Application of Fuzzy Ordinal Peer Assessment in Formative Evaluation

Nicola Capuano[1(✉)] and Francesco Orciuoli[2]

[1] Department of Information Engineering, Electric Engineering and Applied Mathematics, University of Salerno, Via Giovanni Paolo II 132, 84084 Fisciano, SA, Italy
ncapuano@unisa.it

[2] Department of Business Sciences, Management and Innovation Systems, University of Salerno, Via Giovanni Paolo II 132, 84084 Fisciano, SA, Italy
forciuoli@unisa.it

Abstract. Peer assessment has been used for many years as a tool to improve learning outcomes but, only recently, it is becoming an increasingly used support also in students evaluation. Many approaches have been proposed so far to make peer assessment as reliable as possible even in case of incorrect or inaccurate evaluations proposed by students. Among these approaches, Fuzzy Ordinal Peer Assessment (FOPA) relies on ordinal evaluations (rather than cardinal ones) and on the application of models coming from Fuzzy Set Theory and Group Decision Making. FOPA has already demonstrated good results in in-silico experiments. To complement these results, in the work presented in this paper, we experiment the same model in a University context to support formative evaluation. Obtained results show better performance of FOPA with respect to competitor models and a general attitude of peer assessment models to approximate instructor ratings.

1 Introduction

Formative evaluation 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* [1]. An important function of formative evaluation is providing students with continuous feedback, meaning that opportunities for feedback should occur continuously, but not intrusively, as a part of instruction [2].

A feasible way to approach formative evaluation either in classroom and within on-line learning environments is *peer assessment*. In peer assessment, students are required to grade a small number of their peers' assignments as part of their own assignment. The final grade of each student is then obtained by combining information provided by peers [3]. Peer assessment is able to economize teachers' time: an entire classroom can be graded in the time that it would take a teacher to grade just few submissions. Moreover, rather than having a teacher rush through each submission, students are able to take their time to correct just a small subset of them [4].

F. Xhafa et al. (eds.), *Advances on P2P, Parallel, Grid, Cloud and Internet Computing*, Lecture Notes on Data Engineering and Communications Technologies 13,
https://doi.org/10.1007/978-3-319-69835-9_52

The literature reports on many learning benefits connected to peer assessment like the exposure of students to different approaches, the development of their self-learning abilities, the enhancement of critical thinking, etc. [5]. This approach is also capable of easily scaling to any number of students (even in massive contexts like MOOCs) given that the number of assessors naturally grows with the number of students [6].

On the other hand, even if some studies suggest a good correlation between the results of peer assessment and instructor ratings in conventional classrooms and online courses (at least for specific, high structured domains), there is still a general concern on its use as a reliable strategy to approximate instructor marking, even in formative evaluation exercises [7].

To mitigate this issue, several corrected methods have been identified so far. In [8] it has been demonstrated that, asking students to provide ordinal feedback (e.g. "the report x is better than the report y"), allows to obtain better results with respect to asking them to provide cardinal feedback (e.g. "the grade of report x is a B"). Ordinal feedback is easier to provide, more reliable and overcomes the so called *bias problem* occurring when students grade peers on different scales.

Elaborating on these assumptions, in [9] a new ordinal peer assessment model named FOPA, based on *Theory of Fuzzy Sets* and *Group Decision Making*, has been defined. Experimental results with synthetic data (analyzed in the same work) have shown better performance of FOPA in the estimation of students' grades with respect to other peer assessment models. To substantiate this preliminary results, in this paper we experiment the same model in a real University context to support a formative evaluation exercise with classroom students.

The paper is organized as follows: the next section presents related work on peer assessment, Sect. 3 summarizes the defined FOPA model, while Sect. 4 illustrates the experimental setting and discusses the obtained results. Eventually, conclusions are summarized in Sect. 5.

2 Related Work

The main issue of peer assessment, when used as a tool to support formative evaluation, is represented by the lack of accuracy of grades proposed by students that may result in an erroneous feedback. Several approaches have been proposed so far to make peer assessment more reliable. *Calibrated Peer Review* (CPR) [10], proposes a calibration step to be performed by students before starting to assess other students' assignments. During the calibration step, each student rates a set of assignments that have been already rated by the instructor. The discrepancy between students' and instructors' grades is used to weight subsequent assessments.

In [11], three probabilistic models for tuning peer-provided grades are presented. Such models estimate the *reliability* of each assessor as well as her *bias* (i.e., a score reflecting the assessor's tendency to inflate or deflate her grade) based on the analysis of grading performance on special "ground truth" submissions that are evaluated either by the instructor or by a big number of peers. The estimated reliability and bias rate of each student are then used to tune proposed grades. A similar approach has been applied in [12], where a *Bayesian* model is used to calculate the bias of each peer

assessor in general, on each item of an assessment rubric and as a function of the assessor grade assigned by the instructor.

In [13], the ability of an assessor student to correctly rate peer students is assumed to be dependent on the grade obtained by the same student. In other words, final grades to be assigned to students are obtained by weighting the grades proposed by their assessors on the basis of the grades received by the assessors themselves. Given that students' grades recursively depend on other students' grades, an iterative algorithm, named *PeerRank* is proposed for their calculation. In [14], the same model has been improved and applied in formative evaluation within a University course.

In [8] the authors have defined several probabilistic models for obtaining student grades starting from *ordinal feedback* provided by the peers rather than from cardinal one. An experiment with real students have demonstrated that the performance of such models is at least competitive with cardinal models for grade estimation, even though it requires less information from the graders. In [15], the authors have shown also that ordinal peer assessment is highly effective and scalable for student evaluation. They have defined a model for distributing the assignments among peers so that the collected individual rankings can be merged into a global one that is as close as possible to the real ranking.

3 The Defined Model

In a typical peer assessment scenario an *assignment* is given to n different students from a set $S = \{s_1, \ldots, s_n\}$. Each student elaborates her own solution (e.g. an essay, a set of answers to open-ended questions, etc.) generating a *submission*. Each student has then to evaluate m submissions (with $m \leq n$) coming from other students. The assignment of submissions to assessor students is performed in accordance to an *assessment grid*: a Boolean $n \times n$ matrix $A = (a_{ij})$ where $a_{ij} = 1$ if the student s_j is asked to grade the submission of s_i and $a_{ij} = 0$ otherwise.

According to [3], a feasible way to build an assessment grid is starting with an $n \times n$ null matrix and initializing its elements basing on the following equation:

$$a_{\mathrm{mod}(i+j-1,n)+1,i} = 1 \forall i \in \{1, \ldots, n\}, j \in \{1, \ldots, m\} \quad (1)$$

where *mod* indicates the remainder after division of the first term by the second one. The obtained matrix is then shuffled in several iterations by randomly selecting a couple of rows (or columns) $i, j \in \{1, \ldots, n\}$ such that $a_{ij} = a_{ji} = 0$ and swapping them.

After having defined the assessment grid, each student is asked to rank submissions assigned to her. The partial ranks provided by each student are then used to build the overall ranking of submissions and to grade them accordingly. In the next subsections we analyze how this process is performed within a standard ordinal peer assessment model as well as within FOPA.

3.1 Ordinal Peer Assessment

In ordinal peer assessment, each student s_j is asked to define an *ordinal ranking* $\succ_j$ on the subset $S_j = \{s_i \in S | a_{ij} = 1\}$ of her assessees. Being s_k^j the generic element of S_j, with $k \in \{1, \ldots, m\}$, an ordinal ranking takes the following form:

$$s_{p(1)}^j \succ_j s_{p(2)}^j \succ_j \ldots \succ_j s_{p(m)}^j \tag{2}$$

where $p : \{1, \ldots, m\} \rightarrow \{1, \ldots, m\}$ is a permutation function. Equation (2) means that, according to s_j, the submission of the student $s_{p(1)}^j$ is better than that of $s_{p(2)}^j$, etc.

The ordinal ranking $\succ_j$ is undefined for elements not included in S_j so it is a *partial ranking* over S. The partial rankings defined by all students can be collected in a $n \times n$ *ranking matrix* $R = (r_{ij})$ whose generic element r_{ij} is the position of s_i in the ranking $\succ_j$ if $s_i \in S_j$, 0 otherwise. Starting from a ranking matrix, an *aggregation rule* is needed to compute a complete ranking over the whole set of submissions.

Several aggregation rules have been proposed so far by different researchers. A simple and effective rule is the classical *Borda count* [16] 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. Based on the assessment grid A and the ranking matrix R, the Borda score of any $s_i \in S$ can be calculated as:

$$Borda(s_i) = \sum_{j=1}^{n} a_{ij} \cdot \left(m - r_{ij} + 1\right). \tag{3}$$

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

In [15], authors have demonstrated that Borda outperforms other, more complex aggregation rules like *Random Serial Dictatorship* and *Markov chain* inspired models, especially in case of imperfect grading (i.e. when partial rankings are not consistent to the ground truth). In [8], the authors have defined other peer assessment approaches based on models that represent probabilistic distributions over rankings, obtained from the models of *Mallows* [17], *Bradley-Terry* [18] and *Plackett-Luce* [19]. Such models have demonstrated better performance with respect to Borda also in case of imperfect grading and are also capable of detecting meaningful cardinal grades.

3.2 Fuzzy Ordinal Peer Assessment (FOPA)

In [9], an alternative ordinal peer assessment model named *FOPA* is defined. Applying this model, each student $s_j \in S$ is asked to define a *fuzzy ranking* R_j over the subset S_j of her assessees. Such fuzzy ranking is defined as a sequence:

$$s_{p(1)}^j \sigma_1 s_{p(2)}^j \ldots s_{p(k-1)}^j \sigma_{k-1} s_{p(k)}^j \tag{4}$$

with $k \leq m$. Terms in odd positions in the sequence represent elements of S_j while $p : \{1, \ldots m\} \rightarrow \{1, \ldots k\}$ is a k-permutation function. Terms in even positions belong to the set of symbols $\{\gg, >, \geq, \approx\}$ and define a degree of preference between subsequent terms in the sequence (with $\gg$ meaning "is much better than", $>$ "is better than", $\geq$ "is a little better than" and $\approx$ "is similar to"). Each submission appears at most once in the ranking so cycles are not allowed although partial rankings are admitted.

For example, let suppose that the student s_1 has to evaluate the subset of students $S_1 = \{s_2, s_4, s_5, s_6\}$. By proving the ranking $R_1 = (s_4 \gg s_5 \approx s_2 > s_6)$ she states that, according to her opinion, the submission of s_4 is much better than that of s_5 that, in turn, is at the same level of the submission of s_2 that, in turn, is better than the submission of s_6. The main advantage of this approach is that students not only order the submissions from the best to the worst but also express a degree of preference between them. Moreover, it mitigates the *bias* problem given that students provide relative evaluations that consider only a couple of submissions at a time.

According to [9], provided fuzzy rankings are then transformed in fuzzy preference relations, expanded to estimate missing values and then aggregated through ordered weighted averaging. From the aggregated relation, the global score $\phi(s_i)$ is calculated for every $s_i \in S$ and the submissions are ranked accordingly. The *cardinal grade* of each submission is then calculated by asking a reliable expert (e.g. the teacher) to grade the best and the worst submissions (i.e. the first and the last in the final ranking obtained through FOPA) and by normalizing the global scores according to these values.

Let g_{min} and g_{max} be the grades assigned to the best and the worst submissions, the estimated grade g_i for every $s_i \in S$ is obtained as follows:

$$g_i = \frac{(\phi(s_i) - \phi_{min}) \cdot (g_{max} - g_{min})}{(\phi_{max} - \phi_{min})} + g_{min} \tag{5}$$

where ϕ_{min} and ϕ_{max} are the global scores associated to the best and the worst submissions.

4 Experiment and Evaluation

To evaluate the capability of FOPA in supporting formative evaluation in comparison to the other peer assessment models discussed in Sect. 3, we have experimented them within a course on *Computer Skills for Education* of a M.S. degree in Pedagogical Sciences at the University of Salerno. In particular, the experiment was aimed at measuring at what extent each model is able to estimate the grade assigned by the teacher to every student based on imprecise ordinal feedback provided by students themselves. In the next subsections, we describe the experimental setting and, then, we illustrate and analyse the collected data.

4.1 Experimental Setting

The experimental set was composed by first year students taking part in a 20 h course on *Computer Skills for Education* aimed at developing basic competencies on computer architectures, computational thinking and coding. The course, that is part of a 5-year M.S. degree in Pedagogical Sciences, was held through traditional face-to-face lectures and exercises sessions.

The formative evaluation experiment was performed in classroom in two sessions, held in two different days of the same week, with 25 voluntary students. In the first session students have been asked to complete and submit a coding exercise while in the second session students have been asked to assess the submissions coming from a subset of their peers by providing a fuzzy ranking as defined in Sect. 3.

The peer grading task was performed in a blind mode in order that students do not know whom they are assessing. The same submissions have been also assessed by the course teacher to build the ground truth with which to compare the results coming from experimented peer assessment models.

4.2 Data Collection

A total of 11 students over 25 completed the first session by submitting a solution to the proposed exercise while the remaining 14 were not able to complete the task. For this reason, during the second session students were divided in two groups: the first including those that submitted their solution and the second including the remaining ones. Students of the first group (being considered more proficient) were asked to evaluate 5 submissions (over the 11 available) while students of the second group were asked to only evaluate 3 submissions.

To assign the submissions to assessors, two assessment grids have been generated: the first 11×11 grid involved students from the first group both as assessors and as assessees while the second 11×14 grid involved students from the first group as assessees and students from the second group as assessors. In both cases, Eq. (1) was applied to generate the assessment grid.

Only 17 students over 25 completed the second session by providing a fuzzy ranking: 10 coming from the first group and 7 coming from the second one. All provided fuzzy rankings were complete i.e. all assigned submissions were covered by them. The 11 submissions were also evaluated by the teacher in the range [0,30]. The provided fuzzy rankings as well as teacher assigned grades (true grades) are summarized in Table 1.

4.3 Evaluating Peer Assessment Models

We have applied FOPA and the other peer assessment models introduced in Sect. 3 on collected data both to demonstrate the effectiveness of ordinal peer assessment in the estimation of student grades and to compare the results obtained by each model with respect to teacher assigned grades.

The Table 2 shows, for each student, the true grade, the grade estimated by FOPA, those estimated by the models of *Mallow* (MAL), *Score-Weighted Mallows* (MALS),

Table 1. Students' provided fuzzy rankings and true grades assigned by the teacher.

Student	Assessees	Fuzzy rankings	True grade (0–30)
s_1	$\{s_2, s_4, s_7, s_9, s_{11}\}$	$s_4 \geq s_{11} \geq s_9 \approx s_7 \approx s_2$	18
s_2	$\{s_3, s_5, s_6, s_8, s_{10}\}$	$s_3 \geq s_{10} \approx s_5 \gg s_8 \approx s_6$	10
s_3	$\{s_1, s_4, s_6, s_9, s_{11}\}$	$s_4 \gg s_{11} \geq s_9 \geq s_1 \geq s_6$	24
s_4	$\{s_1, s_3, s_5, s_8, s_{10}\}$	$s_{10} \geq s_3 > s_5 > s_1 > s_8$	30
s_5	$\{s_1, s_3, s_6, s_8, s_{11}\}$	$s_3 \gg s_{11} > s_8 > s_1 \gg s_6$	13
s_6	$\{s_2, s_4, s_7, s_9, s_{11}\}$	–	18
s_7	$\{s_1, s_2, s_4, s_6, s_9\}$	$s_4 \gg s_9 > s_1 > s_6 \geq s_2$	10
s_8	$\{s_2, s_5, s_6, s_7, s_{10}\}$	$s_{10} \geq s_5 > s_2 \geq s_7 \approx s_6$	11
s_9	$\{s_3, s_5, s_7, s_8, s_{10}\}$	$s_3 \gg s_{10} > s_8 \geq s_5 \approx s_7$	18
s_{10}	$\{s_2, s_4, s_7, s_9, s_{11}\}$	$s_4 \gg s_{11} > s_9 > s_7 \geq s_2$	28
s_{11}	$\{s_1, s_3, s_5, s_8, s_{10}\}$	$s_3 > s_{10} \gg s_8 \approx s_1 \approx s_5$	26
s_{12}	$\{s_4, s_9, s_{11}\}$	$s_4 \gg s_9 \geq s_{11}$	–
s_{13}	$\{s_4, s_5, s_{10}\}$	$s_4 \gg s_5 \approx s_{10}$	–
s_{14}	$\{s_1, s_5, s_{11}\}$	–	–
s_{15}	$\{s_2, s_6, s_7\}$	$s_7 \gg s_2 \approx s_6$	–
s_{16}	$\{s_1, s_3, s_8\}$	–	–
s_{17}	$\{s_2, s_7, s_{11}\}$	$s_{11} \gg s_7 > s_2$	–
s_{18}	$\{s_2, s_5, s_{10}\}$	$s_{10} \gg s_2 \geq s_5$	–
s_{19}	$\{s_4, s_6, s_9\}$	$s_4 \gg s_9 \geq s_6$	–
s_{20}	$\{s_3, s_8, s_{10}\}$	–	–
s_{21}	$\{s_4, s_8, s_9\}$	–	–
s_{22}	$\{s_3, s_5, s_{10}\}$	–	–
s_{23}	$\{s_2, s_7, s_{11}\}$	$s_{11} \geq s_2 > s_7$	–
s_{24}	$\{s_3, s_6, s_8\}$	–	–
s_{25}	$\{s_1, s_6, s_9\}$	–	–

Bradley-Terry (BT) and *Plackett-Luce* (PL) as defined in [8], and the grade obtained using the Borda count defined by Eq. (3). While FOPA and the Borda count have been implemented in *Matlab*, we used the freely available *PeerGrader* software[1] for the MAL, MALS, BT and PT models.

Equation (5) is used to obtain cardinal grades from the scores associated to each submission. The performance of each model is measured both in terms of *Correctly Recovered Pairwise Relations* (PCRPR) and *Root Mean Square Error* (RMSE). With respect to PCRPR, as it can be seen in Table 2, all models rank the submissions in the same order reaching a 90% of similarity to the ranking made by considering teacher assigned grades. With respect to RMSE, the models behaviour ranges from a minimum error of 2.4, obtained by FOPA, to a maximum error of 2.9, obtained by Borda.

According to such results, we can assert that ordinal peer assessment is a valuable approach to support formative evaluation and is capable of estimating quite accurately

[1] www.peergrading.org.

Table 2. True grades compared to grades obtained with peer assessment methods.

Student	True grade	FOPA	MAL	MALS	BT	PL	Borda
s_1	18.0	15.7	16.0	14.6	14.7	14.3	12.0
s_2	10.0	9.8	9.0	11.1	11.5	10.8	14.0
s_3	24.0	28.0	27.7	26.9	27.1	26.6	25.0
s_4	30.0	30.0	30.0	30.0	30.0	30.0	30.0
s_5	13.0	15.6	18.3	15.9	16.2	15.1	17.0
s_6	18.0	9.0	9.0	9.0	9.0	9.0	9.0
s_7	10.0	11.0	11.3	11.9	12.2	11.7	14.0
s_8	11.0	14.1	13.7	14.9	15.1	14.1	13.0
s_9	18.0	19.6	20.7	19.7	20.3	19.9	18.0
s_{10}	28.0	24.1	25.3	23.6	24.4	23.7	27.0
s_{11}	26.0	23.5	23.0	22.0	22.5	22.1	24.0
Mean	17.9	18.2	18.5	18.2	18.5	17.9	18.5
PCRPR		0.9	0.9	0.9	0.9	0.9	0.9
RMSE		2.4	2.7	2.8	2.8	2.6	2.9

teacher assigned grades, at least in the considered sample. Only small differences can be appreciated with respect to the selected model. In particular, FOPA presents the minimum error but slightly increases the mean grade of the class with respect to teacher assigned grades. Instead, PL shows a slightly greater error rate but it maintains a greater fidelity with respect to the mean grade.

4.4 Additional Experiments

It should be noted that, while FOPA is able to fully interpret collected fuzzy rankings, the other models need to translate them into ordinal rankings before use. In particular, while Borda just interprets the $>$ symbol, MAL, MALS, BT and PT can also interpret the $\approx$ symbol (i.e. they admit ties). The symbols $\geq$ and $\gg$ within fuzzy rankings are so translated in the symbol $>$ before using them with methods different from FOPA. The $\approx$ symbol is also removed with Borda and an artificial random order is introduced between the adjacent symbols.

Given this difference, an additional experiment has been performed to investigate the behavior of FOPA when put under the same conditions of the other methods i.e. when using modified fuzzy rankings rather than the original ones. In such conditions, FOPA ended up with a 2.7 RMSE (with 0.9 PCRPR) so 0.3 points are lost with respect to the preceding settings. So, we can conclude that the contribution of fuzzy symbols is remarkable but not decisive in the estimation of teacher assigned grades.

Two additional experiments have been performed so far to evaluate how the models under examination perform with a reduced set of ranking strings. As said, students have been assigned to two groups, a first group including "more proficient" students and a second group made of "less proficient" ones.

The rows 1–3 of Table 3 show the results obtained by all peer assessment models by considering only fuzzy rankings coming from the group of "more proficient"

Table 3. Performance considering a subset of available fuzzy rankings.

Group	Measure	FOPA	MAL	MALS	BT	PL	Borda
1	Mean	19.1	17.6	18.7	18.9	18.7	17.9
	PCRPR	0.9	0.9	0.9	0.9	0.9	0.8
	RMSE	2.9	3.0	3.0	3.0	2.9	3.6
2	Mean	16.2	18.3	17.3	17.6	17.6	17.9
	PCRPR	0.8	0.6	0.7	0.8	0.8	0.6
	RMSE	4.7	7.8	4.8	4.7	4.8	8.7

students. With a lower amount of data available, all the models result in slightly higher error rates, while keeping the adherence to the teacher ranking almost unaltered. The consideration that can be drawn is that adding evaluations improve the peer grading process even in case of dubious reliability of the new evaluations.

The rows 4–6 of Table 3 show, instead, the results obtained by considering only fuzzy rankings coming from the group of "less proficient" students. As it can be seen, basing on a lower amount of data that, in addition, is of a worst quality, all models result in significantly higher error rates. In particular, Borda and MAL show the higher increase in RMSE (+5.8 for Borda, +5.1 for MAL) while BT shows the lowest one (+1.9). The adherence to the teacher's ranking also lowers drastically with values ranging from 60% to 80%. Nevertheless, also in this case FOPA shows the best performance.

5 Final Remarks

In this paper the results of an experiment aimed at introducing peer assessment as a tool for formative evaluation within a University course on *Computer Skills for Education* are presented. The performance of several peer assessment models in estimating the grades assigned by the teacher are measured and compared. Obtained results confirm that it is possible to estimate with a satisfactory degree of reliability student grades even based on imprecise ordinal feedback provided by students themselves.

The application of alternative aggregation models offers better results with respect to the classical Borda count. In particular, FOPA (the proposed model based on Fuzzy Set Theory and Group Decision Making) outperforms the other models in almost all conditions, at least in the limited framework of the performed experiment. The obtained results are so encouraging and suggest to extend the experience to other courses, both in Sciences and Humanities.

A weakness of the proposed approach is that the contribution of the introduced fuzzy symbols $\geq$ and $\gg$ is remarkable but not decisive to differentiate FOPA performance with respect to other models. One wonders whether using fuzzy numbers (maybe in form of linguistic labels) instead of fuzzy symbols can improve performance by better characterizing the uncertainty coming from students evaluation. This suggests to direct future research activities toward model improvement in this sense. Moreover, in [3, 6] it has been demonstrated that, weighting the grades proposed by assessor students based on their performance, can contribute in improving the reliability of final

grades. To verify this occurrence also for ordinal peer assessment methods, we plan to introduce similar weighting techniques also in FOPA relying on *influence-based* fuzzy models like that introduced in [20].

References

1. Black, P., Wiliam, D.: Assessment for learning in the classroom. In: Assessment and Learning, pp. 9–15. SAGE Publications (2006)
2. Bransford, J.D., Brown, A., Cocking, R.: How People Learn: Mind, Brain, Experience and School. National Academy Press, Washington, DC (2000)
3. Capuano, N., Caballé, S., Miguel, J.: Improving peer grading reliability with graph mining techniques. Int. J. Emerg. Technol. Learn. **11**(7), 24–33 (2016)
4. Sadler, P.M., Good, E.: The impact of self- and peer-grading on student learning. Educ. Assess. **11**(1), 1–31 (2006)
5. Glance, D.G., Forsey, M., Riley, M.: The pedagogical foundations of massive open online courses. First Monday 18(5) (2013)
6. Capuano, N., Caballé, S.: Towards adaptive peer assessment for MOOCs. In: Proceedings of the 10th International Conference on P2P, Parallel, GRID, Cloud and Internet Computing (3PGCIC 2015), Krakow, Poland (2015)
7. Bouzidi, L., Jaillet, A.: Can online peer assessment be trusted? Educ. Technol. Soc. **12**(4), 257–268 (2009)
8. Raman, K., Joachims, T.: Methods for ordinal peer grading. In: Proceedings of the 20th SIGKDD International Conference on Knowledge Discovery and Data Mining (2014)
9. Capuano, N., Loia, V., Orciuoli, F.: A fuzzy group decision making model for ordinal peer assessment. IEEE Trans. Learn. Technol. **10**(2), 247–259 (2017)
10. Carlson, P.A., Berry, F.C.: Calibrated peer review™ and assessing learning outcomes. In: Proceedings of the 33rd International Conference Frontiers in Education (2003)
11. Piech, C., Huang, J., Chen, Z., Do, C., Ng, A., Koller, D.: Tuned models of peer assessment in MOOCs. In: Proceedings of the 6th International Conference on Educational Data Mining (2013)
12. Goldin, I.M.: Accounting for peer reviewer bias with bayesian models. In: Proceedings of the 11th International Conference on Intelligent Tutoring Systems (2012)
13. Walsh, T.: The peerrank method for peer assessment. In: Proceedings of the 21st European Conference on Artificial Intelligence (2014)
14. Albano, G., Capuano, N., Pierri, A.: Adaptive peer grading and formative assessment. J. e-Learn. Knowl. Soc. **13**(1), 147–161 (2017)
15. Caragiannis, I., Krimpas, A., Voudouris, A.A.: Aggregating partial rankings with applications to peer grading in massive online open courses. In: Proceedings of the International Conference on Autonomous Agents and Multiagent Systems, Istanbul (2015)
16. Borda, J.C.: Memoire sur les elections au scrutin. Histoire de l'Académie Royale des Sciences (1781)
17. Mallows, C.L.: Non-null ranking models. I. Biometrika **44**(1), 114 (1957)
18. Bradley, R.A., Terry, M.E.: Rank analysis of incomplete block designs: I. The method of paired comparisons. Biometrika **39**(3), 324 (1952)
19. Plackett, R.L.: The analysis of permutations. Appl. Stat. **24**(2), 193 (1975)
20. Capuano, N., Chiclana, F., Fujita, H., Herrera-Viedma, E., Loia, V.: Fuzzy group decision making with incomplete information guided by social influence. IEEE Trans. Fuzzy Syst. PP (99), 1 (2017)