CRITERIA FOR PEER FEEDBACK COMMENTS IN A MOOC

Birgit R. Krogsstie, Svend Andreas Horgen, Thorleif Hjeltnes
Sør-Trøndelag University College
Department of Informatics and e-learning

Peer feedback is a mechanism that can be incorporated in a MOOC to enhance learning while meeting requirements for scalability to a large number of students. Feedback may however vary in quality with respect to how it serves its pedagogical purpose. The quality of a feedback comment can be characterized by use of relevant criteria. In the study presented in this paper we iteratively developed a set of quality criteria for peer feedback comments. Starting with an initial set of criteria, primarily grounded in theory, we performed an evaluation by having representative users (i.e. academic staff) apply the criteria to feedback comments in an existing MOOC. A web application for applying the criteria to feedback comments was developed for this purpose. Evaluation and revision of the criteria was performed in two rounds. The criteria were also ranked by the users. The evaluation resulted in a set of quality criteria selected based on their ranking and relevance. The evaluation additionally lead to early insights about the quality of the feedback comments in our MOOC.

1. INTRODUCTION

MOOCs (Massive Open Online Courses) are becoming a more established form of providing education, but are still in an early phase in terms of providing effective pedagogy and sustainable business models. There is a need for studies on practical organization of MOOCs; approaches that support students’ learning (e.g. as measured by perceived learner satisfaction and, above all, attrition rate) with a limited use of teaching resources per student.

Peer feedback is a mechanism generally acknowledged to improve reflection and learning among students. Also, having students evaluate each other’s work potentially saves effort on part of the teaching staff. In a MOOC, peer feedback mechanisms may thus be part of a sufficiently scalable pedagogical design.

The Norwegian university college employing the authors of this paper, offer a MOOC for teachers in Norway that has been running since the autumn of 2014. In the MOOC, peer feedback mechanisms have been implemented. Giving feedback on the assignments of peers is mandatory for the MOOC participants to be able to fulfil the course.

From the outset of our research, we are aware that in our MOOC, the current guidelines for content and how the students should provide their feedback are rather vague. More detailed and structured guidelines might lead students to provide more useful comments. Before changing the guidelines or other aspects of the pedagogical design, we would like to collect more insights about the current design, more specifically the quality of the feedback. This will help us decide what to try to avoid and what to try to encourage in an improved design for peer feedback. Criteria for evaluating feedback comments will also enable us to collect comparable data from the MOOC before and after implementation of changes to the design.

We did not find any research instrument covering our needs for collecting data about peer feedback in a MOOC, but there is related work to draw on, as will be discussed in the Background section. The first step in our research should be to develop a set of criteria for evaluating the quality of peer feedback comments in the MOOC, and is the main focus of this paper. The resulting set of criteria should meet our specific needs but also be of general relevance to MOOC research and educational practice.

This paper thus reports research on a MOOC in a Norwegian university college, addressing evaluation criteria to use in determining the quality of peer feedback in the MOOC. We believe that the resulting list of criteria,
along with their rationale, have general value to other current or potential providers of MOOCs and might spur interesting discussions about whether and how to design for peer feedback in MOOCs.

We start the paper by presenting related research on peer feedback. Next we give a brief description of our MOOC and how peer feedback is used in the MOOC. We proceed to describe our research method, and provide an outline of our results, including the full list of criteria. Finally, we discuss our results and limitations to the study, and address what we see as the next steps along this line of research.

2. BACKGROUND

Feedback can be considered broadly as “information provided by an agent (e.g., teacher, peer, book, parent, self, experience) regarding aspects of one’s performance or understanding” (Hattie, J. and Timperley, H. 2007). An information processing perspective on feedback is to consider it as “information with which a learner can confirm, add to, overwrite, tune, or restructure information in memory, whether that information is domain knowledge, meta-cognitive knowledge, belief about self and tasks, or cognitive tactics and strategies” (Winne, P. H. et al 1994. The learner might have an active role in seeking feedback. Furthermore, the learner is active in processing the feedback in light of existing knowledge.

Peer assessment is “an arrangement for learners to consider and specify the level, value, or quality of a product or performance of other equal-status learners” (Topping, K. 2009). The assessment may include qualitative comments and/or marks. The qualitative comments and the process of giving and using them are often denoted peer feedback (Gielen, S. et al 2010). Peer feedback is typically used in formative evaluation, i.e. as a means for improvement. This includes some typical activities: task performance (a product being created by the assessee), feedback provision (one or multiple assessors providing feedback), feedback reception (the assessee can use the feedback to improve the product), and revision (the assessee may improve the product) (Kollar et al 2010). Peer assessment/feedback can be provided among professionals/experts, but in this paper, focus is on peer feedback among students.

In order to provide scalable MOOCs with a large number of students and limited teaching resources, mechanisms for feedback to students are often based on automation and/or some form of peer reviewing.

Peer assessment/feedback may take place in dyads or groups or through other algorithms for assigning assessors to assignments (as an example: upon completion of a task you are assigned as a reviewer to the previous two who just completed the same task). Students may also provide each other with unorganized feedback in discussion forums or blogs.

Peer assessment is increasingly being used for formative assessment, involving the students as active learners (Cheng. W. et al 1999) and reinforcing self-directed learning (Clark, I. 2012). Peer feedback may have a learning benefit both for the assessor (the one providing feedback) and the assessee (the one receiving feedback). The assessor benefits from seeing examples and approaches and from internalising criteria and standards (Topping, K. 1998).

There are many challenges to identifying the effects of feedback on learning, and the complexity inherent to this kind of research as well as a lack of sufficiently rigid studies means that research findings in the area of peer assessment are not completely consistent (Kluger, A. N. et al 1996). Research on peer assessment covers various types of studies (van Zundert, M. et al 2010). Some studies focus on the validity and reliability of peer assessment as compared to a similar assessment by a tutor. Some focus on the quality of the students’ work/performance in the specific domain, which can be linked to students’ learning. Some studies focus on student attitudes to peer assessment. And finally, some studies are “peer assessment skill” studies, focusing on the quality of the feedback. Along the same vein, Gielen and colleagues (2010) consider two main perspectives on peer feedback quality: Feedback quality can be defined in terms of 1) accuracy, consistency across assessors and/or with teacher feedback, or 2) in terms of content and/or style characteristics. By using the second perspective, it is possible to characterize feedback comments based on the comments alone (i.e. without
comparing comments across assessors or between teacher and assessor) and stress characteristics of feedback that have transfer value across domains and settings.

In this paper, our focus is the quality of the feedback, i.e. peer assessment skill. Also, we would like to characterize feedback comments without contextual information, which means we focus on content and/or style characteristics.

Hattie, J. et al (2007) argue that effective feedback answers the following three questions: “Where am I going? How am I going? Where to next?” These questions work at different levels:

- the task level, which is about how well tasks are understood/performenced
- the process level, addressing the main process necessary for understanding/performing the task
- the self-regulation level, which is about self-monitoring, -directing and -regulating of actions, and
- the self level, addressing personal evaluations and affect (normally positive) about the learner

For feedback to have the desired positive effect on learning environment and outcomes, it needs to have the appropriate focus and level (Hattie, J. et al 2007).

Feedback can be more or less effective with respect to learning. Some examples are: Praising task performance is ineffective. Feedback containing information about correct responses is more effective than feedback with information about incorrect ones. Feedback that relates to previous trials is more effective than that which does not (Kluger, A. et al 1996). The quality of peer feedback increases with the competence of the reviewer and with how structured the review is (Strijbos, J. et al 2010). In peer assessment in high-school (students grading each other’s work). Lu, J. et al (2012) found a positive correlation between feedback that pointed to problems and gave suggestions and the performance of the assessor.

Getting feedback from several peers adds value A study on reviewing of students’ writing showed that reviews from multiple peers improved writing quality more than feedback from a single expert (Cho, K. et al 2010).

Gielen, S. et al (2010) apply a set of criteria for evaluating the quality of peer feedback, starting with a comparison of similar criteria used in several other studies. The criteria used in the case study by Gielen et al include:

- Comments related to the assessment criteria ( Appropriateness)
- Explanation of judgement
  - Reference to specific behaviour (Specificity)
  - Justification
- Presence of suggestions for improvement
- Presence of both positive and negative comments (unless no negative possible)
- Presence of thought-provoking questions
- Clear formulation
- (Structure: Not applicable in the study by Gielen et al as they used a pre-structured feedback form)

The findings on peer feedback referred to above is mainly concerned with learning on the individual level. Peer feedback also has a social side that can be considered an aspect of the feedback quality.

In the context of courses at any level of education, it is beneficial to establish learning communities – groups acquiring new knowledge through cooperation and collaboration (Wenger, E. et al 2002). This could benefit students’ learning and motivation as well as attrition rates.

A challenging aspect of MOOCs is that participants in a course generally are geographically distributed. The progression through the course material might also vary among course participants, depending on how the course is organized. Thus, points of contact, social or professional, among students in a MOOC might be irregular and sparse. Mechanisms helping students in a MOOC establish and maintain learning communities should accordingly be encouraged, and peer feedback can play a role in this, helping students establish connections and focusing discussions. The development of a learning community however depends on the establishing of a social space, which is based on social interaction. According to Kollock (1996), social
interaction occurs if there is continuity (i.e., students are likely to meet again in the future), recognisability (i.e. students are able to identify each other), and history (i.e., knowledge of how other students have behaved in the past).

Considering the quality of peer feedback in a MOOC, then, aspects related to social interaction and community building should be included, as they are likely to impact on learning, albeit more indirectly.

Something that might increase opportunities to develop a learning community is an already existing community of practice (Wenger, E. 1998) among the learners – for instance, if the participants in a MOOC belong to the same profession. This provides common ground, shared vocabulary, trust and makes it easier to outline experiences without having to explain the context in great detail.

Feedback comments might include references to a common field of practice. References to work experiences and challenges might be considered as a way of participating in a community of learners through reference to a shared field of practice.

3. CASE: A MOOC FOR ICT IN LEARNING

The MOOC studied in this paper is a 15 ECTS course called “ICT in learning” offered by a Norwegian university college. The primary target group for the MOOC is teachers who would like to improve their knowledge about pedagogical use of ICT in their teaching, through a flexible form of online studies, gaining formal ECTS credits upon course completion. The course is free of charge thus attracting many participants. Despite the peer feedback mechanisms addressed in this paper, the course can be regarded as an xMOOC – i.e. close to a traditional university course with the teacher as the expert (Siemens, G. 2013).

The course comprises 15 modules, each including one mandatory assignment. 10 of the modules require peer assessment, which in detail means doing the tasks required in the assignment text, submitting the assignment, and after some time delay performing a peer assessment of the assignments handed in by a number of fellow students. Likewise, the student receives peer assessments of his own assignment from a similar number of fellow students. The student should read the feedback, and the assignment is now considered complete. In Figure 1 we have illustrated the process from the point of view of a student called S2. S2 completes assignment 1, performs peer assessments of the assignments made by students S1 and S4, and receives peer assessments from students S3 and S5.

![Figure 1: The peer assessment process of an assignment A1 seen from the point of view of Student 2.](image-url)
Since the assessment of assignments in the MOOC is based on peer feedback and multiple choice tests, there is also an exam for every 5 modules in order to have the quality assurance necessary for awarding course credits.

Course staff manually set up peer assessors for each submitted assignment, based on the current inflow of assignments. There are no fixed groups in which students regularly provide feedback to each other, but if students (for instance, colleagues in a particular school) follow a similar time schedule in their studies, it is likely that they will be connected in peer assessment on more than one occasion.

Given the particular type of students in the MOOC (i.e. teachers with their verbal, reflective and pedagogical knowledge and skills) and the rather vague assignment text used to encourage peer feedback, it is interesting to investigate the quality of the actual feedback in the MOOC. From informal inspection of the feedback comments we see that the comments tend to be rather short and shallow with regard to the domain knowledge, which is not desirable from a pedagogical point of view. At the same time, other aspects of participation seem to be captured in the comments, e.g. through references to students’ own practice and through emotionally laden and/or supportive comments. Some of what can be seen in the current feedback comments may be useful and some less so. We would like to address this more systematically.

4. RESEARCH METHOD

The study presented in this paper can be regarded as design science (Hevner A. R. 2007), the object of the design cycles being a set of criteria for evaluating the quality of feedback comments. By developing the criteria, we seek to contribute to research on peer feedback and MOOCs. Also, the criteria will be instrumental to future research intended to improve our MOOC and derive insights about how to support peer feedback in MOOCs more generally. The researchers/authors are experienced teaching staff in distance courses. The steps of the research process are outlined in Figure 2 and further elaborated in the following sections.

![Figure 2: Research process to develop the feedback evaluation criteria.](image)

The set of criteria used in Test 2 and in the ranking are outlined in Table 1. We have broadly categorized the criteria based on their origin, as will be explained below.
4.1. Developing the first version of the criteria

The initial set of criteria were based on three main types of considerations.

First, there are some practical aspects:
- We need a level of detail that makes it easy to recognize the characteristics in a specific feedback comment when performing the evaluation of the comments. These characteristics must allow a comparison of the quality of feedback comments in the present version of the MOOC with the quality of feedback comments in next versions.
- We acknowledge that this is exploratory research, meant for an iterative process that should address ease of use, perceived usefulness among raters as well as interrater reliability. We allow for “plausible sounding” criteria to be included and later discarded, and we start with a larger number of criteria than we aim for in the final version of the set.

Second, the criteria were informed by informal inspection of the feedback comments in the MOOC. The reason for deriving candidate criteria in this way is our assumption that the specific characteristics of our MOOC, its students and the way the request for peer feedback is formulated has resulted in particular qualities that we wanted to capture. The present study does not include a systematic investigation of the connection between these characteristics and the quality of feedback comments, but we would like our criteria to be able to support such investigation in the future. Key characteristics of our MOOC informing the selection of criteria as well as our interpretation of the feedback comments include:
- In the MOOC, instructions to commenters about how to provide peer feedback on an assignment, i.e. what a useful feedback comment should look like, were brief and unstructured. For the first assignment, the instructions were phrased in the following way: “You are to write a few words about what was good, what could have been done better, and possibly other things that you have noted. Write constructively, but be honest (and no one should get offended by the feedback they receive)”. For the following assignments, the students were instructed to provide feedback in accordance with the previous instruction. No template was used for the feedback. (It should be noted that these instructions were made with the knowledge that the students in our MOOC are teachers, who regularly perform assessments in their own profession). We expect the vague instruction to result in shorter and less structured comments than a more elaborate set of instructions might lead to. To capture this, the criteria should capture comment length and structure. Also, the criteria should capture emotional and social aspects (positive and negative).
- The LMS in question (Canvas) has a discussion like mechanism for feedback, where it is possible to reply to a received feedback. However, there is no notification on such replies. When an assessor posts a feedback comment, the assessee is notified, but if the assessee decides to respond to this comment, the assessor is not notified. We believe this to impact on the motivation to provide feedback over time, as commenters will not get a direct response to their comments. Therefore the criteria should be able to gauge how the quality of feedback comments for a particular student changes over time.
- As the students in the MOOC are teachers, they can be expected to be reflective, conscious and knowledgeable of learning. Also, they are all participants in a common field of practice, namely the teaching profession. This might have a positive impact on their motivation for providing feedback. Hence, the criteria should capture references to own experiences, e.g. giving examples or making suggestions, and other signs of encouraging/participating in a professional community.

Third, there is the theory outlined in the Background section. We decided that our criteria should cover the categories proposed by Gielen et al (2010), but adapted to our needs. We also used the levels proposed by Hattie et al (2007) as a checklist to see whether our criteria together covered the four levels.

Below the initial set of criteria is presented, grouped according to the categories of Gielen et al (2010). The criterion numbers are in accordance with the numbering in Table 1.

The Gielen-category Comments related to the assessment criteria (appropriateness) is reflected in the criterion:
7. “The comment contains constructive critique”.

The Gielen-category *Explanation of judgement: Reference to specific behaviour (Specificity)* is reflected in the criteria:

1. “The comment summarizes what the student has done/delivered”
2. “The comment goes in depth, explains in detail”
4. “The comment is specific with regard to what is good and bad about the assignment”

The Gielen-category *Explanation of judgement: explains/justifies the feedback* is reflected in the criteria:

6. “The comment explains/justifies the feedback”
8. “The comment refers to existing knowledge or quality standards (pointing to syllabus, literature or curriculum)”

The Gielen-category *Clear formulation* is reflected in the criterion:

5. “The comment is formulated so that the message is clear”

The Gielen-category *Presence of suggestions for improvement* is reflected in the criterion:

10. “The comment proposes a solution”

The Gielen-category *Presence of both positive and negative comments (unless no positive possible)* is reflected in the criteria:

13. “The comment expresses support/encouragement irrespective of performance” (e.g. “Well done!”)
14. “The comment contains praise linked to what has been performed/achieved”

Negative aspects are reflected in criteria no 2, 3 and 7.

The Gielen-category *Presence of thought-provoking questions* is reflected in the criterion:

9. “The comment asks questions requiring the students to think/explore further”

In addition to the Gielen-categories, we found that the criteria set should also cover the community/social dimension. Therefore, we added to the list a set of criteria capturing exchange of experience from the practice field as well as emotional support:

16. “The comment accounts for the commenter’s own background/status (may be indirect) (e.g.: “I have no experience with this specific tool”)
17. “The comment uses examples from the commenter’s own practice (episodes or more general accounts of how things are done in their classroom/school)”
18. “The comment expresses that the student’s assignment fits with the commenter’s own experience”
19. “The comment expresses that the student’s assignment contradicts the commenter’s own experience”
20. “The comment expresses that the assignment is useful to the commenter and maybe (or certainly) will be used by the commenter in their practice (e.g. “I learnt something here”, “This is something I would like to try next semester”)”
21. “The comment expresses that the assignment is useful for the professional community (teachers, school) in general (and possibly (more concretely) should be used by others)”

Furthermore, we added a criterion covering personal relations:

12. “The comment expresses a personal relationship to the student (e.g. “Hi Kristin”, “Another great piece of work from you, Silje!”)”

The resulting set of 18 criteria was used for the first test, described in the following.

### 4.2. Building a web interface to aid the evaluation of comments

A set of feedback comments from the MOOC course were downloaded from the LMS-system into a MySQL-database. A web interface was developed (using PHP) so that evaluators could effectively apply the criteria to the comments. The results from each evaluator were saved in a database table, allowing for computerized
processing, comparison and layout of the data. Use of a dedicated web interface should result in higher evaluation quality since appropriate contextual information can be presented to the users.

4.3. Test 1

The set of 18 criteria was tested with 3 evaluators, who are the authors of this paper. By use of the web interface, the evaluators applied the criteria to a set of 10 feedback comments (the same set for all evaluators). We used the first 10 comments in the database.

The results of the test were discussed in a meeting between the evaluators, in which the evaluation data were systematically examined. The analysis focused on 1) difficulties in applying the criteria, 2) cases of major discrepancy between the evaluators, and 3) qualities observed in the sample comments that were perceived to be relevant but were not well covered by the current set of criteria.

4.4. Developing the second version of the criteria

Based on the analysis and discussion, the following changes were made to the criteria from Test 1:

- The wording of some of the criteria was adjusted to make the meaning more clear.
- The range of values for each criterion was changed so that in the web interface, all criteria used the range 0..5 on a discrete, visual, analogue scale without a mid point. (The scale resembles a likert scale (Uebersax 2006) but does not provide labels for all the values, only the end points) This allowed for the use of values other than the endpoints in cases of a criterion being met only to a degree. 0 was set to mean “not at all” and 5 “to a very large extent”.
- The sequence of the criteria was slightly changed in order to group together criteria addressing similar aspects.
- The criterion to identify negative comments was divided to be able to distinguish constructive critique (generally good for learning) from having a negative tone (not good for learning). The latter was phrased as 15: “The comment has a negative tone”.
- A new criterion to identify cases in which a part was missing from the assignment, was included as 3: “The comment identifies something as missing from the assignment”. This includes mismatch between the submission and the assignment.
- A new criterion to capture references to the commenter’s own assignment was included as 11: “The comment refers to the commenter’s own submission”. The intention here was to identify comments more or less directly addressing the commenter’s learning process, and/or the participation of commenter and student in a learning community.

The result was a set of 21 criteria for evaluation of feedback comments (see Table 1).

4.5. Ranking of the criteria

Altogether 11 evaluators were invited to participate in the second round of testing. This included the 3 authors who had participated in Test 1. The other evaluators were academic staff at three different higher education institutions.

The evaluators were first asked to rank the criteria by sequentially assigning the numbers 1, 2, 3 etc. based on how important they considered the criterion to be in light of promoting students’ learning. “0” indicated “unimportant”. The ranking was done in individual Excel sheets. After completing the ranking, the evaluator was exposed to the web interface and asked to go forth with test 2.
4.6. Test 2

The web interface for evaluating comments was improved to include more useful information to aid the evaluation process, including an option to display the criteria in short version, full version, and full version with (in some cases) examples. Navigation was also improved with an index showing which comments were not yet evaluated and functionality to change previous ratings and add a note on comments that were considered difficult to evaluate. As shown in Figure 3, entries with such notes get a red flag in the index.

A new test set of 15 comments was drawn quasi randomly from the first 100 comments and stored in the database. The evaluators were asked to perform an evaluation of the 15 comments, using the improved web interface. 9 evaluators performed the task.

### Evaluation Kommentar

**Velkommen hit, Svend. Her kan du rulle tøning for hver kommentar som er gitt på en besvarelse.**

**Du skal evaluere teksten i fint skrift. Se helt nederst på denne siden for mer informasjon om besvarelser som er med i studentverdien.**

**Fyll ut på en skala fra 0-5**
- 0 betyr "ikke i det hele tatt"  
- 5 betyr "svært stor grad"

| Kriterium 1 | 4 | oppsummerer hva studenten har gjort er spesiell/konkret med hensyn til hva som er godadrålig i besvarelseren  
| Kriterium 2 | 3 | påpeker en mangel ved besvarelseren  
| Kriterium 3 | 0 | går i dybden, diskuterer ting i detalj  
| Kriterium 4 | 3 | er klar formulert slik at bruker skal komme godt frem  
| Kriterium 5 | 5 | begynner tilbakemeldingen ("fordi" eller tilsvarende)  
| Kriterium 6 | 0 | inneholder konstruktiv kritikk  
| Kriterium 7 | 0 | refererer til eksisterende kunnskap eller kvalitetstandarder (henvinning til lærestoff/pensum/litteratur, læreplaner,…)  
| Kriterium 8 | 0 | stiller spørsmål som krever at studenten tenker/undersøker videre  
| Kriterium 9 | 0 | gir forslag til løsning  
| Kriterium 10 | 0 | referer til kommentatoren i sitt besvarelse av oppgaven  
| Kriterium 11 | 0 | gir uttrykk for en personlig relasjon til studenten ("Hei Kristin", Nok et godt arbeid av deg, Silje,)  
| Kriterium 12 | 0 |  

![Figure 3: Web interface used to evaluate the quality of feedback comments according to the 21 criteria.](image)

4.7. Analysis of data

Analysis of the data included some simple statistics, within the limits of what the type of data allowed. The processing was done in Excel by use of pivot tables.

In the data from the criteria ranking, values of 0 ("unimportant") in the data were translated to 21, corresponding to a very low priority. We calculated the mean and standard deviation of the ranking across the evaluators (see Table 1). The mean gives an indication of the importance given to each criterion by the evaluators, whereas the standard deviation indicates the degree of inter-evaluator agreement on this importance.

The use of the criteria to evaluate feedback comments through the web interface produced 21 entries (with values in the range 0..5) for each comment from each evaluator. The mean value of each criterion was calculated (see Table 1). This value can be interpreted as an indication of the relevance of the criterion for the comments in our MOOC, and as an indication of the quality of the comments as characterized by the criterion. The standard deviation among evaluators for each criterion was also calculated, along with the
standard deviation for use of each criterion across all comments (see Table 1). These figures indicate how easily the criteria could be consistently applied by the evaluators.

The resulting data were investigated and discussed with special heed to caution in drawing conclusions, given the exploratory stage of our research. The process of filtering out the resulting third version of the criteria set is addressed in the following section.

5. RESULTS AND DISCUSSION

In this section, we discuss the results from the ranking of criteria and the web based evaluation of comments, addressing also some limitations to our study. We then present our final set of criteria.

In Table 1 all the 21 criteria are listed with the wording as used in test 2. (Criteria 3, 7 and 11 were not part of Test 1. The wording of some of the criteria was slightly adjusted after Test 1.) In Table 1 the criteria have been categorised according to which quality they address (see section 4.1), with the following abbreviations:

- CQ: content quality
- P: relating to practice
- EM: emotional expression
- PR: personal relation
- SL: orientation to the student’s learning
- CL: orientation to the commenter’s learning

Criterion 15 (“The comment has a negative tone”) is the only criterion in which a higher value is considered detrimental to learning. For the other criteria, a higher value indicates a type of content assumed to positively impact learning.

In the columns 3 and 4, the mean and standard deviation from the ranking of the criteria are listed. In the last two columns, the mean and standard deviation from the evaluation (web interface summarized) are listed.

5.1. Observations from the ranking

The values from the criteria ranking show that CQ-criteria are generally given higher priority than for instance the ones relating to practice. This might reflect that the evaluators (academics) were asked to rank criteria in light of promoting students’ learning, but not necessarily in a MOOC-based context. The detailed rationale for our particular selection of criteria (as outlined in Section 4) was not made explicit to the evaluators. This may be part of the explanation for the low score of social aspects.

In general, for some criteria we see a substantial deviation among evaluators. This might be related to a disagreement of importance based on the teaching experience of the evaluators. Another explanation can be that it is challenging to rank 21 criteria linearly without some form of grouping, as some of the evaluators pointed out. They would ideally have wanted to apply a likert-scale or to group the criteria within three or so main categories, and then rank the criteria within each category.

Criteria 2, 5, 6, 7 and 9 have a mean rank value of 6 or less, indicating they are very important for learning.

There seems to be a strong agreement on the importance of some of the CQ-criteria (e.g. 2, 6, 9 and 15). These have a relatively low standard deviation (between 2 and 3) meaning that the evaluators have ranked them fairly consistently, but no 15 has a too low score of 19,76.

Based on the ranking process, we therefore have strong reasons to include criteria no 2, 5, 6, 7 and 9 as candidates to our final criteria set.
<table>
<thead>
<tr>
<th>No</th>
<th>Criterion, with quality aspect(s) in square brackets. The comment…</th>
<th>Mean rank</th>
<th>Stddev rank</th>
<th>Mean grade</th>
<th>Stddev grade</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>summarizes what the student has done/delivered [CQ]</td>
<td>11,38</td>
<td>5,88</td>
<td>1,71</td>
<td>1,57</td>
</tr>
<tr>
<td>2</td>
<td>is specific with regard to what is good and bad about the assignment [CQ]</td>
<td>2,00</td>
<td>2,74</td>
<td>2,08</td>
<td>1,68</td>
</tr>
<tr>
<td>3</td>
<td>identifies something as missing from the assignment [CQ]</td>
<td>10,46</td>
<td>5,14</td>
<td>0,93</td>
<td>1,67</td>
</tr>
<tr>
<td>4</td>
<td>goes in depth, explains in detail [CQ]</td>
<td>9,69</td>
<td>4,57</td>
<td>1,49</td>
<td>1,61</td>
</tr>
<tr>
<td>5</td>
<td>is formulated so that the message is clear [CQ]</td>
<td>5,54</td>
<td>4,24</td>
<td>3,25</td>
<td>1,67</td>
</tr>
<tr>
<td>6</td>
<td>explains/justifies the feedback [CQ]</td>
<td>4,38</td>
<td>2,69</td>
<td>1,46</td>
<td>1,42</td>
</tr>
<tr>
<td>7</td>
<td>contains constructive critique [CQ]</td>
<td>4,77</td>
<td>3,19</td>
<td>1,21</td>
<td>1,83</td>
</tr>
<tr>
<td>8</td>
<td>refers to existing knowledge or quality standards (pointing to syllabus/literature/curriculum) [CQ]</td>
<td>9,46</td>
<td>3,26</td>
<td>0,34</td>
<td>1,03</td>
</tr>
<tr>
<td>9</td>
<td>asks questions requiring the student to think/explore further [CQ, SL]</td>
<td>6,00</td>
<td>2,61</td>
<td>1,30</td>
<td>1,67</td>
</tr>
<tr>
<td>10</td>
<td>proposes a solution [CQ, SL]</td>
<td>9,38</td>
<td>3,50</td>
<td>1,03</td>
<td>0,90</td>
</tr>
<tr>
<td>11</td>
<td>refers to the commenter’s own assignment [CQ, SL, CL]</td>
<td>16,38</td>
<td>5,56</td>
<td>0,32</td>
<td>1,91</td>
</tr>
<tr>
<td>12</td>
<td>expresses a personal relationship to the student [PR]</td>
<td>15,00</td>
<td>5,63</td>
<td>0,62</td>
<td>1,46</td>
</tr>
<tr>
<td>13</td>
<td>expresses support; acknowledgement independent of performance [EM, PR]</td>
<td>12,77</td>
<td>4,99</td>
<td>0,52</td>
<td>1,79</td>
</tr>
<tr>
<td>14</td>
<td>contains praise linked to what has been performed/achieved [EM]</td>
<td>9,23</td>
<td>6,18</td>
<td>1,07</td>
<td>1,66</td>
</tr>
<tr>
<td>15</td>
<td>has a negative tone [EM]</td>
<td>19,67</td>
<td>2,81</td>
<td>0,09</td>
<td>0,52</td>
</tr>
<tr>
<td>16</td>
<td>accounts for the commenter’s background/status (may be indirect) [P]</td>
<td>15,15</td>
<td>5,63</td>
<td>0,95</td>
<td>1,71</td>
</tr>
<tr>
<td>17</td>
<td>uses examples from the commenter’s own practice [P]</td>
<td>13,69</td>
<td>4,59</td>
<td>0,72</td>
<td>1,60</td>
</tr>
<tr>
<td>18</td>
<td>expresses that the student’s assignment fits with the commenter’s own experience [P]</td>
<td>13,62</td>
<td>3,71</td>
<td>1,17</td>
<td>1,42</td>
</tr>
<tr>
<td>19</td>
<td>expresses that the student’s assignment contradicts the commenter’s own experience [P]</td>
<td>15,08</td>
<td>4,09</td>
<td>0,07</td>
<td>0,50</td>
</tr>
<tr>
<td>20</td>
<td>expresses that the assignment is useful to the commenter and maybe (or certainly) will be used by the commenter in their practice [P, CL]</td>
<td>12,62</td>
<td>5,06</td>
<td>3,57</td>
<td>1,78</td>
</tr>
<tr>
<td>21</td>
<td>expresses that the assignment is useful for the professional community and possibly should be used by others [P]</td>
<td>12,54</td>
<td>4,88</td>
<td>3,07</td>
<td>1,25</td>
</tr>
</tbody>
</table>

*Table 1: Criteria for evaluating peer feedback comments.*
5.2. Observations from the evaluation

The goal of the evaluation using the web interface was to apply criteria on the actual feedback comments in order to see what criteria the MOOC students implicitly cover in their feedback. Recall that each evaluator used the values 0 to 5 for each criterion for each feedback comment in test set 2, 5 meaning “to a very large extent”.

Table 1 shows that many of the criteria get a low mean score in general even though evaluators rate them as important to learning. 9 of the criteria got a mean grade between values 0 and 1 (no 3, 8, 11, 12, 13, 15, 16, 17, 19) and 8 criteria got a mean grade between values 1 and 2 (no 1, 4, 6, 7, 9, 10, 14, 18). Only criterion 2 got a mean grade between values 2 and 3, and criteria 5, 20 and 21 got a mean grade between values 3 and 4. Notably, the standard deviation across evaluators indicates this scoring is not very reliable.

In order for a criterion to qualitatively describe a feedback comment, it should get a score of at least 3 (more than 50 % on the scale from 0 to 5). From Table 1, it seems that only the criteria 2, 5, 20 and 21 can be added to our final criteria set using this principle for the mean grade. The mean rank (from the Excel-sheets, column 3 in Table 1) supports this choice for the criteria 2 and 5, whereas criteria 20 and 21 have a somewhat higher rank value. We still find these to be among the strongest overall criteria, and therefore add 20 and 21 to the final set (2 and 5 has already been added).

A general impression from the evaluation is that surprisingly many of the criteria lead to low scores. If the average score for a criterion is close to 0, we can reasonably assume that the feedback comments in our test set have little of the quality/ies associated with that criterion. The generally low scores indicate that the feedback comments in our MOOC are not very content rich with respect to learning, which is a topic of further research.

5.3. Developing the final version of the criteria

As noted earlier, we got feedback on the evaluation pointing to the difficulty of linearly ranking as many as 21 criteria. More generally, we acknowledge that it might be difficult to apply this many criteria when evaluating the quality of feedback comments. One important conclusion from the evaluation process is, accordingly, that we should create a shorter criteria list. Sorting the criteria based on the ranking, we get a list of the criteria the evaluators considered most important for learning. Looking at Table 1 we get the following subset, in order: 2, 6, 7, 5, 9, 14, 10, 8, 4 and 3. Most of these are classified as CQ (content quality). They also span all the categories identified in (Gielen et al 2010).

No 20 and 21 from our previous discussion are not included in the set based only on the ranking. We included them as they are highly rated and represent qualities related to practice, which can be motivating for learning.

The feedback comment…
- is specific with regard to what is good and bad about the assignment [CQ, 2]
- explains/justifies the feedback [CQ, 6]
- contains constructive critique [CQ, 7]
- is formulated so that the message is clear [CQ, 5]
- asks questions requiring the student to think/explore further [CQ, SL, 9]
- contains praise linked to what has been performed/achieved [EM, 14]
- proposes a solution [CQ, SL, 10]
- refers to existing knowledge or quality standards (pointing to syllabus / literature / curriculum) [CQ, 8]
- goes in depth, explains in detail [CQ, 4]
- identifies something as missing from the assignment [CQ, 3]
- expresses that the assignment is useful to the commenter and maybe (or certainly) will be used by the commenter in their practice [P, CL, 20]
- expresses that the assignment is useful for the professional community and possibly should be used by others[P,21]

Table 2: The final set of criteria for evaluating peer feedback comments
Given the limitations to our study, our selection of a final set of criteria for evaluation of peer feedback comments are informed not solely by the results of the ranking and evaluation, but also by our more theory- and experience based knowledge of what a high quality peer feedback comment in a MOOC should contain. Based on our discussion, our proposed final set consists of 12 criteria for quality of peer feedback comments in a MOOC, sorted by importance (highest first) (see Table 2). The square brackets contain the criteria category as well as the original criteria numbering.

6. CONCLUSION

In this paper we have presented a set of criteria that can be used to evaluate the quality of feedback comments in a MOOC. We empirically investigated data from peer assessment of assignments in a MOOC for teachers in Norway. The development of the criteria was an iterative process, based on literature study, informal inspection of feedback comments, two rounds of testing and adjustment of the criteria set, development of a web interface to aid the evaluation process, and a ranking performed by academics outside the research team.

Through the evaluation we got an indication of the relevance and applicability of the criteria and the quality of the feedback comments in the MOOC. The criteria ranking strengthened the basis for criteria selection. Our results should be considered in light of the exploratory stage of the research. More testing and the use of better metrics are needed e.g. to address construct validity, which in this paper is only loosely touched upon through the connection between criteria and qualities of peer feedback addressed in related research. From a design science perspective, however, we have useful input for further development of the criteria and the MOOC itself.

The inconsistency among evaluators in applying the criteria indicates a need to reconsider the formulation of the criteria, especially those with the largest deviation among evaluators. Also, we see a need to provide some training to evaluators, including the rationale for the criteria that we included in the final list despite low rating.

The MOOC in our case study is for vocational training, and the students are teaching practitioners. As a foundation for our study we accordingly made the assumption that the students have a good background for addressing many important aspects of a submission when giving feedback to a peer:

- Teachers have practical experience of applying ICT and learning, to varying extents and in varying ways.
- Teachers have arenas in which to try out ideas and newly gained knowledge about ICT and learning.
- Teachers can be assumed to be knowledgeable about learning and reflection and able to critically consider and manage their own learning process.

Overall, the feedback comments in the MOOC score low on many quality criteria, including those that can be linked to the above advantages of teachers as a student group. For instance, the comments only to a small extent relate to work practice. We see a need to be more specific to the students about what a peer feedback comment should address. In this regard, our final set of criteria can be used to inform the development of tools or templates for providing the feedback.

Our MOOC can be regarded as an xMOOC (Siemens, G. 2013), for which the pedagogical model is well supported by the LMS Canvas. The students signing up for the course can expect to be assessed on an individual basis, and the main focus is not – as in a cMOOC - on the generating of networks/communities. There are however pedagogical advantages to cMOOCs, and going in the direction of a cMOOC is a future option for our MOOC. This would require the provision of a broader set of resources and mechanisms for network building.

For the MOOC in largely its current form, we believe that well-functioning peer feedback is an adequate way of supporting learning while adding to the social dimension of MOOC participation. The research presented in this paper contributes towards the measurement of quality of peer feedback. Also, it points to a potential for improvement of peer feedback mechanisms to get comments richer in content likely to aid learning.

As further work we plan to apply the criteria in larger scale testing on the same MOOC. We will focus on the quality of peer feedback and the improvement of support for the provision and use of peer feedback. We will
also refine and expand the explanations and examples associated with the criteria to convey their meaning more clearly to raters. Finally, we plan to apply the criteria to different MOOCs to look into their more generic applicability.

7. REFERENCES


