A Fuzzy Group Decision Making Model for Ordinal Peer Assessment

Nicola Capuano, Vincenzo Loia, *Senior Member, IEEE*, and Francesco Orciuoli

Abstract—Massive Open Online Courses (MOOCs) are becoming an increasingly popular choice for education but, to reach their full extent, they require the resolution of new issues like assessing students at scale. A feasible approach to tackle this problem is peer assessment, in which students also play the role of assessor for assignments submitted by others. Unfortunately, students are unreliable graders so peer assessment often does not deliver accurate results. In order to mitigate this issue, we propose a new model for ordinal peer assessment based on the principles of fuzzy group decision making. In our approach, each student is asked to rank a few random submissions from the best to the worst and to specify, with a set of intuitive labels, at what extent each submission is better than the following one in the ranking. Students' provided rankings are then transformed in fuzzy preference relations, expanded to estimate missing values and aggregated through OWA operators. The aggregated relation is then used to generate a global ranking between the submissions and to estimate their absolute grades. Experimental results are presented and show better performances with respect to other existing ordinal and cardinal peer assessment techniques both in the reconstruction of the correct ranking and on the estimation of students' grades.

Index Terms—Fuzzy set theory, group decision making, massive open online courses, peer assessment

1 Introduction

Since their introduction, *Massive Open Online Courses* (MOOCs) have become increasingly popular with millions of students enrolled, thousands of courses offered and hundreds of educational institutions involved. Today's MOOCs are provided by several organisations around the world. While at the beginning they were taken only by students willing to broaden their education or simply to learn new things, through the so called *credential programs* they are currently interesting also people that want to achieve credits toward a degree or earn credentials to show to prospective employers.

Differently from usual e-learning courses, MOOCs are intended for thousands of simultaneous participants, with some courses offered by *Coursera* and *Udacity* (just to mention some of the most popular providers) exceeding 100,000 registrants. Due to their scale, MOOCs introduce new technical and pedagogical challenges that require overcoming the traditional e-learning model based on tutor assistance to maintain a cheap and unrestricted access to high quality resources. Because of the high numbers of students enrolled and the relatively small number of tutors, in fact, tutor involvement during delivery stages has to be limited to the most critical tasks [1].

- *N. Capuano is with the Department of Information Engineering, Electric Engineering, and Applied Mathematics, University of Salerno, Via Giovanni Paolo II 132, Fisciano, (SA) 84084, Italy. E-mail: ncapuano@unisa.it.*
- *V. Loia and F. Orciuoli are with the Department of Business Sciences, Management and Innovation Systems, University of Salerno, Via Giovanni Paolo II 132, Fisciano, (SA) 84084, Italy. E-mail: {loia, forciuoli}@unisa.it.*

Manuscript received 20 Jan. 2016; revised 18 Apr. 2016; accepted 30 Apr. 2016. Date of publication 10 May 2016; date of current version 16 June 2017. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below. Digital Object Identifier no. 10.1109/TLT.2016.2565476

This requirement especially impacts on assessment tasks aimed at verifying students' proficiency in developed competencies. Given the impossibility for human tutors to follow up with every student and review and grade assignments individually, the use of automated approaches is increasingly required. A typical approach is to use close type questions in exams and assignments so that grading can be done automatically. Unfortunately, automated assessment may result highly unsatisfactory when applied to complex tasks like writing reports, proving mathematical statements, expressing critical thinking, etc. [2]. For these tasks, an approach that is gaining a growing consensus 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 so obtained by combining information provided by peers [3]. Peer assessment has the capability of easily scaling to any size: the number of assessors in fact naturally grows with the number of students. Conversely, the main issue of this approach, is that it relies on grades assigned by unreliable graders (the students themselves) lacking the needed expertise, both didactical and on the specific subject to be assessed [4].

Several approaches (summarized in Section 2) have been proposed so far to improve the reliability of peer assessment. In this paper we discuss *FOPA* (*Fuzzy Ordinal Peer Assessment*): a new peer assessment model based on *Fuzzy Sets Theory* and on the application of *Group Decision Making* (GDM) techniques. In FOPA each student is asked to rank (rather than to grade) few submissions coming from other students and to express preferences (using a set of intuitive labels) between couples of subsequent elements in the ranking. Such preferences are then aggregated among all students and expanded to estimate missing values (i.e., preferences that are not explicitly stated by any student but

that can be inferred by looking at the whole picture). Expanded preferences are then used to generate a global ranking between all the submissions and to estimate their absolute grades.

The paper is organized as follows. The next section presents related work on peer assessment as well as some existing fuzzy-based approach proposed by recent literature. Section 3 defines the ordinal peer assessment problem in formal terms while Section 4 defines background concepts related to fuzzy-based GDM. The FOPA model is exhaustively described in Section 5 while Section 6 evaluates the model with synthetic data and compares obtained results with results coming from other existing approaches. Finally, Section 7 summarizes conclusions and outlines ongoing work.

2 Related Work

Although peer assessment has been studied for several years, it has recently had an increasing interest from the scientific community due to the diffusion of MOOCs. Even if some studies suggest a fair good correlation between the results coming from peer assessment and the grades coming from experienced instructors (at least in conventional classrooms and for specific, high structured, domains), there is still a general concern on the use of peer assessment as a reliable and accurate strategy to approximate instructor grades [5]. For this reason, several approaches, at various stages of development, have been proposed so far to improve its reliability.

The *Calibrated Peer Review* (CPR) proposes a calibration step to be performed by students before starting to assess other students' assignments [6]. During the calibration, each student rates a set of assignments that have been already rated by the instructor. The discrepancy between grades provided by a student and the instructor measures student's accuracy in assessment and it 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 on a peer assessment.

Some probabilistic models for tuning grades obtained in peer assessment are presented in [7]. Such models estimate the reliability of each grader as well as her bias (i.e., the grader's tendency to inflate or deflate the provided assessment) based on the analysis of grading performances on special *ground truth* submissions that have been evaluated either by the instructor or by a big number of peers. Reliability and bias are then used to tune the grades assigned by each student. Similar approaches have been proposed in [8], where a Bayesian model has been adopted to estimate graders' reliability on each item of an assessment rubric and, more recently, in [9] where a hierarchical Bayes model has been used.

In [4] and [10], the ability of an assessor student to correctly rate peer students is assumed to be dependant 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 inspired by *Google PageRank* is proposed for their calculation. The advantage of this approach, compared to the previous ones, is that it does not require any instructor's intervention given that there is no need of a ground truth of professionally graded assignments.

In [11] a different approach, aimed at making the assessment process as simple as possible, has been proposed. Authors have shown how ordinal feedback (e.g., "the report x is better than the report y") is easier to provide and more reliable than cardinal one (e.g., "the grade of report x is a B"). Basing on that assumption, the authors defined several probabilistic models for obtaining student grades starting from partial rankings provided by the peers. An experiment, with data collected from a real course, have demonstrated that the performance of such methods is at least competitive with cardinal methods for grade estimation, even though they require strictly less information from the graders.

In [2], the authors have shown that ordinal peer assessment is a highly effective and scalable solution 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. They have theoretically demonstrated that, given k students, if each correctly ranks the received assignments, the defined aggregation method is able to recover a fraction $1 - \mathcal{O}(1/k)$ of the true ranking (with $\mathcal{O}$ representing the limiting behaviour). They have also demonstrated that the same ordinal peer assessment method is quite robust even when students have imperfect capabilities as graders.

According to the latter works cited, ordinal peer assessment methods seem to be more promising with respect to cardinal ones. In particular, they overcome the problem that students may be grading on different scales: by letting students propose ordinal statements rather than cardinal grades, there is no need to develop a scale from each student onto the peer grading algorithm. For these reasons, we have chosen this approach as a basis for our fuzzy model and we have compared our results both with ordinal methods presented in [2] and [11] that with standard average-based cardinal methods.

With respect to the application of *Fuzzy Set Theory* to peer assessment, some experiment has been already performed so far. In [12], the students of a class have been asked to express a grade, in terms of a fuzzy value in $[0, 1]$, for each assignment coming from the other students in the same class. The final grade of each assignment is then obtained by averaging the proposed grades, weighted with respect to expertise levels assigned by the teacher.

In [13] the authors have proposed a framework aimed at enhancing the effectiveness of peer assessment by letting students expressing grades as fuzzy membership functions with respect to a set of assessment criteria. The proposed grades are then adapted basing on assessors' learning styles (through defined heuristics) and differences among peers are reconciled through agent negotiation based on fuzzy constraints.

A more recent work [14] have proposed another approach: each student of a class evaluates the assignments coming from peers in terms of linguistic labels mapped to interval *Type-2* fuzzy sets. The final grade of each assignment is then obtained by aggregating the grades proposed by peers,

weighted with respect to expertise levels assigned by teachers. Obtained results are then re-mapped on the linguistic labels to obtain a final literal grade.

It should be noted that the reported experiments based on fuzzy sets, are mainly thought for small classes, to encourage students to participate in the evaluation of their learning, so enhancing their reflective and critical thinking, rather than to provide reliable grades for student assignments. Conversely, the main aim of our work is to improve the reliability of peer assessment in massive contexts with simple tools that minimize the instructor's involvement.

3 The Ordinal Peer Assessment Problem

In a typical peer assessment scenario an *assignment* is given to n different students $S = \{s_1, \ldots, s_n\}$. Each student elaborates an own solution (e.g., an essay, a set of answers to open-ended questions, etc.) generating a *submission*. Each student has then to grade m different submissions (with $m \leq n$) coming from other students (maybe based on an assessment rubric).

The assignment of submissions to assessor students is performed in accordance to an *assessment grid*: a Boolean $n \times n$ matrix G where $G_{i,j} = 1$ if s_j has to grade the submission of s_i while $G_{i,j} = 0$ otherwise. The matrix G has the following properties:

1) the sum of the elements in each row and column is equal to m (i.e., each student grades and is graded by m other students);
2) the sum of the elements in the main diagonal is equal to 0 (i.e., nobody evaluates himself).

The easiest way to build the assessment grid is by filling it at random with an algorithm preserving the above properties. A feasible algorithm [4] starts with a null matrix and initialises its elements basing on the following equation:

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

The obtained matrix is then shuffled in several iterations by randomly selecting a couple of rows (or columns) i and j so that $G_{i,j} = G_{j,i} = 0$ and swapping them.

Students have then to review submissions according to the assessment grid, i.e., each student s_j reviews the assignments coming from students in $S_j = \{s_i |\ G_{i,j} = 1\}$. In *ordinal peer assessment* this means that each student s_j is required to define a ranking $\succ_j$ over S_j. For example, the following ranking:

$$s_{i1} \succ_j s_{i2} \succ_j \ldots \succ_j s_{i|S_j|} \quad (2)$$

means that, according to s_j, the submission of s_{i1} is better than that of s_{i2} etc. A ranking $\succ_j$ is undefined for elements not included in S_j so it is a *partial ranking* over S. All defined partial rankings can be collected in a *ranking matrix* R where $R_{i,j}$ 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 able to compute a complete ranking over the whole set of submissions. A simple and effective aggregation rule is the classical *Borda count* [15] 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:

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

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

The performances of an aggregation rule are measured in terms of *percentage of correctly recovered pairwise relations* with respect to the ground truth. In [2] it has been demonstrated that, in case of perfect grading (i.e., when partial rankings defined by students are consistent to the ground truth), Borda recovers an expected fraction of $1 - \mathcal{O}(1/\sqrt{m})$ of correct pairwise relations independently of the total number of submissions.

In the same work, authors demonstrated that Borda outperforms other, more complex aggregation rules like *Random Serial Dictatorship* [16] and *Markov chain* inspired methods [17] especially in case of imperfect grading (i.e., when partial rankings defined by students are not consistent to the ground truth). In [11] authors have defined additional methods for ordinal peer assessment based on models that represent probabilistic distributions over rankings, obtained extending the models of *Mallow* [18], *Bradley-Terry* [19] and *Plackett-Luce* [20]. Such methods obtain better results with respect to Borda also in case of imperfect grading and are also capable of detecting meaningful cardinal grades. For these reasons they have been chosen, together with Borda, as the baseline against which to compare the techniques defined in this research.

4 Fuzzy Group Decision Making

A peer assessment problem can be seen as a special case of *Group Decision Making*. In a typical GDM problem, a set of experts has to define a ranking among a finite set of alternatives. Each expert expresses her own preferences about all the alternatives in form of *preference ordering* (alternatives are ordered from the best to the worst) or *degree of preference* (a utility value is assigned to each alternative). Expert preferences are then aggregated and a ranking over all alternatives is calculated.

In particular, ordinal peer assessment can be seen as a GDM problem with *preference ordering* where:

1) experts and alternatives belong to the same set (i.e., students grade the submissions made by other students);
2) each expert only ranks a small subset of alternatives (i.e., only few submissions are graded by each student);
3) experts' opinion is not fully reliable (students are far to be perfect graders).

These properties (in particular the last two) suggest to refer to GDM approaches that take into account the uncertainty resulting from imprecisions and lack of knowledge in experts' evaluations like those based on the *fuzzy set theory*. The next sections explain basic concepts and techniques related to fuzzy-based GDM while the Section 5 describes how, in this research, such techniques have been adapted and extended for ordinal peer assessment.

4.1 Fuzzy Preference Relations

Fuzzy sets were introduced in [21] as an extension of classical sets. While in a classical (crisp) sets, each element can either belong to or not belong to a set, fuzzy sets allow various degrees of membership of an element to a set, ranging from 0 (no membership) to 1 (full membership). More formally, if X is a collection of objects, a fuzzy set F in X is a set of ordered pairs $F = \{(x, \mu_F(x)) \mid x \in X\}$ where $\mu_F(x)$, called *membership function*, maps X to the membership space $[0, 1]$.

The author has also defined the concept of *fuzzy relation*: a relation where various degrees of association strength between elements are allowed. Given two collections of objects X and Y, a fuzzy relation R from X to Y is a fuzzy subset of $X \times Y$, i.e., $R = \{((x, y), \mu_R(x, y)) \mid (x, y) \in X \times Y\}$. Fuzzy relations can be used in GDM problems to let experts define imprecise and vague preferences between pairs of alternatives.

Given a finite set of alternatives $A = \{a_1, \ldots, a_n\}$, a *Fuzzy Preference Relation* (FPR) $P \subset A \times A$ defines the degree of preference of each alternative over another one. According to [22], P is characterised by the following membership function for $i, j \in \{1, \ldots, n\}$:

$$\mu_P(a_i, a_j) = \begin{cases} 1 & \text{if } a_i \text{ is definitely preferred to } a_j, \\ x \in (0.5,\ 1) & \text{if } a_i \text{ is slightly preferred to } a_j, \\ 0.5 & \text{if there is no preference,} \\ y \in (0,\ 0.5) & \text{if } a_j \text{ is slightly preferred to } a_i, \\ 0 & \text{if } a_j \text{ is definitely preferred to } a_i. \end{cases} \tag{4}$$

A FPR P can be conveniently represented as a $n \times n$ matrix where $P_{i,j} = \mu_P(a_i, a_j)$. In addition, a FPR can have the following properties:

1) $P_{ii} = 0.5 \ \ \forall i \in \{1, \ldots, n\}$, i.e., any alternative a_i is never preferred to itself;
2) $P_{i,j} + P_{j,i} = 1 \ \ \forall i \in \{1, \ldots, n\}$ (additive reciprocity), i.e., for any i and j, if a_i is preferred to a_j then a_j is evenly non preferred to a_i.

A FPR showing both properties is said to be *consistent*.

4.2 Aggregation of Fuzzy Preference Relations

In GDM, each expert belonging to a group $E = \{e_1, \ldots, e_m\}$ expresses her own preferences about any couple of alternatives. This results in a set of *individual* FPRs $\{P^1, \ldots, P^m\}$ where each P^i is defined by the expert e_i. A *collective* FPR P is then generated by aggregating all available individual FPRs. To this purpose, several aggregation rules have been proposed.

The *Ordered Weighted Average* (OWA) family of operators [23] are the most widely used aggregators. An OWA operator of dimension m is a function $OWA : [0, 1]^m \to [0, 1]$ associated with a set of weights $W = \{w_1, \ldots, w_m\}$ with $w_i \in [0, 1]$ and $\sum_i w_i = 1$. Let $X = \{x_1, \ldots, x_m\}$ be the list of values to aggregate, the OWA operator is defined as:

$$OWA(x_1, \ldots, x_m) = \sum_{i=1}^{m} w_i \cdot y_i \tag{5}$$

were y_i is the ith largest value in X (so each element of X is not associated with a weight; rather a weight is associated with a position in the ordered set of values). Through OWA, the global preference $P_{i,j}$ for every pair of alternatives can be obtained as: $P_{i,j} = OWA(P^1_{i,j}, \ldots, P^m_{i,j})$. Extending the notation to matrices we have $P = OWA(P^1, \ldots, P^m)$.

The behaviour of OWA strictly depends on the used weight vector. In [24] a weight configuration has been defined to let OWA assumes the behaviour of *soft majority*: i.e., the global preference between each pair of alternatives is defined according to the *majority* of experts' opinions.

Using OWA to compute the collective FPR P does not guarantee that $P_{i,j} + P_{j,i} = 1$ for any i and j (so P may be *inconsistent* w.r.t the definition given in Section 4.1) [25]. For this reason other aggregators have been proposed by other authors like in [22] where the *Simple Additive Weighting* (SAW) operator is used in place of the OWA. Nevertheless, for applications where the consistency of the collective FPR is not a constraint, OWA can be considered a fair choice.

4.3 Ranking of Alternatives

After having aggregated all the individual FPRs, the available alternatives must be ranked from the best to the worst by associating a degree of preference $\phi(a_i)$ to any $a_i \in A$. Several measures have been proposed so far to quantify the degree of preference of an alternative. In [24] the *Quantifier Guided Dominance Degree* (QGDD) has been proposed to calculate the dominance that one alternative has over all the others in a *fuzzy majority* sense as follows:

$$\phi_{QGDD}(a_i) = OWA_Q\big(P_{i,j};\ j = 1, \ldots, n;\ j \neq i\big). \tag{6}$$

In the same paper a *Quantifier Guided Non-Dominance Degree* (QGNDD) has been proposed to calculate the degree in which a given alternative is not dominated by a fuzzy majority of the remaining ones as follows:

$$\phi_{QGNDD}(a_i) = OWA_Q\left(\begin{matrix} 1 - max\{P_{j,i} - P_{i,j}, 0\}; \\ j = 1, \ldots, n;\ j \neq i \end{matrix}\right). \tag{7}$$

The two measures can be used alternatively or combined. Instead, in [22] the degree of preference of each alternative is calculated in terms of *Net Flow* as follows:

$$\phi_{NF}(a_i) = \sum_{j=1, j\neq i}^{n} P_{i,j} - \sum_{j=1, j\neq i}^{n} P_{j,i} \tag{8}$$

where the first summation is the *leaving flow* i.e. the total degree of preference of a_i over all the other alternatives while the last summation is the *entering flow* i.e. the total degree of preference of all the other alternatives over a_i. It does not exist a measure absolutely better than the others so the selection of the right measure to use is often based on application dependent heuristics.

5 Fuzzy Ordinal Peer Assessment

In this section we describe *FOPA* (*Fuzzy Ordinal Peer Assessment*): a new approach to peer assessment that exploits and adapts fuzzy-based GDM techniques outlined in Section 4. As described in Section 3, in ordinal peer assessment, each student belonging to a set $S = \{s_1, \ldots, s_n\}$ has to rank the submissions coming from m other students according to an

TABLE 1
Meaning and Preference Degrees of Ranking String Symbols

Symbol	Meaning	Preference degree
$\gg$	The previous is much better than the next	0.85
$>$	The previous is better than the next	0.65
$\geq$	The previous is a little better than the next	0.58
$\approx$	The previous and the next are at the same level	0.50

assessment grid G. To do that, each student s_j defines a partial ranking on the set $S_j = \{s_i |\ G_{i,j} = 1\}$. In our approach, instead of defining a partial ranking, preferences between the elements of S_j are expressed in terms of a FPR P^j. Individual FPRs, coming from all the assessor students, are then aggregated and a global ranking of submissions is calculated.

The main advantage of FOPA is that students not only order the submissions from the best to the worst but also express a variable degree of preference between submissions. As demonstrated in Section 6, this allows to obtain better performances when reconstructing the global ranking and, also, allows to obtain a reliable cardinal grade for each submission. Moreover, with respect to cardinal peer assessment (CPA), the proposed approach mitigates the *bias* problem given that students provide relative rather absolute evaluations that consider only a couple of submissions at a time.

On the other hand, one drawback of this approach is that the definition of a FPR may result too complex and time-consuming with the risk of introducing errors and inconsistencies impacting assessment performances and nullifying the advantages described so far. To overcome this issue, students are not asked to directly define FPRs but to specify a simpler *ranking string* as explained in Section 5.1. Ranking strings are then converted in FPRs and used for subsequent processing.

Another issue that must be considered is that each individual FPR defined by a student only covers the preferences among m of the n elements of S (with $m \ll n$) while it is undefined for the other ones. This prevents the straightforward application of the aggregation techniques described in Section 4.2. To overcome this issue, a two-steps initialization method for undefined elements has been defined and described in Sections 5.2 and 5.4. Aggregation rules are then described in Section 5.3 while the methods for the ranking of submissions and the assignment of absolute grades are explained in Section 5.5.

5.1 Ranking Strings and Fuzzy Preference Relations

A *ranking string* defined by a student s_j is a finite sequence $R_j = (s_{i1}\ \sigma_1\ s_{i2}\ \sigma_2 \ldots\ s_{im-1}\ \sigma_{m-1}\ s_{im})$ made of $2m - 1$ terms, where the elements in odd positions represent the students evaluated by s_j i.e., so that $s_{ik} \in S_j$ for $k \in \{1, \ldots, m\}$, while the elements in even positions, i.e., σ_k for $k \in \{1, \ldots, m-1\}$, belong to the set of symbols $\Sigma = \{\gg, >, \geq, \approx\}$ and define a degree of preference between the preceding and the subsequent element according to Table 1.

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 following ranking string:

$$R_1 = (s_4 \gg s_5 \approx s_2 > s_6) \tag{9}$$

the student 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. To ensure consistency it is needed that each element of S_j cannot be included more then once in the ranking string (i.e., cycles are unallowed).

Starting from a ranking string R_j, it is possible to generate a partial FPR $P^j \subset S \times S$ by associating a fuzzy preference degree $d(\sigma)$ to each symbol $\sigma \in \Sigma$ (see the last column of Table 1 for a feasible set of values) in this way:

1) $P^j_{ik,ik+1} = d(\sigma_k)$ for any substring $(s_{ik}\ \sigma_k\ s_{ik+1})$ of R_j;
2) $P^j_{ik+1,ik} = 1 - d(\sigma_k)$ for any substring $(s_{ik}\ \sigma_k\ s_{ik+1})$ of R_j;
3) $P_{ii} = 0.5$ for $i \in \{1, \ldots, n\}$;

where the first statement transforms the degrees of preference embedded in the ranking string R_j in values of the membership function of P^j, while the second and third statements are aimed at ensuring the consistency of P^j according to the definition given in Section 4.1.

It should be noted that, after that initialisation, only a fraction of $(n + 2m - 2)/n^2$ elements of P^j are defined. For example, starting from the ranking string reported by the equation (9) and supposing that $n = 6$ (i.e., six students has to be assessed in total), the following partial FPR is generated according to the previous rules:

$$P^1 = \begin{pmatrix} 0.50 & - & - & - & - & - \\ - & 0.50 & - & - & 0.50 & 0.65 \\ - & - & 0.50 & - & - & - \\ - & - & - & 0.50 & 0.85 & - \\ - & 0.50 & - & 0.15 & 0.50 & - \\ - & 0.35 & - & - & - & 0.50 \end{pmatrix} \tag{10}$$

where the symbol – indicates an undefined cell.

Nevertheless, in real contexts, hundreds of students (thousands in MOOCs) has to be evaluated in total (so n becomes very large) while each student can be requested to evaluate only a small number of other submissions (so m remains small). This means that every P^j becomes a sparse matrix with only few elements defined. A feasible expansion technique enabling to obtain missing values is so needed as described in the next section.

5.2 Expansion of Individual Preferences

In GDM each expert should express a preference between each pair of alternatives. When this is not possible, as in the peer assessment case, a method enabling the estimation of missing values must be adopted. In [26] a method to estimate the missing preferences of each expert is described basing on the information available in the individual FPR of the same expert. A different approach is defined in [27] where missing preferences of each expert are estimated using information coming from other experts. We propose a two-step approach relying on both sources of information.

In a *first expansion step*, performed soon after having collected an individual FPR P^j, some of the missing values of P^j are estimated in a way they are consistent to those declared by the same student. The expanded individual FPR (that is still incomplete) is then aggregated with other FPRs as described in Section 5.3. In a *second expansion step* (described in Section 5.4) the collective FPR is expanded again to obtain last missing values. It is worth noting that, while the first step uses the information provided by each student to expand her own preferences, the second step uses information coming from all the students to expand the collective FPR.

With respect to the first expansion step, to detect missing values in an individual FPR, it is needed that the new values do not contradict the values stated by the student, according to a given property. In [28], the *additive transitivity* is demonstrated to be the most suitable property for this purpose. A preference relation P has additive transitivity if its elements satisfy [29]:

$$P_{i,j} + P_{j,k} + P_{k,i} = 1.5 \quad \forall i, j, k \in \{1, \ldots, n\}. \tag{11}$$

Intuitively, the additive transitivity can be seen as an extension, to three alternatives, of the additive reciprocity defined in Section 4.1. A FPR that respects the additive transitivity property is said to be *additive consistent*.

If we combine the definitions of additive reciprocity and additive transitivity we obtain that, when P is additive consistent, an unknown element can be obtained by combining other known elements of P. Unfortunately, user defined FPRs are not always additive consistent. Even in this case we can use the additive transitivity property to identify missing values that are as consistent as possible with the existing (maybe partially inconsistent) FPR values. This is done through a set of estimators for each unknown value. In particular, given an unknown value $P_{i,j}$, we can define the following estimators:

$$e_{k,1}(P_{i,j}) = P_{i,k} + P_{k,j} - 0.5 \quad \forall\, k : P_{i,k} \text{ and } P_{k,j} \text{ are defined} \tag{12}$$

$$e_{k,2}(P_{i,j}) = P_{k,j} - P_{k,i} + 0.5 \quad \forall\, k : P_{k,i} \text{ and } P_{k,j} \text{ are defined} \tag{13}$$

$$e_{k,3}(P_{i,j}) = P_{i,k} - P_{j,k} + 0.5 \quad \forall\, k : P_{i,k} \text{ and } P_{j,k} \text{ are defined} \tag{14}$$

A missing value $P_{i,j}$ can be so estimated by averaging the values obtained by all available estimators as follows:

$$e(P_{ij}) = \frac{\sum_{k=1}^{n} \left(M_{ik}M_{kj}e_{k,1}(P_{ij}) + M_{kj}M_{ki}e_{k,2}(P_{ij}) + M_{ik}M_{jk}e_{k,3}(P_{ij})\right)}{\sum_{k=1}^{n} \left(M_{ik}M_{kj} + M_{kj}M_{ki} + M_{ik}M_{jk}\right)} \tag{15}$$

where M is a *mask matrix* so that $M_{i,j} = 1$ if $P_{i,j}$ is defined while $M_{i,j} = 0$ otherwise. It is possible that no estimators are available at all for some $P_{i,j}$. In such cases the $e(P_{i,j})$ returns the indefinite form $0/0$ so it remains undefined.

The generation of missing values is done in several iterations. In each, all possible new values are generated relying on defined ones. If in a given iteration no new values are generated, then the process stops. For example, by applying the equations (12), (13), (14), (15) to the sample FPR reported in (10), we obtain the following expanded version of P^1:

$$P^1 = \begin{pmatrix} 0.50 & - & - & - & - & - \\ - & 0.50 & - & 0.15 & 0.50 & 0.65 \\ - & - & 0.50 & - & - & - \\ - & 0.85 & - & 0.50 & 0.85 & 1.00 \\ - & 0.50 & - & 0.15 & 0.50 & 0.65 \\ - & 0.35 & - & 0.00 & 0.35 & 0.50 \end{pmatrix}. \tag{16}$$

Given n students and m assignments per student, this expansion step defines a fraction of $(m^2 - m + n)/n^2$ elements of P^j for any student s_j.

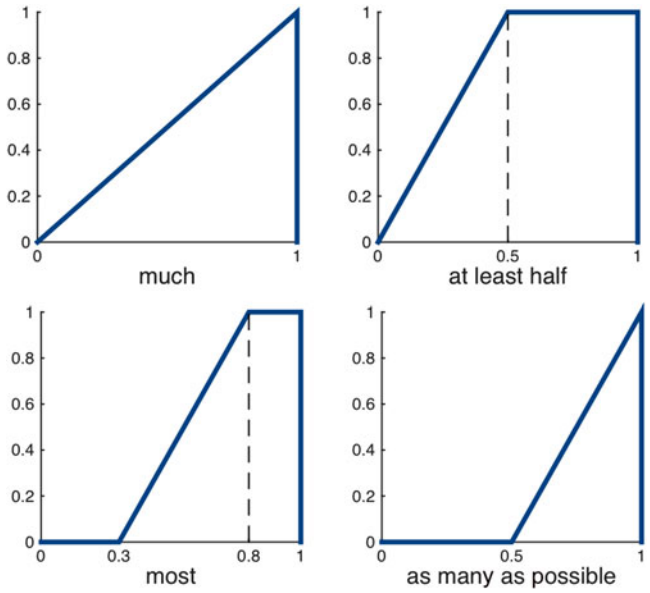

Fig. 1. Example of proportional fuzzy quantifiers.

5.3 Aggregation of Preference Relations

After all individual FPRs have been collected and partially expanded, an aggregation step is needed to build the collective FPR. As reported in Section 4.2 OWA is a widely used aggregator for FPRs, provided that a feasible weight vector is selected. To assign a specific behaviour to OWA, in [24] it has been proposed to initialize the weight vector starting from a non-decreasing proportional fuzzy quantifier.

Fuzzy quantifiers are defined in [30] as imprecise representations of the amount of items satisfying a given predicate. A *proportional fuzzy quantifier* Q is as fuzzy subset of the unit interval [0, 1] where, for any $x \in [0, 1]$, $\mu_Q(x)$ represents the degree to which the proportion x is compatible with the meaning of the quantifier it represents. A *non-decreasing proportional fuzzy quantifier* satisfies the additional property that $\mu_Q(x_1) \geq \mu_Q(x_2)$ for every x_1 and x_2 so that $x_1 > x_2$. Examples of non-decreasing proportional fuzzy quantifier are *much*, *at least half*, *most* and *as many as possible* (see Fig. 1).

The membership function of a *non-decreasing proportional fuzzy quantifier* can be represented through the following equation:

$$\mu_Q(x) = \begin{cases} 0 & \text{if } x < a, \\ \frac{x-a}{b-a} & \text{if } a \leq x \leq b, \\ 1 & \text{if } x > b. \end{cases} \tag{17}$$

with $a, b, x \in [0,1]$ e.g., the parameters (a, b) of the quantifiers shown in Fig. 1 are: (0, 1), (0.5, 1), (0.3, 0.8) and (0.5, 1) respectively. Starting from the selected quantifier, the weights of an OWA operator of dimension n can be initialized according to the following expression [23]:

$$w_k = \mu_Q\left(\frac{k}{n}\right) - \mu_Q\left(\frac{k-1}{n}\right); \quad k \in \{1, \ldots, n\}. \tag{18}$$

After having selected a *non-decreasing proportional fuzzy quantifier* Q and calculated the weight vector according to equations (17), (18), given a set of individual FPRs $\{P^1, \ldots, P^n\}$, the global preference $P_{i,j}$ between every pair of alternatives can be obtained as $P_{i,j} = OWA_Q(P^1_{i,j}, \ldots, P^n_{i,j})$, where OWA_Q indicates the OWA operator initialised with the weights coming from the quantifier Q.

When dealing with incomplete FPRs, as in our case, it is needed to exclude undefined elements from the computation by slightly reviewing the summation in equation (5) and combining it with equation (18) so obtaining elements of the collective FPR as follows:

$$P_{i,j} = \sum_{k=1}^{|def(P_{i,j})|} \left(\mu_Q\left(\frac{k}{|def(P_{i,j})|}\right) - \mu_Q\left(\frac{k-1}{|def(P_{i,j})|}\right)\right) \cdot y_k(def(P_{i,j})); \tag{19}$$

for $i, j \in \{1, \ldots, n\}$, where $\mathrm{def}(P_{i,j})$ indicates the subset of defined elements in $\{P^1_{i,j}, \ldots, P^n_{i,j}\}$ while $y_k(\mathrm{def}(P_{i,j}))$ is the k-th largest defined value in the same set.

For example, let P^2 and P^3 be individual FPRs generated and expanded, respectively, from the ranking strings $R_2 = (s_1 \geq s_6 \approx s_5 \geq s_3)$ and $R_3 = (s_4 > s_1 \geq s_5 > s_6)$:

$$P^2 = \begin{pmatrix} 0.50 & - & 0.65 & - & 0.58 & 0.58 \\ - & 0.50 & - & - & - & - \\ 0.35 & - & 0.50 & - & 0.43 & 0.43 \\ - & - & - & 0.50 & - & - \\ 0.43 & - & 0.58 & - & 0.50 & 0.50 \\ 0.43 & - & 0.58 & - & 0.50 & 0.50 \end{pmatrix}$$
$$P^3 = \begin{pmatrix} 0.50 & - & - & 0.35 & 0.58 & 0.73 \\ - & 0.50 & - & - & - & - \\ - & - & 0.50 & - & - & - \\ 0.65 & - & - & 0.50 & 0.73 & 0.88 \\ 0.43 & - & - & 0.28 & 0.50 & 0.65 \\ 0.28 & - & - & 0.13 & 0.35 & 0.50 \end{pmatrix} \tag{20}$$

the collective FPR P obtained by aggregating P^1 (shown in equation (16)) with P^2 and P^3, through OWA initialised with the quantifier *most* (see Fig. 1), is shown below:

$$P = \begin{pmatrix} 0.50 & - & 0.65 & 0.35 & 0.58 & 0.64 \\ - & 0.50 & - & 0.15 & 0.50 & 0.65 \\ 0.35 & - & 0.50 & - & 0.43 & 0.43 \\ 0.65 & 0.85 & - & 0.50 & 0.78 & 0.93 \\ 0.43 & 0.50 & 0.58 & 0.20 & 0.50 & 0.61 \\ 0.34 & 0.35 & 0.58 & 0.05 & 0.36 & 0.50 \end{pmatrix}. \tag{21}$$

5.4 Expansion of Collective Preferences

After having aggregated individual preferences as described in Section 5.3, it could happen that some values of the collective FPR P still remain undefined. In fact when none of the assessor students has expressed a preference for a given couple of submissions i and j, then the corresponding values $P_{i,j}$ and $P_{j,i}$ of the collective FPR can't be calculated. In most cases it does suffice to execute the expansion algorithm described in Section 5.2 again on the collective FPR to obtain a fully defined FPR.

Nevertheless, in some cases, especially for $n \gg m$ and when some student provides a partial ranking string (or some student does not provide preferences at all), some elements of P can't be calculated even after the additional expansion step described so far. It is the case when for some couple of values $i, j \in \{1, \ldots, n\}$ it does not exist a $k \in \{1, \ldots, n\}$ so that at least one of the couples $(P_{i,k}, P_{k,j})$, $(P_{k,i}, P_{k,j})$ or $(P_{i,k}, P_{j,k})$ is fully defined. In [31] authors refer to this case as an *ignorance situation* and suggest an approximate technique to estimate missing values.

The technique uses some seed values to initialize the estimation process that is based, again, on the additive consistency property. In accordance with this approach, we first assume indifference for any undefined value by setting $P_{i,j} = 0.5 \ \forall \ i, j \mid M_{i,j} = 0$, where $M_{i,j}$ has the same meaning as in equation (15). Then, we apply again the estimators defined by equations (12), (13), (14), and (15) to define consistent values for them. In particular, given that in this case all elements of P result as defined (with actual or seed values), the equation (15) can be simplified as follows for each $i, j \mid M_{i,j} = 0$:

$$e(P_{i,j}) = \frac{\sum_{k=1}^{n} \big(e_{k,1}(P_{i,j}) + e_{k,2}(P_{i,j}) + e_{k,3}(P_{i,j})\big)}{3n}. \tag{22}$$

For example, by applying the expansion steps defined in this section to the collective FPR defined in (21), the following complete version of P is obtained:

$$P = \begin{pmatrix} 0.50 & 0.59 & 0.65 & 0.35 & 0.58 & 0.64 \\ 0.41 & 0.50 & 0.65 & 0.15 & 0.50 & 0.65 \\ 0.35 & 0.35 & 0.50 & 0.11 & 0.43 & 0.43 \\ 0.65 & 0.85 & 0.89 & 0.50 & 0.78 & 0.93 \\ 0.43 & 0.50 & 0.58 & 0.20 & 0.50 & 0.61 \\ 0.34 & 0.35 & 0.58 & 0.05 & 0.36 & 0.50 \end{pmatrix}. \tag{23}$$

5.5 Global Ranking and Absolute Grades Calculation

Once the collective FPR has been obtained, it is possible to calculate an absolute degree of preference $\phi(s_i)$ for each submission $s_i \in S$ according to one of the measures defined in Section 4.3. For example, by applying the equation (8) to the sample collective FPR shown in (23), we obtain the following preference degree vector:

$$\phi_{NF}(S)^T = (0.63, \ -0.28, \ -1.68, \ 3.23, \ -0.33, \ -1.58); \tag{24}$$

The global ranking between the alternatives is then computed by ordering all the submission in decreasing order of their preference degree. In the case of equation (24), the final ranking over S is:

$$s_4 \succ s_1 \succ s_2 \succ s_5 \succ s_6 \succ s_3 \tag{25}$$

Starting from the obtained preference degrees it is also possible to calculate the absolute grade of each submission, provided that an ordinal assessment is made by a reliable expert (e.g., the teacher) to the best and the worst submissions (i.e., the first and the last in the final ranking). Let be e_{min} and e_{max} the grades assigned to the best and the worst submission, the estimated grade $\tilde{e}(s_i)$ for every student $s_i \in S$ can be obtained through normalization as follows:

$$\tilde{e}(s_i) = \frac{(\phi(s_i) - \phi_{min}) \cdot (e_{max} - e_{min})}{(\phi_{max} - \phi_{min})} + e_{min} \tag{26}$$

where ϕ_{min} and ϕ_{max} are the minimum and the maximum preference degrees in $\phi(S)$. By applying the equation (25) on the sample data in (24) with $e_{min} = 2$ and $e_{max} = 9$, we obtain the following grades vector:

$$\tilde{e}(S)^T = (5.3,\ 4.0,\ 2.0,\ 9.0,\ 3.9,\ 2.2) \tag{27}$$

that assigns an absolute grade to each submission.

6 Experiments

To demonstrate the effectiveness of FOPA and to compare it with different approaches, we have performed several experiments with synthetic data. In all the experiments, 100 students are supposed to have submitted a solution to a given assignment. The submission of each student s_i has a *true grade* $e(s_i)$ belonging to [0, 10] assigned according to a normal distribution: $e(s_i) \sim \mathcal{N}(6,\ 2)$ so centred in 6 with a standard deviation of 2.

Each student has then to evaluate the submissions of m peers (with m constant or variable according to the specific experiment) matching an assessment grid G defined as specified in equation (1). Students are imperfect graders so, according to [7], we have modelled such imperfection with two parameters:

- a *bias* term $b \geq 0$ that reflects a tendency of an assessor student to either inflate or deflate her assessment (i.e., high biases describe lenient assessors while low biases describe stringent ones);
- an *unreliability* term $u \geq 0$ that reflects how far, on average, a grader's assessment tends to land with respect to the corresponding true grade (i.e., a low unreliability describes a proper attitude to distinguish between good and bad submissions).

Basing on these two parameters, the *perceived grade* $\tilde{e}_j(s_i)$ of the student s_i from the assessor student s_j, is defined according to the following probability distributions:

$$\tilde{e}_j(s_i) \sim \mathcal{N}\big(e(s_i) + b_j,\ u\big) \text{ so that } b_j \sim \mathcal{N}(0, b). \tag{28}$$

Each student s_j is supposed to define a *ranking* $\succ_j$ over S_j (the set of peers to be evaluated) by ordering peers decreasingly on their perceived grades. Such ranking is changed in a *ranking string* $R_j = (s_{i1}\ \sigma_1\ s_{i2}\ \sigma_2 \ldots\ s_{im-1}\ \sigma_{m-1}\ s_{im})$ adding ranking symbols according to the difference between the perceived grades of two subsequent submissions. To do that we have used the following empirical rule:

$$\sigma_k = \begin{cases} \approx & \text{if } \tilde{e}_j(s_{ik}) - \tilde{e}_j(s_{ik+1}) < 0.5; \\ \geq & \text{if } 0.5 \leq \tilde{e}_j(s_{ik}) - \tilde{e}_j(s_{ik+1}) < 1; \\ > & \text{if } 1 \leq \tilde{e}_j(s_{ik}) - \tilde{e}_j(s_{ik+1}) < 2; \\ \gg & \text{if } \tilde{e}_j(s_{ik}) - \tilde{e}_j(s_{ik+1}) \geq 2. \end{cases} \tag{29}$$

We have observed that FOPA is quite insensitive to the exact values adopted as thresholds for symbols selection, provided that an adequate distance is left between subsequent values. For this reason, we have chosen thresholds that can be easily understood by assessor students.

It is worth noting that, while the first three symbols are applied in fairly narrow ranges, $\gg$ is applied to a virtually larger range (2,10]. However, since each student grades several submissions, the application range of $\gg$ is, in practice, much narrower. For this reason, and also thanks to the FPR aggregation step, the selection of the application range for $\gg$ does not cause any flattening of the grades to the mean.

Starting from synthetic data generated in this way, the global ranking and the absolute grades have been estimated according to the methodology defined in Section 5 and compared to true grades (and related rankings). This has allowed us to determine the performances of FOPA in revealing the ground truth also in presence of noisy data (taking into account different values for bias and reliability) and in comparisons to other methods (described in Section 3) as well as to cardinal peer grading. The details and the results of the performed experiments are discussed in the next sections.

6.1 Experiment 1: Optimal Parameters Setting

This experiment is aimed at discovering the best settings for the parameters used by FOPA. This is done by measuring the performances obtained with respect to global ranking both in case of perfect grading (i.e., when students make no errors when assessing other students) that in the more realistic case of imperfect one. The results obtained with different settings are then compared to discover the most promising settings to be used in next experiments.

According to Sections 5.5 and 4.3, the global ranking among submissions can be obtained using several measures, namely the *Quantifier Guided Dominance Degree*, the *Quantifier Guided Non-Dominance Degree* and the *Net Flow* (NF). Given that, a first parameter to set is the ranking measure to adopt, i.e., the one that offers the best performances for the specific problem. Moreover, according to Section 5.3, the aggregation of alternatives based on OWA can be done starting from several fuzzy quantifiers like *much*, *at least half*, *most* and *as many as possible*. Another parameter to set is so the quantifier to apply.

To identify which setting offers the best performances, we have executed the experiment described so far with 100 students and four assignments to be evaluated by each (so $m = 4$). When generating perceived grades, we have set $b = 0$ and u ranging from 0 (perfect grading) to 3 (average difference of 3 between the true grades and the perceived ones). For each value assigned to u we have repeated the experiment 1,000 times and mediated the obtained results in terms of *Percentage of Correctly Recovered Pairwise Relations* (PCRPR) according to the following equation:

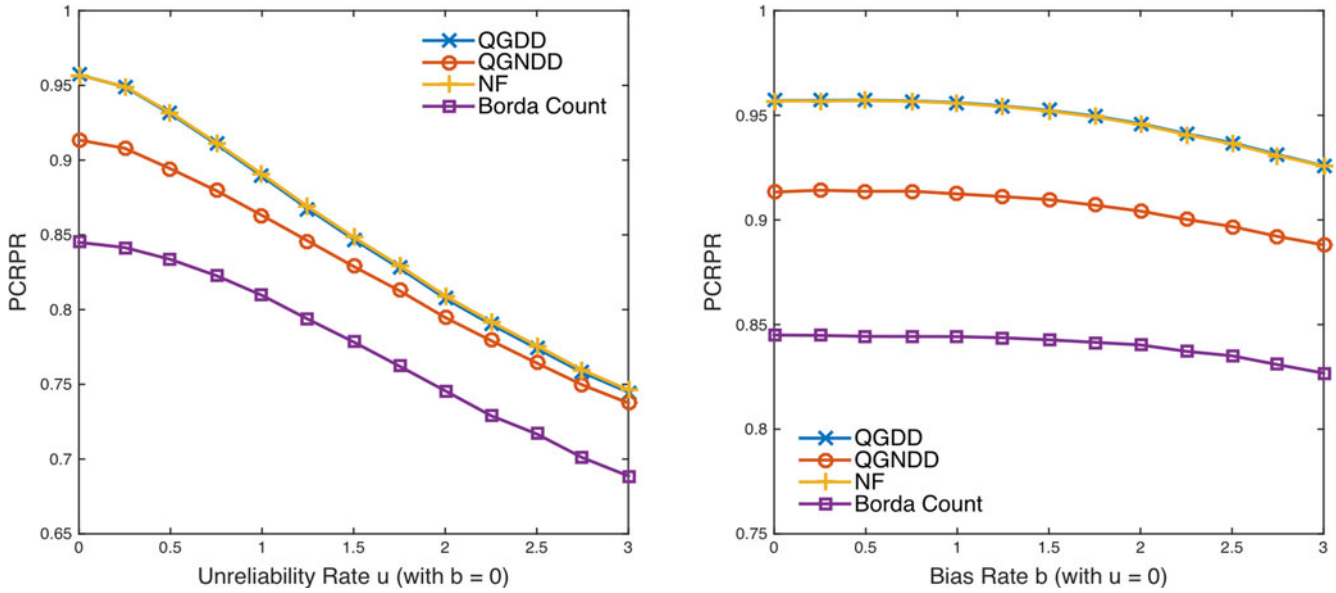

Fig. 2. Performances of the QGDD, QGNDD, and NF ranking measures compared with the Borda count in terms of PCRPR (higher is better).

$$PCRPR = 2 \cdot \frac{\left|\left\{(s_i, s_j) \mid s_i, s_j \in S;\ i \neq j;\ s_i \succ s_j;\ s_i \tilde{\succ} s_j\right\}\right|}{n \cdot (n-1)} \tag{30}$$

where $\succ$ and $\tilde{\succ}$ are, respectively, the real and the estimated rankings among the students belonging to S and n is the number of students, i.e., the cardinality of S. Then, we have repeated the process by setting $u = 0$ and b ranging from 0 (no bias at all) to 3 (average bias of 3).

The Fig. 2 plots the results in terms of PCRPR, obtained by FOPA, changing the applied ranking measure among QGDD, QGNDD and NF, against the unreliability rate u (on the left) and the bias rate (on the right). The plots show that, among the available measures, two obtain the best performances with any value of u and b: QGDD and NF. In case of perfect grading (i.e., when $u = b = 0$), they show a PCRPR of 95.7 percent, that is far beyond the 84.5 percent obtained by the Borda count. Both measures demonstrate a fair robustness to unreliability but, the improvement with respect to the Borda count, decreases when u increases. Moreover, it should be noted that all the methods are very robust with respect to the bias with average variations of less than 1 percent in terms of PCRPC for each increase of 1 grade in bias. Nevertheless, this is a common advantage of ordinal grading methods.

On the other hand, FOPA results to be insensitive with respect to the selection of the OWA quantifier for the aggregation step: the same results are in fact obtained regardless of the adopted one. The same level of insensitivity has been also detected by changing the fuzzy quantifier adopted within the QGDD and QGNDD measures. For this reason, the results obtained changing the quantifier are not plotted.

6.2 Experiment 2: Comparison with Other Ordinal Peer Assessment Methods

This experiment is aimed at measuring the performances of FOPA with respect to the other methods for ordinal peer assessment described in Section 3 in case of perfect and imperfect grading. To do that, we have executed the experiment described in 6 with 100 students and four assignments to be evaluated by each. When generating perceived grades, we have set $b = 0$ and u ranging from 0 to 3. For each value assigned to u we have repeated the experiment 1,000 times and mediated the obtained results in terms of PCRPR, calculated according to equation (30).

Then, for each iteration and experimented method, the obtained scores have been transformed in grades through the equation (26), setting e_{min} and e_{max} equal, respectively, to the minimum and the maximum true grade. Then the *Root Mean Square Error* (RMSE) between the grades estimated through each experimented method and the true grades have been calculated according to the following equation and mediated over the 1,000 iterations for each value of u:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} \left(e(s_i) - \tilde{e}(s_i)\right)^2}{n}} \tag{31}$$

where $e(s_i)$ and $\tilde{e}(s_i)$ are, respectively, the real and the estimated grades assigned to a student $s_i \in S$ and $n = |S|$.

The Fig. 3 (on the left) plots the results in terms of PCRPR, obtained by FOPA (adopting the *Net Flow* aggregation measure) compared with the models of *Mallow* (MAL), *Bradley-Terry* (BT), *Plackett-Luce* (PL) and *Borda*. An additional model named *Score-Weighted Mallows* (MALS) defined in [11] as an improved version of the Mallow model has been also tested. The same figure (on the right) plots the results in terms of RMSE of the same models after having transformed the scores in grades as described so far. To experiment the methods described in [11] we have used the *PeerGrader* software[1] made publicly available by the authors. Differently from Borda, these methods also deal with the case in which an assessor expresses indifference between submissions. According to equation (29) we have declared indifference between s_i and s_j when $|\tilde{e}_j(s_i) - \tilde{e}_j(s_j)| < 0.5$.

1. www.peergrading.org

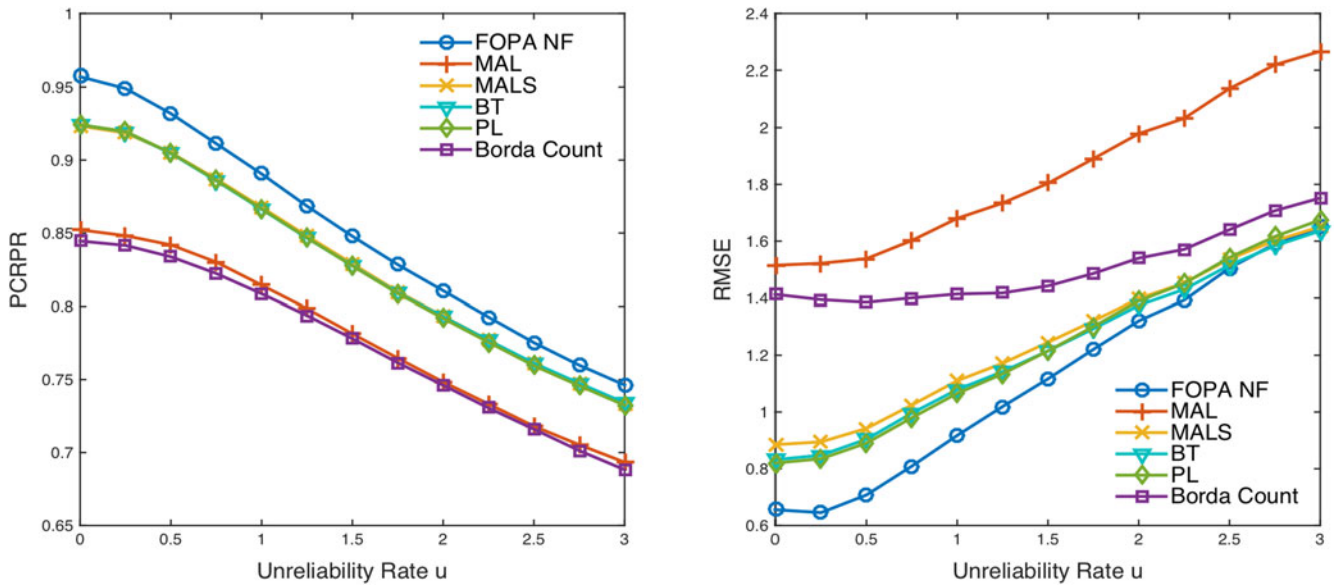

Fig. 3. Performances of FOPA against MAL, MALS, BT, and PL in terms of PCRPR (higher is better) and RMSE (lower is better).

Among the introduced methods, MALS, BT and PL show similar PCRPR values while PL performs a little better than the other two in terms of RMSE, at least with $u < 1.5$. The performance of MAL are worst and comparable with those of Borda in terms of PCRPR while, with respect to RMSE, MAL reaches a higher error rate even with small unreliability rates. Nevertheless, it should be noted that, as explained in [11], MAL (as Borda) are not conceived for obtaining cardinal grades and this is the reason why the authors have improved MAL defining MALS.

The plots show that FOPA outperforms the other methods both in terms of PCRPS that in terms of RMSE. When considering PCRPC, FOPA gains about 4 percent against MALS, BT and PL in case of perfect grading (from 92.4 to 95.7 percent) but the improvement decreases when u increases until about 2 percent for $u = 3$ (from 73.2 to 74.6 percent). When considering RMSE, FOPA is able to lower the mean error of about 0.2 grades in case of perfect grading (from 0.82 of PL to 0.65 of FOPA) while this difference tends to nullify when increasing the unreliability until $u = 3$.

6.3 Experiment 3: Comparison with Cardinal Peer Assessment

This experiment is aimed at measuring the performances of FOPA (and some other ordinal approaches) in comparison to *Cardinal Peer Assessment*. In CPA, any assessor student s_j directly proposes, for any assessee $s_i \in S_j$, a cardinal grade equal to the perceived grade $\tilde{e}_j(s_i)$ defined by equation (28). The final estimated grade $\tilde{e}(s_i)$ for each s_i is then obtained by averaging all grades proposed by the peers with the following equation:

$$\tilde{e}(s_i) = \frac{1}{m} \sum_{j:\ s_i \in S_j} \tilde{e}_j(s_i). \tag{32}$$

To compare FOPA and CPA we have executed the experiment described in 6 with 100 students and four assignments to be evaluated by each. When generating perceived grades, we have considered both b and u ranging from 0 to 3. For each values setting, we have repeated the experiment 1,000 times and mediated the obtained results in terms of RMSE, calculated according to equation (31).

The Fig. 4 plots the results in terms of RMSE, obtained by FOPA (with *Net Flow*), by the *Plackett-Luce* method (PL), by Borda and by CPA while ranging the bias rate from 0 to 3. The plot on the left considers the case when assessor students are perfectly reliable ($u = 0$) while the plot on the right considers a moderate level of unreliability ($u = 1$). As it can be seen, CPA is very sensitive to the bias rate compared with ordinal approaches. In both cases CPA introduces a lower error with respect to FOPA until the bias rate reaches a given threshold, variable according to the unreliability rate (about 1.4 for $u = 0$, about 1.7 for $u = 1$). After the threshold, the gap in term of RMSE between CPA and FOPA increases until a difference of about 0.60 grades for $u = 0$ and $b = 3$ and about 0.43 grades for $u = 1$ and $b = 3$. It is worth noting that, in all cases, FOPA outperforms the other ordinal methods.

To provide a comprehensive view of the behaviour of FOPA and CPA, the Fig. 5 plots the 3d surfaces of the RMSE curves obtained ranging u and b from 0 to 3. Clearly the error level in FOPA mainly depends on the unreliability rate, while the error in CPA quite evenly depends on the unreliability and the bias rates. With medium-low bias and medium-high unreliability rates, CPA is a little better than FOPA. Conversely, with medium-high bias and medium-low unreliability, FOPA is quite better than CPA.

It is worth noting that CPA requires, by each assessor student, an amount of information significantly higher with respect to ordinal approaches. Given this complexity, as shown in Section 1, in real contexts cardinal feedback is less reliable with respect to the ordinal one, even when assessors are at the same level of knowledge and experience. In light of this, the performed experiment ultimately benefits CPA because it assumes, for each iteration, the same level of bias and unreliability between cardinal and ordinal feedback. Nevertheless, the performances obtained by FOPA are comparable and in some cases better than those obtained by CPA.

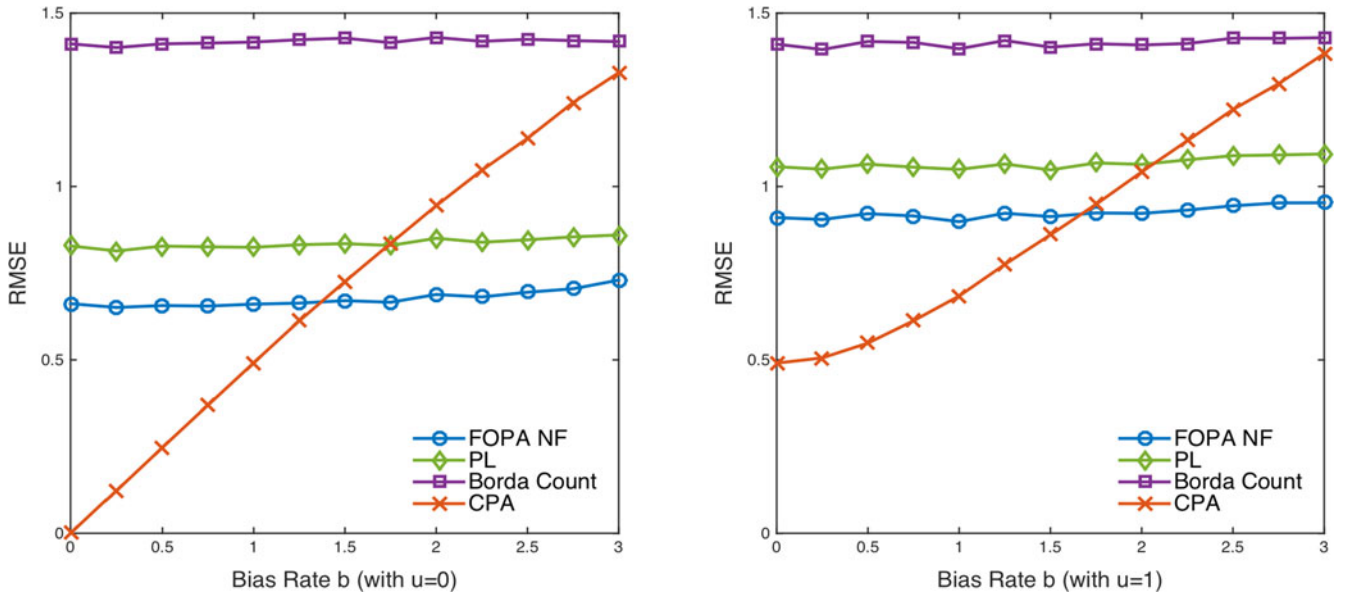

Fig. 4. Performances of FOPA, PL, and Borda against cardinal peer assessment in terms of RMSE (lower is better) when $u = 0$ (left) and $u = 1$ (right).

6.4 Experiment 4: Selection of the Number of Assessors

The number m of submissions that each student has to evaluate is one of the main parameters that must be defined to setup a peer assessment session. On one hand, this number should be kept as small as possible to avoid overloading the students, with the risk that they do not respond adequately to the exercise providing rough, partial or void estimations. On the other hand, according to the definition of assessment grid provided in 3, this number corresponds to the number of assessors for each submission. In this respect, m should be kept as big as possible to have sufficient information to estimate the final ranking and grades.

To determine how the selection of m impacts on the performance of FOPA, we have executed the same experiment described so far with 20 and 200 students and a number of assignments to be evaluated by each student variable from 2 to 20. When generating perceived grades, we have set $b = 0$ (the previous experiments have shown that FOPA is insensitive to the bias) and u variable from 0 (perfect grading) to 3. For each setting we have repeated the experiment 1,000 times and mediated the obtained results in terms of RMSE, calculated according to equation (31).

The Fig. 6 (left) plots the results obtained by FOPA (with *Net Flow*) with 20 students and m ranging from 2 to 20. A first thing to observe is that, while for high unreliability rates ($u \geq 2$) an increase of m always determines a decrease of the whole error level, for low unreliability rates ($u < 2$) an increase of m determines a decrease of the RMSE only until a given threshold. After the threshold, adding more assessors, results in an increase in the RMSE. This can be explained by the fact that, while using ranking strings for assessing the submissions, a noise is introduced in the model (in fact, ranking strings can be seen as approximated FPRs). Such noise increases when the strings length increases (so when m increases) but it is balanced by the additional information obtained with more assessors.

In the (unrealistic) case of perfect grading (when $u = 0$), all assessors have exactly the same perception of the student grades so, after a given threshold, adding more assessors does not increase the quantity of available information until the extreme case of $m = n$, when all the assessor students provide exactly the same ranking string. So in these cases the noise introduced by ranking string approximation remains unbalanced and the error increases. This is evenly true in settings with low unreliability rates ($u < 2$) and with more students to evaluate (Fig. 6, right) even if the threshold becomes higher and higher.

With respect to the selection of m, it should be noted that, apart the unrealistic case where $u = 0$, the curves plotted on the left and on the right side of Fig. 6 have a similar trend. Regardless of the number of students and of the unreliability rate u, we notice a steep decrease of the RMSE while moving from two to three assessors and a smoother decrease for subsequent values of m. By looking at the right part of the plots we see that, when $u = 1$, the RMSE start to increase for $m > 16$ while, even for $u > 1$, the decrease in RMSE obtained adding a new assessor is less then 0.02 grades. Such reflections suggest to select a number m of submissions to be assessed per student so that $3 \leq m \leq 16$

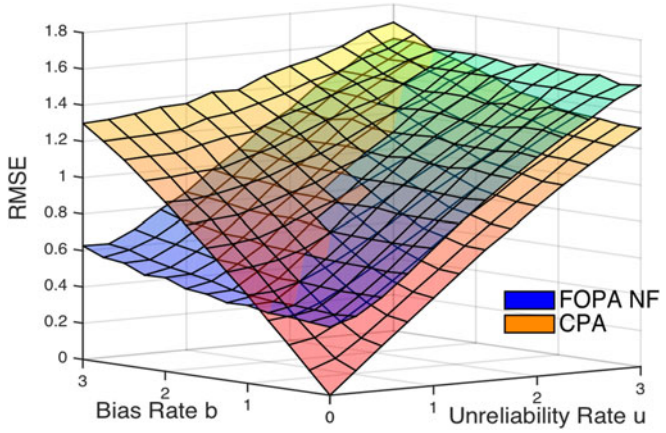

Fig. 5. Performances of FOPA and cardinal peer assessment in terms of RMSE (lower is better) ranging both the bias and unreliability rates.

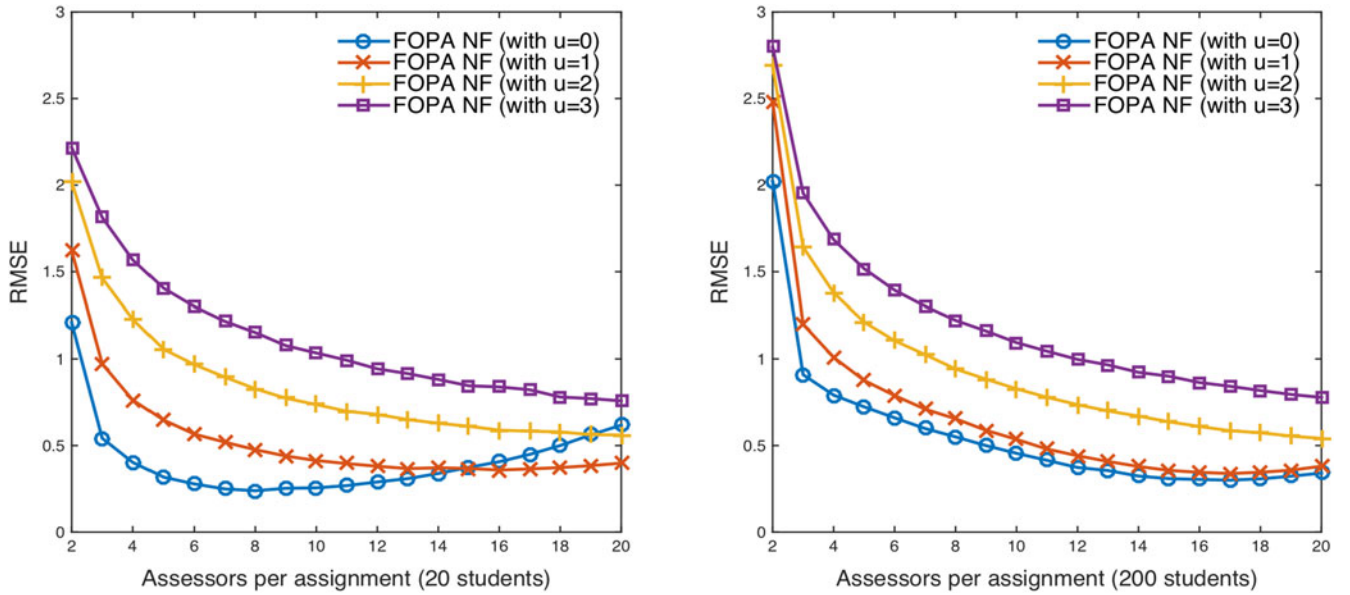

Fig. 6. Performances of FOPA in terms of RMSE (lower is better) with different values for u, ranging m from 2 to 20, with $n = 20$ (left) and $n = 200$ (right).

regardless of the total number of students involved and on the expected degree of unreliability and bias.

7 Final Remarks

In this paper a model for ordinal peer assessment has been defined as a special case of a GDM problem and solved by adopting a fuzzy-based approach. The defined model has been compared with other existing ordinal and cardinal peer assessment models and has shown better performances (both in the reconstruction of the real ranking that in the estimation of students' real grades) in several experiments with synthetic data generated from realistic probability distributions.

In order to hide the complexity of the model based of FPRs, students are asked to specify simple ranking strings that order submissions from the best to the worst and specify, with a set of intuitive symbols, at what extent each submission is better than the following one in the ranking. A natural extension of this approach is to use linguistic labels (mapped on fuzzy numbers) rather then symbols (mapped on fuzzy values) to specify the preferences between two subsequent submissions in the partial ranking provided by each student. This would complicate the model a bit but may enable a better representation of the vagueness inherent in the assessments made by students. Specific approaches to fuzzy GDM based on linguistic assessment, like in [32], could be adapted to deal with this case.

Another feasible extension to the proposed model is to integrate techniques coming from cardinal peer assessment, like those described in [4], [7], [8] and [10] to detect the reliability of each grader and use this data to weight the effect of the feedback provided by that student in the aggregation step. To this purpose, additive weighting aggregators (like SAW) should be preferred to OWA.

Further extensions can be conceived by seeing the problem not just as a GDM problem but as a *Consensus Process*: a negotiation process developed iteratively and composed by several rounds, where the experts (in this case the assessor students) accept to change their preferences following the advice provided by the model. According to the approach described in [33], consensus measures can be calculated on FPRs to identify the level of agreement between the different opinions on every alternative (submission). If such measures are over a given threshold, then FPRs are aggregated and alternatives are ranked as seen in Section 4. Otherwise, a set of suggestions are generated for the experts (assessor students) to solicit the modification of some preferences in order to improve the overall consensus.

The application of a consensus process for ordinal peer grading requires an additional step for advise generation but is capable of improving the overall results of the evaluation by forcing students to reflect on proposed rankings if they strongly disagree with the preferences expressed by a fuzzy majority of peers. On the other hand, it makes the peer grading process slower because additional "refining" steps are needed to improve provided rankings. So, the suitability of such technique, especially in massive contexts, should be carefully evaluated.

Finally, it is worth noting that, despite that it has been conceived for peer assessment in MOOCs, FOPA can be easily adapted in other contexts where alternatives must be evaluated taking into account the opinion of several assessors but each assessor has only a partial view of the whole picture. For example, in a *Conference Review Process* many submissions must be ranked (to chose the best ones to invite for presentation and/or to be awarded) basing on a set of (possibly unreliable) experts, each of them reviewing just a relatively small number of submissions of the whole set. Other examples are the *Employee Reward and Recognition Systems* set up by companies to motivate their best employees. Here employee performances must be ranked according to suggestions coming from managers, each of them evaluating just the subset of employees involved in projects he manages.

REFERENCES

[1] D. G. Glance, M. Forsey, and M. Riley, "The pedagogical foundations of massive open online courses," *First Monday*, vol. 18, no. 5, 2013, http://firstmonday.org/article/view/4350/3673.

[2] I. Caragiannis, G. A. Krimpas, and A. A. Voudouris, "Aggregating partial rankings with applications to peer grading in massive online open courses," in *Proc. 14th Int. Conf. Auton. Agents Multi-agent Syst.*, 2015, pp. 675–683.

[3] C. Alario-Hoyos, M. Perez-Sanagust, C. Delgado-Kloos, H. A. Parada, and M. Munoz-Organero, "Delving into participants' profiles and use of social tools in MOOCs," *IEEE Trans. Learn. Technol.*, vol. 7, no. 3, pp. 260–266, Jul.-Sep. 2014.

[4] N. Capuano and S. Caballé, "Towards adaptive peer assessment for MOOCs," in *Proc. 10th Int. Conf. P2P, Parallel, GRID, Cloud Internet Comput.*, 2015, pp. 64–69.

[5] L. Bouzidi and A. Jaillet, "Can online peer assessment be trusted?," *Educ. Technol. Soc.*, vol. 12, no. 4, pp. 257–268, 2009.

[6] P. A. Carlson and F. C. Berry, "Calibrated peer review and assessing learning outcomes," in *Proc. 33rd Int. Conf. Frontiers. Edu.*, 2003, vol. 2, pp. F3E-1–F3E-6.

[7] C. Piech, J. Huang, Z. Chen, C. B. Do, A. Y. Ng, and D. Koller, "Tuned models of peer assessment in MOOCs," in *Proc. 6th Int. Conf. Edu. Data Mining*, 2013, pp. 153–160.

[8] I. M. Goldin, "Accounting for peer reviewer bias with Bayesian models," presented at the 11th Int. Conf. Intelligent Tutoring Systems, Chania, Greece, 2012.

[9] M. Uto and M. Ueno, "Item response theory for peer assessment," *IEEE Trans. Learn. Technol.*, no. 1, p. 1, doi:10.1109/TLT.2015.2476806.

[10] T. Walsh, "The peerrank method for peer assessment," in *Proc. 21st European Conf. Artificial Intelligence*, 2014, pp. 909–914.

[11] K. Raman and T. Joachims, "Methods for ordinal peer grading," in *Proc. 20th SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2014, pp. 1037–1046.

[12] J. Ma and D. Zhou, "Fuzzy set approach to the assessment of student-centred learning," *IEEE Trans. Edu.*, vol. 43, no. 2, pp. 237–241, May 2000.

[13] C. H. Lan, S. Graf, K. R. Lai, and Kinshuk, "Enrichment of peer assessment with agent negotiation," *IEEE Trans. Learn. Technol.*, vol. 4, no. 1, pp. 35–46, Mar. 2011.

[14] K. C. Chai, K. M. Tay, and C. P. Lim, "A new fuzzy peer assessment methodology for cooperative learning of students," *Appl. Soft Comput.*, vol. 32, pp. 468–480, 2015.

[15] J. C. Borda, "Memoire sur les elections au scrutin," Histoire de l'Académie Royale des Sciences, pp. 657–664, 1781, http://asklepios.chez.com/XIX/borda.htm.

[16] A. Abdulkadiroglu and T. Sonmez, "Random serial dictatorship and the core from random endowments in house allocation problems," *Econometrica*, vol. 66, no. 3, p. 689, 1998.

[17] C. Dwork, R. Kumar, M. Naor, and D. Sivakumar, "Rank aggregation methods for the Web," in *Proc. 10th Int. Conf. World Wide Web*, 2001, pp. 613–622.

[18] C. L. Mallows, "Non-null ranking models. I," *Biometrika*, vol. 44, no. 1, p. 114, Jun. 1957.

[19] R. A. Bradley and M. E. Terry, "Rank analysis of incomplete block designs: I. The method of paired comparisons," *Biometrika*, vol. 39, no. 3, p. 324, 1952.

[20] R. L. Plackett, "The analysis of permutations," *Appl. Statist.*, vol. 24, no. 2, p. 193, 1975.

[21] L. A. Zadeh, "Fuzzy sets," *Information and Control*, vol. 8, no. 3, pp. 338–353, Jun. 1965.

[22] Y.-M. Wang and Z.-P. Fan, "Fuzzy preference relations: Aggregation and weight determination," *Comput. Ind. Eng.*, vol. 53, no. 1, pp. 163–172, 2007.

[23] R. R. Yager, "Families of OWA operators," *Fuzzy Sets Syst.*, vol. 59, no. 2, pp. 125–148, 1993.

[24] F. Chiclana, F. Herrera, and E. Herrera-Viedma, "Integrating three representation models in fuzzy multipurpose decision making based on fuzzy preference relations," *Fuzzy Sets Syst.*, vol. 97, no. 1, pp. 33–48, Jul. 1998.

[25] F. Chiclana, F. Herrera, E. Herrera-Viedma, and L. Martinez, "A note on the reciprocity in the aggregation of fuzzy preference relations using OWA operators," *Fuzzy Set. Syst.*, vol. 137, no. 1, pp. 71–83, Jul. 2003.

[26] E. Herrera-Viedma, F. Chiclana, F. Herrera, and S. Alonso, "Group decision-making model with incomplete fuzzy preference relations based on additive consistency," *IEEE Trans. Syst., Man Cybern., Part B (Cybern.)*, vol. 37, no. 1, pp. 176–189, Feb. 2007.

[27] S. H. Kim, S. H. Choi, and J. K. Kim, "An interactive procedure for multiple attribute group decision making with incomplete information: Range-based approach," *Eur. J. Operational Res.*, vol. 118, no. 1, pp. 139–152, Oct. 1999.

[28] E. Herrera-Viedma, F. Herrera, F. Chiclana, and M. Luque, "Some issues on consistency of fuzzy preference relations," *Eur. J. Operational Res.*, vol. 154, no. 1, pp. 98–109, Apr. 2004.

[29] J. Ma, Z.-P. Fan, Y. P. Jiang, J.-Y. Mao, and L. Ma, "A method for repairing the inconsistency of fuzzy preference relations," *Fuzzy Sets Syst.*, vol. 157, no. 1, pp. 20–33, 2006.

[30] L. A. Zadeh, "A computational approach to fuzzy quantifiers in natural languages," *Comput. Math. Appl.*, vol. 9, no. 1, pp. 149–184, 1983.

[31] S. Alonso, E. Herrera-Viedma, F. Chiclana, and F. Herrera, "Individual and social strategies to deal with ignorance situations in multi-person decision making," *Int. J. Inf. Technol. Decision Making*, vol. 8, no. 2, pp. 313–333, 2009.

[32] F. Herrera, S. Alonso, F. Chiclana, and E. Herrera-Viedma, "Computing with words in decision making: foundations, trends and prospects," *Fuzzy Optimization Decision Making*, vol. 8, no. 4, pp. 337–364, 2009.

[33] E. Herrera-Viedma, S. Alonso, F. Chiclana, and F. Herrera, "A consensus model for group decision making with incomplete fuzzy preference relations," *IEEE Trans. Fuzzy Syst.*, vol. 15, no. 5, pp. 863–877, Oct. 2007.

Nicola Capuano received the academic qualification as an associate professor. He is a scientific officer at the University of Salerno. His research interests include computational intelligence, intelligent tutoring systems, and knowledge representation. He was a project manager and consultant for private companies and public organisations in research and development initiatives. He is author of about 100 research papers published in international journals, books, and conference proceedings. He serves as an editor and scientific referee for international journals and conferences.

Vincenzo Loia (SM'08) is a full professor of computer science at the University of Salerno. He is the author of more than 190 research papers in international journals, books, and conference proceedings. His research interests include soft computing and agent technology for technologically complex environments and Web intelligence applications. He is the coeditor-in-chief of *Soft Computing* and the editor-in-chief of *Ambient Intelligence and Humanized Computing*. He serves as editor for 14 other international journals. He has been the chair of the Emergent Technologies Technical Committee of the IEEE Computational Intelligence Society, where he is currently the chair of the task force on intelligent agents. He is a senior member of the IEEE.

Francesco Orciuoli is an associate professor in computer science at the University of Salerno. He has been involved in several European and Italian Research and Development projects about information and communication technologies and, in particular, on technology Enhanced learning, enterprise information management, and e-commerce. At the present, he is focusing his research activities on semantic technologies and computational intelligence. He is author of several scientific papers on these topics published on international journals and conference proceedings.