A PREPARATORY ANALYSIS OF PEER-GRADING FOR A DIGITAL HUMANITIES MOOC

Kaplan, Frédéric
EPFL, Switzerland

Bornet, Cyril
EPFL, Switzerland

Category: Short Paper
Session: 2
Date: 2014-07-09
Time: 11:00:00
Room: 415 - Amphimax

Introduction

Over the last two years, Massive Open Online Classes (MOOCs) have been unexpectedly successful in convincing large number of students to pursue online courses in a variety of domains. Contrary to the "learn anytime anywhere" moto, this new generation of courses are based on regular assignments that must be completed and corrected on a fixed schedule. Successful courses attracted about 50 000 students in the first week but typically stabilised around 10 000 in the following weeks, as most courses demand significant involvement. With 10 000 students, grading is obviously an issue, and the first successful courses tended to be technical, typically in computer science, where various options for automatic grading system could be envisioned. However, this posed a challenge for humanities courses. The solution that has been investigated for dealing with this issue is peer-grading: having students grade the work of one another. The intuition that this would work was based on some older results showing high correlation between professor grading, peer-grading and self-grading (Wagner et al. 2011[1], Topping 1998[2]). The generality of this correlation can reasonably be questioned. There is a high chance that peer-grading works for certain domains, or for certain assignment, but not for others. Ideally this should be tested experimentally before launching any large-scale courses.

EPFL is one of the first European schools to experiment with MOOCs in various domains. Since the launch of these first courses, preparing an introductory MOOC on Digital Humanities was one of our top priorities. However, we felt it was important to first validate the kind of peer-grading strategy we were planning to implement on a smaller set of students, to determine if it would actually work for the assignments we envisioned. This motivated the present study which was conducted during the first semester of our masters level introductory course on Digital Humanities at EPFL.

Method

56 students were asked to produce a blogpost in which they had to identify trends in Digital Humanities based on the online proceedings of the DH2013 conference in Nebraska[3].

Students had to choose three abstracts from the conference, summarise and compare them, then use the Wordpress blog dh101.ch to publish their post. Students were informed that their post would be graded by the professor but also by other students. Following the usual Swiss grading system, the grade range was from 0 to 6. The students were informed that only the grade given by the professor would count for their semester results but that 10% of their semester results depended on whether they took the peer-reviewing seriously.

The grading criteria was presented in detail to the class at the same time. Students had to check whether the blog post followed the guidelines of the assignment (discussing three articles, identifying a trend) (4 points); whether the English was correct and clearly understandable (+0.5); whether the keywords and post layout were adapted to its content (+0.5); whether the post was not just a summary of the three articles but really compared them, and, more subjectively, whether the post's content was well discussed (+0.5) and the identified trend interesting (+0.5). The students were also asked to verify that the blog post did not contained plagiarised content.

Each student had to anonymously grade five randomly chosen blog posts. In order to simplify the task and to reduce the risk of manipulation errors, we developed a simple dedicated Wordpress app[4] to organise this process. Students used their Wordpress account to log in and by doing so accessed a page

listing the five posts that were assigned to them, as well as one checkbox per criteria to be checked. This assignment process was done beforehand in an automated way: each paper was assigned once to the professor and randomly to five students, but under the constraint that no student could get their own paper and no more than five papers in all. Although we are aware that more sophisticated systems exist for assigning work in peer-reviewing processes (e.g. The Caesar system developed at MIT [5]), we assumed that this random assignment process was relevant in this context, given the uniform nature of the content to be graded. The professor graded all the blog posts without any information on the results of the peer-grading process.

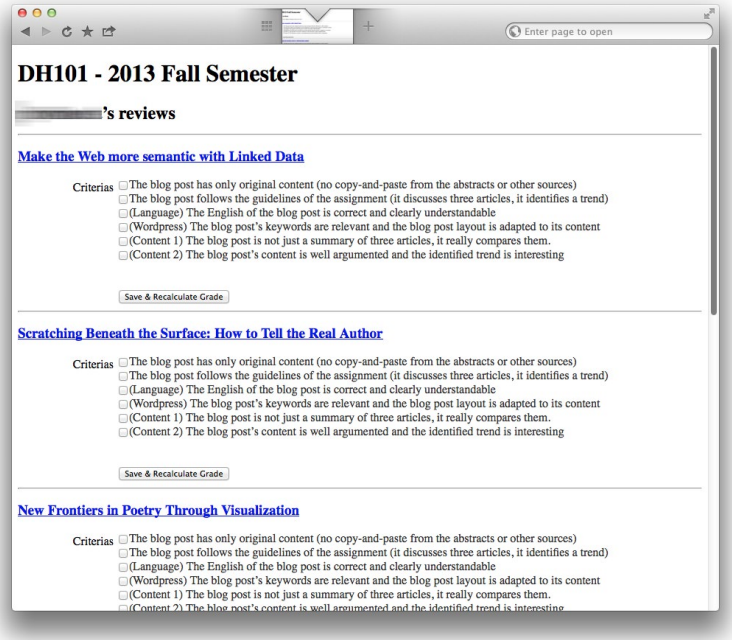

Fig. 1: Wordpress app for peer-grading

Results and Discussion

52 blogs posts were produced and published on the public website dh101.ch. 47 students completed the peer review grading, and all of them graded all of their five papers. Three students accessed the website without grading any papers, and two didn't even attempt to log in.

The process drew a lot of interest and questions from the students. The criteria of the grading grid were questioned by many students, in particular the one linked with the more subjective evaluation of the posts (several students thought this was unfair). Other studies have shown that students' writing and understanding in core courses can be improved through peer ranking (Rhett A. et al 2005[6]). Although it is difficult to measure, it seems that the precise explication of the grading criteria imposed by the peer-review system had a positive effect on the quality on the posts when compared to the last year's course, when this system was not included.

We measured a strong correlation between the average of the peer-graded marks and the mark of the professor ($r(50) = 0.39$, $p < .01$).

Figure 2 shows the level of matching between the professor's grade and the peer graded marks after normalisation to the closest half point. In 38% of the case, the peer-graded mark is the same as the professor's, in another 38% of the cases the mark shows a 0.5 difference. The remaining 24% of cases show a larger difference. These latter cases mostly correspond to situations where either the professor or one of the students concluded that the post did not respect the instructions for the exercise and therefore gave a sanctioning mark (0 in case of plagiarism, 3 in the case of uncompleted post).

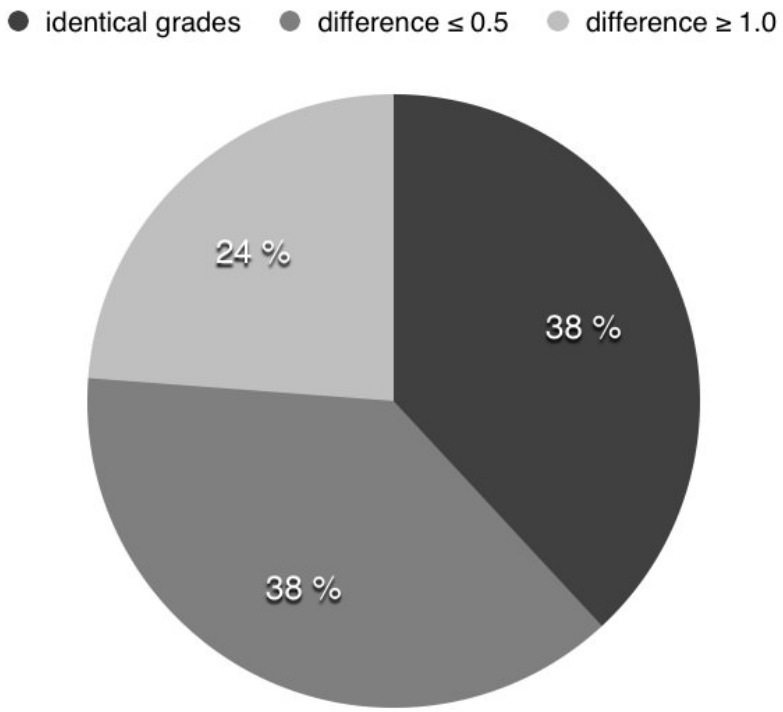

Fig. 2: Level of matching between professor and student grades

This relationship can also be shown by breaking down the ratios of validated criteria:

Figure 3 shows the average grades given by students (sorted ascending), the squared mean error intervals for each of them and the professor's gradings (**x** marks), confirming that 76% of the papers received an average grading within 0.5 to the professor's.

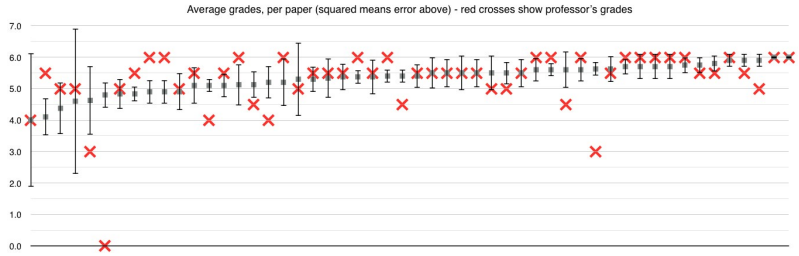

Fig. 3: Average grades, per paper

Figure 4 shows a comparison between the grading criteria used by the students and the professor's grading. The distribution is visually similar (we cannot perform a more detailed statistical analysis because of the sample size).

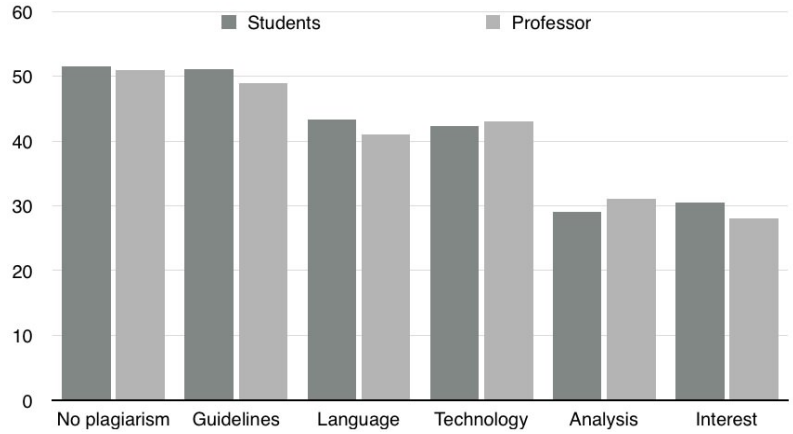

Fig. 4: Criterias Distribution

Figure 5 presents the correlation between the student's grade and their own grading. Although it is not obvious at first sight, there is a marginal significant negative correlation between these two variable datasets ($r(50) = -0.26$, $p = .07$). This could suggest that students who wrote good papers are more critical of their peers.

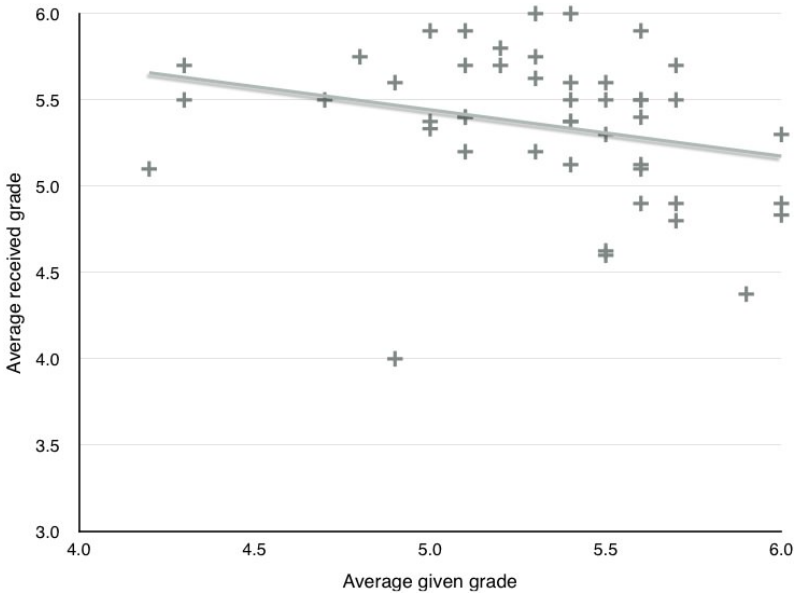

Fig. 5: Correlation between received and given grades

Figure 6 presents the grades in the order published (the first posts are the ones submitted the earliest by the students). The professor graded the posts following this order. As expected, no correlation exists between the order and the grades in the students' gradings, as they were assigned randomly. However, there seems to be a tendency towards lower grades in the professor's grading sequence. This could be explained if a correlation existed between the quality of the post and their publication time (the best students would publish the first). However, this correlation was not found in the student's grading. This could suggest a potential temporal bias observed through the fact that the last evaluations were tendentially lower that the first ones (the professor becoming more critical).

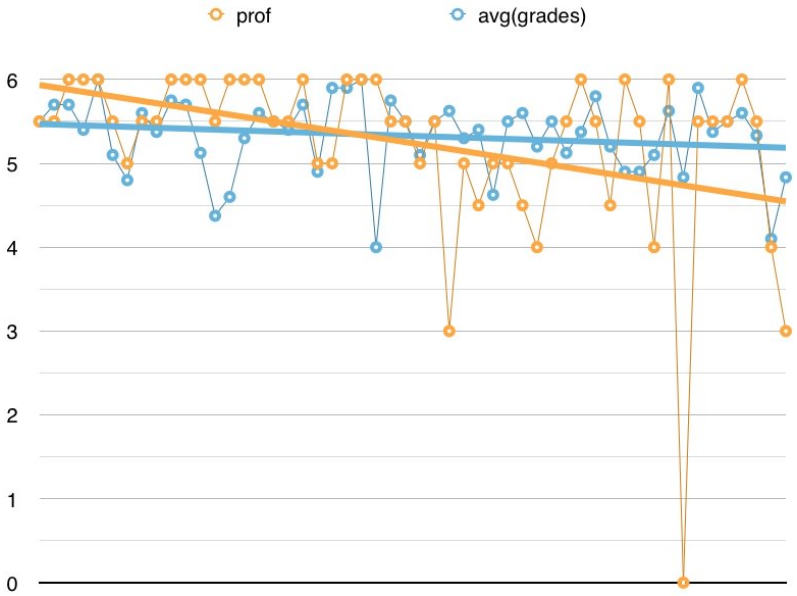

Fig. 6: Evolution of grades with time

Conclusion

This article presents a preliminary study for a Digital Humanities MOOC conducted on a class of 56 students. In the present context, we can conclude that peer-grading is highly correlated to the professor's grading. In about ¾ of the cases, the grades obtained by this method are within a range of 0.5 points of the professor's grading. Qualitative observations tend to show that the quality of the posts increased when compared to the previous year, likely because students had to reflect on the grading criteria and were cautious of producing good work when this work would be evaluated by their peers. In addition, our study may suggest some possible temporal biases in the way the professor grades a long sequence of work, reinforcing the idea that peer-grading may not only be an interesting positive alternative to traditional grading but also that it may, in some case, be less biased.

Nevertheless, these preliminary results, dependent of the particular context of this study, should be extrapolated with care and would not eliminate the need to conduct regular quality evaluation in the context of a MOOC on Digital Humanities. Indeed, these results do not guarantee that the same peer grading method could scale to a 10 000 student MOOCs without problem. As the number of students

increases, and the cultural backgrounds and linguistic competencies diversify, part of the behavioural homogeneity that we observed in this prestudy may no longer be valid. For this reason, this research should definitely be completed with *a posteriori* study testing the efficiency of peer-grading with a similar method in a randomly chosen set of learners of the entire MOOC.

Acknowledgments:

The authors would like to thanks Andrea Mazzei for fruitful discussions about the analysis of the results of this preliminary study.

References

1. **Wagner, M.L., Churl Suh, D., Cruz, S.** (2011). *Peer- and Self-Grading Compared to Faculty Grading*. American Journal of Pharmaceutical Education 75, 130.

2. **Topping, K.** (1998). *Peer Assessment between Students in Colleges and Universities*. Review of Educational Research 68, 249–276.

3. *DH2013 Abstracts*, dh2013.unl.edu/abstracts/

4. *Wordpress Login API*, codex.wordpress.org/Plugin_API/Action_Reference/wp_login

5. **Tang, M.** (2011), *Caesar: A social code review tool for programming education*, Master project, MIT

6. **Rhett A. et al** (2005), *Using peer ranking to enhance student writing*, IOP Science