point of view.

Since, it is evident that peer assessment is a very viable assessment method in MOOCs, hence, the need for scouting all the available systems and studies becomes paramount in importance as it could be beneficial for future developments as well as provide a good comparison of available tools. In this study, we look at the peer assessment in general and the state of art of peer assessment in the MOOC era along with perceived benefits and challenges of peer assessment. We also look at different tools for peer assessment and the way they

try to address the challenges and drawbacks of peer assessment.

The remainder of this paper is structured as follows: Section II introduces peer assessment and its pros and cons. Section III is a review of the related work. Section IV describes the research methodology and how we collected the research data. In Section V, we review and discuss the current research based on several dimensions. Section VI summarises the results of our findings. Section VII presents challenges and future perspectives in peer assessment. Finally, Section VIII gives a conclusion of the main findings of this paper.

II. PEER ASSESSMENT

In recent years, student assessments have shifted from the traditional testing of knowledge to a culture of learning assessments [10]. This culture of assessment encourages students to take an active part in the learning and assessment processes [11]. Peer assessment is one of the flag bearers in this new assessment culture. Peer assessment also known as Peer grading, is defined by Topping as "an arrangement in which individuals consider the amount, level, value, worth, quality or success of the products or outcomes of learning of peers of similar status" [12].

Peer assessment has been leveraged in a wide range of subject domains over the years including natural sciences, social sciences, business, medicine and engineering [13]. According to Somervell [14], at one end of the spectrum peer assessment may involve feedback of a qualitative nature or, at the other, may involve students in the actual marking process. This exercise may or may not entail previous discussion or agreements over criterion. It may involve the use of rating instruments or checklists, which may have been designed by others before the exercise, or designed by the user group to meet its particular needs [15]. The use of peer assessment not only reduces the teacher workload; it also brings many potential benefits to student learning. These benefits include a sense of ownership and autonomy, increased motivation, better learning and high level cognitive and discursive processing [13], [16].

Despite these potential benefits, peer assessment still has not been able to have strong backing from either teachers or students [17]. Both parties have pre-conceived notions of low reliability and validity on their minds when discussing peer assessment [18], [19]. A number of possible factors have been identified for the lack of effectiveness of peer assessment in MOOCs. These factors include the scalability issue, diversity of reviewers, perceived lack of expertise, lack of transparency and fixed grading rubrics [8].

There have been many studies on the effectiveness and usefulness of peer assessment but these studies focus on a certain context and tool, which covers the aspects related to the context of the study. The aim of this paper is to examine the available literature and tools for peer assessment, provide a systematic analysis by reviewing them according to different aspects critical to their usage in the MOOC platform, and provide a bigger picture of the research domain. We will try to highlight the challenges of peer assessment and then provide some viable solutions to overcome these challenges.

III. RELATED WORK

Peer assessment in MOOCs is still an emerging field, hence, we did not find any research directly related to our work. Luxton-Reily [20] made a systematic comparison of a number of online peer assessment tools in 2009, but the study was conducted with limited dimensions for comparing the tools. The study examined tools including legacy systems, and divided the tools in different categories; namely generic, domain specific and context specific. The study identifies the problem that majority of online tools have been used in computer science courses, and most of the tools could not be used outside the context in which they were developed. The context limitations of the tools are the biggest hindrance preventing them from being widely adopted, which gives rise to the need for more general-purpose tools. Luxton-Reily also stressed the need to investigate the quality of the feedback provided by students [20].

Apart from this, another study identifies a number of approaches taken by different peer assessment tools to address the concerns of the involved stakeholders [21]. These approaches include connectivist MOOCs where the onus is on getting superior results through collaboration and not focusing on correctness. Rather, the course is designed in a way to encourage and welcome diverse perspectives from participants. Another approach is the use of calibration like in Calibrated Peer Reviews (CPR), where raters have to evaluate a number of training submissions before they get to evaluate submissions from their peers [22]–[24]. Other approaches highlighted in the study involve making use of a Bayesian post hoc statistical correction method [25]–[27] and to create a credibility index by modifying and refining the CPR method [21].

In comparison to the above-mentioned studies, our study adds a wide range of latest tools and analyzes them over several dimensions based on cognitive mapping approach. The study further provides critical discussion according to each dimension and suggests new areas for future work.

IV. METHODOLOGY

The research methodology used for this study is divided in two parts; namely, identification of eligible studies followed by a cognitive mapping approach to find certain criterion for categorizing and analyzing peer assessment tools.

A. Identification of Eligible Studies

We applied the significant research method of identifying papers from internet resources in our study [28]. This method was carried out in two rounds.

Firstly, we conducted a search in 7 major refereed academic databases. These include Education Resources Information Center (ERIC), JSTOR, ALT Open Access Repository, Google Scholar, PsychInfo, ACM publication,

IEEEXplorer, and Wiley Online Library. We used the keywords (and their plurals) "Peer Assessment", "Peer Review", "Open Assessment", "Assessment in MOOC", and "Peer Assessment in MOOC". As a result, 87 peer-reviewed papers were found.

In the second round, we identified a set of selection criteria as follows:

1- Studies must focus on using peer assessment preferably in a MOOC setting.

2- Studies that focus on design of peer assessment systems or that detail the setting in which peer assessment should be carried out were included.

3- Studies focusing on peer assessment in a manual setting were excluded.

4- Tools older than 10 years have not been included in the study, however the tools having current support are included.

This resulted in a final set of 48 research papers/studies on peer assessment in MOOCs and we extracted a list of 17 peer assessment tools that were used in these studies. These tools include Peer Studio [29], Cloud Teaching Assistant System (CTAS) [30], IT Based Peer Assessment (ITPA) [31], Organic Peer Assessment [32], EduPCR4 [33], GRAASP Extension [34], Web-PA [35], SWoRD (Peerceptiv now) [36], Calibrated Peer Reviews (CPR) [23], [22], [24], Aropä [37], Web-SPA [38], Peer Scholar [39], [40], Study Sync [41], [42], Peer Grader [43] and L²P (Lehr und Lern Portal, RWTH Aachen) Peer Reviews [8]. We also took a look into some open systems providing peer assessment capabilities that could be used in MOOCs as well, namely: TeamMates [44] and TurnItIn [45].

B. *Cognitive Mapping Approach*

Cognitive mapping is a method that enables researchers to classify and categorize things into several dimensions based on the research questions [46]. The study provides an example of using cognitive mapping to elicit mental models of emotions in the work place by conducting a series of interviews at an office and then code these interviews into maps. These maps were then analyzed to uncover the relationship between the job conditions and the outcomes associated with different kind of emotional experiences at work [46].

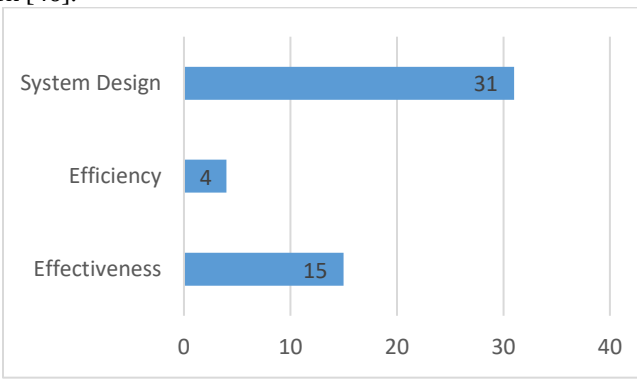

Figure 1. Peer Assessment classification map.

For the sake of our study, we scouted the literature available on peer assessment to form a directed cognitive map for each study identifying main ideas related to peer assessment. These maps were then analyzed for distinct clusters of concepts, grouping similar terms and ideas. After analyzing the clusters, we were able to identify certain dimensions namely: system design, efficiency and effectiveness (see Figure 1), which were all part of the discussed peer assessment systems. These dimensions provide an easy and efficient way to assess different peer assessment tools/studies.

In order to capture the information gained from the literature analysis, we created a detailed field diagram (see Figure 2), which has been partitioned into three categories and ten sub-categories. It is worth mentioning here that some of the sub categories could be mapped to multiple main categories and in such scenarios, we used the best match for better classification.

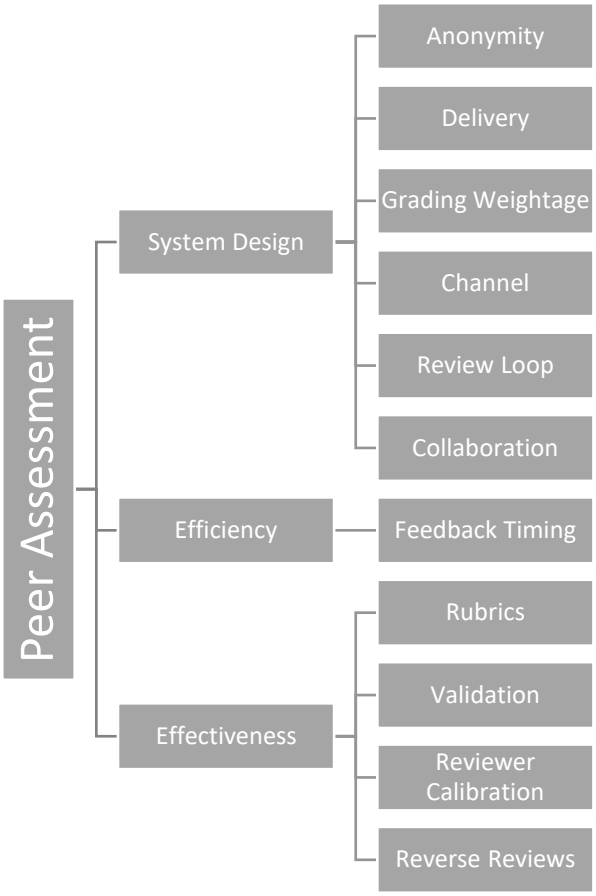

Figure 2. Peer Assessment cognitive map.

Apart from these dimensions, we identified a number of challenges as well for peer assessment from the literature review. The list of challenges of peer assessment includes transparency, credibility, accuracy, reliability, validity, diversity, scalability, and efficiency. *Transparency* refers to the fact that the assessee is aware of how the review process works and has confidence in it. The *credibility* refers to the issue, whether the reviewing person has sufficient knowledge in the subject area and is capable enough of providing credible feedback. *Accuracy* is closely linked to credibility in the sense that if the reviewer has a good mastery of the subject then his/her reviews would tend to be more accurate. *Reliability* is the consistency of grades that different peers would assign to the same assessment, or in other words as inter-rater reliability. Whereas, *validity* is calculated as a correlation coefficient between peer and instructor grades, assuming that the instructor grade is considered a trustworthy benchmark [30]. *Diversity* refers to the different educational backgrounds of the assessors. *Scalability* is inherent to open learning environments with a large number of participants. And last but not the least, *efficiency* related to feedback timing. Studies have shown that the earlier the learners get feedback to their work, the more time they have to improve the final product. Reducing the time, it takes to get feedback to a draft submission automatically allows for a better final product.

The peer assessment dimensions identified earlier try to address some of the challenges presented here in a number of ways, which will be discussed later with the discussion of each peer assessment dimension in the following sections.

V. DISCUSSION

This section deals with the critical analysis of the peer assessment literature based on the cognitive mapping dimensions derived in the previous section. For the critical discussion part, we look at the identified dimensions and then discuss the way in which certain tools cater to that dimension (if at all).

A. System Design

A lot of effort has been put into the design of peer assessment systems, design of certain features provided by the system and the manner in which they are implemented. Nearly 70% of the studies deal in one way or the other with system or a feature design in peer assessment.

In the following sections, we discuss some key features of peer assessment systems and the way they are realized by different tools.

1) Anonymity: Anonymity is a key feature that is to be kept in mind while designing any peer assessment system, as it safeguards the system against any type of bias (gender, nationlity, friendship etc.) to play a factor in the assessment from peers. There are three levels of anonymity namely, single blind: assessor knows the assessee but the assessee has no idea of the assessor, double blind: both assessor and assessee are unaware of each other and finally no anonymity in which the identity of both the assessor and assessee is known to each other. Most of the systems reviewed in this study follow the principle of double blind reviews for the sake of bias free reviews, however, TurnItIn [45] and Study Sync [41] only implement the single blind reviews.Whereas, organic peer assessment [32] has no mention of the feature at all.

Anonymity in peer assessment is also important as it helps increase the reliability of reviews to some extent by removing bias from the system.

2) Delivery: This feature entails the delivery mode of the review, whether it is delivered indirectly (as is the case in most of the MOOC courses), or directly face to face (could be a situation in a bMOOC). All the reviewed systems only support indirect feedback at the moment. Moreover, a study [8] conducted in a bMOOC platform found that students feel more free to voice their evaluations in an indirect way rather than delivering it directly to the assessee. This helps them to give their honest feedback and enables them to be more fair in their assessment.

The in-direct delivery of reviews also serves the purpose of addressing the challenge of accuracy of reviews, as students do not have to worry about giving their feedback face to face to their peers and they can provide honest assessment of peer's work.

3) Grading Weightage: Almost two third of the reviewed systems assign a pre-defined weightage to the review from the peers in the overall grade. This means that the final grade is calculated by combining the grade from the peers and the instructor and assigning certain weightages to each of them. L^2P Peer Reviews [8] implements a novel way of assigning weightage to the reviews from peers, by allowing the teacher to define the weightage per peer review task. This also enables the system to bypass reviews from peers, as the teacher could assign a zero weightage to student reviews.

Moreover, the systems that do not give any weightage to student reviews still use these reviews in order to help the teacher in giving their assessment of the task. The teacher could use the student review as an input to write their own review for the submission.

4) Channel: It is a general principle, the more the merrier/better. Researchers believe that the same holds true for the assessment reviews, as more reviews help the assessee to have multiple inisghts about their work and learn from them instead of a single point of view being forced upon them [8]. However, this also means that every reviewer/reviewing group has to review a greater number of submissions from their peers, which puts extra burdern on the students.

The peer assessment system could handle the channel requirement in two ways, namely single channel: where every submission is reviewed by exacly one peer or peer group, or multi-channel: where the number of reviewers varies and is grater than one. All the reviewed systems provide multi-

channel feedback support for the reviews, except the L²P Peer Reviews module, which only offers single channel reviews at the moment [8].

A study conducted at Stanford and University of California, proposed a process of selecting an appropriate number of reviewers needed for each submission by making use of an automated system. Initially the student grade is predicted by a machine learning algorithm, which then estimates the confidence value. This value is used to determine the required number of peer graders. These graders then identify the attributes of the answer with the help of a rubric. Finally, other peer graders then verify whether these attributes actually exist or not. If the results of these peer graders are similar then final score will be generated and if it is not the case then re-identification of attributes takes place by one more peer grader [47]. This automated process aims at putting manageable load on peers by trying to reduce the number of peers required for each submission.

The multi-channel review also paves the way for the peer assessment system to calculate inter-rater reliability from the difference between peer reviews for any submission.

5) Review Loop: The purpose of this feature is to allow the students to work on their assignments in multiple iterations in order to improve the final product and have a better learning outcome. Although, researchers claim it to be a very important feature for any peer assessment tool, only a handful of the reviewed tools actually implement more than one review loops. These systems include PeerStudio [29], EduPCR4 [33], Peerceptiv [36], Aropä [37], Web-SPA [38] and Peer Grader [43].

Peer grader is unique in this respect as it allows for a communication channel between the author and the reviewer to help the authors improve their submissions. The assessor can provide their review that is directly available to the assessee, and then assessee can then in turn improve their original submission untill the deadline. Essentially, it makes use of the single review loop in an efficient way to accommodate multiple loops [43].

6) Collaboration: Collaboration means the ability of the tool to allow students to form and work in small groups. This leads to sharing of ideas inside the group and promotes a healthy learning enviroment. Although many MOOC platforms make use of discussion forums and wikis for enabling collaboration and idea sharing between the students, but we found that only a few systems actually allow the students to form groups and submit their work in groups.

Team mates [44] is an open source tool that allows the students to form smaller groups/teams and submit their work. Also L²P Peer reviews [8], makes use of a separate module Group Workspace in their learning management system to manage student groups. This separate module allows students to work collaboratively in their own workspace online with document sharing and chat functionalities. The peer reviews tool communicates with this module to get the group information for the students and allows group submissions and reviews of the assignments. The L²P Peer reviews tool, also offers the option of individual submissions and feedback, which are available in the individual assignment settings, so the teacher could decide whether to have individual or group work possibilities per assignment.

B. Efficiency

In this section, we list the features that contribute to the overall efficiency of the system. These features allow the system to be more efficient for its users and help them get the most value out of the system. The dimensions discussed here directly relate to the challenge of efficiency of peer assessment systems.

1) Feedback Timing: Research has shown that the optimal timing of a feedback is early in the assessment process, as it gives the learners more time to react and improve their work. Peer Studio, a tool used in Coursera MOOC platfrom proposes an effective way to reduce the review response time. The learners are required to review work from two peers to get an early feedback to their submission. Also, the learners can submit their work any number of times for a peer review and get the review by reviewing others. The system assigns the reviews to reviewers based on certain criteria that includes the users who have submitted their work for review, currently online users etc. A study conducted on the usefulnes of the system concludes that the students in the Fast Feedback condition did better than the No Early Feedback condition group. It also states that on average students scored higher by 4.4% of the assignment's total grade, hence proving the usefulnes of early feedback. The study also claims to have average feedback times of 20 minutes and 1 hour in MOOC and in person classes respectively [29].

C. Effectiveness

Several researchers in TEL have explored how to design effective peer assessment modules with a higher level of user satisfaction. We identified certain features that contribute to the effectiveness of the reviews being provided by the peers, which are discussed in the following sections.

1) Rubrics: Rubrics provide a way to define flexible task specific questions that could include descriptions of each assessment item to achieve fair and consistent feedback for all course participants.

Their have been certain studies that focus on establishing methods to enhance the effectiveness of peer assessment by asking direct questions for the peer to answer, in order to assess the quality of someone's work [8]. This eables the reviewer to easily reflect on the quality of submitted work in a goal oriented manner. Hence, a flexible rubric system becomes a must have feature for any good peer assessment system.

In our study, we found that majority of the reviewed systems offer this feature in one way or the other with a

notable exception of Peer Grader. While many tools allow the teachers to define rubrics for tasks, Peerceptiv offers a shared rubric library that allows for templating the rubrics for re-use and editing [36].

Another variation of the use of rubrics in the systems is the way they are handled in peer studio tool. The tool allows the teachers to define rubrics and then enforces the students to answer these questions in a better way by using a technique they call scaffolding comments [29]. The system does this scaffolding by making use of short tips for writing comment below the comments box. The tool provides helpful tips to the reviewers like "Is your feedback actionable?" or it may ask reviewers to "Say more…" when they write "great job!" etc. to enforce the reviewers to write more meaningful comments.

Rubrics are an efficient and effective way of introducing transparency to the peer assessment process, as all the course participants could see the criteria/questions for the evaluation of their submissions. The rubrics also address the challenge of diversity in course participants to some extent. The participants are provided with the same rubrics, which lays a benchmark for them to evaluate the peer submissions in a similar manner.

2) Validation: A number of studies have been carried out on the validation aspect of the reviews provided by peers, i.e., on methods to make sure that the feedback provided the peers is valid and of a certain value. Luo et al. [13] conducted a study, specifically on Coursera platform to evaluate the validity of the reviews from peers. In their study they propose that increasing the number of reviewers and giving prior training to the reviewers on how to review the work of others are some techniques used to bolster the validity of the reviews.

Similar studies focus on other ways to achieve validiy of the reviews, like Peerceptiv measures the validation of reviews to a submission by simply calculating the agreement rate between different reviewers. It takes score difference, consitency and the spread of scores into consideration for evaluating the validity of reviews. Although, this is a minimalistic approach but it still provides a good starting point for other measures to be carried out, to judge the validity of reviews in detail [36].

The validation dimension identified here is linked to several challenges of peer assessment including reliability, accuracy, validity and credibility. By validating the assessment provided by the peers, the peer assessment tool could address these challenges and ensure quality feedback for all course participants.

3) Reviewer Calibration: Calibrated peer reviews [24] along with some other studies carried out in MOOCs [48] propose a different method to achieve system effectivness, namely, reviewer calibration. In this method, the reviewers are required to grade some sample solutions that have been pre-graded by the instructor to train them in the process of providing reviews. The reviewers are not allowed to review the work of their peers, unles they achieve a certain threshold in the review of the sample submission and only then can they review the work of their peers. In the end, the system takes into account the calibration accuracy of the reviewer by assigning weightage to each submitted review.

The calibration of reviewers before the actual review phase increases the level of accuracy of their reviews and also makes it easier to identify credible reviewers from all course participants.

4) Reverse Reviews: Another interesting method to verify the effectiveness of the reviews is to use the reverse review method. Peer Grader [43] and EduPCR4 [33] tools make use of this method to allow the original authors of the reviewed submissions to rate the reviews they received from their peers. The students can specify, whether the review helped them in improving their submission, or was of a certain quality, or helped them understand the topic clearly. This review is then taken into consideration at the time of calculation of the final grade, so the peers who provided better reviews have a chance to better their assignment score.

Aropä varies from other tools in this aspect, as it manages the reverse reviews by giving this option to teachers instead of students [37]. This way teachers could judge the credibility of the review and take it into consideration before providing their own review.

The reverse reviews, also provide an easy and efficient way of creating a credibility index for the course participants, which could be used in later assignments to help the teachers in the grading process.

VI. SUMMARY

Table 1 shows a summary of evaluation of different tools against the dimensions identified in Section IV. The table shows that nearly all the tools reviewed in our study follow a similar system design varying slightly based on the context in which they are used. The only major discrepancy in most tools is their inability to allow students to work in groups (for assignment submission and reviews). Another pattern emerging from studying the table is that more and more tools are giving weightage to the student reviews in the overall grade of the students. This means that the teachers must be sure about the validity and quality of the student reviews, and the system must provide features for its insurance.

Another useful observation is the usage of assessment rubrics by the tools to help students in the process of reviewing their peers. As identified by Yousef et al. [8], rubrics are an easy way to provide learners with task specific questions, allowing the achievement of fair and consistent feedback for all course participants.

In the comparison for the validation, we mention all the tools for whom a study has been conducted for the validation of peer reviews. It does not specify that the tool has some in-built validation mechanism for the reviews provided by peers.

Table 1. A systematic comparison of peer assessment tools

	System Design						Efficiency			Effectiveness		
Tools	**Anonymity**	**Delivery**	**Grading Weightage**	**Channel**	**Review Loop**	**Collaboration**	**Time/Rapid Feedback**	**Rubrics**	**Validation**	**Reviewer Calibration**	**Reverse Reviews**
Peer Studio [29]	Double	InDirect	Yes	Multiple	Multiple	No	Yes	Yes	Yes	No	No
CTAS [30]	Double	InDirect	Yes	Multiple	Single	-	No	Yes	Yes	No	No
ITPA [31]	Yes	InDirect	No	Multiple	Single	-	No	Yes	Not measured	No	No
Organic PA [32]	No	InDirect	No	Multiple	Single	-	No	No	Yes	No	No
EduPCR4 [33]	Double	InDirect	Yes	Multiple	Double	-	No	Yes	Not measured	No	Yes
GRAASP extension [34]	No	InDirect	Yes	Multiple	Single	-	No	Yes	Yes	No	No
Web-PA [35]	Yes	InDirect	Yes	Multiple	Single	Yes	No	Yes	Not measured	No	No
SWoRD/Peerceptiv [36]	Double	InDirect	Yes	Multiple	Double	Yes	No	Yes	Yes	No	No
CPR [22]-[24]	Double	InDirect	Yes	Multiple	Single	No	No	Yes	Yes	Yes	No
Aropä [37]	Yes	InDirect	Yes	Multiple	Double	-	No	Yes	Yes	No	Yes
Web-SPA [38]	Yes	InDirect	No	Multiple	Double	Yes	No	Yes	Yes	No	No
Peer Scholar [39][40]	Double	InDirect	Yes	Multiple	Single	No	No	Yes	Yes	No	No
Study Sync [41][42]	Single	InDirect	No	Multiple	Single	No	No	Yes	Yes	No	No
Peer Grader [43]	Double	InDirect	Yes	Multiple	Double	No	No	No	Yes	No	Yes
L²P Peer Reviews [8]	Double	InDirect	Yes	Multiple	Single	Yes	No	Yes	Yes	No	No
Team Mates [44]	Double	InDirect	No	Multiple	Single	Yes	No	Yes	Not measured	No	No
TurnItIn [45]	Single	InDirect	No	Multiple	Single	No	No	Yes	Yes	No	No

Table I also highlights an important trend in the field of peer assessment for MOOCs. It shows that most systems are moving on from the basic system design and looking for ways to improve the efficiency and effectiveness of the system. This leads to the use of more innovative ways to ensure the quality of reviews provided by peers, and a focus to find ways on improving the overall user experience and learning. The main reason behind this trend is to decrease the workload on the teachers while addressing the challenges of peer assessment making sure that students get the most out of the course.

VII. CHALLENGES AND FUTURE VISION

MOOCs with their large number of participants pose a challenge when it comes to assessment and feedback, and peer assessment offers a viable solution to the problem. However, peer assessment itself faces many challenges including scalability, reliability, quality and validation. Several studies have focused on overcoming these limitations, as outlined in the previous sections but there is still a lot of room for improvement.

The challenges faced by peer assessment are inherent from the challenges of open assessment in general [49], and the field of learning analytics offers a number of techniques to overcome these challenges. In this section, we try to offer some solutions from the field of learning analytics, which could be used to overcome certain peer assessment challenges.

1) Scalability: The massive number of participants in the MOOC courses requires the feedback provided to students to be scalable as well. This requires the use of certain measures to decrease the time required by the teacher to provide useful feedback to the student submissions. Although, peer assessment tries to lessen the teacher's burden but still the teacher has to be in the loop to ensure quality feedback. To overcome this issue of scalability, we could make use of clustering techniques in a number of ways. We could cluster similar submissions together and in case of peer assessment, the similar reviews (including rubric answers) could also be clustered together to form a single unit. The teacher could easily grade the clusters, in turn, saving valuable time. A similar approach has been used in scaling short answer questions grading with satisfactory results . The study in [50], found out that using clustering to scale feedback not only saves time but it also helps teachers to develop a high-level view of students' understanding and misconceptions.

Another solution to the problem of scalability could be the use of word clouds by extracting important parameters from the submitted work of students. This could help the teacher by providing an overview of the submission and giving a fair idea about the contents. Hence, a teacher could decide if the submission requires in depth review or they could grade based on the provided information.

Further, it can be helpful to leverage statistical methods and visualization techniques (e.g., dashboards) to support teachers in getting a good overview on the provided feedback in a visual manner.

2) Reviewer Credibility/Reliability: There have been cases identified in peer assessment studies, where students do not take the process of reviewing others work seriously. This leads to invalid reviews and casts a doubt over the credibility of the reviews being provided to students. In this scenario, the teacher must be in the loop to ensure valid reviews. One solution to this could be to rate the reviewers using the reverse reviews method and maintain a ranking of reviewers based on these reverse reviews. This way, we could identify possible bad reviewers and they could be screened out for further reviews or they could be urged to provide better reviews. This could lead to the use of predictive analytics methods to predict the accuracy of reviewers based on knowledge in the subject area, received ratings and feedback history etc.

Another approach, could be to use the peer rank method, similar to the page rank method for ranking online search results [51]. The peers are rated based on the ratings they received for their own submissions. The idea behind this approach is that in a usual scenario, the student getting a better grade should have a better grasp of the concept and hence, it is safer to predict that he/she is able to provide better feedback on the topic.

3) Validity: We have already seen the usage of calibration to improve the validity of the reviews. Raman and Joachims make use of a statistical method in their study to ensure the validity of the reviews. They use Bayesian ordinal peer grading to form an aggregated ordering for all the submissions in a course room. The difference in ranking from different peers is also taken into account to ensure the effectiveness and validity of reviews [25].

Another approach could be the usage of semi-automated assessment, as is the case in automatic essay grading systems. The system considers the grade from one human reviewer and the automated assessment grade. If the difference in grades from both sources is greater than a certain threshold, then the system asks for an additional review from a human grader [52]. This technique can be applied to the peer assessment, and if the disagreement between the review from peer and the automated assessment is significant, the system could mark the submission for grading by the teacher or ask for a review from some other peer as well.

4) Quality: Rubrics provide an easy way of improving the quality of the reviews by providing certain questions that a student has to answer in the review process [8]. The peer assessment system could further enhance this by providing a way for the teacher to specify common mistakes that students make, so that the reviewer could look for them in the submission and in turn, improve the quality of the review.

5) System Configuration: Another improvement to the peer assessment tools could be to allow the user to configure different settings from a central location rather than making

it a part of system design that could not be altered. Majority of peer assessment systems in use today have pre-defined configuration in features like anonymity, review loops, grading weightage, collaboration etc. These pre-configured settings make it difficult for the tool to be used in a more generic way and in different contexts. Also, a large number of these tools are only used in computer science courses as the teachers could tailor make a tool for their specific needs and use it in their course. These domain specific tools make it impossible for the peer assessment to be used in different disciplines of study uniformly. Hence, a tool that allows its users to configure all these settings could be a lot more useful across different domains and have a higher acceptance rate from users all over the world.

VIII. Conclusion

Peer assessment is a rich and powerful assessment method used in technology-enhanced learning (TEL) to improve learning outcomes as well as learner satisfaction. In this paper, we analysed the research on peer assessment published in the MOOC era, and the tools that could be used to provide peer assessment capabilities in a MOOC. A cognitive mapping approach was used to map the selected studies on peer assessment into three main dimensions namely: system design, efficiency and effectiveness. Furthermore, we identified the challenges of peer assessment and linked them to the system dimensions, which try to overcome these challenges.

The following is a summary of the main findings in our study as well as aspects of peer assessment that need further research, according to each dimension.

A. System Design

The analysis of the peer assessment research showed that majority of the systems are designed on similar lines to each other, differing in only a small number of features or the way these features are implemented. Despite these possible differences in implementation, the general idea for different system features remains the same across different tools. However, several features concerning system design need a better acceptance across these tools: (1) Collaboration: The tools should allow the students to work in a collaborative environment and submit their assignments and even review in groups. This could help ease the burden on individual students and the sharing of knowledge would in turn help them achieve better learning objectives. (2) Review Loops: In our opinion, all peer assessment tools should provide at least double review loops, to give students more chances of improvement and in doing so we leverage the peer assessment model in an effective way to achieve better overall results.

B. Efficiency

Studies have established the positive effect of timely feedback on student performance but the assessment tools are lagging far behind in this regard. In our opinion, more tools should focus on efficient ways to decrease the feedback time, and focus on more innovations to make the process more efficient.

C. Effectiveness

Several methods are being used in peer assessment to increase effectiveness of the reviews and in turn the learners' satisfaction with peer assessment. Although, rubrics, reviewer calibration and reverse reviews are good ideas to improve the effectiveness of the reviews; more and more research must be put into measuring the validity of the reviews provided by peers. Future research needs to find out new ways to record validity of reviews and improvements to this validity.

The systematic comparison of peer assessment tools also reveals certain patterns and trends across the analysed tools. It points out the fact that most tools are quite similar in system design, and the way they carry out the peer assessment process. The difference arises in the way they apply validation and effectiveness techniques to the peer reviews. The study also highlights the shift in focus from basic system design to innovative ways of improving the quality and effectiveness of the reviews provided by peers. It also lists a few techniques that are being used in different peer assessment tools to ensure quality and effectiveness.

The study concludes with providing a list of open challenges in the peer assessment process/systems and proposes certain techniques that could be applied to address these challenges. The proposed solutions include a number of techniques from the field of learning analytics including statistics, prediction, visualizations, and data mining techniques that could prove useful in improving the peer assessment process/tools.

References

[1] U. Wahid, M. A. Chatti, and U. Schroeder, "A Systematic Analysis of Peer Assessment in the MOOC Era and Future Perspectives," in *Proceedings of the Eighth International Conference on Mobile, Hybrid, and On-line Learning, elml 2016*, pp 64-69.

[2] T. Liyanagunawardena, S. Williams, and A. Adams, "The Impact and reach of MOOCs: A developing countries' perspective," *eLearning Pap.*, no. 33, 2013.

[3] G. Siemens, "MOOCs are really a Platform. Elearnspace (2012).".

[4] G. Siemens, "Connectivism: A learning theory for the digital age," 2014.

[5] A. M. F. Yousef, M. A. Chatti, U. Schroeder, M. Wosnitza, and H. Jakobs, "MOOCs-A Review of the State-of-the-Art," in *Proc. CSEDU 2014 conference*, vol. 3, pp. 9–20.

[6] J. R. Frederiksen and A. Collins, "A systems approach to educational testing," *Educ. Res.*, vol. 18, no. 9, pp. 27–32, 1989.

[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," in *Design Thinking Research*, Springer, 2015, pp. 131–168.

[8] A. M. F. Yousef, U. Wahid, M. A. Chatti, U. Schroeder,

and M. Wosnitza, "The Effect of Peer Assessment Rubrics on Learners' Satisfaction and Performance within a Blended MOOC Environment," in *Proc. CSEDU 2015 conference*, vol. 2, pp. 148–159.

[9] R. O'Toole, "Pedagogical strategies and technologies for peer assessment in Massively Open Online Courses (MOOCs)," 2013.

[10] A. Planas Lladó, L. F. Soley, R. M. Fraguell Sansbelló, G. A. Pujolras, J. P. Planella, N. Roura-Pascual, J. J. Suñol Martínez, and L. M. Moreno, "Student perceptions of peer assessment: an interdisciplinary study," *Assess. Eval. High. Educ.*, vol. 39, no. 5, pp. 592–610, 2014.

[11] S. Lindblom-ylänne, H. Pihlajamäki, and T. Kotkas, "Self-, peer-and teacher-assessment of student essays," *Act. Learn. High. Educ.*, vol. 7, no. 1, pp. 51–62, 2006.

[12] K. Topping, "Peer assessment between students in colleges and universities," *Rev. Educ. Res.*, vol. 68, no. 3, pp. 249–276, 1998.

[13] H. Luo, A. C. Robinson, and J.-Y. Park, "Peer grading in a mooc: Reliability, validity, and perceived effects," *Online Learn. Off. J. Online Learn. Consort.*, vol. 18, no. 2, 2014.

[14] H. Somervell, "Issues in assessment, enterprise and higher education: The case for self-peer and collaborative assessment," *Assess. Eval. High. Educ.*, vol. 18, no. 3, pp. 221–233, 1993.

[15] N. Falchikov, "Peer feedback marking: developing peer assessment," *Program. Learn.*, vol. 32, no. 2, pp. 175–187, 1995.

[16] T. Papinczak, L. Young, and M. Groves, "Peer assessment in problem-based learning: A qualitative study," *Adv. Heal. Sci. Educ.*, vol. 12, no. 2, pp. 169–186, 2007.

[17] W. Cheng and M. Warren, "Peer and teacher assessment of the oral and written tasks of a group project," *Assess. Eval. High. Educ.*, vol. 24, no. 3, pp. 301–314, 1999.

[18] N. Falchikov and J. Goldfinch, "Student peer assessment in higher education: A meta-analysis comparing peer and teacher marks," *Rev. Educ. Res.*, vol. 70, no. 3, pp. 287–322, 2000.

[19] O. McGarr and A. M. Clifford, "'Just enough to make you take it seriously': exploring students' attitudes towards peer assessment," *High. Educ.*, vol. 65, no. 6, pp. 677–693, 2013.

[20] A. Luxton-Reilly, "A systematic review of tools that support peer assessment," *Comput. Sci. Educ.*, vol. 19, no. 4, pp. 209–232, 2009.

[21] H. Suen, "Peer assessment for massive open online courses (MOOCs)," *Int. Rev. Res. Open Distrib. Learn.*, vol. 15, no. 3, 2014.

[22] M. E. Walvoord, M. H. Hoefnagels, D. D. Gaffin, M. M. Chumchal, and D. A. Long, "An analysis of calibrated peer review (CPR) in a science lecture classroom," *J. Coll. Sci. Teach.*, vol. 37, no. 4, p. 66, 2008.

[23] A. Russell, O. Chapman, and P. Wegner, "Molecular science: Network-deliverable curricula," *J. Chem. Educ.*, vol. 75, no. 5, p. 578, 1998.

[24] P. A. Carlson and F. C. Berry, "Calibrated peer review/sup TM/and assessing learning outcomes," in *fie*, 2003, pp. F3E1–6.

[25] K. Raman and T. Joachims, "Bayesian Ordinal Peer Grading," in *Proceedings of the Second (2015) ACM Conference on Learning @ Scale*, pp. 149–156.

[26] C. Piech, J. Huang, Z. Chen, C. Do, A. Ng, and D. Koller, "Tuned models of peer assessment in MOOCs," *arXiv Prepr. arXiv1307.2579*, 2013.

[27] I. M. Goldin, "Accounting for peer reviewer bias with bayesian models," in *Proceedings of the Workshop on Intelligent Support for Learning Groups at the 11th International Conference on Intelligent Tutoring Systems*, 2012.

[28] A. Fink, *Conducting research literature reviews: from the Internet to paper*. Sage Publications, 2013.

[29] C. Kulkarni, M. S. Bernstein, and S. Klemmer, "PeerStudio: Rapid Peer Feedback Emphasizes Revision and Improves Performance," in *Proceedings from The Second (2015) ACM Conference on Learning@ Scale*, pp. 75–84.

[30] T. Vogelsang and L. Ruppertz, "On the validity of peer grading and a cloud teaching assistant system," in *Proceedings of the Fifth International Conference on Learning Analytics And Knowledge*, 2015, pp. 41–50.

[31] K. Lehmann and J.-M. Leimeister, "Assessment to Assess High Cognitive Levels of Educational Objectives in Large-scale Learning Services," 2015.

[32] S. Komarov and K. Z. Gajos, "Organic Peer Assessment," in *Proceedings of the CHI 2014 Learning Innovation at Scale workshop*.

[33] Y. Wang, Y. Liang, L. Liu, and Y. Liu, "A Motivation Model of Peer Assessment in Programming Language Learning," *arXiv Prepr. arXiv1401.6113*, 2014.

[34] A. Vozniuk, A. Holzer, and D. Gillet, "Peer assessment based on ratings in a social media course," in *Proceedings of the Fourth International Conference on Learning Analytics And Knowledge*, 2014, pp. 133–137.

[35] P. Willmot and K. Pond, "Multi-disciplinary Peer-mark Moderation of Group Work," *Int. J. High. Educ.*, vol. 1, no. 1, p. p2, 2012.

[36] J. H. Kaufman and C. D. Schunn, "Students' perceptions about peer assessment for writing: their origin and impact on revision work," *Instr. Sci.*, vol. 39, no. 3, pp. 387–406, 2011.

[37] J. Hamer, C. Kell, and F. Spence, "Peer assessment using arop{ä}," in *Proceedings of the ninth Australasian conference on Computing education-Volume 66*, 2007, pp. 43–54.

[38] Y.-T. Sung, K.-E. Chang, S.-K. Chiou, and H.-T. Hou, "The design and application of a web-based self-and peer-assessment system," *Comput. Educ.*, vol. 45, no. 2, pp. 187–202, 2005.

[39] D. E. Paré and S. Joordens, "Peering into large lectures: examining peer and expert mark agreement using peerScholar, an online peer assessment tool," *J. Comput. Assist. Learn.*, vol. 24, no. 6, pp. 526–540, 2008.

[40] S. Joordens, S. Desa, and D. Paré, "The pedagogical anatomy of peer-assessment: Dissecting a peerScholar assignment," *J. Syst. Cybern. Informatics*, vol. 7, no. 5, 2009.

[41] B. McCrea and M. Weil, "On Cloud Nine: Cloud-Based Tools Are Giving K-12 Collaboration Efforts a Boost," *J. (Technological Horizons Educ.*, vol. 38, no. 6, p. 46, 2011.

[42] D. L. White, "Gatekeepers to Millennial Careers: Adoption of Technology in Education by Teachers," *Handb. Mob. Teach. Learn.*, p. 351, 2015.

[43] E. F. Gehringer, "Electronic peer review and peer grading in computer-science courses," *ACM SIGCSE Bull.*, vol. 33,

no. 1, pp. 139–143, 2001.

[44] G. Goh, X. Lai, and D. C. Rajapakse, "Teammates: A cloud-based peer evaluation tool for student team projects," 2011.

[45] S. Draaijer and P. van Boxel, "Summative peer assessment using 'Turnitin' and a large cohort of students: A case study," 2006.

[46] S. McDonald, K. Daniels, and C. Harris, "Cognitive mapping in organizational research In C. Casssell & G. Symon (Eds.), Essential guide to qualitative methods in organizational research (pp. 73-85)." London: Sage, 2004.

[47] C. E. Kulkarni, R. Socher, M. S. Bernstein, and S. R. Klemmer, "Scaling short-answer grading by combining peer assessment with algorithmic scoring," in *Proceedings of the first ACM conference on Learning@ scale conference*, 2014, pp. 99–108.

[48] J. Wilkowski, D. M. Russell, and A. Deutsch, "Self-evaluation in advanced power searching and mapping with google moocs," in *Proceedings of the first ACM conference on Learning@ scale conference*, 2014, pp. 109–116.

[49] A. M. Chatti, V. Lukarov, H. Thüs, A. Muslim, F. A. M. Yousef, U. Wahid, C. Greven, A. Chakrabarti, and U. Schroeder, "Learning Analytics: Challenges and Future Research Directions," *eleed*, vol. 10, no. 1, 2014.

[50] M. Brooks, S. Basu, C. Jacobs, and L. Vanderwende, "Divide and Correct: Using Clusters to Grade Short Answers at Scale," in *Proceedings of the First ACM Conference on Learning @ Scale Conference*, 2014, pp. 89–98.

[51] T. Walsh, "The peerrank method for peer assessment," *arXiv Prepr. arXiv1405.7192*, 2014.

[52] H. Chen and B. He, "Automated Essay Scoring by Maximizing Human-Machine Agreement.," in *EMNLP*, 2013, pp. 1741–1752.