Participatory Evaluation and Grading: Experiencing with Teachers

Carlo Giovannella

[1] ISIM_Garage, Dept. of History, Cultural Heritage, Education and Society
University of Rome Tor Vergata, Rome, Italy
carlo.giovannella@uniroma2.it

(submitted on May 2015, last revised version accepted on November 2016)

Abstract. In this paper we describe: a) the *participatory evaluation and grading* (PG/E) method - principles and algorithm - that has been implemented within the on-line learning environment LIFE in two, untrusted (UPG/E) and trusted (TPG/E), versions; b) the outcomes of its use as part of the activities carried on to retrain K12 and high school teachers. The analysis of the outcomes of this case study - preliminary to the transfer of the method into the schools - evidenced a high level of appreciation and acceptance, as well as some indications for further improvements.

Keywords: peer grading, peer evaluation, trust, LIFE, teacher training

1 Introduction

Peer reviewing - i.e. the evaluation of a work operated by people owning similar competences (peers) of the author/s, often called also peer evaluation - is considered by the members of scientific communities as one of the engines of the continuous growth of our knowledge and, in fact, when sustained by an ethical attitude (mainly of the peers), it leads to a rapid improvement of ideas and an opening up of new perspectives. It can be considered a social practice, self-sustained by the contribution of all members of a given community, but unfortunately it is not fully free from malpractices These latter are often induced by the desire to generate control groups, achieve prominent positions in the society and/or by the stiff competition to access grants. Luckily nowadays one has the possibility to detect such malpractices, as well as the plagiarism, in a quite short time with a detection probability that increases with the extension of all dimensions involved in the process - number of people, geographical and time extensions, etc. - and with the continuously increasing power of the searching engines available on internet.

Despite of the problems mentioned above, from a pedagogical perspective peer evaluation - possibly including also a peer grading (PG/E: *peer evaluation and grading*) - is believed capable to strengthen the educational process and provide advantages that are by far much larger than its cons (see for example ref. [38] and references therein). Because of this, since more than two decades peer evaluation is experiencing a growing interest among educators and researchers [1-7, 32-35 and references therein]. Indeed *peer evaluation and grading* is believed by the greatest part of the scholars to foster the acquisition of:

a) an important amount of LIFE skills [41] considered very relevant for the XXI century education; LIFE skills can be grouped, as shown in [8], in *individual*, *social* and *management* skills and include critical analysis and understanding, self confidence, autonomy, self-motivation, ethics and respect, reliability, performance monitoring, ability to judge, etc.;

b) the ability to evaluate and self-evaluate her/his own learning path (also useful for lifelong learning [31]) and self-regulation [9].

PG/E is also believed capable to augment the individual level of employability [30] and to improve learning performances [26-28, 36-37].

Moreover, in the case of very large classes - e.g. crowded university courses and MOOCs - the application of social reviewing practices (like peer reviewing) implying the sharing of the efforts among all actors of the learning process, is believed to represent a possible, if not the only, way to assure scalability of learning processes and of the associated evaluation activities with the increasing number of the students [10]. Actually, the reduction of the tutors/teachers efforts will be achievable only if peer reports would achieve a reasonable quality level, would be trustable and of course if, at the same time, cheating is contrasted by effective actions.

These are the reasons why many scholars have focused their efforts in producing systems and algorithms able to determine the reliability of reports and reviewers and to assign the most suitable peers [1, 11-14], e.g. by studying variance, correlation between teacher and students grades, by using fuzzy logic, etc. However, although the intelligence of the algorithms used is increasing with the time, it is still quite difficult to assure the trustability of the reports [15], especially when peers have to evaluate open essays or when they have to monitor and evaluate complex learning activities and processes. In fact, less deterministic are the outcomes expected from a learning activity, more blurred its focus and larger the dispersion of the grades (even when a detailed rubric is provided [16, 17]) so that, in some cases, the number of tests required to achieve a reasonable level of trustability may increase well above any manageable figure. Most of the educational settings, thus, because of the limited number of students, are not suitable for the application of automatic approaches to assure a *trustable peer evaluation and grading*.

Peer evaluation and grading, however, are pedagogically relevant for all learning settings and processes and promise to contribute to improve the quality of the evaluation and feedback [29] in any kind of learning setting. There exist, thus, good reasons to undertake an effort to design and implement trustable peer reviewing processes applicable also in cases of small and medium size classes, open essays and complex activities, i.e. the elements characterizing the learning setting considered in this work. Because of this, few years ago, we started a research program that, instead of focusing on the development of algorithms capable to determine the level of trustability of a report, aims at implementing a different approach capable to make peer evaluation and grading usable in any learning setting. We called its untrusted form *participatory evaluation and grading* (UPG/E) [18,19], while when the trust level of peers is taken into account we called it *trusted participatory evaluation and grading* (TPG/E) [21]. It is worthwhile to stress that both differ from the so called *participatory examinations* [20].

In the following, first we describe briefly UPG/E and TPG/E algorithms and, as well, the process to carry on UPG/E and TPG/E within the on-line learning environment LIFE [22]. Then we report on the use of this method as part of the activities realized to re-train K12 and high school teachers.

2 UPG/E and TPG/E: principles and algorithms

Since the pedagogical background that inspired the development of UPG/E and TPG/E approaches has been already extensively described in [18,21] in the following of this sections, we provide the reader with a short summary of the main cues involved in their development and describe in details the algorithms that have been implemented.

First of all it is important to stress that the UPG/E, unlike the peer evaluation and grading (PG/E), considers also the grade given by the teacher/tutor’s, Tg. This latter is taken as a starting reference against which one can measure the ability of the students to evaluate themselves and their peers. Of course we are fully aware about the impossibility to define an absolute reference but the participatory nature of the procedure is expected to mitigate the influence of any starting reference on the final grade.

Concretely the approach requires to work out the mean of the peers’ grades and use it to correct the teacher/tutor’s grade. In practice, one has to take the difference between the mean of the peer grades and the teacher/tutor grade and add it to this latter, after a weighting procedure, where the weight - w1 or w1’ - allows to take into account the sign of the difference: w1 if positive and w1’ if negative. The weight w1 and w1’ may coincide or be different, usually w1’ is lower than w1 to realize a procedure more favorable to the students. Summarizing:

$$[\sum_n |(Tg-Spg_n)|/n]*w1(\text{or } w1')$$

where Spg_n is the grade assigned by the nth peers and n the number of peers that evaluated the work.

Of course the teacher's grades are not known by students, who are only aware of the mechanism used to determine the final grade.

As stated in par. 1 the adoption of a peer evaluation should support the acquisition of a set of skills that usually are not owned by students. Because of this, in fact, it is not unusual to come across students' evaluation and grading more or less influenced by sympathies and antipathies toward their peers, or compiled in a superficial manner due to the hurry to close the task. Such deprecable behaviors require the elaboration of a strategy to contrast them and, thus, we have introduced the following factor

$$\pm \{[\sum_{n'}|[(Tg_{n'})-(Sg_{n'})|]/[(\sum_{i}\sum_{n'}|(Tg_{i\,n'})-(Sg_{i\,n'})|/i]\}*w2(\text{or } w2').$$

It allows to work out the sum of the absolute value of the difference between the grades assigned by a given student to their peers, $Sg_{n'}$, and those given by teacher/tutor to the same peers, $Tg_{n'}$. This difference is then divided by the mean of the absolute value of all distances among the grades assigned by all i students to their peers and those given by teacher/tutor to the same peers. When the ratio is higher that a given threshold, Th, the sign minus (-) is taken to produce a negative feedback. The overall result is that a negative quantity, weighted with a weight w2, is subtracted to the student grade.

On the other hand, to make the procedure more appealing for the students, also a mechanism to reward the best performances has been provided: when the ratio is lower than Th, the sign plus (+) is taken to produce a positive quantity to be added to the student grade, after weighting with a weight w2'. The weight w2 and w2' may coincide or be different, usually w2' is greater than w2.

In the first version of the UPG/E [18,19] Th was taken as a portion of the standard deviations of the distribution of the distances among the grades assigned by all i students to their peers and those given by teacher/tutor to the same peers. A choice that was justified by the fact that usually a data set composed by independent items is expected to distribute normally around its mean value. Later, however, we realized that such choice may cause problems in presence of systematic errors (e.g. when students might all agree to assign to their peers grades that do not differ very much among them but that are much higher than those assigned by the teacher/tutor). As a consequence we decided to modify, as described above, the procedure and introduce the threshold value Th [21].

Finally, a third factor, was introduced to account for the ability of the students to self-evaluate her/his own work. The logic adopted to define this factor is the same used in the case of the second factor.

$$\pm \{[|Tg-Sg_{self}|]/[(\sum_{i}|(Tg_{i})-(Sg_{self\,i})|/i]\}*w3(\text{or } w3')$$

The only differences are that: a) the numerator considers only the absolute value of the difference between the grade assigned by a given student to its own work, Sg_{self}, and that given by the teacher/tutor, Tg; b) the sum in the denominator runs only on the number of the student i.

As for the second factor we have to define a threshold, Th', that determine the sign taken by the expression and, as well, two weighting factors that, again, may coincide or be different, w3 and w3'. Usually w3' is greater than w3.
The combination of these three factors produces a very robust algorithm that in principle, as shown in previous studies, allows also to drop the anonymity required by the double blind reviewing process without affecting the overall results [18,19]. The final grade, G, is obtained by modifying Tg, the teacher/tutor grade, according to the following formula:

$$\begin{aligned} G &= Tg + [\textstyle\sum_n|(Tg-Spg_n)|/n]*w1(\text{or } w1') \\ &\pm \{[\textstyle\sum_{n'}|[(Tg_{n'})-(Sg_{n'})|]/[(\textstyle\sum_i \textstyle\sum_{n'}|(Tg_{i\,n'})-(Sg_{i\,n'})|/i]\}*w2(\text{or } w2') \\ &\pm \{[|Tg-Sg_{self}|]/[(\textstyle\sum_i|(Tg_i)-(Sg_{self\,i})|/i]\}*w3(\text{or } w3') \end{aligned} \tag{1}$$

There are no universal guidelines to define weights and thresholds, however, as already reported in previous works [21], as rule of thumbs, we usually set weights and thresholds in such a way that the mean gain for students (i.e. the average increase of G with respect to Tg) is slightly greater than 10% (14% in this case study, see table 2). This because you have to provide a win-win perspective to encourage students to take part and be seriously involved in the UPG/E (as shown also in [40]). Such approach, depending on the cohort, may leads to a maximum individual gain ranging between 20% and 25% and a maximum loose ranging between 5% and 20%. Looses, however occurs only for a very limited number of students (few unities). In our multi-annual use of the UPG/E (since a.y. 2009-2010): (i) w1 ranged from 0.2 to 0.3 (0.3 in the present case study) while w1' has been always set to 0.1; (ii) w2 and w2' have always been set to, respectively 0.5 and 0.125; (iii) w3 and w3' experienced the larger variability, depending on how much the students tend to overestimate their own work, and usually range respectively between 0.5 and 0.1 (0.2 in the present case study) and 0.25 and 0.05 (0.1 in the present case study) with their ratio maintained always around 2; (iv) the value of the thresholds Th e Th' have been always taken, respectively, as the value of the ratio of $[(\sum_i \sum_{n'}|(Tg_{i\,n'})-(Sg_{i\,n'})|/i]$ and $[(\sum_i|(Tg_i)-(Sg_{self\,i})|/i]$ (see(1)) minus a bonus that ranged between 1 and 0.5 (0.5 in the present case study); (v) in addition, sometimes, we have also introduced a penalty to discourage the failure to deliver the peer evaluations.

Despite the robustness of the algorithm, we detected a lack of trust by the students in their peers and, as well, of confidence in themselves. Such feelings have been observed also in [10, 23, 39] and are not unexpected. In fact much the same as in the practices of scientific peer review, it may happen also to students not to trust in their peers or feel themselves inadequate to review a given work. However, differently from the scientific peer review procedures where you can kindly reject the assignment if you do not feel appropriate, in the *participatory evaluation and grading* students cannot refuse their assignments and, thus, are forced to face with their own level of

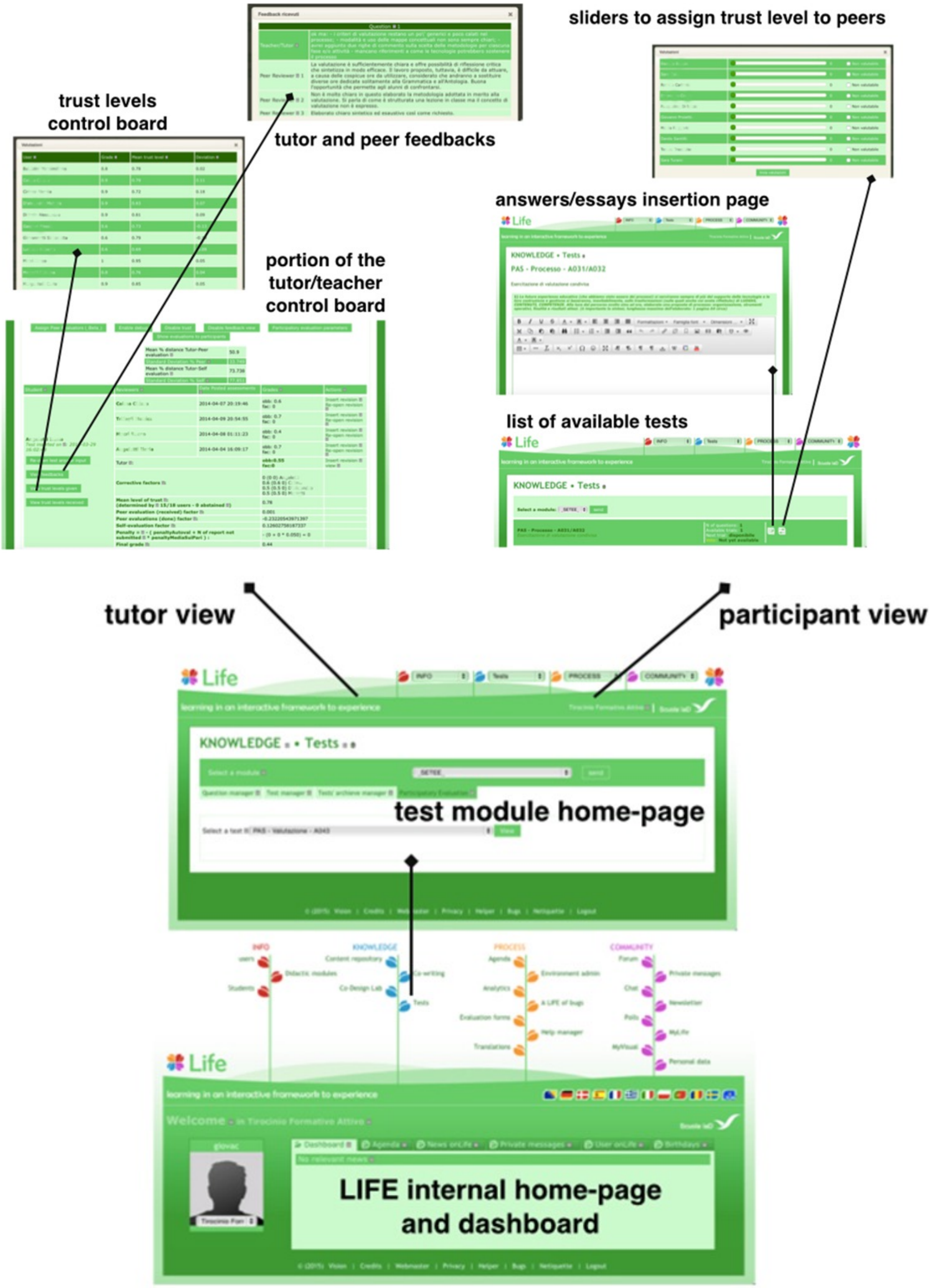

Fig. 1. Screenshots taken from LIFE to show how UPG/E and TPG/E procedures have been implemented in the on-line learning environment (for sake of simplicity only the most relevant functionalities have been shown in the figure).

understanding of the subject and with the responsibility to contribute to the determination of the grade of their peers and, as well, of their own grade. A sense of inadequacy can possibly explain the tendency of the students to be stricter with their best peers and more generous with the weakest ones, as shown by the comparison with teacher/tutor grades [25].

Due to these observations we have developed a "trusted" version of the "participatory evaluation and grading", TPG/E where students are asked to rate, anonymously, the level of trust they have in each of their peers, using a value ranging from 0 to 1. Students are free not to rate those peers that they do not think to know well enough. All values are then averaged to produce the mean trust level that the community of students have in each one of its member, Trust_in. These values, then, are used to weight the grades assigned by the peers within the first corrective factor of formula (1):

$$[\sum_n|(Tg-Spg_{n*}Trust\_in)|/n]*w1(\text{or } w1')$$

As a consequence the modified algorithm to be used in a TPG/E is the following:

$$\begin{aligned} &G_T = Tg + [\sum_n|(Tg-Spg_{n*}Trust\_in)|/n]*w1(\text{or } w1') \\ &\pm \{[\sum_{n'}|[(Tg_{n'})-(Sg_{n'})|]/[(\sum_i \sum_{n'}|(Tg_{i\,n'})-(Sg_{i\,n'})|/i]\}*w2(\text{or } w2') \\ &\pm \{[|Tg-Sg_{self}|]/[(\sum_i|(Tg_i)-(Sg_{self\,i})|/i]\}*w3(\text{or } w3') \end{aligned} \tag{2}$$

Actually if the learning process contemplates more than one TPG/E run, then, one could improve further the algorithm by including a corrective mechanism that takes into account the individual evaluation and self-evaluation performances and, accordingly, redetermine the initial individual level of trust, to obtain a more accurate value of Trust_in.

The overall effect of using UPG/E or TPG/E algorithms and approaches was quite positive as demonstrated, for example, by the improved students' ability to evaluate and self-evaluate their works. In fact we observed a progressive decrease of the mean distance between students' grades and teacher/tutor grades: in the best case, during a course of Physics [18], we observed an initial mean distance of 64% and a final mean distance of 5% that was achieved in three UPG/E runs. Readers interested in further details and analysis are referred to [18,19].

3 Carrying on UPG/E and TPG/E in LIFE

The algorithms and the process have been fully implemented within the on-line learning environment LIFE [22], as part of its test module. LIFE, in fact, is a modular, multilingual, community-based learning environment that has been designed to support design based processes and that implements a certain number of advanced analytics [18]. It has been realized in PHP, uses a MySQL database and is available as open service, upon request.

To start an UPG/E or TPG/E you are requested, as first step, to store in the archive the questions out of which you want to compose your test. Then, you have to define all test parameters - e.g. the number of the peer reviewers, time window allocated to the process, etc. - and finally you can start the process and enable the students you selected to take part in the participatory evaluation and grading process.

Once enabled, students can answer the questions and attribute anonymously the level of trust to their peers (TPG/E). A control board, see fig. 1, allows the teacher/tutor to follow the process and check for the assignment of the trust rating, in order to solicit the laggards. In case of mistakes occurred during the input the rating process can be re-opened either upon student request or because of the detection of an abnormal situation. As a variant of the process the rating of the trust level can be disabled (UPG/E).

Once that the deadline to insert the answers has expired the teacher/tutor is enabled to assign randomly the peer evaluators. If needed, a manually refinement of the assignments is allowed. When the peers assignment has been confirmed (a check box should be checked) all, teacher/tutor and students, are allowed to insert their written feedbacks and grades (evaluation of peers' work and self-evaluation). Again, in case of input mistakes, much the same as in the case of the trust rating, the insertion of feedbacks and grades can be re-opened.

Finally the teacher/tutor has to insert the values of all weights and parameters required by the algorithm described in the previous section. The final grade, then, is automatically calculated by the LIFE module and displayed in the control board, see fig.1. When the TPG/E (or UPG/E) has ended the teacher/tutor can enable students to visualize both grades and/or textual feedbacks, given by either the teacher/tutor and peers. No rebuttal procedure is allowed at present.

4 Peer evaluation and grading in action as part of teachers training

4.1 Training frame of reference

We are deeply convinced that the acquisition of an adequate evaluation literacy, and as well of LIFE skills (see par. 1) should start at K12 school and progressively develop along high school and university attendance. Accordingly we have decided to include the *peer evaluation and grading* within the set of activities proposed by the courses organized at the University on behalf of the MIUR (Ministry of Education and Research) to retrain K12 and high school teachers having at least three years of teaching experience. Unfortunately such course suffered from a quite critical organization and required to the attendants a concentrated effort in parallel to standard school and family duties. Because of this we decided to propose *participatory evaluation and grading* as a voluntary activity to be carried on almost at the end of the training process. Although such training process will be fully described in a forthcoming paper we wish to summarize briefly here below the sequence of activities that have been proposed to the participants. They were part of the module focused on the design of technology enhanced educational processes, aimed at fostering also the acquisition of an adequate design literacy: a) participation in a preliminary surveys based on a set questionnaires; b) 8 hours of introductory face-to-face lectures; c) 5 collaborative (between 15 and 25 participants per group) web explorations and brainstormings focused on the transformation that each component of a learning ecosystem - spaces, contents, competences, processes, monitoring and evaluation - is undergoing due to the irrepressible technological development of the last 10-15 years; d) elaboration of SWOT analysis and matrices to resume the outcomes of the 5 brainstormings; e) intergroup brainstormings on the outcomes of the groups' brainstormings; f) elaboration of a short essay (see below); g) voluntary activity: *participatory evaluation and grading* of the essays; h) voluntary activity, design of a participatory evaluation and grading process to be carried on with the students at school; i) participation in a final survey based on a questionnaires; l) debriefing about the training experience. Usually, as part of the training processes we design and deliver, we use to involve participants also in tasks aimed at stimulating their creativity (e.g. between task (e) and task (f)) but this was not the case due to the shortage of time.

The proposal of *participatory evaluation and grading* as a voluntary activity allowed also to filter out K12 and high school teachers who were not motivated enough, i.e. teachers who were attending the retraining course only to obtain the needed certification. As a results only 12% of the teachers attending the retraining course decided to experience also *participatory evaluation and grading* activity, that is 56 teachers belonging mainly, but not exclusively to the so called classes A043 (Italian, Geography and History for K12 schools) A050 (Italian for high schools), A061 (Art History for high schools). Overall the percentage of voluntary participation is not so unexpected since it correspond to that of innovators that one may encounter in the Italian schools. Somewhat surprising is the absence of teachers belonging to the scientific classes, but along the years we experienced their quite strong conservative attitude and resistance against innovation and usage of on-line learning technologies. For example, in our experience, Math teachers tend to act against the penetration of collaborative approaches and new technologies into the learning processes, with the exception of the geogebra software.

4.2 The participatory evaluation and grading process

To take part in the *participatory evaluation and grading* activity, the participants have been requested to produce, in one week, a short essay, no longer than two A4 pages. The participants had the possibility to choose the theme of the essay among the following two: a) a technology enhanced learning process: organization, supporting tools, objectives and expected results ("learning process" in the Group column of tables 1 and 2); b) a technology enhanced evaluation process: identification of a competence and description of the monitoring process aimed at assessing and self-assessing its acquisition ("evaluation" in the Group column of tables 1 and 2). The participants were asked to take into account and make use of the outcomes of the brainstormings. Following their choices, participants were grouped in 4 homogeneous groups (composed by teachers belonging to the same teaching class: overall 40 participants) and 2 heterogeneous groups (composed by teachers belonging to different teaching classes: overall 16 participants). The number of the group members ranged from 6 to 16. The 4 homogeneous groups took part in a TPG/E while the other 2 groups, due to the lack of mutual acquaintance, took part in a simple UPG/E. Once that all short essays were uploaded the participants had 5 days to produce their double-blinded reports: 3 peer evaluations and grading and one self-evaluation and grading. No one of the participants had previous experience with the proposed evaluation and grading method and only 5% of the participants had experienced previously peer-assessment with their students, although in quite rough manner.

At the end of the evaluation process results have been disclosed and, as written above, a questionnaire has been distributed and a final debriefing organized.

After the *participatory evaluation and grading* experience, always on voluntary basis, teachers have been asked to work in small group to design a peer-grading/evaluation process to be delivered in their classrooms. This aspect of the training process, however goes beyond the purpose of this work and will be discussed in a forthcoming paper.

4.3 Outcomes: trust level and grading

As described above, the members of the 4 homogeneous groups were required to rate the trust level they had in their peers. Means, standard deviations and range of the rated values, reported in Tab. 1, are fully in line with those reported in [21] except for group A061.

Because of this, at first glance, one may induced to think that the participants have reached a considerable degree of confidence with their peers, equivalent to that of students attending the third year of a bachelor course (students at the first year tend to rate the level of trust in their peers with a considerably lower value). In turns it may also means that the overall learning process was quite successful because it succeeded in creating very cohesive communities. Even too much, if one considers the figures characterizing group A061. This latter, in fact, is a group within which each member seems to trust fully in all the others. Actually it represents a borderline case, we have never faced before, where the effect of the trust factor is overwhelmed by that of the group cohesion, so that all reports assume the same relevance (no effect of the weighting procedure).

Tab. 1. Mean Trust_in values (standard deviation in brackets) calculated by averaging Trust_in rates attributed by all participants belonging to the same groups (first column: in brackets the number of the participants). Last column: range of variability of the rated Trust_in level within each single group.

Group	Mean_Trust_in	Range
A050_Evaluation (6)	0,74 (0,09)	0,60 ÷ 0,85
A043_Evaluation (16)	0,78 (0,07)	0,70 ÷ 0,95
A043_LearningProcess (11)	0,71 (0,07)	0,60 ÷ 0,85
A061_LearningProcess (7)	0,95 (0,02)	0,93 ÷ 0,98

However, if we examine more in depth the data and try to correlate the mean values of the trust level characterizing each participant with her/him evaluation performance we get Fig. 2 that, diversely from what we have observed in [21], does not show any relevant meaningful correlation, R=0.12. This means that the each-other frequentation occurred during the retraining course - actually few months - was not long enough to achieve a sufficient mutual knowledge. As a consequence we have to conclude that the values of trust were possibly influenced by professional camaraderie. The only exception is represented by group "A043 Learning Process" for which the correlation between trust levels rated by the peer and the evaluation performances is quite evident, see Fig. 3: R=0.65.

As far as grades assigned by the participants to peer and to themselves (self evaluation) we observed, as usual for the first experience with *participatory evaluation and grading* [18, 19, 21], an over-evaluation of the peers (ranging from 25% to 51%) and an over self-evaluation (ranging from 19% to 73%) with respect to teacher/tutor evaluations (Tg). The worst performances were those of the two A043 groups with over-estimation percentages very similar to those of university students at their first experience with T *participatory evaluation and grading*. The group that performed at best were the mixed groups and A050 group.

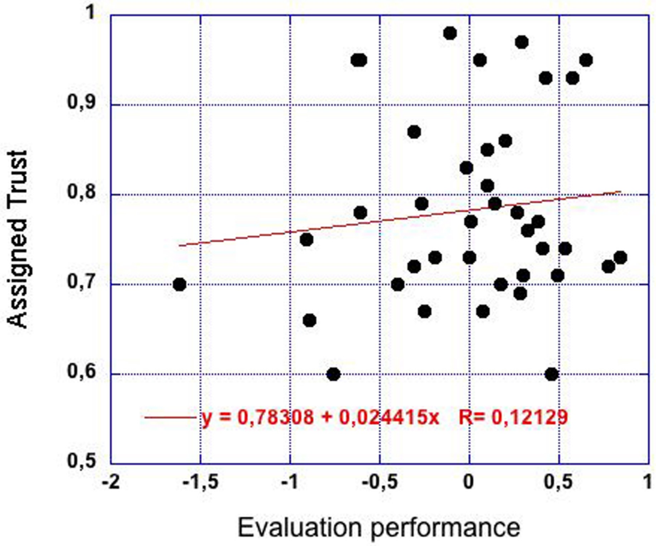

Fig. 2. Individual evaluation performance (measured as the normalized distance in the evaluation of peers and in the self evaluation with respect to the grades given by the tutor's, Tg, see formula 1 and 2) vs. mean trusted level rated by peers

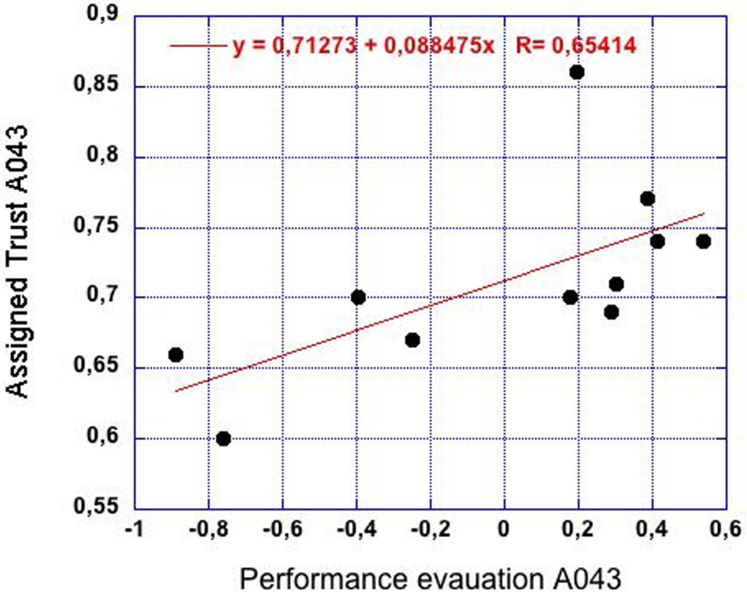

Fig. 3. Individual evaluation performance vs. mean trusted level attributed by peers. Only for group "A043 Learning Process"

As far as G, final grade worked out according to the peer evaluation and grading algorithm, we can observe that the mean grade does not change very much from group to group, with a range of variability equal to 13% of the full scale, and so do standard deviations that are very similar for all groups. Again not all groups shown identical behaviors: we observed, in fact, that the members of some groups tended systematically to grade the essays of their peers higher than tutor/teacher while for other groups the grade values were more balanced and distributed around those given by the tutor/teacher.

Tab. 2. Mean grade and standard deviation characterizing each group (in brackets the number of the participants), together with its internal grade of variability (Range). Last column: average increment of the grade (%) observed in each group with respect to the grade given by the tutor.

Group	Mean_grade	Range	Mean gain %
A050_Evaluation (6)	0,79 (0,30)	0,23 ÷ 1,09	3
A043_Evaluation (16)	0,66 (0,33)	0,0 ÷ 1,29	15
Mix_Evaluation (10)	0,73 (0,27)	0,34 ÷ 1,08	13
A043_LearningProcess (11)	0,68 (0,26)	0,24 ÷ 1,07	13
A061_LearningProcess (7)	0,78 (0,26)	0,39 ÷ 1,02	21
Mix_LearningProcess (6)	0,70 (0,30)	0,21 ÷ 1,00	17

In any case the evaluation performances of the participants show a strong correlation (R=0,58) with the test performances, see Fig. 4. This correlation is even stronger (R=0,85) when the grade given by the tutor/teacher are modified according to the outcomes of the peer evaluation and grading process, a result that shows how a participatory process can contribute to the improvement of the grading.

Fig. 4, thus, confirms a trend that was already observed by the author in a completely different learning setting [21]. This figure also shows that although some participants may get their grade decreased, the largest part of them, due to the win-win mechanism, gained a bonus that ranges between 3% and 21% of the tutor/teacher evaluation, see also Table 2.

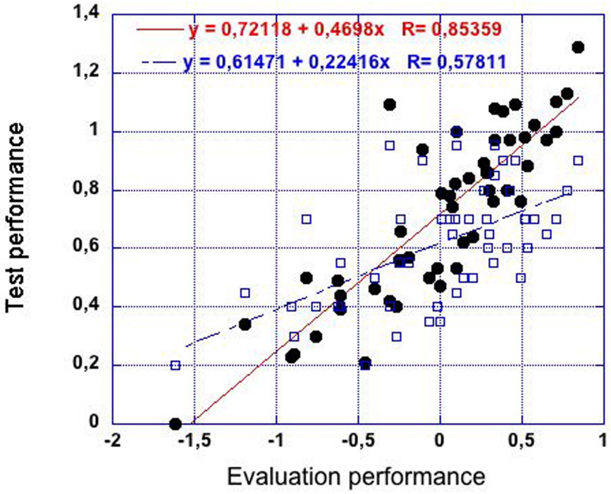

Fig. 4. Individual evaluation performance vs. test performance. Comparison between tutor / teacher grade (open square) and final grade, G, (black dot).

4.4 Questionnaire Outcomes

At the end of the *participatory evaluation and grading* activity and of the overall retraining process we have requested to the participants to fill anonymously a couple of on-line questionnaires.

For all questions related to the *participatory evaluation and grading*, see Table 3, we used a 1 to 5 scale with the exception of one question for which we used a 0% to 100% scale. For all questions, participants had the possibility to enrich the quantitative evaluation with a comment and, thus, to add considerations also on aspects not covered by the proposed questions.

Overall the participants have been very positive with TPG/E and UPG/E, much more than students [21] by at least 20%-30% of the full scale. They think that it helps to feel more involved in the training process and to acquire higher responsibility and ethical behavior (in particular with respect to peers). Slightly to a less extend they think that *participatory evaluation and grading* can also foster content deepening and meta-reflection. Much the same as the students, they think that TPG/E and UPG/E are demanding activities.

The participants of this case study - K12 and high school teachers - still think, after the experience, that tutor's evaluation and grading are more reliable than theirs (mainly due to a larger experience and a better vision of the overall didactic process) but to a less extend with respect to university students (around 10% less of the full scale). They think that their reports should be weighted at 60% ± 35% of the tutor's reports, i.e. that one should use a value of w1 egual to 0.6. Actually we used w1 = 0.3, a value that it is, anyway, within one standard deviation from teachers' expectation. In the case of students we got a weight expectation of about 35% ± 15%. while we used w1 = 0.2; a value that was again within one standard deviation from students' expectation.

Tab. 3. Outcomes of the questionnaire: mean values and standard deviations. A 5 levels scale has been proposed for all questions, with the exception of one question where a percentage scale, 0%-100%, has been used.

Question	Mean ± SD
How much have you enjoyed the participatory grading/evaluation (PG/E)?	4,1 ± 0,9
How much PG/E allowed you to self-assess the level of your preparation with respect to the content of the training course ?	3,5 ± 1,0
How much PG/E stimulated you to deepen the content of the training course (e.g. identify gap in understanding, weaknesses of your work and mitigate them) ?	3,8 ± 0,9
How much PG/E made your feel more involved in the training process ?	4,1 ± 0,9
How much PG/E helped you to acquire a higher level of care towards your training process ?	4,3 ± 0,9
How much PG/E helped you to acquire a higher level of care towards your peers ?	4,2 ± 0,9
How high is the effort required for PG/E ?	4,0 ± 0,9
In your opinion how much the teacher's grades/evaluations are more reliable than those formulated by your peers ?	3,4 ± 1,2
How much the use of the level of "trust" to weigh the grades given by your peers makes, in your opinion, PG/E more reliable?	3,3 ± 1,1
How much the grades you gave should be weighted with respect to those given by the teacher ?	60% ± 35%
In your opinion, how much an upgrade of the trust level given to your peers as a function of their evaluation performances will make TPG/E more reliable ?	3,1 ± 0,9
How much would help a rubric to give your grades ?	4,3 ± 0,9

Coherently the above outcomes teachers think, although to a less extend with respect to university students, that the introduction of the trust rating procedure and the redefinition of the trust level with the evaluation performance of the participants can improve the trustability of the overall process. On the other hand, like students, also teachers think that the availability of a detailed rubric would have been very useful to

express more reliable judgments in a shorter time. This is also the main concern that emerges from the comments that on all other aspects are, in general, very positive.
In addition, participants underlined the capability of *participatory evaluation and grading* to foster responsibility, meta-reflection and metacognition that are the basis of the self-regulation. Very appreciated was their involvement with the double role of evaluator and object of the evaluation. Also appreciated was the possibility to compare her/his own work with that of the peers and to receive comments capable to offer different perspectives on their own essay. Some of the participants, finally, declared their intention to apply *participatory evaluation and grading* in their classroom and one, in particular, wrote "*students will feel closer to the point of view of the teacher who often tends to be considered as a hostile entity judging from the highest of their authority*"

The final debriefing was not specifically dedicated to the *participatory evaluation and grading* but concerned the whole didactic process. Because of this not many additional elements specific to TPG/E or UPG/E emerged. The most interesting one was the request to enable school teachers/tutors to personalize the evaluation scale for special cases, like students affected by specific learning disabilities (i.e. dyslexia, dysgraphia, dyscalculia, attentional deficit, etc.) or having special educational needs (i.e. foreigners). Of course the request of this new functionality is less relevant for university or professional learning/training courses.

5 Conclusions

Despite the growing interest in peer reviewing, the diffusion of this evaluation approach into the schools requires first of all a dissemination action aimed at involving K12 and high school teachers and to favor the acquisition of an adequate evaluation literacy. The case study reported in this article demonstrates how participatory evaluation practices are likely to be largely appreciated by teachers when the proposed approach, thanks also to the fluidizing action of the technologies, is able to support more objective assessments and, at the same time, offers the possibility to verify the outcomes of the process.
This case study shows also that TPG/E or UPG/E allow to identify the peculiar behaviors and characteristics of each group of students and that, at the same time, are capable to strengthen and make more objective the grading process, thanks to a win-win approach. It also demonstrates how the participation of the tutor in the evaluation process is deemed essential, at least in an initial phase during which students must be accompanied in the acquisition of a more objective approach to the evaluation of peers.

The significant appreciation by the teachers, most of which seem very keen to transfer the method into the schools let us hope for a progressive diffusion of the TPG/E or UPG/E.

This article, thanks to the detailed description of the algorithms needed to implement the TPG/E or UPG/E and of the methodologies to apply them, can be considered a reference for all those wish to challenge themselves in participatory evaluation practices, possibly also to improve them.

6 References

1. N. Falchikov and J. Goldfinch, “Student peer assessment in higher education: a meta-analysis comparing peer and teacher marks”, Review of Educational Research, 70(3), 2000, pp. 287-322.
2. E.F. Gehringer, “Strategies and mechanisms for electronic peer review” in Frontiers in Education Conference (FIE 2000), Vol. 1., 2000, pp. F1B/2–F1B/7.
3. K.D. Strang, “Exploring summative peer assessment during a hybrid undergraduate supply chain course using Moodle”, in Proceeding of 30th ascilite Conference, 2013, pp. 840-853.
4. C. Bauer, K. Figl, M. Derntl, P.P. Beran, S. Kabicher, “The Student View on Online Peer Reviews” SIGCSE Bull., 41, 2009, pp. 26–30.
5. J. Pearce, R. Mulder, C. Baik, "Involving students in peer review. Case studies and practical strategies for university teaching", CSHE 2009, 2009 retrieved on 9 January 2014 http://www.cshe.unimelb.edu.au/
6. R. Beach, "Showing students how to access: demonstrating techniques for response”, in the writing conference", in Writing and response: Theory, practice, and research", pp. 127-148. Urbana, IL: National Council of Teachers of English, 1989.
7. K. Topping, "Peer assessment between students in colleges and universities", Review of Educational Research, 68(3), 1998, pp.249- 276.
8. C. Giovannella, V. Baraniello, “Smart City Learning”, IJDLDC, vol. 3(4), 2013, pp. 1-15
9. B.J. Zimmerman, “Self-regulated learning and academic achievement: An overview” Educational Psychologist, 25, 1990, pp. 3-17.
10. C. Piech, J. Huang, Z. Chen, C.Do, A. Ng, D.Koller, “Tuned Models of Peer Assessment in MOOCs”, in Proceedings of the 6th International Conference on Educational Data Mining, Memphis, Tennessee, 2013, retrived on 9 january 2014 http://www.stanford.edu/~cpiech/bio/papers/tuningPeerGrading.pdf
11. J. F. Hair, W. C. Black, B. J. Babin, R. E. Anderson, R. L. Tatham, Multivariate Data Analysis. Upper Saddle River, NJ: Prentice-Hall, 2006.
12. A. Zhang and P. Blakey, “Peer assessment of soft skills and hard skills”. Journal of Information Technology Education, 11, 2012, pp. 155-168.
13. A. Dollisso and V. Koundinya, “An integrated framework for assessing oral presentations using peer, self, and instructor assessment strategies”. NACTA Journal, 55(4), 2011, pp. 39-44.
14. R.M. Crespo García, A. Pardo, C. Delgado Kloos, “An adaptive strategy for peer review”, in prceeding of Frontiers in Education, IEEE publisher, vol. 2, 2004, pp. F3F–7–13.
15. S. K. Green and R.L. Johnson (Eds.), Essential Characteristics of Assessment (Vol. 6): NY: Mcgraw-Hill, 2010 .
16. G. G. Bitter and J. M. Legacy, Using Technology in the Classroom, NY: Pearson, 2008.
17. A. Bayat and V. Naicker, “Towards a learner-centred approach: interactive online peer assessment”, South African Journal of Higher Education, 26(5), 2012, pp. 891-907.
18. C. Giovannella, S. Carcone, A. Camusi, “What and how to monitor complex educative experiences. Toward the definition of a general framework”, IxD&A Journal, N. 11-12, 2011, pp. 7-23
19. C. Giovannella, D. Camusi, “Participatory grading in a blended course on Multimodal Interface and Systems”, IxD&A Journal, N. 13-14, 2012, pp. 84-91
20. J. Shen, M. Bieber M. S,R. Hiltz S. R., "Participatory Examinations in Asynchronous Learning Networks: Longitudinal Evaluation Results", Journal of Asynchronous Learning Networks, No. 3,

2005, pp. 93-113.

21. C. Giovannella, F. Scaccia, "Technology-Enhanced 'Trusted' Participatory Grading", ICALT 2014, IEEE publisher, pp. 347-349
22. LIFE (Learning in an Interactive Framework to Experience) http://www.mifav.uniroma2.it/inevent/events/isim/index.php?s=131&a=165#
23. W. Cheng, M. Warren, "Having second thoughts: student perceptions before and after a peer assessment exercise", Studies in Higher Education, 22 (2), 1997, pp.223-239.
24. D. Sluijsmans, G. Moerkerke, F. Dochy, J. Van Merrienboer, "Peer assessment in problem based learning.", Studies in Educational Evaluation, 27 (2), 2001, pp,153-173.
25. P.M. Sadler, E. Good, "The Impact of Self-and Peer-Grading of Student Learning", Educational Assessement, 11(1), 2006, pp. 1–31.
26. N.M. Trautmann, "Interactive learning through web-mediated peer review of student science reports", Educ. Technol. Res. Dev., 57, 2009, pp. 685–704.
27. N.J. Pelaez, "Problem-based writing with peer review improves academic performance in physiology", Adv. Physiol. Educ., 26, 2002, pp. 174–184.
28. K. D. Strang, "Measuring self-regulated e-feedback, study approach and academic outcome of multicultural university students", International Journal of Continuing Engineering Education and Life-Long Learning, 20(2), 2010, pp. 239-255.
29. A. Walker, "British psychology students' perceptions of group work and peer assessment", Psychology Learning and Teaching, 1(1), 2001, pp. 28-36.
30. D. Boud & N. Falchikov, "Aligning assessment with long-term learning. Assessment and Evaluation in Higher Education", 31(4), 2006, pp. 399-413.
31. D. Boud, Enhancing learning through self-assessment. London: Kogan Page, 1995.
32. K.J. Topping, "Peer assessment between students in colleges and universities", Review of Educational Research 68, 1998, pp. 249-276.
33. I. van den Berg, W. Admiraal , A. Pilot, "Design principles and outcomes of per assessment in higher education", Studies in Higher education, 31, 2006, pp. 341-356.
34. D. Boud & N. Falchikov, Rethinking assessment in higher education: Learning for the longer term. Abingdon: Routledge, 2007.
35. F. Dochy, M. Segers, D. Sluijsmans, "The Use of Self-, Peer and Co-assessment Higher Education: a review", Studies in Higher Education Volume 24, No. 3, 1999, pp. 331-350.
36. H. Andrade & Y. Du, "Student responses to criteria-referenced self-Assessment", Assessment and Evaluation in Higher Education, 32(2), 2007, pp. 159-181.
37. C. Rust, B. O'Donovan, M. Price, "A social constructivist assessment process model: how the research literature shows us this could be best practice", Assessment and Evaluation in Higher Education, 30(3), 2006, pp. 233-241.
38. J. W. Strijbos, D. Sluijsmans, "Unravelling peer assessment: Methodological, functional, and conceptual developments", in Learning and Instruction, Elsevier, 20, 2010, pp.265-269.
39. P. Mupa, O. Chabaya, C. Chiome, R.A. Chabaya, "Peer Assessment in Higher Education: The Roadmap for Developing Employability Skills in Potential Job Seekers", International J. Educational & Research, 1(2), 2013, PP. 62-69.
40. D. Boud, R. Cohen, J. Sampson, "Peer learning and assessment. Assessment and evaluation in higher education", 24(4), 1999, pp. 413-426.
41. http://disco-tools.eu/disco2_portal/terms.php & http://en.wikipedia.org/wiki/Life_skills