Does peer grading work? How to implement and improve it? Comparing instructor and peer assessment in MOOC GdP

Rémi Bachelet, Drissa Zongo, Aline Bourelle

Does peer grading work? How to implement and improve it? Comparing instructor and peer assessment in MOOC GdP

Rémi BACHELET, Drissa ZONGO, Aline BOURELLE
École Centrale de Lille, BP 48, F-59651 Villeneuve d'Ascq Cedex, France
remi.bachelet@ec-lille.fr, drizongo@gmail.com, alinebourelle@msn.com

Abstract: Large scale peer assessment is arguably the most critical innovation required for development of MOOCs. Its core principle is to involve students in the evaluation and feedback process of correcting assignments. However, it has been criticized for being less rigorous than instructor assessment, too demanding on students and not reliable or fair due to student biases. This paper is drawn from data and practical hands-on experience from MOOC GdP2, in which assignments were both graded by instructors and by peers. Using data from 4650 papers, each graded by 3-5 peers and by an instructor, we test hypotheses and discuss a series of questions: How to train MOOC students to grade their peers? Is peer grading as accurate as instructor grading? What data pre-processing is to be used prior to testing hypotheses on peer grading? Which grading algorithm is best for processing peer-produced data? Is anonymity in peer assessment preferable to increased student interaction? We also present the improved peer grading systems we implemented in MOOC GdP 3 and 4 thanks to this research.

Keywords: Massive open online courses, MOOC, peer assessment, peer grading.

Introduction

Despite recent attempts at improving platforms and course quality, the pedagogy of most MOOCs today rely mostly on two things: video courses and quizzes. The strength of these technologies comes from how easily they can be deployed on a massive scale, but their weakness lies in their inability to process, grade and provide feedback for complex and open-ended student assignments and overall supporting activities for critical thinking. We believe that the solution to this challenge is not machine grading (Burrows, Gurevych & Stein, 2014), but peer grading, and more broadly, peer assessment. More than a promising solution to evaluate assignments on a massive scale, peer assessment holds the promise of major learning benefits, developing student autonomy and, in Bloom's taxonomy, higher levels of learning (Bachelet 2010, Sadler & Good 2006).

But is it really possible to implement peer evaluation? Does it provide reliable results? How to minimize student's workload while retaining accurate peer grading? In this paper, we shall address a series of questions through experience of four MOOC GdP sessions and quantitative data from the second session of the MOOC:

1. Qualitative/experience testing: How to train MOOC students to grade their peers and provide them with constructive feedback as well?
2. Quantitative data/hypothesis testing: Is peer grading as accurate as instructor grading?
3. Quantitative data/hypothesis testing: What data pre-processing is to be used to test hypotheses on peer grading?
4. Quantitative data/hypothesis testing: Which grading algorithm is best for processing peer-produced data in order to deliver grades which are similar to instructor grading?
5. Quantitative data/hypothesis testing: How many peer grades are required to provide an accurate final grade?

Amongst the four sessions of the "fundamentals of project management" MOOC, the second session is the best fit for exploring these issues, mostly because:

- We have a significant dataset: 1011 to 831 assignments were submitted each week, for 5 weeks, 4650 assignments total.
- A large variety of assignments : 1/a concept map, 2/a series of slides comparing alternative projects, 3/a meeting report with an allocation of tasks, 4/ specifications and functional analysis of a project and 5/a Gantt chart
- Last but not least, a comparison of instructor and peer grading was readily available: there were *both* 3-5 peer grades and one instructor/AT grading (considered as the "true score") for each student assignment.

Context

MOOC GdP1 was the first xMOOC to be organized in France, it was developed from an existing Open Course Ware (OCW) website (Bachelet, R. 2012a) and experience of running a distance learning course (8 editions, 400 laureates from 2010 to 2012, see Bachelet, R. 2012b). It was set up with almost no financial resources using a personal home studio and run by an open team of volunteers. Three external actors brought support: École Centrale

de Lille sponsored the project, various Free Google services were used (Google groups, Drive, Hangouts, YouTube) and the US Company Instructure hosted the course on the Open source MOOC application Canvas. Enrollment opened January 11, 2013, course started March 18 and offered two individual tracks: Basic and Advanced. A "Team project" track was also proposed after the individual tracks, it ran 14 projects and was instrumental in recruiting volunteers supporting the next editions of the MOOC.

MOOC GdP2 followed the same principles, but was one week longer, adding a functional analysis course. Beside the free tracks, it became for the first time possible to be awarded European University Credits (ECTS) by Centrale Lille, by taking either a webcam (ProctorU) or an on-site table exam (AUF campuses in two developing countries). Technical support for Canvas was provided by the French startup Unow. All non-foreign Centrale Lille first year engineering students (approximately 200 of the 1011 students of this track) were enrolled in the advanced track, as required by their curriculum.

Basic and advanced tracks started September 16, 2013. MOOC GdP2 was 5 weeks long, plus one week dedicated to preparing the final exam. A one week add-in SPOC session on Project Evaluation was organized from November 4 for employees of Agence Universitaire de la Francophonie, and the Team project track of the MOOC started mid-October (sharing project ideas, putting the teams together) and ended mid-December with a final video presentation and report.

Attendance was as follow - enrolled: 10848, completing at least one quiz for the basic track: 5711, submitting at least one deliverable for the advanced track: 1011.

Success rates (Bachelet, R. 2013) were:

- Entry rate, 53%: this ratio is the number of students submitting at least one short quiz divided by total enrollment in the MOOC, thus there was a 47% population of "no-show" and auditors.
- Basic track pass rate, 61%: calculated as follow - the number of students submitting at least one short quiz divided by the number of students awarded with basic track "pass" (i.e. an average of 70% of maximum points and 60% of maximum points on the final exam)
- Advanced track success rate 78%: the number of students submitting at least one assignment divided by advanced track "pass" (i.e. basic track "pass", plus an average of 70% of total maximum points).

Advanced track: organization and assignments

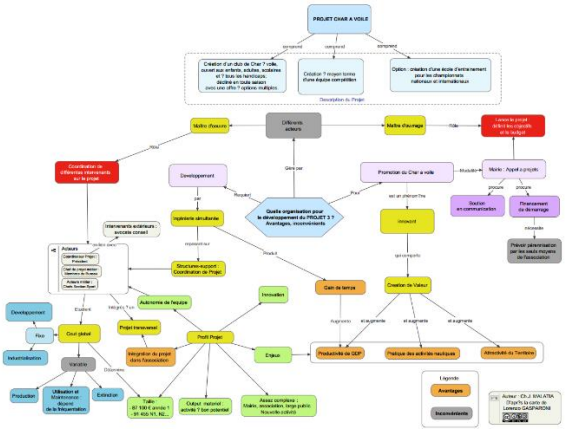
Students undertaking the advanced track were required to submit an assignment each week, in relation to a topic covered in the week's course. After that, each student was randomly allocated 4 assignments submitted by other students. He/she had to: 1/provide feedback to the student (comments could be global, specific to a grading category or "comment bubbles" inserted on the pdf assignment) and 2/grade from 0 to 100 (70/100 being the expected "pass" grade) according to a 4-items rubric. Instructions for evaluation as well as other systems were provided to help grading and feedback, which will be described more thoroughly below. Students had 5 days to complete the peer evaluation process after which it was closed and the teaching team started awarding final grades.

Assignments were: 1/a concept map, 2/a series of slides comparing possible projects, 3/a meeting report with an allocation of tasks, 4/ specifications and functional analysis and 5/a Gantt chart (figures 1 to 3)

Each assignment required 4-9 hours of work, depending on previous skills and knowledge. Plagiarism was detected using Euphorus (1-3% of submissions, later removed from our dataset).

Table 1: Assignments overview (from Bachelet, R. Petit, Y. (2014).

	# of papers submitted	average final grade	standard deviation of grades
paper 1 : concept map	997	61.6	20.71
paper 2 : investment study (slides)	1011	69.5	21.5
paper 3 : meeting report	944	72.5	16.24
paper 4 : functional analysis	867	69.6	20.96
paper 5 : Gantt chart	831	74.2	22.56



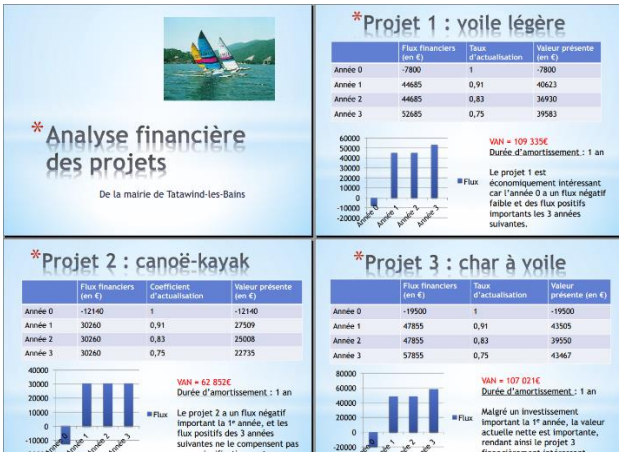
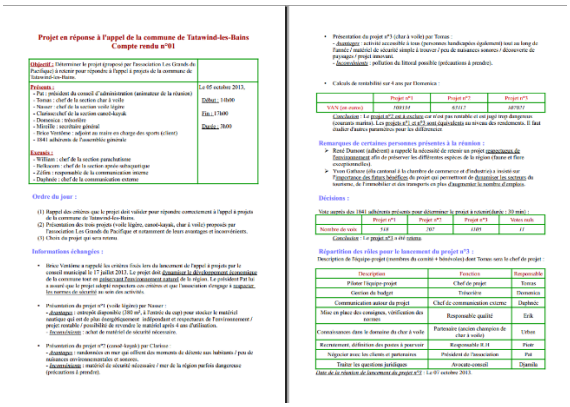


Fig 1.1 Week 1 assignment: concept map Fig 1.2 Week 2 assignment: investment study.



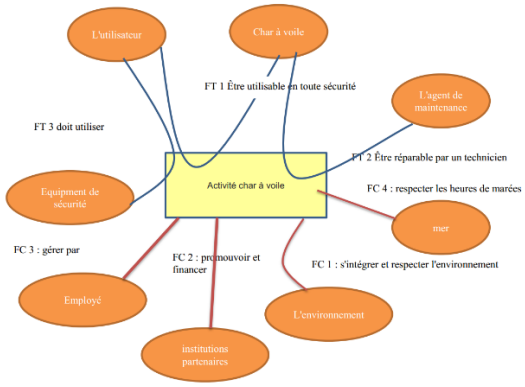


Fig 2.1. Week 3 assignment: meeting report Fig 2.2 Week 4 assignment: functional analysis.

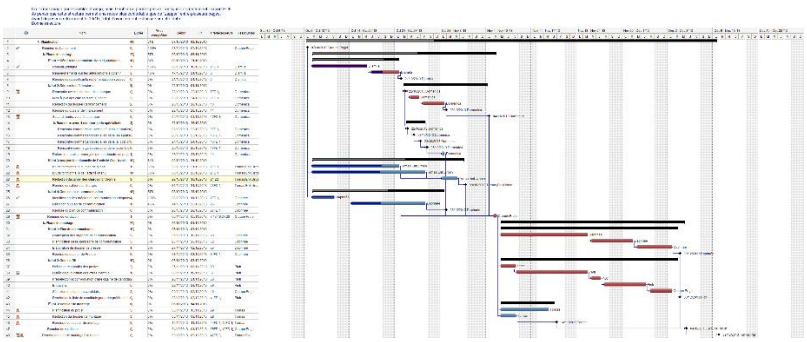


Fig 3. Week 5 assignment: Gantt chart.

How to train and empower MOOC students so that they can grade their peers and provide them with constructive feedback?

Several systems were developed, some generic to MOOCs and peer grading in general, and some dedicated to grading specific assignments of MOOC GdP2.

Generic peer Evaluation training:

- Peer grading and evaluation was claimed from the start as a major requirement of the advanced track ("learning agreement" presented in the teaser and "introduction to advanced track" video).
- After submission of the first essay, a video was dedicated to explaining the rationale and importance of peer assessment in MOOCs and how to write motivating and constructive feedback.

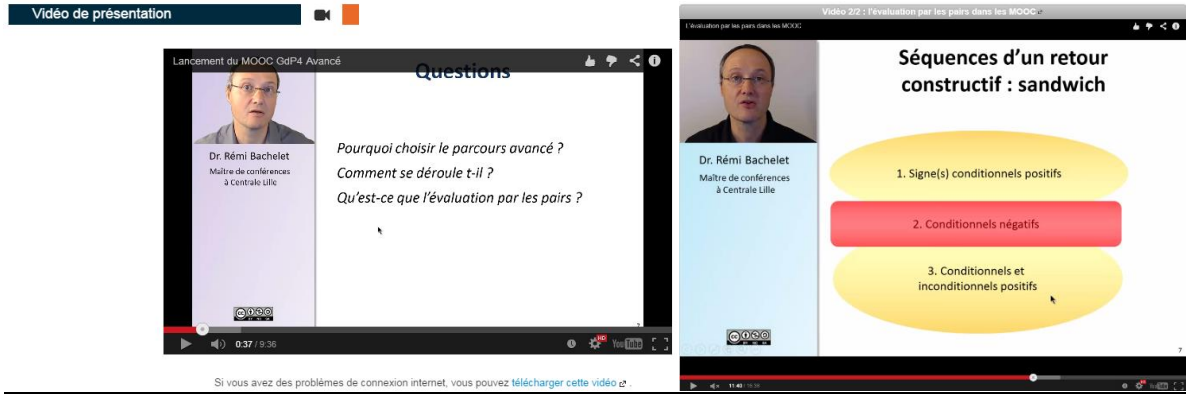


Fig 4.1 "Introducing peer evaluation in the advanced track" fig 4.2 "Generic peer Evaluation training".

Specific peer Evaluation training:

In addition, specific resources were available to guide peer evaluators in assessing each assignment: a benchmark version of the assignment, a video (sometimes), an interactive grading rubric and a discussion thread (1649 total forum posts on assignments and peer evaluation in GdP2). Guidelines on how to use the platform for peer grading were also provided.

Anonymity

Unlike most other studies of peer assessment in MOOCs, while anonymity is possible in MOOC GdP (using a pseudonym as one's platform profile), it is rarely used. On the contrary, students are encouraged to exchange messages via the Canvas platform message system, even only for a "thank you" to their evaluators. Messages make it also possible to ask an evaluator to revisit the assigned grade or to give additional feedback. Though we have no data on messages exchanged on these occasions, we believe that there is much more to gain by fostering social contacts between students than there are risks of conflicts or bias in grading. Biases in grading can be controlled using either instructor grading (MOOC GdP 1 and 2) or a grading process and algorithm (MOOC GdP 3 and 4). Though the legitimacy of student grading is often questioned. Conflicts between persons are rare: Instructor staff can be called upon if student's interactions get sour: we had to intervene on only one such occasion in the 4 editions of MOOC GdP (more than 10.000 assignments assessed by 3-4 peers).

Critics point out that there is a bias assessing someone you can identify (besides googling the person, you can know him/her first hand, especially for the 200 Centrale Lille students). To minimize biases, the final grade was issued by an instructor in GdP1 and 2 while a control and a reward system was designed for GdP 3 and 4 (described below).

What data pre-processing is to be used to test hypothesis on peer grading?

Before starting quantitative analysis of grades, we studied the data distribution and consistently found the same pattern (figures 5.1 and 5.2):

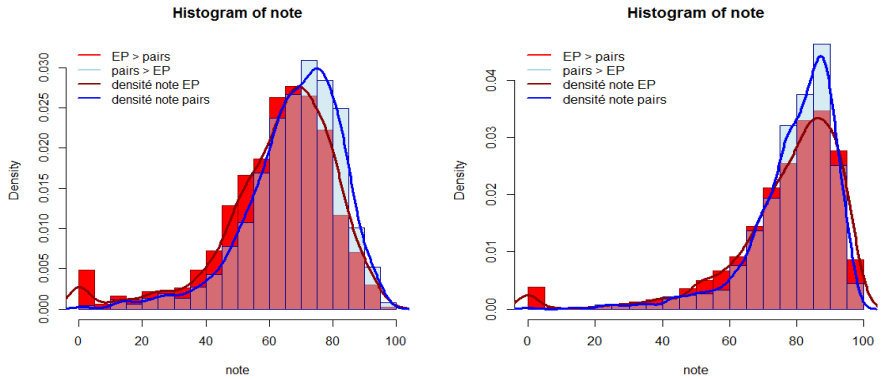


Fig 5.1 and 5.2 Data Pattern of peer and instructor grades Assignments 1 and 5
Red filing: instructors assign higher grades, blue filling if not. Red and blue lines are grades density

Two problems were highlighted:

1. A peer/instructors grading discrepancy with the "0" grade.
2. The shape of the grades distribution does not seem to be a Gaussian/Normal distribution ("bell curve").

The discrepancy with the "0" grade is illustrative of plagiarism: peers issue a "normal grade", while instructor's grade is 0. It is simple to fix: we pre-process and filter these grades out. The second issue: the shape of the grades distribution not being a normal distribution, is confirmed visually (data distributions is "left-skewed"), as well as by a Shapiro-Wilk normality test. This constitutes a major hurdle since all standard statistical tests for hypotheses work under the assumption that data fits a normal distribution.

We used EasyFit (Schittkowski, K. 2002) to test a match with a series of known distributions, without finding a single simple best fit for all assignments. As a consequence and in order get a closer fit to a normal distribution, we decided to go for a basic pre-processing variable change matching the simplest fit, a Log-normal distribution: *ln(100-grade)*. Figure 6 displays the very same data as above, after the change of variable and filtering out plagiarism.

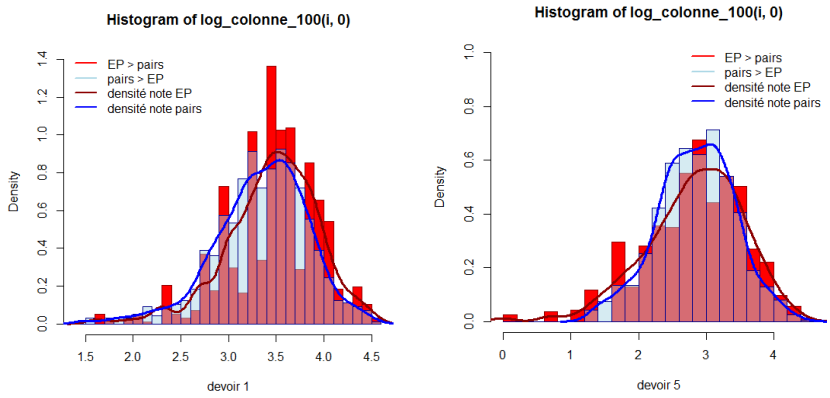


Fig 6.1 and 6.2 Filtered Data Pattern of peer and instructor grades for assignments 1 and 5

While the new data distribution is not really a normal distribution, it is now much closer to it. Using this new dataset to conduct confirmatory data analysis improves the validity of our study.

Is individual peer grading as accurate as instructor grading?

Literature on peer grading offers contrasting views on this issue (Bachelet 2010, Doiron, G. 2003, Sadler & Good 2006). In Sadler & Good (2006), peer grades are highly correlated with instructor grades r = 0.905, while we find an average correlation of 0,772 with a p-value <2.2e-16 (Pearson correlation, data processing with R version 3.1.1). In our dataset, with consideration to the difference between instructor and peer grade results are as follow: 36% of grades are extremely similar (±5 points on a scale of 100), 60% very similar (±10 points), 85% dissimilar (±20 points off).

Now what precision can reasonably be expected? It is widely recognized that different instructors grading an essay does not give the same grade. But what actual precision does instructor grading provide? While we did not find research in English on this question, French education science research on grading (*docimologie*) can provide us with a benchmark on the convergence when different instructors grade a given essay. According to Suchaut, B. (2008)[i]: 5, 10 and 20% convergence of individual instructor's grades to average grade is: 39%, 65% and 91%. While this is better than the 36%, 60% and 85% we get with student grading, we need to test whether processing several peer grades instead of using only one will perform better.

Which grading algorithm is best for processing peer-produced data in order to deliver grades which are similar to instructor grading?

Beyond "one shot" individual grading, MOOCs allows for improving grading accuracy by using several peer grades. Our goal with students taking part in MOOC GdP2 was to require 5 peer assignments grades, but with platform allocation issues and defections they were 0-6 grades given, and on average 3.57 to 4.67 peer grades for each assignment (with 6 to 11% of students not taking part in peer evaluation at all).

Is it possible to improve the precision of grading by using not one, but several peer grades? Before that, we need to determine the best algorithm for processing peer grades. While sophisticated methods exist, like Calibrated Peer Reviews, Credibility index (Suen, H. 2014), or Bayesian post hoc stabilization (Piech, C. et al. 2013), we shall stick to the most common and understandable methods which are grades *average* and grades *median*. Each has strong and weak points (average uses all available data, while median is "robust" as it filters out extreme values…). While using the average is the usual way to go, Coursera MOOC platform uses median. In order to find the best method, we study two "error functions": the difference either of the average or the median of students grades with instructor grades. Our research shows using the average as slightly more accurate (1 point on a 100-points scale), see figures 7.1 and 7.2 below.

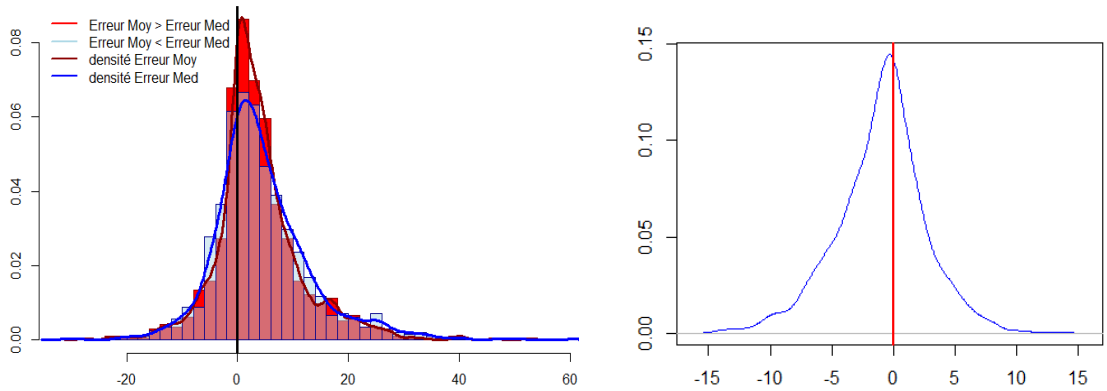


Fig 7.1 Error function of average (brown) median (blue) 7.2. Difference (ErrorWithAverage – ErrorWithMedian)

How efficient is the average peer grade algorithm?

With consideration to the difference between instructor and peer grade and using all the available peer data 56.6% of average grades are extremely similar to instructor grades (±5 points on a scale of 100), 82.6% very similar (±10 points) and so, approx. 17% of assignments are graded more than 10 points off). Figures 8.1 and 8.2 show a comparison of average peer grades and instructor grades: it highlights the fact that 1/ the peer average algorithm gives very slightly higher grades (2.6 points on a 100 scale) and 2/ 13.9 average std deviation.

This claim is supported by literature on instructor grading consistency, (Suchaut B. 2008) which shows that instructors grading a series of identical assignments have a consistency measured by standard deviation of 15. In this regard, this peer evaluation system performs better than instructors.

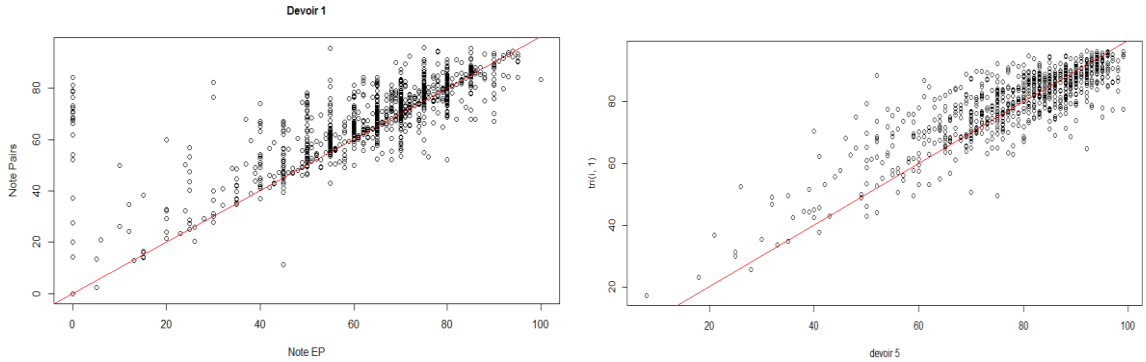


Fig 8.1. Average peer grades vs. instructor grades in Assignment 1 and Fig 8.2 in Assignment 5
x axis is instructor grade, y is average of students grades

How many peer grades are required to correctly estimate "best grade"?

Does grading precision improve with the number of peer grades we take into consideration? How many are required to have an optimum end grade? To study this question using R version 3.1.1, we make a series of "what if" simulations in which 1, 2, 3, 4 or all peer grades are used to calculate the average peer grade. Then checking what proportion of the final grade is close to instructor's grades.

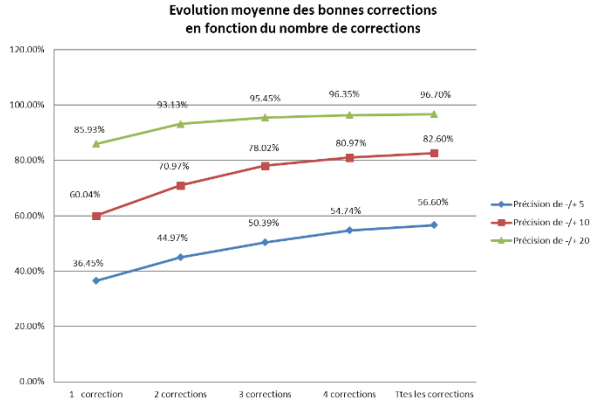


Fig 9. : 5, 10 and 20% accuracy of average peer grades to instructor grades vs. number of peer grades

Data shows that taking into account more peer grades consistently improve the accuracy of grading, though with decreasing returns.

Although in this case peer grading quickly performs better (with two peers), than instructor's grading (as stated before according to Suchaut, B. 2008 5, 10 and 20% convergence of individual instructor's grades to average grade is 39%, 65% and 91%), it is our opinion that the best "return for work" appears with 3-4 peer grades, as our graph shows a quick improvement going from 1 to 3 peer grades.

Improving peer evaluation monitoring and grades processing in MOOC GdP3 and 4

This study was conducted with GdP2 data, but since then we have experimented new avenues of work and improved the design and management of the advanced track:

- During MOOC GdP3, we estimated students grading bias with a benchmark assignment and calculated final grades using the average of bias-corrected grades. We also saved a lot of teaching-staff time by drastically reducing instructor grading in favor of peer grading: instead of using 100% instructor grading, we implemented "targeted instructor grading". Based on a series of warning indicators (such as "too few peer grades", "non-consensus amongst peers") we were able to make progress, going from 100% instructor grading to 10% instructor grading. We also initiated a messaging system before they grade the next assignment, providing feedback to peer graders that were clearly "off", grading to high or too low.
- In MOOC GdP4, using the results presented in this paper, we were able to estimate the quality of grades issued by peers and act on this information: a dedicated VBA/Excel application (Benbitour M. H., Bachelet, R. Zongo D. 2014) was developed to provide feedback on whether each grade was correct, high or low and reward accurate grading with bonus points (or issue negative points when grossly inaccurate grades we given or when grading work was simply "not done" see figure 10). We were able to track whether peer grading improved with time and successive assignments. We also detected assignments with less peer grades and set up an extra peer assignment attribution session in order to insure the availability of at least 3 peer grades for each assignment. Also, we implemented self-evaluation of submissions (using the additional self-grade in a new algorithm). Adding self-evaluation is an important advance, as Sadler & Good (2006) point out it is the best source for learning (even better than peer evaluation itself). In the new system, developed for Canvas in association with Unow, students were asked to get a fresh look at their own work and grade it after having evaluated at least 3 other student's assignments.

[Public] DEVOIR1 Bonus-malus [MOOC GdP4] : v

mode d'emploi :goo.gl/qrr7f9

Student ID	Bonus de Précision	Malus de Précision	Malus d'évaluation	Bonus-Malus final	Note remise 1	Note remise 2	Note remise 3	Note remise 4	Note remise 5
52931	0	0	12	-12	83				
42965	6	7	0	-1	66	42	50	65	62
45091	5	15	0	-10	73	74	70	92	74
53849	0	0	15	-15					
35114	4	7	0	-3	72	87	78	53	51
55807	8	4	0	4	80	66	83	51	42
28613	0	0	15	-15					
55857	5	12	0	-7	70	55	82	73	67
43748	11	0	0	11	80	69	63	53	75
48163	6	7	0	-1	50	70	45	49	67
46998	0	11	0	-11	44	35	55	29	42
55696	7	2	0	5	57	59	42	51	51
34116	10	0	0	10	54	65	78	53	34
50785	9	2	0	7	82	75	71	76	59
51505	8	2	0	6	50	36	65	38	63
50634	9	4	0	5	63	24	51	65	70
48499	4	0	6	-2	73	55	26		
45714	7	5	0	2	51	43	54	54	21
28669	12	0	0	12	72	60	61	68	58
52783	11	0	0	11	76	30	40	45	73
50148	9	0	0	9	67	66	65	50	49
7352	5	15	0	-10	66	69	60	88	78

Fig 10: Application of this research – providing feedback and rewarding peer grading in MOOC GdP4

Conclusions and recommendations to researchers and MOOC designers

While it has promising potential, peer grading is not easy to implement: not only does it require careful work, beta testing and setting up a new case study each session, but also dedicated data processing and expertise in the instructors and technical teams. Carefully setting up the peer assessment protocol itself is paramount: deadlines, reminders, targeted messages, deciding how each student gets feedback on the quality of the grades that were assigned to him/her and. Rewards for accurate grading must be provided, along with penalties when grading is incorrect or has not been submitted. Moreover, although the instructor team is not required to grade all papers anymore, monitoring and manual grading is still required on occasions.

Our final recommendations to practitioners are:

- Researchers should first look closely at peer grades distribution: it needs to be filtered and a variable change can bring considerable improvement in hypothesis testing validity.
- How many assignments should a student be required to grade? We recommend four as the best compromise to obtain a quality final grade, accounting for peers who could drop out of the process and to leave time to work on issuing one additional grade, which would be a self-assessment.

- What algorithm should be preferred? Use average if grading data has been correctly checked and filtered. Otherwise, median is probably the best way to go.
- The optimum threshold to switch from automatic peer grading to instructor grading remains to be determined. We used three criteria: less than 2 peer grades, non-consensus (i.e. peer grades standard deviation >20) and presence of a 0 grade. Thus in GdP4, respectively 10%, 9% and 1.6% of assignments 1, 2 and 3 were graded manually (Bachelet, R. Carton P-A. 2015).

Before concluding our paper, we have to stress one more time the major importance of using quality learning materials, good peer training and sound incentives so that the conditions for peer assessment are optimal. Hard work and dedication of instructors, as well as MOOC platform ergonomics and functionalities are also paramount. It is regrettable that all these vital elements in peer grading are overlooked in most studies: having "lots of data" and "sophisticated algorithms" is important, but secondary to setting up the best possible conditions for peer evaluation and peer learning.

Software and data sources all retrieved 15/02/2015

Bachelet, R. (2012a). Cours et MOOC de gestion de projet : formations en vidéo, ppt, pdf et modèles de documents from: http://gestiondeprojet.pm

Bachelet, R. (2012b). Formation à distance en gestion de projet, Retrieved, January 11, 2015, https://sites.google.com/site/coursgestiondeprojet

Bachelet, R. (2013). Analytics et taux de réussite du MOOC GdP 2 (septembre-décembre 2013) http://goo.gl/Ia7Syw

Bachelet, R. Carton P-A. (2015). Analytics et taux de réussite du MOOC GdP 4 (septembre-décembre 2014) http://goo.gl/RC8JLV

Bachelet, R. Petit, Y. (2014). Gradebook of MOOC GdP 2 Extracted from Canvas MOOC platform, December 20, 2014

Benbitour M. H., Bachelet, R. Zongo D. (2014). Traitement des notes des évaluations par les pairs du MOOC GdP3 et 4, Excel Visual Basic Spreadsheet http://goo.gl/1UGNTX

Schittkowski, K. (2002). Easy-Fit: a software system for data fitting in dynamical systems. Structural and Multidisciplinary Optimization, 23(2), 153-169. http://www.mathwave.com

References

Bachelet, R. (2010). Le tutorat par les pairs in *Le tutorat par les pairs" in Accompagner des étudiants*, direction Verzat C. Villeneuve L. Raucent B. De Boeck, ISSN 0777-5245

Bachelet, R. (2014). Les MOOC, analyse de dispositifs, Évaluation par les pairs, Atelier n°1 : Les MOOC : analyse de dispositifs médiatisés et d'usages par des apprenants Colloque TECFA e-learning 3.0, Université de Genève, 17-18 octobre 2014 http://gestiondeprojet.pm/mes-contributions-sur-les-mooc

Bachelet, R., Cisel, M. (2013). Evaluation par les pairs au sein du MOOC ABC de la Gestion des projets: une étude préliminaire. Atelier MOOC, EIAH, Toulouse http://goo.gl/6JHYOv

Cisel, M., Bachelet, R. Bruillard E. (2014). Peer assessment in the first French MOOC: Analyzing assessors' behavior. Proceedings of International Educational Data Mining Society. http://goo.gl/2YNPhw

Doiron, G. (2003). The value of online student peer review, evaluation and feedback in higher education. *CDTL Brief, 6*(9), 1-2. http://www.cdtl.nus.sg/Brief/Pdf/v6n9.pdf

Piech, C., Huang, J., Chen, Z., Do, C., Ng, A., & Koller, D. (2013). Tuned models of peer assessment in MOOCs http://arxiv.org/pdf/1307.2579.pdf

Sadler, P. M., & Good, E. (2006). The impact of self-and peer-grading on student learning. Educational assessment, 11(1), 1-31 http://goo.gl/giznLS

Steven Burrows, Iryna Gurevych, Benno Stein. (2014) The Eras and Trends of Automatic Short Answer Grading. International Journal of Artificial Intelligence in Education. http://goo.gl/3YsplI

Suchaut B. (2008) La loterie des notes au bac : un réexamen de l'arbitraire de la notation des élèves. Document de travail de l'IREDU 2008-03. 2008. https://halshs.archives-ouvertes.fr/halshs-00260958v2

Suen, H. (2014). Peer assessment for massive open online courses (MOOCs). The International Review Of Research In Open And Distributed Learning, 15(3). Retrieved from http://goo.gl/E3PsHJ

[i] Though experimental conditions are not the same, Suchaut (2008) is the only benchmark we could find as to the convergence of grades when different instructors grade the same assignment. It was carried out on 6 assignments, each graded by 32 to 34 instructors, submitted for national exam *baccalauréat* in economics and social sciences. This is not the best possible benchmark: given the resources, this experiment should be carried out in conditions closer to MOOC peer grading, i.e. same student paper, same rubric and grading training, and same online platform.