Is Peer Review an Appropriate Form of Assessment in a MOOC? Student Participation and Performance in Formative Peer Review.

Sarah E. M. Meek[1#], Louise Blakemore[2] and Leah Marks[1]

[1] *School of Medicine, University of Glasgow, Glasgow, U.K.*

[2] *National Heart & Lung Institute, Imperial College London, U.K.*

[#]*Corresponding Author*

Full Contact Details and Short Biographical Note for Authors:

- Dr Sarah Meek is a University Teacher in the School of Medicine at the University of Glasgow, and was co-lead educator on the University of Glasgow's first MOOC. Her research interests include Threshold Concepts in medical education, MOOCs, and blended learning. Contact: Room 507, Wolfson Medical School Building, University of Glasgow, University Avenue, Glasgow, G12 8QQ, U.K. Tel: +44 (0)141 330 7549. E-mail: Sarah.Meek@glasgow.ac.uk

- Dr Leah Marks Meek is a University Teacher in the School of Medicine at the University of Glasgow, and was co-lead educator on the University of Glasgow's first MOOC. Her research interests include student use of feedback, peer review in offline and online settings, and the use of MOOCs in medical education. Contact: University Dept of Pathology and Medical Genetics, Laboratory Medicine, South Glasgow University Hospitals, 1345 Govan Road, Glasgow G51 4TF, UK Tel: +44 (0) 141 354 9457. E-mail: leah.marks@glasgow.ac.uk

- Dr Louise Blakemore is a Teaching Fellow in Genomic Medicine. Contact: National Heart & Lung Institute, Imperial College London, Guy Scadding Building, London SW3 6LY, U.K. Tel: +44 (0)20 7594 9996. E-mail: l.blakemore@imperial.ac.uk

Abstract

Many aspects of traditional higher education must be reconceptualised for massive open online courses (MOOCs). Formative and summative assessment of qualitative work in particular requires novel approaches to cope with the numbers involved. Peer review has been proposed as one solution, and has been widely adopted by major MOOC providers but there is currently little evidence about whether it is appropriate in the MOOC context, or under what conditions. Here, we examine student participation, performance and opinions of a peer review task in a biomedical science MOOC. We evaluate data from approximately 200 student topic summaries and 300 qualitative peer reviews of those summaries, and compare these to student demographic data (gender, age, employment status, education, national language) and to performance in multiple choice tests (MCQs). We show that higher performance in the written topic summary correlated with both higher participation in the peer review task, and with writing higher quality peer reviews. Qualitative analysis of student comments revealed that student opinion on the usefulness of the peer review task was mixed: some strongly believed it benefitted their learning, while others did not find it useful or did not participate. We suggest instructional design strategies to improve student participation and increase learning gain from peer review in the MOOC context.

Keywords

Peer Review, Peer Assessment, MOOC (massive open online course)

Introduction

Assessment and feedback in massive open online courses

As massive open online courses (MOOCs), gain increasing prominence in education, it is important to reconceptualise several important ideas, including the definition of curriculum, and how achievement is defined and measured (DeBoer et al., 2014). MOOCs often have thousands of students, with the most popular course to date enrolling over 440,000 (https://about.futurelearn.com/press-releases/futurelearn-delivers-the-largest-mooc-ever-as-nearly-400000-learners-convene-for-english-language-learning/). Because of these numbers, there is a risk that MOOCs become one-way information transfer, and novel methods are needed to complete the teach-learn-assess/feedback cycle (Suen, 2014).

Most MOOCs on the UK-based FutureLearn platform are not assessed summatively, instead offering Certificates of Completion or Participation for a small fee. Summative assessment is more widely used on the major platforms such as Coursera and EdX. The use of automated multiple choice questions (MCQ) to test knowledge and understanding throughout a course is common on all platforms. However, it can be time-consuming and challenging to write MCQs that measure high-level cognition such as analysis, synthesis, and evaluation (Suen, 2014). In particular, assessment of qualitative work requires novel approaches to cope with the numbers involved. It is logistically challenging for staff to mark the large number of student assignments. Qualitative assessment must therefore come either from other students enrolled on the course, or from further developing automated processes to mark written work. As MOOCs are beginning to be taken by students for credit (e.g. http://blog.coursera.org/post/42486198362/five-courses-receive-college-credit), the evaluation of suitable forms of assessment for this type of online learning environment is essential. Formative feedback, in particular qualitative comments, is equally problematic. In MOOCs with elements of social learning such as the one in this study, staff and other students can reply to individual learners in discussion activities, which can provide useful feedback both to those learners and to others reading it, but is not systematic. Staff can also post generic feedback after assessments, and automated replies to responses in MCQs can be used to provide feedback and direct students to additional resources.

Principles of peer assessment

The benefits of using peer assessment and review in teaching include the capacity to give feedback to large numbers (Topping, 2009), and the transferable skills gained by students

during the process (Morris, 2001). Students gain from reviewing as well as from their own work being reviewed (e.g. Demiraslan Çevik, 2015). However, although many studies do show clear benefit to peer assessment and review, the evidence for their effectiveness on learning overall is mixed, possibly because of the variability of assessment tasks used (Topping, 2010). In terms of student perception, many students believe that it improves their learning, increases motivation, and helps them develop critical thinking and empathy (Falchikov, 1986; Falchikov and Goldfinch, 2000; Hanrahan and Isaacs, 2001). However, some students find marking others' written assignments difficult, uncomfortable and time-consuming (Hanrahan and Isaacs, 2001). It is unclear whether these benefits and drawbacks also apply in MOOCs.

In this study, we use qualitative peer review. This is less studied than quantitative peer assessment (Topping, 1998), and as such has less support from the literature. However, qualitative feedback, while harder to show reliability and validity, can provide much more detailed information about how a student can improve their work in future. Cho & Cho (2011) show that peer feedback without grading improves student performance, and suggest that quality comments focusing on meaning rather than surface features are more likely to have this beneficial effect.

Peer review in massive open online courses

In the classroom, peer review is used not only to improve subject learning, but also to develop transferable skills such as critical analysis, communication skills, and teamwork (Morris, 2001). However, arguably, solving the issue of assessment and feedback with large student numbers, and improving the student experience (and thus retention), have been the main motivation for peer review in MOOCs to-date, as most of the literature focuses on these aspects.

MOOC entry is by definition open, and as such participants come from a wide variety of backgrounds. Overall, however, participants tend to be young, well-educated, employed, and mainly from developed countries (Christensen et al., 2013). The potential range of backgrounds and abilities may mean that learners are not in fact true peers i.e. they may have unequal subject knowledge and language proficiency. Furthermore, different learners likely have their own, possibly very different, intended learning outcomes, illustrated by different behaviour patterns such as 'browsers' versus 'committed learners' (Koller et al., 2013); or 'lurkers', 'passive participants' and 'active participants' (Milligan et al., 2013). As such, it is possible that some learners may not engage with peer assessment at all, and may drop out if it is a course requirement.

Coursera, a major MOOC provider, has developed a peer review system involving both summative grading and qualitative comments. This uses multiple peer reviewers, with students each commenting on 4 other pieces of work. Peer assessment grades on a geography MOOC correlated strongly with tutor marks, but inter-rater reliability was poor (Luo et al., 2014). Similarly, Admiraal et al. (2014) found low inter-reviewer reliability in peer assessment on three humanities MOOCs, with low or moderate reliability of peer assessments as a predictor of performance in weekly quizzes and final exams. However, several groups are now developing statistical systems to increase the reliability of peer assessment, such as UCLA's Calibrated Peer Review software. This allows peer reviewers to 'calibrate' their grading based on instructors' grading of several sample essays. However, only summative scores can be calibrated (not formative comments) and multiple peer reviewers are required (Balfour 2013, Suen 2014, Diez et al. 2015). Shah et al. (2013) found that ordinal peer grading was more robust than cardinal grading in a MOOC. In contrast to the peer review approach, EdX uses an Automated Essay Scoring (AES) system trained on instructor-scored student essays to grade freeform answers (Mitros et al., 2013). In general, however, AES is not yet able to predict human scores for writing more complex than short essays with very specific focus (Balfour 2013).

In a MOOC on Design, Kulkarni et al. (2014) found that although peer and staff grades were highly correlated, only 43% of students gave grades within 5 % of the tutor grade, and 66 % within 10%, even after calibration had been performed. Notably, however, this moderate agreement of staff and peer grades was similar for the in-person version of the course. This suggests that consistency may be problematic regardless of whether peer review is in-person or online. Qualitative feedback was not analysed in the above studies, but there are anecdotal reports of problems of wide variability in peer feedback quality in a writing MOOC (e.g. Gibbs, 2012). Staff oversight of reviewer comments is a possible solution, but would be challenging with a large class size.

In summary, there is currently little evidence to determine whether peer review is useful in MOOCs, and under what conditions. Here, we examine student participation, performance and opinions in a peer review exercise in a biomedical science MOOC. We evaluate data from 203 student summaries and 314 qualitative peer reviews along with demographic data, to ascertain the suitability of this peer review as an assessment and feedback tool for MOOC students.

Methodology

Background

Data was gathered from a six-week MOOC entitled 'Cancer in the 21st Century: The Genomic Revolution', which ran in May-June 2014 on the FutureLearn platform. There were over 7000 students enrolled on this course. Of these, around 2500 were active learners (i.e. completed at least one learning activity), 203 students submitted a written summary assignment, and 192 students participated in peer review (producing 314 peer reviews).

Student task - structure, instructions and scaffolding

i) Written summary. The student summary assignment and the peer review took place in the final week of the course. Students were introduced to the topic of epigenetics through a video and a short article. They were also invited to watch a video titled 'Using a Scientific Literature Database' to help them prepare for the written assignments. Students were then asked to write ~300 words in answer to the question, 'What do we know about how epigenetic regulation goes wrong in cancer, and what types of targeted treatment could arise from our knowledge of epigenetic de-regulation in cancer?' Students were asked to report which resources they used.

Links to four open-access articles were provided to assist the students with this task, although the instructions stated that students were free to search for and use any resource they felt appropriate.

ii) Peer review. Various instructional techniques can be used to improve the quality of peer reviewers' comments, such as using directed questions or sentence openers (Gielen et al., 2010). Asking reviewers to make global judgements, based on well-understood criteria, increases the validity and reliability of peer feedback (Falchikov and Goldfinch 2000). Structuring feedback using a template may be beneficial (Gielen & de Wever, 2015; Ashton and Davies, 2015). In this study, each student's written summary was passed on to another learner electronically, who reviewed it and provided qualitative feedback under three specific subheadings:

1. What did you like about the author's work?

2. Had they carried out research using reliable resources and had good use been made of these?

3. How might the author improve the communication of their key ideas?

Students could peer review as many written summaries as they chose to, with numbers ranging from 1-7.

Data collection

Demographic data was collected from students who chose to fill in pre- and post-course surveys. Using a unique identifier code, this data was linked anonymously to 79 students who participated in both the written assignment and the peer review process. Demographic data analysed in relation to summary and peer review performance included age, gender, previous level of education, whether students were currently employed or in full time education, field of employment and country of residence.

Qualitative comments from students were gathered from the 'comments' section associated with the peer review task. Comments specifically pertaining to the peer review exercise were also gathered from the post-course survey, and from the comments section at the end of the week and course. Comments related to the topic of the week in general (epigenetics) or the course as a whole were not included.

Thematic analysis

Student comments were subject to content analysis using NVivo software to identify themes arising. The Miles and Huberman approach to analytic induction was used, with successive rounds of coding (Punch, 2005). Coding started with descriptive nodes for each comment (such as ‘word count was too low’ and ‘peer review task consolidated my knowledge’). These were then grouped into more inferential nodes such as ‘benefits of being reviewed’ and ‘problems with assignment’, and finally further grouped into overall ‘positive’ and ‘negative’ nodes). Coding categories were verified independently by two researchers, to increase reliability.

Staff assessment of student written summaries and peer reviews

Each student summary was graded (A-E) independently by two reviewers. Clarity, accuracy, originality (based on a Turnitin report), ability to adhere to the word count and the use of references were considered. The reviewers compared grades, discussed differences and assigned a final agreed grade to each report. The rubric used by markers is available on request from the authors.

Marking criteria for peer reviews

Gielen et al. (2010) summarise criteria used to define ‘good’ qualitative feedback, such as the presence of thought-provoking questions or comments, and constructive suggestions. Whether feedback is ‘good’ can also be defined by its accuracy and consistency with expert feedback, or in terms of its style and general content. Marking criteria were developed to

evaluate the quality of peer reviews, which included whether the students had attempted to provide constructive criticism and whether instructors perceived their feedback on the summary to be valid. To do this, each peer review was matched to the original written summary. In some cases more than one student had provided a peer review for an individual summary (highest number of peer reviews per student was 7; lowest zero). Six of the final summaries were not peer reviewed. Peer review marking was anonymised. Each peer review was marked by two academics who later discussed and agreed upon a final grade (0-3).

Analysis of Peer Review Grading

First, for all 203 written summaries we compared student proficiency in the written summary exercise (grade obtained A-E) with the number of peer reviews completed. We also compared written summary proficiency with the quality of peer review comments each student wrote (grade obtained 0-3) using a t-test.

We then compared the demographics of the 79 students who completed both written summaries and peer reviews, versus the overall demographics of the 747 students who participated the pre-course survey.

For subsequent analysis, we compared the demographics (age, gender, previous education, whether in full time work, and whether or not English was a national language in the country of residence) of those who passed versus those who failed the written summary assignment. Because of the relatively small student numbers (79), each demographic category was split into only two groups (e.g. age: under 45 and over 45)

To determine whether success in the written summary was associated with any of the demographics, students were divided into 2 groups – (pass, grades A-C) and (fail, grades D and E) - and Chi squared analysis was performed.

Next, to determine whether peer review quality was associated with any of these same demographics, the average peer review grade was calculated for each category of demographic and t-testing was performed.

Finally, we compared performance in the peer review tasks (written summary and the subsequent peer review of other's summaries) with performance on quizzes within the MOOC, to determine whether there was any correlation in success between these tasks. If a student submitted an incorrect answer they could retry questions until they found the correct answer. Therefore, each student's average grade was calculated by dividing the number of questions answered correctly on the first attempt by the total number of questions answered. The average number of retries was calculated by subtracting the total number of unique questions answered by the total number of answers submitted by a student.

Findings

Peer Review Quality and Grades

Firstly, the overall quality of peer reviews written by the students was high (Figure 2). The majority of students' reviews were an attempt to give at least a reasonable standard of feedback. Notably, a high number of students (44 of 192) were allocated 'summary non-submissions' (SNS) to review. These were 'dummy' written summaries entered by students for various reasons that contained no relevant content. This may have adversely affected their perception of the peer review process and influenced whether they undertook additional peer reviews. Students who were allocated this type of summary to review were not allocated a grade.

Secondly, students who did well in the written summary task gave significantly higher quality peer reviews than those who did not (Table 1). Furthermore, students who did well in the written summary task were more likely to complete multiple peer reviews (Figure 3). This may indicate that more able students have higher levels of engagement with Peer Review.

Student Demographics, Participation and Performance

Demographic data was collected from 747 students who participated in the pre-course survey, of whom 79 participated in both the Written Summary assignment and Peer Review (Table 2).

Table 2 shows that a smaller proportion of females than males participated in the peer review task, compared to the total number of each gender at the start of the MOOC. peer review participation was also biased in favour of those who had a university-level education, which could indicate that those without higher education experience may find the task too challenging or intimidating. Younger students (age 18-25) and those in full-time education, categories which likely overlap, also showed proportionally reduced participation in the peer review exercise. These students may not want to take part in additional academic tasks outwith their non-MOOC courses. Lastly, the retired and those unavailable for work showed increased peer review participation, which may indicate time pressure as a factor influencing participation.

Which factors predict written summary and peer review success?

Of the 79 students who undertook the peer review, we compared the demographics of those who passed (38 of 79) versus those who failed (41 of 79) the written summary assignment (Table 3). For each of these categories, the dependent variable was whether the student was deemed to have written a satisfactory (grades A-C) or unsatisfactory (grades D-E) written summary. Chi squared analysis was carried out for each condition. This showed that working full time and the level of prior education significantly altered the likelihood of students passing the summary assignment. Students in full-time employment were almost twice as likely to fail the assignment. Students with a doctorate were significantly more likely to pass the summary assignment (Table 3).

Of the 79 students who undertook the peer review, we compared the demographics (as in previous section) with the average grade awarded by academic staff to the peer review (Table 4). peer review quality was not correlated with age, gender, employment status, or level of previous education. However, students in countries where English is a national language gave slightly higher quality peer reviews, with an average peer review grade of 2.5 versus 2.2 .

Performance in peer review versus performance in other assessments

Performance in the summary assignment was compared to performance in the formative MCQs throughout the course using students' unique FutureLearn Identifier numbers, for the 79 students who completed the written summary exercise (Table 5). Students who were proficient in the summary exercise (grades A-C) were significantly more likely to achieve a higher average grade in these MCQs and to require fewer 'retries' to obtain the correct answer than those with a grade D or E in the summary exercise.

There were no significant differences in engagement with MCQs when students were grouped by demographic category, with the exception of age. Students aged 26 to 45 completed significantly fewer formative MCQs throughout the course than those in other age groups. The average number of MCQs completed (out of a total of 65) by students aged 26 to 45 was 57, in comparison to 63 by students in all other age categories ($p=0.002$).

We then examined the quality of peer reviews written by students in comparison to their engagement with the MCQs. Students who wrote high quality peer reviews were significantly more likely to achieve a higher average grade in the multiple choice questions (Table 6). Interestingly, these students were also likely to require significantly fewer attempts to achieve the correct answer in the formative MCQs. There was no apparent link between the quality of peer reviews written by students and the average number of MCQs that they completed throughout the course, consistent with Admiraal et al. (2015).

Student Perceptions of Peer Review

Student perceptions of the peer review task were investigated by qualitative analysis of their comments in the comments sections of the course and in the post-course survey. Overall, students made a roughly equal number of positive (93) and negative (105) comments, with a few specific suggestions to improve (6). Three participants named the peer review task as their favourite part of course; 30 named it their least favourite. Two people specifically asked for more writing assignments.

Perceived benefits of peer review

Twenty-three comments identified specific positive aspects of having work reviewed by peers (the other 70 comments may have also referred to other aspects of the task). These included stating that the reviewer's comments were useful or encouraging (7 occurrences), and that they confirmed the writer's understanding (3) and appropriateness of depth (2). In keeping with this, seven students used the comments thread to thank their reviewer.

Participants valued the varied backgrounds of the reviewers (2), indicating that some students value different viewpoints. Writing for review by others was seen as an additional and valued challenge (2). Twelve comments mentioned specific positive aspects of reviewing others' work. Seeing others' insights and viewpoints was seen as a benefit of being a reviewer (6), including how others dealt with specific aspects (e.g. word limit) and seeing an example of good writing. Three students noted that the act of reviewing can confirm your understanding of the topic. One student commented that they had learned the utility of peer review itself from this task.

Commonest negative aspects mentioned by participants

Of those students (13) who disliked peer review specifically, almost half of these comments (6) relate to receiving poor quality reviews. Six students stated that they thought peers were not qualified to review work, and / or that they felt uncomfortable reviewing peers. Only one person had a previous bad experience of peer review.

Discussion

Main findings. We show that students can benefit from the inclusion of a peer review writing exercise in a MOOC. While previous reports indicate high variability in peer feedback quality (Gibbs, 2012), this study highlights the overall high standard of qualitative peer reviews completed by students on a biomedical science MOOC. Performance in the written task, peer review and MCQs was correlated, suggesting there are a group of students who perform well across the board. However, student opinion on the usefulness of the peer review task was mixed: some strongly believed it benefitted their learning, while others did not find it useful and a significant number did not participate. Certain groups may require more support or scaffolding of the task - for example those with less academic experience or whose first language is not English.

Engagement with the peer review exercise. Not all MOOC participants chose to participate in the peer review exercise. From analysis of the demographics of students at the start of the course in comparison to those involved in peer review, it seems that groups less likely to engage were females, those in full-time education, those with no previous degree and those aged 18-35 (Table 2). The latter three groups are likely overlap, and it is possible that those who are already studying formally may not wish to complete additional academic tasks. Retired learners were more likely to participate (Table 2e), which suggests that time constraints may be a key factor in determining participation. Another key group less likely to participate was students with no prior university education. Although we cannot tell why this was the case, we speculate that students without a degree may need more support with peer review activities. It is vital to investigate this possibility further if MOOCs are to fulfil their aim of inclusiveness. However, it is important to note that overall student participation in MOOC activities is variable, with several distinct patterns of engagement depending on their own goals (Milligan et al., 2013; Koller et al., 2013). Indeed, it is difficult to judge what level of engagement with peer review is desirable, as students set their own learning goals. It is therefore conceivable that some students may choose not to do the peer review and still be happy with their MOOC experience, depending on their own learning aims. The data in this study does not provide evidence for or against this possibility, but it is important to explore this further in the future. Awarding virtual ‘badges’ as rewards to learners has been shown to result in higher levels of engagement with peer review in an online setting (O’Connor & McQuigge, 2013).

A sub-group of students perform well throughout the course, including Peer Review. Students who achieved the highest mark in the quality of their peer reviews had an MCQ average grade of approximately 10% higher than those with the lowest quality peer reviews. Similarly, students who were proficient in the written summary exercise were also achieved a 10% higher average MCQ grade than those who failed the summary exercise, and required

fewer ‘retries’ to pick the correct MCQ answer. Students who performed well in the writing exercise also completed more MCQs and more peer reviews than poorer students, indicating more engagement. These results are consistent with Comer et al (2014), who show that more proficient students tend to engage and participate well, and complete the course. Due to the marking criteria for the quality of peer reviews, students who have a better understanding of the course content are likely to have achieved a higher average peer review grade, since an understanding of both the subject area and of what was required of students in the peer review process was required for what we deemed to be high quality peer reviews. Interestingly, students with lower quality peer reviews required more attempts to select the correct MCQ answers, but did not complete significantly fewer MCQs. This emphasises that a good understanding of the subject area may be important for positive engagement in the Peer Review task.

Student perceptions of peer review in the MOOC. Students appreciated both having their work reviewed and reviewing others’ work. These findings are similar to literature on perceived benefits of peer review in other contexts, where many students say peer review improves their learning and increases motivation (Falchikov, 1986; Falchikov and Goldfinch, 2000; Hanrahan and Isaacs, 2001). In a Chemistry MOOC, students said they learned from providing feedback, demonstrated what they had learned, and increased their understanding (Comer et al., 2015). Although many students liked the peer review exercise, a significant number of students viewed it negatively. This is consistent with the mixed student perceptions of peer review found in the literature (Hanrahan and Isaacs, 2001; Vickerman, 2009). Surprisingly, only a few students disagreed with peer review *per se*, stating that staff should mark. This contrasts with Sluijmans et al. (2001), who found that only 19% of students agreed that students are capable of assessing each other fairly and responsibly in a classroom setting. It may be that MOOC participants are particularly open to collaborative tasks, as part of the MOOC ethos is participant collaboration, compared to on-campus students.

Implications for peer review design in MOOCs. Most of the negative comments were about the practicalities of the exercise e.g. word limit and time constraints, and appropriateness of a writing task. These issues could be addressed by explaining more clearly to students the rationale for the exercise and format.

i) Mixed abilities. Some students said the task was too difficult, but these numbers were consistent with numbers for the MOOC overall (27% said course was too challenging in post-course survey). This implies that the peer review task is at the correct level for the course, but that some participants found the course too advanced. MOOC staff should consider the level at which they want to pitch a course, and ensure that course advertising is consistent with this. FutureLearn is currently developing functionality to allow groups within MOOCs:

streaming based on interests, ability or background, might better cope with a mixed student group. This may be particularly important given the possibility of regressive collaboration in peer feedback, where students are persuaded by peers to change from an appropriate to an inappropriate understanding (Sainsbury & Walker 2008).

ii) Instructions, signposting, and scaffolding

A significant proportion of summaries were not a genuine attempt at a peer review ('Summary Non-Submissions', SNS). It was perhaps not clear to students that they could opt out of the peer review exercise. In order to mark this section 'complete', many students (50 out of 313, 16%) therefore submitted random characters or an explanation of why they did not submit the assignment. These 'summaries' were then allocated to other students to peer review, giving some students an unnecessarily negative peer review experience. FutureLearn has since implemented technology that allows students to request an alternative assignment to review in such cases. Alternatively, a separate 'opt out' stream that allows students who do not wish to complete the task to progress separately (Ossiannilsson et al., 2015), or a quality control step (manual or automated) to filter out summaries that do not contain academic content, could also be used. Estévez-Ayres et al (2013) have developed an algorithm for review allocation that minimises the number of students who don't receive peer feedback.

Both review quality and performance quality increase with practice (Gielen & de Wever 2015, Cheng and Hou 2015). Multiple peer review tasks may therefore be needed to see the most benefit. Similarly, extensive scaffolding or teaching about how to write peer reviews and make the most of feedback improves peer review in classroom settings (e.g. Taylor *et al* 2015). However, this may be neither feasible or desirable in the context of a short (6-week) MOOC. In this study, in a stand-alone peer review exercise with three questions to scaffold the peer reviews, most peer reviews were judged by staff to be of high quality (Figure 2), implying that this level of scaffolding was sufficient for students who chose to take part.

Given that MOOCs' remit is in part to widen access to education, we would like to find ways to better engage, and improve performance by, less able or less resourced students. Although there was no relationship of prior education level with peer review quality and no relationship below doctoral level with written summary grade (Table 4), it is important to note that students with no university education were less likely than university-educated students to participate in the peer review task (Table 2c). It may be that peer review is better suited to more able or experienced students, or that more guidance or support is needed for particular groups of students to benefit from it. Alternatively, less able or experienced students may benefit from additional instructions and scaffolding for peer review. Most of the negative comments about the peer review task were about its design and rationale (e.g. why the format was a written summary or criticism of the word limit): we therefore recommend that the rationale for the task be clearly communicated to students, as emphasised in Planas Llado et al. (2014). Finally, MOOC participation is generally international, and our results suggest that students whose first language is not English may need additional support in peer review writing tasks.

Limitations of study.
First, the use of post-course surveys and comments sections means that this analysis reflects the views of learners who choose to comment, and / or who have completed the course. Second, the demographic data collected in the pre-course questionnaire may not reflect the demographic mix immediately prior to the peer review exercise. Third, we were only able to make inferences regarding whether English was a student's first language from the country they live in: students whose first language is English may reside in countries where English is not a national language, and vice versa. Finally, the conclusions we have drawn apply to formative peer review of a written assignment, in a MOOC with a mix of didactic and social learning, and may or may not hold true for other contexts. For instance, English proficiency may not be as important in, for example, a computer programming task; or students may be more accepting of peer review in MOOCs with more social learning.

Future Directions. Further investigations could address whether peer review in the MOOC context produces learning gains directly, by examining performance before and after the exercise, or by specifically asking students about perceived learning gains. Cromer et al. (2014) found that peer review can facilitate learning gains in a MOOC.

Disclosure Statement

The authors declare no potential conflict of interest.

References

Admiraal, W., Huisman, B. and Van de Ven, M. 2014. "Self- and Peer Assessment in Massive Open Online Courses." *International Journal of Higher Education* 3 (3). doi:10.5430/ijhe.v3n3p119.

Admiraal, W., Huisman, B. and Pilli, O. 2015. "Assessment in Massive Open Online Courses." *The Electronic Journal of E-Learning* 13 (4): 207–16.

Ashton, S. and Davies, R.S. 2015. "Using Scaffolded Rubrics to Improve Peer Assessment in a MOOC Writing Course." *Distance Education* 36 (3): 312–34. doi:10.1080/01587919.2015.1081733.

Balfour, S.P. 2013. "Assessing Writing in MOOCs: Automated Essay Scoring and Calibrated Peer Review (TM)." *Research & Practice in Assessment* 8 (Summer 2013): 40–48.

Cheng, K, and Hou, H. 2015. "Exploring Students' Behavioural Patterns during Online Peer Assessment from the Affective, Cognitive, and Metacognitive Perspectives: A Progressive Sequential Analysis." *Technology, Pedagogy and Education* 24 (2): 171–88. doi:10.1080/1475939X.2013.822416.

Cho, Y.H., and Cho, K.. 2011. "Peer Reviewers Learn from Giving Comments." *Instructional Science* 39 (5): 629–43. doi:10.1007/s11251-010-9146-1.

Christensen, G., Steinmetz, S., Alcorn, B., Bennett, A., Woods, D. and Emanuel, E. 2013. "The MOOC Phenomenon: Who Takes Massive Open Online Courses and Why?" http://ssrn.com/abstract=2350964 orhttp://dx.doi.org/10.2139/ssrn.2350964.

Comer, D.K., Clark, C.R., and Canelas, D.A. 2014. "Writing to Learn and Learning to Write across the Disciplines: Peer-to-Peer Writing in Introductory-Level MOOCs." *International Review of Research in Open and Distributed Learning* 15 (5). http://www.irrodl.org/index.php/irrodl/article/view/1850/3066.

DeBoer, J., Ho, A.D., Stump, G.S. and Breslow, L. 2014. "Changing 'Course': Reconceptualizing Educational Variables for Massive Open Online Courses." *Educational Researcher* 43 (2): 74–84. doi:10.3102/0013189X14523038.

Demiraslan Çevik, Y. 2015. "Assessor or Assessee? Investigating the Differential Effects of Online Peer Assessment Roles in the Development of Students' Problem-Solving Skills." *Computers in Human Behavior* 52 (November): 250–58. doi:10.1016/j.chb.2015.05.056.

Diez, J., Luaces, O., Alonso-Betanzos, A., Troncoso, A. and Bahamonde, A. 2013. "Peer Assessment in MOOCs Using Preference Learning via Matrix Factorization." In *Proceedings of NIPS Workshop on Data Driven Education*. Lake Tahoe.

Estévez-Ayres, I., Crespo-García, R.M., Fisteus, J.A. and Delgado Kloos, C. 2013. "An Algorithm for Peer Review Matching in Massive Courses for Minimising Students' Frustration." *Journal of Universal Computer Science* 19 (15): 2173–97. doi:10.3217/jucs-019-15-2173.

Falchikov, N. 1986. "Product Comparisons and Process Benefits of Collaborative Peer Group and Self Assessments." *Assessment & Evaluation in Higher Education* 11 (2): 146–66. doi:10.1080/0260293860110206.

Falchikov, N., and Goldfinch, J. 2000. "Student Peer Assessment in Higher Education: A Meta-Analysis Comparing Peer and Teacher Marks." *Review of Educational Research* 70 (3): 287–322. doi:10.3102/00346543070003287.

Gibbs, L. 2012. "Peer Feedback: The Good, the Bad and the Ugly." http://courserafantasy.blogspot.co.uk/2012/08/peer-feedback-good-bad-and-ugly.html.

Gielen, M., and De Wever, B. 2015. “Structuring the Peer Assessment Process: A Multilevel Approach for the Impact on Product Improvement and Peer Feedback Quality: Structuring the Peer Assessment Process.” *Journal of Computer Assisted Learning* 31 (5): 435–49. doi:10.1111/jcal.12096.

Gielen, S., Peeters, E., Dochy, F., Onghena, P. and Struyven, K. 2010. “Improving the Effectiveness of Peer Feedback for Learning.” *Learning and Instruction* 20 (4): 304–15. doi:10.1016/j.learninstruc.2009.08.007.

Hanrahan, S.J., and Isaacs, G. 2001. “Assessing Self- and Peer-Assessment: The Students’ Views.” *Higher Education Research & Development* 20 (1): 53–70. doi:10.1080/07294360123776.

Koller, D., Ng, A., Do, C. and Chen, Z. 2013. “Retention and Intention in Massive Open Online Courses: In Depth.” *EDUCAUSE Review Online*, June. http://er.educause.edu/articles/2013/6/retention-and-intention-in-massive-open-online-courses-in-depth.

Kulkarni, C.E., Socher, R. Bernstein, M.S. and Klemmer, S.R.. 2014. “Scaling Short-Answer Grading by Combining Peer Assessment with Algorithmic Scoring.” In *Proceedings of the First ACM Conference on Learning @ Scale Conference*, 99–108. ACM Press. doi:10.1145/2556325.2566238.

Luo, H., Robinson, A.C. and Park, J. 2014. “Peer Grading in a MOOC: Reliability, Validity, and Perceived Effects.” *Online Learning: Official Journal of the Online Learning Consortium; Vol 18, No 2 (2014)*, June. http://olj.onlinelearningconsortium.org/index.php/jaln/article/view/429.

Milligan, C., Littlejohn, A. and Margaryan, A. 2013. “Patterns of Engagement in Connectivist MOOCs.” *Journal of Online Learning and Teaching* 9 (2). http://jolt.merlot.org/vol9no2/milligan_0613.htm.

Mitros, P.F., Paruchuri, V., Rogosic, J. and Huang, D. 2013. “An Integrated Framework for the Grading of Freeform Responses.” In *Proceedings of the 6th International Learning International Networks Consortium Conference*. Cambridge, MA. http://linc.mit.edu/linc2013/proceedings/Session3/Session3Mit-Par.pdf.

Morris, J. 2001. “Peer Assessment: A Missing Link between Teaching and Learning? A Review of the Literature.” *Nurse Education Today* 21 (7): 507–15. doi:10.1054/nedt.2001.0661.

O’Connor, E.A., and McQuigge, A.. 2013. “Exploring Badging for Peer Review, Extended Learning and Evaluation, and Reflective/Critical Feedback Within an Online Graduate Course.” *Journal of Educational Technology Systems* 42 (2): 87–105. doi:10.2190/ET.42.2.b.

Ossiannilsson, E., Altinay, F. and Altinay, Z. 2015. “Analysis of MOOCs Practices from the Perspective of Learner Experiences and Quality Culture.” *Educational Media International* 52 (4): 272–83. doi:10.1080/09523987.2015.1125985.

Planas Lladó, A., Feliu Soley, L., Fraguell Sansbelló, R.M., Arbat Pujolras, G., Pujol Planella, J., Roura-Pascual, N., Suñol Martínez, J.J., and Montoro Moreno, L. 2014. “Student Perceptions of Peer Assessment: An Interdisciplinary Study.” *Assessment & Evaluation in Higher Education* 39 (5): 592–610. doi:10.1080/02602938.2013.860077.

Punch, K.F. 2005. *Introduction to Social Research: Quantitative and Qualitative Approaches*. 2nd ed. London ; Thousand Oaks, Calif: SAGE. p197-202.

Sainsbury, E.J., and Walker, R.A. 2008. “Assessment as a Vehicle for Learning: Extending Collaboration into Testing.” *Assessment & Evaluation in Higher Education* 33 (2): 103–17. doi:10.1080/02602930601127844.

Shah, N.B., Bradley, J., Parekh, A., Wainwright, M.J. and Ramchandran, K.. 2013. “A Case

for Ordinal Peer-Evaluation in MOOCs." In *Proceedings of Neural Information Processing Systems (NIPS): Workshop on Data Driven Education*. Lake Tahoe. http://lytics.stanford.edu/datadriveneducation/papers/shahetal.pdf.

Sluijsmans, D.M.A., Moerkerke, G., van Merriënboer, J.J.G., and Dochy, F.J.R. 2001. "Peer Assessment in Problem Based Learning." *Studies in Educational Evaluation* 27 (2): 153–73. doi:10.1016/S0191-491X(01)00019-0.

Suen, H.K. 2014. "Peer Assessment for Massive Open Online Courses (MOOCs)." *The International Review of Research in Open and Distance Learning; Vol 15, No 3 (2014)*, June. http://www.irrodl.org/index.php/irrodl/article/view/1680/2904.

Taylor, S., Ryan, M. and Pearce, J.. 2015. "Enhanced Student Learning in Accounting Utilising Web-Based Technology, Peer-Review Feedback and Reflective Practices: A Learning Community Approach to Assessment." *Higher Education Research & Development* 34 (6): 1251–69. doi:10.1080/07294360.2015.1024625.

Topping, K.J. 1998. "Peer Assessment Between Students in Colleges and Universities." *Review of Educational Research* 68 (3): 249–76. doi:10.3102/00346543068003249.

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

Topping, K.J. 2010. "Methodological Quandaries in Studying Process and Outcomes in Peer Assessment." *Learning and Instruction* 20 (4): 339–43. doi:10.1016/j.learninstruc.2009.08.003.

Vickerman, P. 2009. "Student Perspectives on Formative Peer Assessment: An Attempt to Deepen Learning?" *Assessment & Evaluation in Higher Education* 34 (2): 221–30. doi:10.1080/02602930801955986.

Tables and Figures

Table 1. Comparison of written summary grade with peer review grade.

* denotes a significant difference in the average peer review grade of students who passed their written summary assignment compared to those who failed (p=0.019).

Written Summary Grade	**Average peer review grade (0-3)**
Pass (A-C)	2.7
Fail (D-E)	2.2*

Table 2. Demographics of those who participated in peer review exercise compared to overall student demographics at start of course.

(a) Gender

	Start of course (%)	**Peer Review (%)**
Male	27	42
Female	69	57

Note: Where the percentages do not add up to 100% this is because some respondents to the survey have chosen not to answer this particular question.

(b) Country of residence

	Start of course (%)	**Peer Review (%)**
UK	79	76
USA	7	3
India	3	4
Spain	3	1
Australia	2	4
Greece	2	3
New Zealand	1	1
Canada	1	3
Nigeria	1	1
Germany	1	1

(c) Level of prior education

	Start of course (%)	**Peer Review (%)**
No university	24	13
Undergraduate Degree	42	46
Masters	26	29
Doctorate	8	10

(d) Age

	Start of course (%)	**Peer Review (%)**
<18	4	0
18-25	22	11
26-35	16	18
36-45	12	14
46-55	15	18
56-65	16	23
>66	9	15

(e) Employment status

	Start of course (%)	**Peer Review (%)**
Full time employed	35	34
Part time employed	17	15
Full time education	20	10
Unemployed	7	5
Retired	17	27
Not available to work	3	9

Table 3. The effect of demographics on written summary quality

Countries in which English is considered a national language were Australia, Canada, New Zealand, Nigeria, UK and USA English is not considered a national language in Columbia, Cyprus, Germany, Greece, India, Nepal, Pakistan, Portugal, Russia, Saudi Arabia, Serbia, and Spain. P-values are given for statistically significant results.

Where the total numbers are not 79 this means that some participants did not answer this question.

	Demographic Category	**Passed (A-C)**	**Failed (D/E)**	**Significance (Chi squared analysis)**
Age	Under 45	13 (38%)	21 (62%)	NS
	Over 45	25 (57%)	19 (43%)	
Gender	Female	22 (49%)	23 (51%)	NS
	Male	15 (45%)	18 (55%)	
Working?	In full time work	8 (30%)	19 (70%)	**p = 0.018***
	Not in full time work	30 (58%)	22 (42%)	
Education	Doctorate	7 (88%)	1 (12%)	**p = 0.017***
	No doctorate	29 (43%)	39 (57%)	
National language	English	33 (53%)	30 (47%)	NS
	Non English	5 (31%)	11 (69%)	

Table 4. The effect of demographics on peer review quality.

P-values are given for statistically significant results.

	Demographic Category	**Average peer review grade**
Age	Under 45	2.4
	Over 45	2.4
Gender	Male	2.5
	Female	2.4
Working?	In full time work	2.4
	Not in full time work	2.4
Education	Doctorate	2.5
	No doctorate	2.4
National language	English	2.5
	Non English	2.2
		p = 0.046*

Table 5. Comparison of written summary grade with engagement in formative MCQs throughout course

† denotes a statistically significant difference (p=0.0009) in the average MCQ grade in comparison to students with a written summary grade of A-C. ‡ denotes a statistically significant difference (p=0.0016) in the average number of MCQ retries in comparison to students with a written summary grade of A-C.

Written Summary Grade	**Average MCQ grade (in %)**	**Average number of MCQ retries (attempts to achieve correct answer)**	**Average number of MCQs completed (total 65)**	**Number of students achieving summary grades**
A-C (Pass)	84	12	63	38
D and E (Fail)	75†	20‡	60	41

Table 6. Comparison of the quality of peer reviews written by students and their performance and engagement with the formative MCQs.

The average number of MCQs answered is indicated, out of a maximum of 65. A peer review grade of three is the highest. Students who were not allocated a summary to review and those that could not be matched to their peer reviews (8 in total) were excluded from this analysis. † denotes a statistically significant difference (p=0.020) between average MCQ grades in comparison to students who were awarded an average peer review grade of three. ‡ denotes a statistically significant difference in the average number of MCQ retries (p=0.002) in comparison to students who were awarded an average peer review grade of three.

Average peer review quality	Average MCQ grade (in %)	Average number of MCQ retries	Average number of MCQs completed (total 65)	Number of students achieving peer review grade
3	82	13	61	41
2	79	17	62	25
0 and 1	72†	25‡	62	5

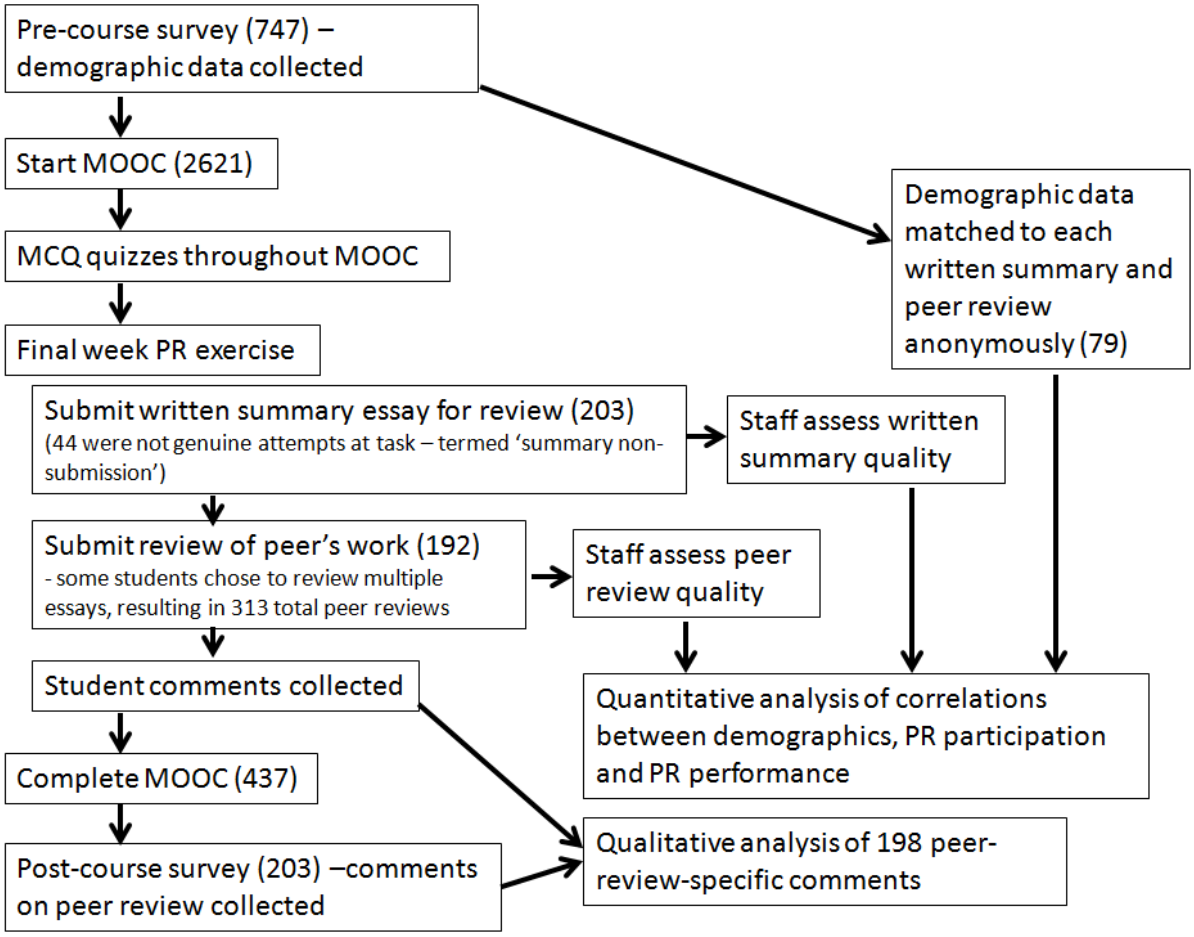

Figure 1. Flow chart of assessment in MOOC (student numbers in brackets) and research study design.

Figure 2. Peer review quality, as judged by staff

Grading criteria available on request from authors. SNS= summary non-submission.

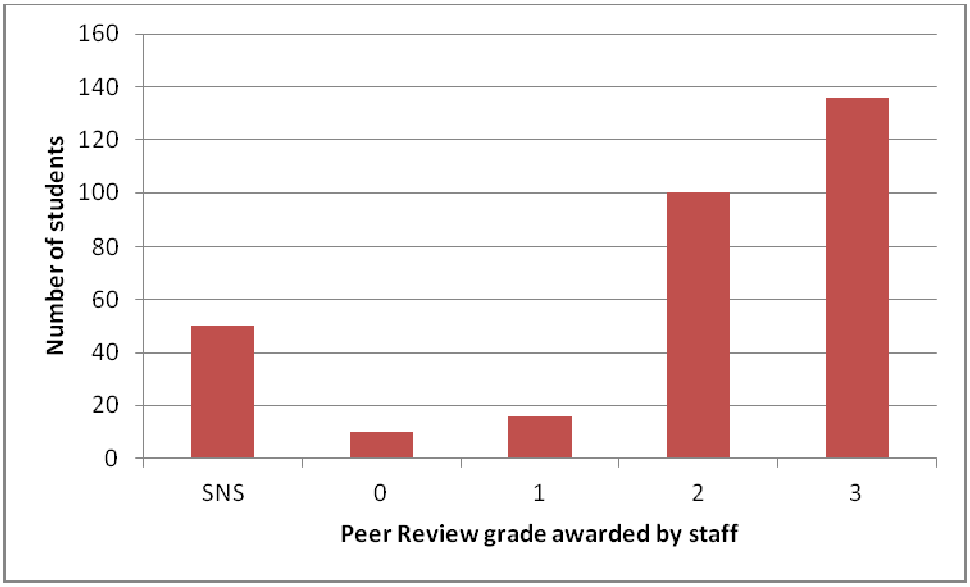

Figure 3. Students who conducted two or more peer reviews were more likely to be proficient in the reading summary assignment.

Students who were awarded a pass grade A, B or C by MOOC academics were deemed proficient in the reading summary written exercise. Students who were awarded a fail grade D or E in this exercise were deemed to be not proficient.