Je-LKS
Journal of e-Learning and Knowledge Society
The Italian e-Learning Association Journal
PEER REVIEWED PAPERS
Vol. 13, n.1, 2017
ISSN: 1826-6223 | eISSN: 1971-8829

ADAPTIVE PEER GRADING AND FORMATIVE ASSESSMENT

Giovannina Albano,
Nicola Capuano,
Anna Pierri

Dept. of Information, Electrical Engineering and Applied Mathematics
University of Salerno, Italy
galbano@unisa.it - ncapuano@unisa.it - apierri@unisa.it

Keywords: Formative assessment, Peer grading, Peer rank, Fuzzy ordinal peer assessment.

Peer grading is a process whereby students are required to grade some of their peers' assignments as part of their own assignment. Peer grading is capable of improving students' learning outcomes, metacognition and critical thinking and, at the same time, it can support formative assessment, saving teacher's time and providing fast feedback, especially for large classes. In this paper we report the results of an experiment where a technology supported peer grading exercise has been assigned to students within a University course on calculus and linear algebra. To improve the reliability of students' grades, several approaches have been experimented and the obtained results have been compared to grades coming from the teacher. Moreover, we attempted to understand how the peer grading task has contributed to reinforce the development of student's explanation and argumentation processes.

for citations:
Albano G., Capuano N., Pierri A. (2017), *Adaptive Peer Grading and Formative Assessment,* Journal of e-Learning and Knowledge Society, v.13, n.1, 147-161. ISSN: 1826-6223, e-ISSN:1971-8829

1 Introduction

Explanation, argumentation and proof are mathematics activities that assume a main role in teaching and learning mathematics, in particular of linear algebra. Indeed, University students, especially in their first semesters, often lack specific mathematical learning and working techniques that are necessary to develop and apply mathematical notions, definitions, theorems and proofs. For this reason, the need to implement a feasible assessment strategy that contributes to improve students' learning is widely recognized (William, 2007)

In this paper, we attempted to understand if peer grading can be used for this purpose. Such educational practice foresees that, given an assignment coming from the teacher, students are asked not only to complete and submit it, but also to grade a small number of assignments submitted by their peers and provide additional feedback. The proposed grades are then combined and final grades are obtained and assigned to the students themselves.

The literature reports on many learning benefits for peer-assessors like the exposure to different approaches, the development of self-learning abilities, the enhancement of critical thinking, etc. (Glance *et al.*, 2013), also reinforcing the development of student's explanation and argumentation processes. Moreover, even if it relies on grades assigned by intrinsically unreliable graders (the students), the application of peer grading also presents logistics advantages in saving teacher's time and providing fast feedback to the class (Sadler & Good, 2006).

The paper is organized as follows. The next section presents theoretical background on different assessment strategies including formative and peer assessment. Section 3 presents the methods we have experimented for the aggregation of grades coming from peer assessment. Section 4 illustrates the experimental setting and the applied methodology. Section 5 discusses the results of the experiment performed with real students in a University class. Finally, conclusions are summarized in Section 6.

2 Theoretical background

Several studies support the fact that, the focus on assessment for learning, may produce substantial improvement in the performance of students (William, 2007). In (Black & Wiliam, 2006) authors recognize that *"assessment in education must, first and foremost, serve the purpose of supporting learning"*. This is true in general and, especially, in mathematics where students need to develop their knowledge about specific topics while developing a reflective practice, that includes self-assessment.

Formative assessment is a teaching method where *"evidence about student achievement is elicited, interpreted, and used by teachers, learners, or their peers, to make decisions about the next steps in instruction that are likely to be better, or better founded, than the decisions they would have taken in the absence of the evidence that was elicited"* (*Ibidem*).

An important function of formative assessment is providing students with "continuous feedback", meaning that opportunities for feedback should occur continuously, but not intrusively, as a part of instruction (Bransford *et al.,* 2000). In particular, the attention should be focused on two questions that arise when designing assessment tools (Thomas *et al.*, 2011): who should be doing the assessment and how can the results be measured to ensure that they are valid (Gibson & Dunning, 2012).

Assessment can be also useful to develop students' *argumentation* skills where argumentation is seen as an intentional explication of the reasoning used during the development of a mathematical task (Seeryet *et al.*, 2012). Different assessment strategies can be adopted for this purpose. Among others, the written report seems to be a privileged tool to monitor students' learning, indeed it can help students to reflect upon their work. As reported in (Forman *et al.*,1998), *"intensive approach to argumentative skills, relevant for mathematical argumentation, seems to be possible through an interactive management of students' approach to writing"*.

Peer grading can be used to support both the development of students' argumentation skills and the provision of formative assessment. It encourages students to clarify, review and edit their ideas, through the focus of peer feedback. At the same time, it requires students to provide either feedback or grades to their peers on a product or a performance, based on the criteria of excellence for that product or event which students may have been involved in determining (Sadler & Good, 2006).

Nevertheless, while using peer grading in formative assessment, the main issue is represented by the lack of accuracy of grades proposed by other students that may result in an erroneous feedback. Several approaches have been proposed so far to make peer grading more reliable. One of the most used is the *Calibrated Peer Review* (CPR), (Carlson & Berry, 2003), that proposes a calibration step to be performed by students before starting to assess other students' assignments.

During the calibration, each student rates the same small set of assignments that have been already rated by the instructor. The discrepancy between grades provided by a student and by the teacher measures her accuracy in assessment and is then used to weight subsequent assessments provided by that student. The more accurate is an assessor the more weight is given to her judgment in the peer grading task.

The main drawback of CPR and similar methods is that they require additional work for the calibration step. Moreover, they do not take into account the progresses that students make over time until a new calibration step is performed. For this reason, researchers are developing new approaches able to automatically tune peer grades based on different parameters (Capuano & Caballé, 2015). In the next section we summarize some of these methods that have been used in this work to support formative assessment based on peer grading.

3 Peer Grading Methods

In a typical peer grading scenario, an *assignment* is given to n students. Each student elaborates her own solution generating a *submission* and has then to grade m different submissions (with $m < n$) coming from other students. The assignment of submissions to assessor students is done in accordance to an *assessment grid*: a Boolean $n \times n$ matrix A where $A_{i,j} = 1$ means that the student j has to grade the student i. In (Walsh, 2014) some properties of the assessment grid are analysed and several (random and smart) methods aimed at building such grid are described.

The grades proposed by the students are then collected in the *grades matrix* G where $G_{i,j}$ is the grade proposed by the assessor student j for the assessee student i so that $0 \leq G_{i,j} \leq 10$. In an ideal peer grading setting, every student performs the grading task so, the *final grade* g_i of each student i is obtained starting from the matrix G, by averaging all the grades obtained by peers (a matrix row) with the following equation:

$$g_i = \frac{1}{m}\sum_{j=1}^{n} G_{i,j} \quad \forall\, 1 \leq i \leq n. \tag{1}$$

In order to improve the reliability of final grades, the *PeerRank* method has been proposed in (Raman & Joachims, 2014). Such method weights the grade that each assessor student gives to another student by her own grade i.e. it uses the grade of a student as a measure of her ability to grade correctly. In other words, the grade g_i of a student i is so that:

$$g_i = \frac{\sum_{j \to i} G_{i,j} \cdot g_j}{\sum_{j \to i} g_j} \tag{2}$$

where both summations are performed over all students j having evaluated i (indicated with $j \to i$) i.e. so that $A_{i,j} = 1$.

Given that the grades of all assessor students are themselves weighted ave-

rages of grades obtained by their own assessors, an iterative process is needed to calculate the final grade of each student. Such process has been improved in (Capuano & Caballé, 2015) where the *F-PeerRank* rule has been proposed to apply a super-linear modifier to the grades proposed by peer assessors in order to minimize the contribution of low skilled student while maximising that of high skilled ones.

In the same paper, the *BestPeer* rule is also proposed to assign the maximum influence only to the best grader for each student and no influence at all to any other proposed grade. The method performs particularly well when at least one good grader is available for each assignment. This constraint may be satisfied generating assessment grids that, taking into account past grading performances, balances reliable graders among students.

In (Raman & Joachims, 2014), authors have shown that ordinal feedback (e.g. "the report *x* is better than the report *y*") is easier to provide and more reliable than cardinal one. Given an assessment grid A, each student s_j can so define a ranking on the assignments coming from students in $\{s_j \mid A_{i,j} = 1\}$. The defined rankings can be collected in a ranking matrix R where s_i is the position of S_i in the ranking defined by i if $s_i \in S_j$, 0 otherwise.

Starting from the ranking matrix, a simple and effective way to compute a complete ranking over the set of submissions is the classical *Borda* count (Borda, 1781) where the partial ranking provided by each assessor is interpreted as follows: m points are given to the submission ranked first, $m-1$ points to the one ranked second, etc. The *Borda* score of the submission coming from s_i is calculated as follows:

$$Borda(s_i) = \sum_{j=1}^{n} A_{i,j} \cdot (m - R_{i,j} + 1). \quad (3)$$

The global ranking is then computed by ordering all the submissions in decreasing order of their Borda scores.

In (Capuano *et al.*, 2016) an alternative ordinal peer assessment method named *FOPA* (Fuzzy Ordinal Peer Assessment) is discussed. In such method, each student is asked to rank few random submissions from the best to the worst and to specify, with a set of intuitive labels from the set $\{\approx,\geq,>,>>\}$, at what extent each submission is better than the next one in the ranking. Provided rankings are then transformed in fuzzy preference relations, expanded to estimate missing values and aggregated through ordered weighted averaging. The aggregated relation is then used to generate a global ranking between the submissions and to estimate their absolute grades.

4 Experimental Methodology

To evaluate the capability of peer grading in supporting learning activities, we have experimented the methods summarized in the previous section within a University course on mathematics involving about 200 students. The experiment was aimed at answering two experimental questions:

1. at what extent peer grading is a valuable support for formative assessment?
2. at what extent peer grading is also capable of improving students' learning outcomes?

In the next subsections, we describe the experiment setting and report about collected data. In the next section we analyse the data with the aim of providing an answer to the experimental questions here reported.

4.1 Experimental setting

The experimental set was composed by first year students taking part in a two trimester intensive module of mathematics within a 3-year B.Sc. degree in Computer Engineering. In particular, the focus was on the second module, which concerned topics from calculus and linear algebra.

The module was made of eight hours per week in face-to-face traditional lectures/exercises sessions, supported by a standard e-learning system (a *Moodle* implementation) which provided the students with additional learning resources and communication tools. The experiment has been performed with voluntary students. In particular, in a class of about 200 students 43 students have decided to participate.

The peer grading exercise was implemented through the *workshop* component of Moodle allowing students to assess each other's work related to a specific topic of the course and based on criteria established by the teacher. The submission consists of plain text and optional attached files including the answer of the student with respect to a specific topic assigned by the teacher. The workshop activity consists of six sequential phases (shown in Fig.1):

- *planning:* the teacher decides the grading strategy and the allocation method;
- *setup*: the teacher creates assessment forms and instructions and configures settings;
- *submission*: students submit their own work and submissions are allocated to reviewers;
- *assessment*: students review each other's work according to criteria established by the teacher;

- *grading evaluation*: student grades are calculated;
- *closing*: students can see their own grades, peer reviews and other feedback.

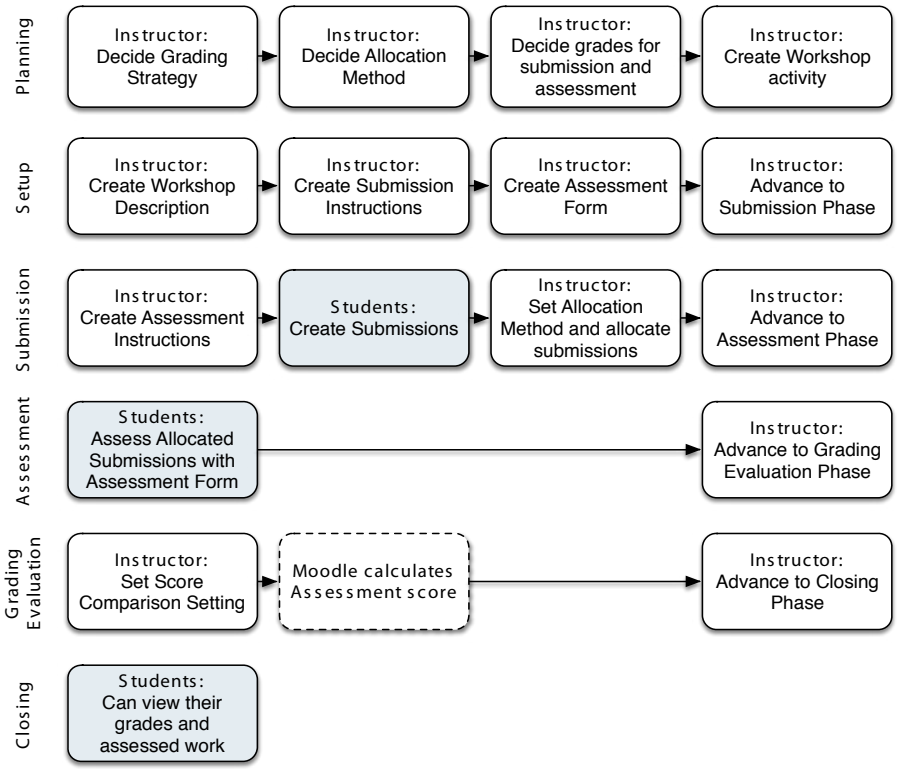

Fig. 1 - Workflow implemented by Moodle workshop component

The workshop component calculates and assigns the final grade to each student through the average rule (equation 1). Optional weighting of grades is possible if the instructor wishes to contribute a peer assessment. The workshop also assigns a second grade evaluating the ability of each student in assessing others' work. To compare the applied method to the other methods summarized in section 3 we have disregarded the second assigned grade. Moreover, the peer grading task has been performed in a blind mode in order that students do not know whom they are assessing.

4.2 Data collection

A total of 43 students have participated in the experiment providing a

submission for the 4 questions making up the assignment. Then, 3 other submissions to be graded have been assigned to each student through a randomly filled assessment grid so that:

- the sum of the elements in each row and column is equal to 3 (each student was planned to grade and be graded by 3 other students);
- the sum of the elements in the main diagonal is equal to 0 (i.e. none has evaluated himself).

A subset of 26 students decided to participate also to the (optional) assessment step by grading peer submissions. A total of 24 students completed the grading task for all the 4 questions of the 3 assigned submissions while 2 students provided only partial marks. This resulted in a total of 304 assigned grades with an average of 1.8 grades per question.

Unfortunately, while some students received all expected evaluations, 3 students did not receive any evaluation at all. While this singularity has little impact on the *Average* method (provided that *m* in equation 1 is settled to the number of available votes for each submission rather than to the number of expected votes), the impact on *PeerRank* and *F-PeerRank* methods is higher.

As seen in section 3, such methods weight the grades provided by each assessor by her own grade. So, grades provided by ungraded students have no value at all. This impacts recursively on the grades of the assessed students and on those of the students assessed by them. To avoid this problem, we have assigned dummy grades to ungraded students and used them throughout the algorithm iterations. Dummy grades, initially set to the average grade of the class, have been removed after all class grades have been calculated.

To use ordinal methods like *Borda* and *FOPA* on students' cardinal input we have generated the ranking for each assessor by simply ordering the submissions she graded from the best to the worst (according to the proposed grade). To feed the FOPA method, we have also specified at what extent each submission is considered better than the next one in a student ranking.

In particular, for each pair of subsequent submissions, a feasible symbol $\sigma_i \in \{\approx, \geq, >, >>\}$ has been selected, starting from the grades g_i and g_{i+1} assigned by the student to those submissions, according to the following equation (Capuano *et al.*, 2016):

$$\sigma_i = \begin{cases} \approx & \text{if } g_{i+1} - g_i < 0.5; \\ \geq & \text{if } 0.5 \leq g_{i+1} - g_i < 1; \\ > & \text{if } 1 \leq g_{i+1} - g_i < 2; \\ \gg & \text{if } g_{i+1} - g_i \geq 2. \end{cases} \quad (4)$$

To evaluate the effectiveness of peer grading as a tool for formative assessment, we have also asked the teacher to provide her grades for all the available submissions. Teacher grades have been collected separately and did not affected the peer grading process

5 Experimental Results

In this section we report the results obtained with respect to both experimental questions summarized in section 4 i.e. at what extent peer grading is a valuable support for formative assessment and at what extent it is also capable of improving students' learning outcomes in terms of development of explanation and argumentation processes. A specific sub-section is dedicated to each question.

5.1 Effectiveness w.r.t. formative assessment

To evaluate the effectiveness of peer grading as a tool for formative assessment, we have applied the methods described in section 3 to the data collected through the *Moodle* workshop and have compared the obtained grades to those calculated by *Moodle* (adopting a standard *Average* rule) as well as to those assigned by the teacher.

Table 1 compares the results obtained using the *Average* rule with those obtained by alternative rules summarized in section 3. Performances have been measured in terms of Root Mean Square Error (RMSE) between the grades estimated through each experimented rule and the grades assigned by the teacher (where each grade is expressed in a scale ranging from 0 to 10). The RMSE is calculated according to the following equation:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(g_i^r - g_i)^2}{n}} \tag{5}$$

where g_i is the grade assigned by the teacher to the i-th student,g_i^r is the estimated grade assigned to the same student thorough the the rule r and n is the number of students. In the table we use *AVG* for Average, *PR* for *PeerRank*, *FPR* for *F-PeerRank* and *BP* for *BestPeer*. Parameters for *PR*, *FPR* and *BP* have been set as suggested in (Capuano & Caballé, 2015).

Table 1
PERFORMANCE OBTAINED ON EXPERIMENTAL DATA

Question	RMSE per Method					
	AVG	PR	FPR	BP	Borda	FOPA
1	3.73	3.65	3.64	3.54	4.20	3.70
2	4.54	4.04	4.00	4.38	3.92	4.14
3	4.04	3.22	3.18	4.01	3.27	2.95
4	**4.19**	**3.80**	**3.71**	**4.21**	**3.95**	**3.92**
Mean	**4.12**	**3.68**	**3.63**	**4.03**	**3.84**	**3.68**

The first thing that can be noted is that grades coming from students are very unreliable if compared with grades assigned by the teacher. This may be due to the fact that the data comes from the first experience of the class with a peer-grading exercise and it has been performed at the very beginning of the course. Moreover, only about the 60% of all students have also participated in the assessment step resulting in a lack of data to be used by aggregation methods.

The positive thing is that the proposed alternative methods reach a lower error with respect to the baseline *Average* method provided by *Moodle*. In particular *F-PeerRank* outperforms the other methods on average and in almost all the single cases. It is also notable that *FOPA* reaches similar results by relying only on a subset of the information used by *F-PeerRank* (just the ranking of submissions is used rather than the assigned ordinal grades).

Given the low participation in the assessment task, two additional analyses have been performed on collected data to evaluate the behaviour of grading methods when the amount of available information increases. Given the availability of teacher's evaluations for all submissions, we have measured how the performance of all the methods changes by considering, in addition to grades coming from students, increasingly large subsets of grades coming from the teacher.

Both analyses were made in 43 steps (one for each submission). At each step, 4 additional grades coming from the teacher were considered, one for each question of a new submission (the priority was given to submissions with the fewer amount of available evaluations).

In the first analysis, the teacher was considered as a common student evaluating some of the available submissions. For each question, a new column filled of 0 has been so added to both the assessment grid and the grades matrix. At each step an element i of this row was turned to 1 in the assessment grid and the corresponding element of the grades matrix was set as the grade assigned by the teacher to the i-th submission.

An additional row was also added to both matrices to set dummy grades

assigned by other students to the teacher (used by *PeerRank*, *F-PeerRank* and *BestPeer* methods). In particular, the new row has been filled of 1 (apart for the last element, set to 0) in the assessment grid and filled of 10 (apart the last element, set to 0) in the grades matrix. The teacher is so considered as graded 10 by all other students.

The Fig.2 shows how the RMSE of the proposed methods changes while adding new grades from the teacher. As it can be seen, *BestPeer* and *FOPA* obtain the best performance while *F-PeerRank* shows an error which is always below than that made by the standard *Average* method. The *PeerRank* rule is better than the *Average* one until 17 added grades, then it results to be a bit worse. *Borda* is quite better than *Average* until 11 added grades, then it becomes quite worse.

Although *BestPeer* and *FOPA* seem to show a similar behaviour, it should be noted that the performance of *BestPeer* is boosted by the dummy grade of 10 assigned to the teacher. Given that it returns the grade assigned by the best grader, in almost all cases, when available, it returns the grade assigned by the teacher. Instead *FOPA* makes no assumption on the grades obtained by graders so it can be considered as the most reliable rule among those experimented.

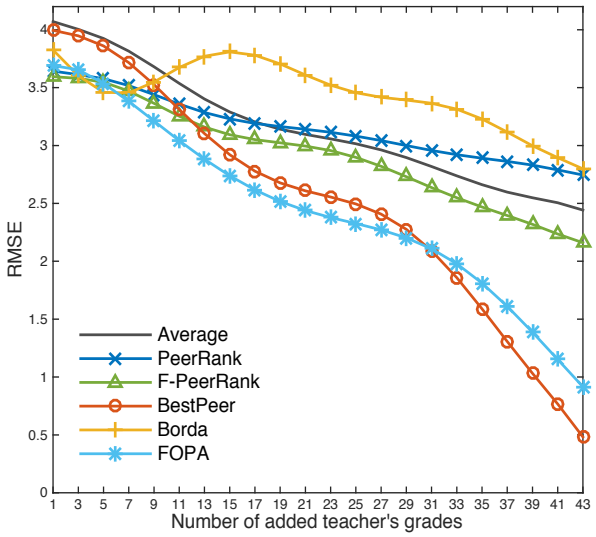

Fig. 2 - Performance in terms of RMSE of the defined methods considering increasingly large subsets of grades coming from the teacher. Case 1: the teacher is seen as a common student

It should be also noted that the results of *Borda* are quite penalised by the fact that, to uniform scores, they have been normalized by the total number of assessment made by each assessor. So, while the number of teacher's grades increases, their weight with respect to the other decreases.

The second analysis is similar to the first one, except that the teacher is considered as a "super" student, whose grades, if available, are preferred over the grades provided by common students. In fact, while the first analysis is aimed at determining how the described methods behave with additional available grades, the second one is aimed at determining if they can reach even better performances by asking to the teacher to fill the gaps in the data.

The Fig.3 shows how the RMSE of the proposed methods changes while adding new grades from the teacher. Also in this case *BestPeer* and *FOPA* show the best performances: *FOPA* wins until 33 added grades, then *BestPeer* is better. In this case, the differences among methods remains almost constant while in the previous case they increase with the number of available grades. Also in this case, the results of *Borda* are penalised for the same reasons explained above.

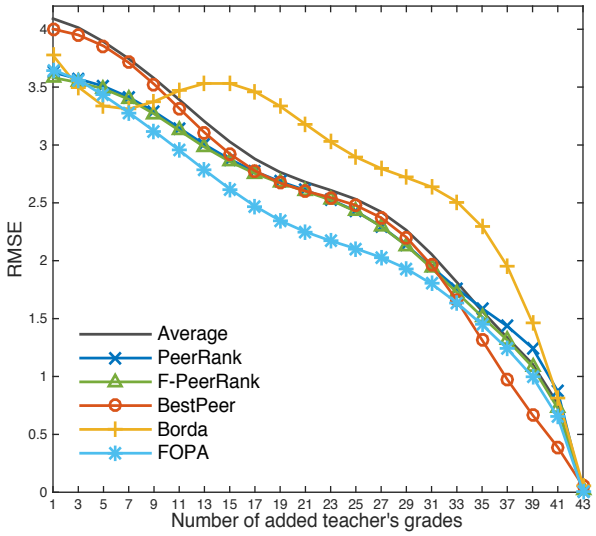

Fig. 3 - Performance in terms of RMSE of the defined methods considering increasingly large subsets of grades coming from the teacher. Case 2: the teacher is seen as a super student

Based on experimental data we can affirm that it is possible to improve the results of peer grading through the application of alternative methods with

respect to the standard *Average* rule. In particular, *FOPA* is the method that is able to provide the best results with less information (a ranking is needed rather than ordinal grades). Moreover, the results of *FOPA* improve more than the others with increasing amount of information available.

When only few unreliable evaluations are available (as in the analysed case) the use of peer grading as support for formative assessment is questionable. The results obtained, even when corrective algorithms are used, are quite far from grades assigned by the teacher. As seen, such results can be improved by asking the teacher to fill the gaps in the data but significant improvements in terms of RMSE are obtained only with a large number of additional quality grades.

5.2 Effectiveness w.r.t. learning outcomes

To evaluate the effectiveness of peer grading as a tool for improving the learning outcomes, we have had an open interview with the tutor that has oversaw the online activities of the students. She sees *peer grading as a good strategy for filling knowledge gaps through a different perspective* and suggests its application also on the subsequent topics of the course.

Formative assessment is in fact a process observable over a long period of time and the proposed methodology is capable of catching information over time. Accordingly to the *Meta-Didactic Transposition Model* (Arzarello *et al.*, 2014), the observations are seen as windows open on the classroom at key moments, accompanied by teachers' auto-reflections. Moreover, according to the Theory of Didactic Situations (Brosseau, 1997). they can be performed from the introduction to the institutionalization of knowledge.

The involved tutor also thinks that the method enables to review learnt topics in a collaborative way. In fact, peer grading sees an involvements of students both as assessors of their own learning and as resources to other students. One of the key components of engaging students in the assessment of their own learning is providing them with descriptive feedback as they learn. Descriptive feedback provides students with an understanding of what they are doing well, links to classroom learning, and gives specific input on how to reach the next step in the learning progression. In such sense the tutor recognize that peer grading can contribute not only to improve the students' understanding of key concepts of linear algebra but, also, to reinforce the development of explanation and argumentation processes.

Conclusions

In this paper we have presented the results of an experiment aimed at introducing peer grading within a University course on calculous and linear algebra

to both support formative assessment and improving learning outcomes.

We have demonstrated that it is possible to improve the accuracy of peer grading through the application of alternative methods with respect to the standard Average rule. On the other hand, when only unreliable evaluations are available the reliability of obtained grades can be still low. Nevertheless, also in this case, the results can be improved by asking the teacher to fill the gaps in the data by providing high quality evaluations for a selected subset of submissions.

To understand how peer grading can contribute to reinforce the development of student's explanation and argumentation processes, we made an interview with the tutor for the on-line course activities that returned a positive preliminary feedback. Further investigations, correlating final summative grades obtained by course participants to peer grading activities, will be made within the same course to obtain additional quantitative evidences supporting this thesis.

The obtained results suggest to extend the experience to other university courses, both in Sciences and Humanities. Moreover, it is also conceivable the application of the discussed assessment tool to massive on-line courses, provided by universities and other educational organisations and intended for thousands of simultaneous participants. Given the high numbers of enrolled students and the relatively small number of tutors, such courses make great use of automated assessment approaches. Among them, peer grading, made more reliable with the application of the discussed methods, may become the elective assessment tool thanks to its capability of easily scaling to any size.

REFERENCES

Arzarello, F., Cusi, et. al. A., Garuti, R., Malara, N., Martignone, F., Robutti, O., Sabena, C. (2014), *Meta-Didactical Transposition: A theoretical model for teacher education programmes*. The Mathematics Teacher in the Digital Era, 347-372, Springer.

Black, P., Wiliam, D. (2006), *Assessment for learning in the classroom*. In J. Gardner (Ed.), Assessment and learning, 9-25, SAGE Publication.

Borda, J. C. (1781), *Memoire sur les elections au scrutin*, Histoire de l'Académie Royale des Sciences.

Bransford J.D., Brown A. & Cocking R. (2000), *How People Learn: Mind, Brain, Experience and School*. National Academy Press.

Brosseau,G. (1997), *Theory of Didactical Situations in Mathematics*. Kluwer.

Capuano, N., Caballé, S. (2015), *Towards Adaptive Peer Assessment for MOOCs*. Proc. of the 10th Int. Conf. on P2P, Parallel, GRID, Cloud and Internet Computing (3PGCIC 2015), 64-69, IEEE.

Capuano, N., Caballé, S., Miguel, J. (2016), *Improving Peer Grading Reliability with Graph Mining Techniques*. Int. J. of Emerging Technologies in Learning, in press.

Capuano, N., Loia, V., Orciuoli, F. (2016), *A Fuzzy Group Decision Making Model for Ordinal Peer Assessment*. IEEE Trans. on Learning Technology, PrePrints, doi:10.1109/TLT.2016.2565476.

Carlson P. A., Berry F. C., (2003), *"Calibrated Peer Review™ and Assessing Learning Outcomes"*, in proc. of the 33rd Int. Conf. Frontiers in Education.

Douek, N., Pichat, M. (2003), *From oral to written texts in grade I and the approach to mathematical argumentation*. Proceedings of PME XXVII, 341-348.

Forman, E., Joerns, J., Stein, M., Brown, C. (1998), *You're going to want to find out which and prove it. Collective argumentation in a mathematics classroom.* Learning and Instruction, 8(6), 527-548.

Gibson, P. A., Dunning, P. T. (2012), *Creating quality online course design through a peer-reviewed assessment*. Journal of Public Affairs Education, 18(1), 209-228.

Glance, D., Forsey, M., Riley, M. (2013), *The pedagogical foundation of massive online courses*, First Monday, 18 (5).

Raman K., Joachims, T. (2014), *Methods for ordinal peer grading*, Proc. of the 20th SIGKDD Int. Conf. on Knowledge Discovery and Data Mining.

Seery, N., Canty, D., & Phelan, P. (2012), *The validity and value of peer assessment using adaptive comparative judgment in design driven practical education*. International Journal of Technology & Design Education, 22(2), 205-226.

Thomas, G., Martin, D., & Pleasants, K. (2011), *Using self- and peer-assessment to enhance students' future-learning in higher education*. Journal of University Teaching & Learning Practice, 8(1), 1-17.

Sadler, P. M., Good, E. (2006), *The Impact of Self- and Peer-Grading on Student Learning*. Educational Assessment 11(1), 1-31.

Walsh, T. (2014), *The PeerRank Method for Peer Assessment*. Proc. of the 21st European Conf. on Artificial Intelligence.

William, D. (2007), *Keeping learning on track*. In F. Lester Jr. (Ed.), 2nd handbook of research on mathematics teaching and learning, 1053-1098, Information Age Publishing.