International Journal of Engineering Education Vol. 34, No. 3, pp. 968–980, 2018
Printed in Great Britain

0949-149X/91 $3.00+0.00

Metrics for Estimating Validity, Reliability and Bias in Peer Assessment*

RAFAEL MOLINA-CARMONA, ROSANA SATORRE-CUERDA, PATRICIA COMPAÑ-ROSIQUE and FARAÓN LLORENS-LARGO
Cátedra Santander-UA de Transformación Digital, University of Alicante, Ctra. San Vicente del Raspeig s/n, 03690, Alicante, Spain.
E-mail: [rmolina, rosana.satorre, patricia.company, faraon.llorens]@ua.es

Peer assessment is a widespread way of evaluating and rating the quality of a work in the field of education. Although it results to be a very effective learning instrument, it is subjected to possible problems of reliability, validity and some potential biases. Most works that study and try to solve these problems are focused on specific cases and the statistics for measuring reliability, validity or bias are global, that is, they give a measure of these values for the whole process, but they do not allow an individual study. In this work the approach is different. It proposes some metrics for reliability and validity of each reviewer, as well as an approximation to the possible biases that may appear in the assessment process, so that the review process can be itself assessed. An analogy between the work of a reviewer in a process of peer assessment and the operation of an automatic classifier is proposed. This has allowed us to leverage the usual measures in evaluating the quality of automatic classifiers to establish the quality of peer assessment. The reviewers are characterized by obtaining their confusion matrices and six new indicators: success rate (which estimates the validity); agreement degree (as a measure of reliability); assessment median and its interquartile range (for the estimation of central tendency and restriction of range biases); and average distance to diagonal and its standard deviation (to determine possible leniency and harshness biases). This method provides indicators of the reviewer's task and the detection of different profiles, so that the teacher can assess the work of the students as reviewers and introduce some correction mechanisms in the final assessment of the works. A practical example of application to an engineering degree is provided to illustrate the potential of the method.

Keywords: peer assessment; success rate; agreement degree; reliability; validity; bias; confusion matrix; automatic classification

1. Introduction

Peer assessment is a widespread way of evaluating and rating the quality of a work in several fields [1–3]. It is traditionally used in the review process of papers or in the evaluation of research projects and it is considered a reliable and effective method [4]. In the editorial process in journals and conferences it is undoubtedly the most habitual method of evaluation for these works. Publishers make their decisions using the reviews made by recognized experts in the area. It is assumed to be one of the most effective ways to maintain high quality standards of science. However, this type of review has been subjected to criticism and it has some known drawbacks, such as the problem of reliability and validity of the revisions, as well as other aspects such as the potential biases that can be introduced in the review process [1].

Peer assessment has also been applied widely in education, resulting to be a very effective learning instrument [5–7]. Topping [8] defines peer assessment in the context of education as "an arrangement in which individuals consider the amount, level, value, worth, quality, or success of the products or outcomes of learning of peers of similar status". In general, peer review is based on subjecting the work to the review of experts of the area in which it is framed. Thus, an expert review, usually anonymously, the work of his or her colleagues who, in turn, can become reviewers of his or her own work. In recent times, due to the emergence of massive learning environments, such as Massive Open Online Courses (MOOCs), this type of evaluation has received even greater attention, given the impossibility of a personalized assessment by the teacher [2].

Researchers acknowledge the positive features of peer assessment [2, 5–7]. While evaluating the work of their peers, students consolidate their own knowledge and develop specific abilities such as critical thinking [9], and the teachers appreciate not only the work done by the students but their ability to evaluate the work of other students. Moreover, social interaction among learners is acquiring more importance with the increasing use of technology for learning and the massive access to this technology. Two examples are MOOCs, supported by large communities of learners whose work is assessed using peer review, and Personal Learning Environments (PLE) where peer assessment is intensely used too [3].

Nevertheless, some other authors have detected some problems related to the use of peer assessment. For a peer assessment to be effective, it is necessary to provide well-defined criteria for evaluation and to

* Accepted 12 January 2018.

train students to fairly assess the work of their classmates. However, the reality is that in any review there is a subjective component that must be taken into account. The ideas that the reviewer has about what is right or wrong, the degree of knowledge of the area where the work is framed, or the reviewer's dedication may affect the result of the assessment. From this fact a question arises: is it possible to establish objective criteria for reviewers, analyzing issues such as the level of confidence of the reviewer, the casuistry of success/failure relative to other colleagues, and so on? One of the major concerns among researchers about this question is to ensure the reliability and validity of peer assessment, as well as detecting and avoiding possible biases [1, 8, 10]. Reliability is the degree of coincidences in evaluations by different students on a process or product; validity is the level of similarity related to the assessment made by the teacher or expert; bias is the inclination to present a partial perspective when assessing the work.

The main contribution of this paper is to propose an analogy between the process of peer assessment and the operation of automatic classifiers. The process of peer review can be seen as a classification process, in which several classifiers (the reviewers), from a given input (the work to revise), should assign a particular class (e.g., a grade or a discrete class such as "accept", "accept with changes" or "reject") based on certain classification algorithms (criteria that have been established for the assessment). From this perspective, it would be interesting to use some of the tools traditionally used to assess the accuracy of a classifier. Among them, the confusion matrices may be mentioned [11]. A confusion matrix visualizes the distribution of errors made by a classifier using a contingency table. In addition to the tools that are commonly used in classifiers other factors may be incorporated to that metric, such as measures of the reliability, the validity and the different biases.

This work aims to contribute to improve the quality of peer assessment processes and to provide a more objective assessment of the work of the reviewers. In addition, an example of application to an engineering degree is provided so that it allows discovering the potential of this improved peer review method as an assessment strategy in the field of engineering. Section 2 is devoted to present previous works that identify the advantages and problems of applying peer assessment in education and the previous research that propose metrics to evaluate the quality of the peer assessment. The proposed measures are presented in section 3, including an analogy of peer review and automatic classifiers, so that the findings in the field of automatic classification can be leveraged, and the proposal for new measures. The results of their application in a particular case of an engineering degree and a discussion about these results are presented in section 4. Finally, conclusions are in section 5.

2. Background

2.1 Peer assessment in education

Assessment is one of the key elements for an effective teaching-learning process. Teachers want the assessment to measure different skills and concepts but they usually have limitations in terms of resources and time [12]. In fact, time is the critical factor. Peer evaluation can help to solve this problem because of its advantages. To begin with, in crowded classes peer assessment can be much quicker than teacher's assessment. In addition, a student can devote more time to an evaluation than can dedicate the teacher so it can be more detailed [13]. This type of assessment improves some students' skills too. Reviewing the work of other students is an excellent opportunity to deepen the subject [14]. It also encourages student's autonomy, cooperation and productivity [5]. Moreover, reading the responses of the peers and taking the time to understand their point of view help develop cognitive empathy capabilities. Looking ahead on their professional future, peer review is a good way to develop these skills. Unfortunately, this type of evaluation is not free of dangers: lack of consistency, tendency to award everyone the same mark, risk to undervalue of overvalue the works, or increment in the teacher's workload in case of additional reviews. This is why this type of evaluation has been subjected to criticism and the problems of reliability, validity and potential biases have been receiving the spotlight.

The technological development has favored the massive access of the students to learning platforms, such as the so-called MOOCs [2] or the social learning environments [15, 16]. The large amount of data to be handled and students following these studies precludes the direct assessment by teachers. Instead, they must rely on technology to implement alternative assessment systems. An example is the case of automatic evaluation, under development nowadays, and another example is peer assessment. But it is not just a question of amount of data but also about the influence of social environment in learning. Interaction between peers is a rich source of learning. Seeking help from peers, as well as coaches and teachers, of course, is a self-regulated learning behavior. Peer assessment and feedback empower students to be self-regulated learners and promote motivation. Students can develop skills such as reflecting on and justifying

what they have done. As such, one way to help students become self-regulated and life-long learners is for teachers to provide the students with a supportive social learning environment that incorporates feedback techniques such as peer assessment [16]. Again, some threats are detected when using peer review in MOOCs and other massive access learning environments. The main detected problem is the fact that students generally do not trust peer assessment, since there is no teacher mediation or guidance for peer assessment. This problem is, in fact, the problem of validity of the assessment. Another big concerns regarding peer assessment are grading bias and rogue reviewers, that causes some authors even to consider whether there may be students are not eligible to assess peers [17].

It can be concluded from the previous paragraphs, regardless of the application field, that the main problems of peer review in which most researchers agree are the problems of reliability, validity and bias. In the next section these problems are analyzed in more detail and some of the solutions that have been proposed are presented.

2.2 Reliability, validity and bias

Reliability means consistency of judgments made by several reviewers on the same original. Validity is, on the contrary, the degree of agreement between the assessments made by the reviewers and by the teacher or expert, which is supposed to be fair and accurate. Bias is the systematic tendency for assessments to be influenced by anything other than the work being measured [18].

Although there are some works that propose studies of reliability and validity, Topping [8] points out that they compare peer assessments with assessments made by professionals rather that with those of other peers or the same peers over time. Falchikov and Goldinch [19] present a meta-analysis from the works of several authors, and obtain conclusions about the fields and levels in which the reliability and validity is higher, as well as a set of recommendations for practitioners for implementing peer assessment based on the conclusions of this meta-analysis.

Different measures have been used to calculate reliability and validity. The most common measure is the correlation coefficient between the average grades given by the students and the teachers [20–23]. Other not so common measures are the proportion of students who give a grade in a range of confidence on the teacher's [24], the use of a T-test for comparing the means between the grades of students and teachers [25] and the analysis of variance (ANOVA) to determine the reliability between reviewers [26].

Bias is prone to appear when there are human decisions. The problem of bias has attracted the attention of many researchers. Tversky and Kahneman [27] studied how heuristics are applied in the process of decision-making and the biases that arise in this process, establishing a cognitive basis to explain this human behavior. Saal [28] and Thiry [18], on their behalf, applied these concepts in the process of peer assessment, having made interesting studies compiling the causes and consequences of biases, among other factors, in peer rating in multi-rater feedback systems. The most important causes of bias found in performance ratings are: Halo (tendency to rate a person the same or almost the same on all items), Similarity (tendency to favorably assess the work of individuals who are similar in characteristics unrelated to the ones that are assessed such as age or race), Central tendency (tendency to always give midrange grades regardless of actual quality of the work), Leniency (tendency to give mostly high ratings), Harshness (tendency to be severe in their judgments), Restriction of range (the extent to which obtained ratings discriminate among different performance levels), First impression (tendency to allow one's first impression of the rate to influence ratings), Reliance on stereotypes (tendency to maintain ratings stable over time when the individual is well known in the group) and Fear of retaliation (when there are later rating opportunities)

The problem of bias has been studied from the point of view of psychology. For instance, the Social Relations Model (SRM) [29] is a tool to conceptualize and to analyze dyadic processes (interpersonal phenomena) that accounts for the complexities of the interpersonal perception and behaviors of two individuals (the perceiver and the target) by decomposing them into three independent components: a general tendency of the perceiver (perceiver effect), a general tendency of the target (target effect), and a specifically relational perception that is independent of these two main effects (relationship effect). This model is used in other research studies such as the one of Thompson [30] that uses the SRM model and ANOVA to study what is the tendency for raters to give similar ratings to each ratee (rater effect), what is the tendency among raters to agree with other raters (ratee effect) and what is the variance unaccounted for by the rater and ratee effects (rater by ratee interaction).

All these works focus on specific cases, trying to obtain the reliability and validity of the assessments that students obtain in their subject and under the established conditions, as well as determining how biases are conditioning the final rates. They are, in general, focused on determining the quality of the assessments to improve them but few studies are

devoted to the task of the reviewers as an evaluable task.

Moreover, in most cases, the statistics that are obtained for measuring reliability or validity are global, that is, they give a measure of these values for the whole process, but they do not allow an individual study of the degree of agreement of each review and the assessment of the other students or the teacher. Finally, the most habitual statistical measures in these studies are the correlation coefficient and the variances.

The proposed approach is radically different. It allows a case-by-case study and, therefore, not only the grade of the revised work is obtained but also a measure of the reliability and validity of the reviewer's evaluation of that work [31], as well as an approximation to the possible biases that may appear in the assessment process. A double evaluation is done: the work assessment and review process assessment.

3. Proposal

3.1 Peer assessment and automatic classifiers

In peer assessment, each participant must assess the work of the other students, assigning a grade or a category. In other words, from the input provided by the work to evaluate, the reviewer produces an output in the form of classification. To make this classification, the reviewers must have a set of criteria (e.g., a rubric) so that they could carry out their task in such an objective manner as possible.

This evaluation process can be likened to that performed by automatic classifiers. An automatic classifier is a computer model that assigns an individual, characterized by a set of variables, one label among several possible labels associated with different classes. The algorithm used for classification establishes the criteria to make this assignment. Beyond the obvious differences between the two processes, both have an individual to be classified. In the case of an automatic classifier, the individual is characterized by a set of variables of different type that can be handled automatically. In the case of a human evaluator, the element that characterizes the individual is the work to be reviewed, so the available information is much richer but less structured and difficult to automate. Anyway, from these inputs a classification algorithm must be applied, based on computational methods in one case, and based on a rubric and a subjective task of applying this rubric in the other. As a result of the algorithm, in both cases it outputs a label that identifies the class in which the individual is classified.

The key question is: what is the quality of the classification? In the case of computer models, researchers have spent much effort in seeking ways to compare classifiers attending the successes and failures that occur in the classification. Since this type of metric is just based on the results but not on the technical characteristics of the algorithm, would it be possible to apply it in the case of a human classification? This is the hypothesis of this work.

Two of the simpler and more habitual measures to evaluate the quality of a classifier are its accuracy or success rate and its error rate. They are actually complementary measures since accuracy = 1 – error rate.

Although accuracy is a very popular metric and has the virtue of a single value representing a measure of the quality of a classifier, it has the drawback of assuming that the cost of a misclassification is the same in any case. Let us take an example to explain the problem: Suppose a classifier that makes a disease diagnosis, that is, from a set of values related to diagnostic tests or symptoms, it classifies the patients indicating if they affected or not by the disease. The classifier has a success rate of 95%, i.e., it fails only in 5% of cases. The question is: these misclassifications, are they referred to patients who have the disease but are classified as healthy, or to healthy patients who are classified as sick? Obviously, the cost of misclassification cannot be the same, since in this case a conservative classifier that classifies every sick patient as sick even at the cost of worsening the accuracy is preferable to a more accurate classifier that considers sick patients as healthy.

Other more complete measures that analyze other aspects of classifiers have been presented. Suppose the case of a multiclass classifier with n possible classes. Formally, for each individual or sample the classifier estimates a label X among a set of possible labels, each one representing a different class. The actual class to which the individual belongs is known and is represented as x, to distinguish the real classes (lowercase) from that estimated by the classifier (uppercase). The results are usually represented in a confusion matrix or contingency table $\boldsymbol{M}$, which is an $n \times n$ square matrix, where n is the number of classes. The confusion matrix is constructed placing the actual classification in the columns and the estimated classification in the rows. In Table 1, for instance, the results of a three-class classifier are represented, where the

Table 1. Confusion matrix for a multiclass classifier

	a	*b*	*c*	***FP***
A	8	1	0	**1**
B	2	11	1	**3**
C	0	1	9	**1**
FN	**2**	**2**	**1**	***TP* = 28**

first column indicates that there are 10 individuals in class a, 8 of which were labeled as A, 2 as B and 0 as C.

The confusion matrix is the basis of many common metrics. There may be several possible outcomes for a confusion matrix $\boldsymbol{M}$, some of them are for a particular class x, and other are global for the whole classifier:

- All the correctly labeled samples (for example, samples classified as A and actually belonging to a class) are considered true positives (TP). True positives for a given class x are placed at the main diagonal $TP_x = M_{xx}$. The global amount of true positives is calculated summing up all the true positives: $TP = \sum_{\forall i} M_{ii}$.
- All the samples labeled as belonging to a class but not actually belonging to it (for example, samples classified as A but not belonging to class a) are considered false positives (FP). False positives for a given class x are calculated summing up all the elements at row X but the one at the main diagonal: $FP_x = (\sum_{\forall i} M_{xj}) - M_{xx}$.
- All the samples actually belonging to a class but not correctly labeled (for example, samples actually belonging to class a but not classified as A) are considered false negatives (FN). False negatives for a given class x are calculated summing up all the elements at column x but the one at the main diagonal: $FN_x = (\sum_{\forall i} M_{ix}) - M_{xx}$.

The following metrics for a given class x can also be calculated from the confusion matrix:

Precision: $Prec_x = TP_x/(TP_x + FP_x)$

Sensitivity: $Sens_x = TP_x/(TP_x + FN_x)$

F-score: $FScore_x = \frac{2}{1/Prec_x + 1/Sens_x}$

Precision and sensitivity metrics are particularly interesting for the proposed measures. The class precision indicates the proportion of individuals classified as belonging to this particular class that actually belong to it (although it says nothing about individuals of the class that are misclassified). For its part, the class sensitivity indicates the proportion of individuals belonging to the class that are classified as belonging to it (although it says nothing about individuals belonging to other classes that are classified as belonging to this one). An ideal classifier must have a precision and a sensitivity of 1 for every class. This situation seldom occurs, so these metrics are very useful to choose the most suitable classifier depending on the objective to achieve, privileging the precision over the sensitivity or vice versa.

Furthermore, it is possible to obtain global indicators for the whole classifier such as precision and sensitivity averages. There are several ways to obtain such indicators but a simple and widely accepted way is to use the so-called micro-average method, where the average is obtained for all individuals and all classes [32].

To assess the classifier quality and to calculate all these measures, the actual classification of individuals is assumed to be known, that is, it is necessary to have a canonical reference classification to compare with, the so-called gold standard test. For example, in character recognition systems the classification made by humans can be used as reference classification. This canonical classification should be as perfect as possible, however in practice the perfect classification is usually not available. For example, in the case of peer assessment, the grade of a work (i.e., the class it belongs to) is always subjective and a perfect classification cannot be established, but the grade given by the teacher or an expert can be considered as the gold standard. In the case of peer review of papers, the final decision of the editor can be used as canonical classification, or even incorporate other bibliometric indicators to determine the impact of the publication and assume that this impact is a measure of the quality of the contribution [33].

3.2 Measures to assess peer assessment

Once established the parallelism between the review process and a classification process, the measures to evaluate the goodness of the peer assessment process can be designed. A starting point may be to use the tools that are commonly used to estimate the accuracy of a classifier: the confusion matrix and the related metrics.

First, let us define the elements that will be used in the proposed measures:

- Let $\boldsymbol{W}$ be the set of works to be assessed and w_i each work.
- Let $\boldsymbol{R}$ be the set of reviewers that assess the works and r_j each reviewer.
- Let $\mathbf{C}$ be the set of canonical classification of the works and c_i the canonical classification of work w_i, i.e. c_i is the class to which work w_i is considered to belong to (for instance, the assessment made by the teacher).
- Let $\mathbf{A}$ be the set of assessments and a_{ij} the assessment of work w_i made by reviewer r_j, i.e. a_{ij} is the label given by reviewer r_j to work w_i.
- Let n_j the number of assessments made by every reviewer r_j.

The use of confusion matrices allow us to define metrics for the estimation of the validity, the reliability and different types of biases (restriction of range, central tendency, leniency bias and harshness bias) [18, 28].

3.2.1 Success rate

The success rate for a reviewer r_j, SR_j, is defined as:

$$SR_j = \frac{\sum_{\forall\, a_{ij}} s_{ij}}{n_j} \tag{1}$$

where

$$S_{ij} = \begin{cases} 1, & \text{if } a_{ij} = c_i \\ 0, & \text{in other cases} \end{cases} \tag{2}$$

that is, S_{ij} is 1 if the assessment of work w_i made by reviewer r_j (i.e., a_{ij}) is the same as the canonical classification c_i and so, SR_j is the proportion of assessment made by reviewer that are the same as the one done by the teacher (the canonical assessment). These measures can take values in the interval [0,1] so that $SR_j = 1$ means a complete success in the assessment (all the works are correctly classified) and there are no false positives neither false negatives. A value of $SR_j = 0$ means that every assessment made by the reviewer is different from the canonical classification.

This measure has a similar meaning to that of validity, with the difference that validity is an overall measure for all revisions, and SR is a particular measure for each reviewer.

3.2.2 Agreement degree

The agreement degree of reviewer r_j with the other reviewers, AD_j, is defined with the following equation:

$$\frac{AD_j = \sum_{\forall\, a_{ij}, a_{ik}} L_{ijk}}{n_j} \tag{3}$$

where

$$L_{ijk} = \begin{cases} 1, & \text{if } a_{ij} = a_{ik} \\ 0, & \text{in other cases} \end{cases} \tag{4}$$

that is, L_{ijk} is 1 if the assessment of work w_i made by reviewer r_j (i.e., a_{ij}) is the same as the assessment of the work made by reviewer r_k (i.e., a_{ik}) and so, AD_j measures the proportion of assessments in which the reviewer coincides with the other reviewers for a given work. As in the case of SR, AD can take values in the interval [0,1] so that $AD_j = 1$ means a complete agreement in the assessment with the other reviewers of the same work and $AD_j = 0$ means that every assessment made by the reviewer is different from that of his or her peers.

In this case, the agreement degree establishes a measure similar to that of reliability but, as in the case of the success rate, it is an individual measure for each reviewer and not a global one.

3.2.3 Assessment median and its interquartile range

The assessment median of a reviewer (AM_j) and its interquartile range (AIR_j) represent the central value and the dispersion of the assessments of a reviewer. Other central position and dispersion measures, such as the mean and the standard deviation could be used instead, but median and the interquartile range are preferred because they are not skewed by extreme values. Moreover, they can be calculated even for ordinal data, in which values are ranked relative to each other but are not measured absolutely. This is the case of categorical classes for which an order can be established.

The AM_j and the AIR_j allow the estimation of two possible biases of the reviewer: central tendency and restriction of range. Restriction of range bias, or the tendency to rate every work with the same grade because a lack of discriminability among different performance levels, is present when there is a low value for the AIR_j, since there is a low dispersion among the values. Moreover, if a low value of AIR_j is combined with an AM_j that is near the center of the interval of possible rates, it can be considered that a central tendency bias, that is, the tendency to always give midrange grades regardless of actual quality of the work, is observed. In case of low interquartile range and median near the extremes of the interval of possible rates, we can conclude a certain tendency to overrate or underrate the works. However, since there is no reference to the actual quality of the works it cannot be established whether there are some biases or not. In this case, the following measures are much more meaningful.

3.2.4 Average distance to diagonal and its standard deviation

The average distance to diagonal for a reviewer, $\overline{DD}_j$, is defined as:

$$\overline{DD}_j = \frac{\sum_{\forall\, i}(a_{ij} - c_i)}{n_j} \tag{5}$$

that is, it is the mean of the differences of the assessments made by the reviewer and the canonical assessment. In other words, it is the average distance from the estimated classification to the diagonal of the confusion matrix, where the canonical classification is placed. Every distance is positive if the estimation is higher that the canonical rate and negative if the estimation is lower. To this measure, the standard deviation for the reviewer, s_{DDj}, is calculated as usual:

$$S_{DDj} = \sqrt{\frac{\sum_{\forall\, i}(a_{ij} - \overline{DD}_j)^2}{n_j - 1}} \tag{6}$$

The combination of $\overline{DD_j}$ and s_{DDj} helps us to determine possible leniency and harshness bias. A high positive value of $\overline{DD_j}$ is an indicator of leniency bias, since the reviewer has a clear tendency to overrate the work of his peers. In case of low negative values, this metric indicates some harshness bias, since the tendency is to underrate the works, compared to the canonical assessment. Moreover, when the dispersion, s_{DDj} is low, the tendencies are even more pronounced.

4. Results and discussion

To illustrate the application of the proposed measures, a subject of the Computer Engineering Master that uses a system of peer assessment for some of its aspects has been used. There were 24 students enrolled in the course, distributed in 5 groups of 4 or 5 students. Every student should assess the work of every group but his or hers, i.e., every student assesses 4 works, but every work is assessed by 19 or 20 students. Students must evaluate the work of their peers assigning a grade, which is not categorical (Excellent/Good/Fair/Poor), but numerical, having grades between 0 (fail) and 10 (excellent with honors), being 5 the minimum grade to pass. In practice, since no peer dares to rate low grades, the grades are always between 5 and 10. The final grade for every work is calculated as the average of all grades. This value is taken as reference classification.

The previously defined metrics has been applied to try to evaluate the assessment process: for each reviewer r_j, the six metrics are calculated: its success rate, SR_j, its agreement degree with the other reviewers, AD_j, its assessment median and interquartile range, AM_j and AIR_j, and its average distance to diagonal and standard deviation, $\overline{DD_j}$ and s_{DDj}. Some charts, to better understand the results are also presented.

4.1 Success rate and agreement degree

Figure 1 shows the success rate and the agreement degree for the 24 students of the subject. Although the number of assessments each reviewer makes is different, the formulation of the metrics is normalized in interval [0,1] so that the comparison between them is possible. The values of SR_j and AD_j are represented on the vertical axis and the identifiers of each reviewer r_j on the horizontal axis. Reviewers are ordered in ascending order of SR_j to make the chart easier to interpret.

Figure 1 allows us to visualize at a glance the values of SR_j and AD_j for each reviewer r_j and the differences in behavior between them. We can observe a certain tendency of increase of AD_j as SR_j increases, which only makes sense. The value of SR_j is obtained by comparing the evaluation of the reviewer r_j with the canonical classification c_i, which is itself based on the evaluations of the other reviewers. Therefore, it is normal that there is some degree of agreement. However, although the trend is this, we can observe that there are many cases in which this is not exactly so. This makes us think that the measures do not have a high degree of dependence and they both provide relevant information.

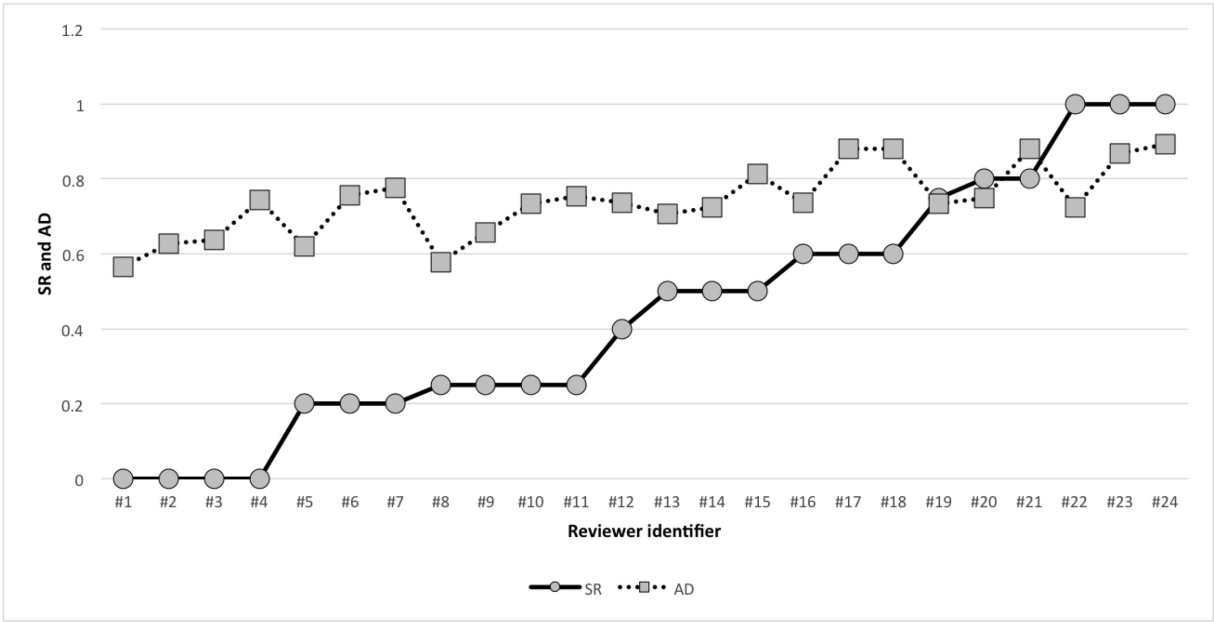

Fig. 1. Success rate and agreement degree for each reviewer.

4.2 Assessment median and interquartile range

The AM_j and the AIR_j are used to estimate central tendency (tendency to always give midrange grades) and restriction on range (lack of discriminability). A box and whiskers graph has been chosen for the graphical representation. These charts are adequate to represent variables such as those used in this study, since they present information about the central tendency, the dispersion and the symmetry of the data. The upper and lower ends of the whiskers indicate, respectively, the maximum and minimum values above or below which the values are considered atypical (outliers). The upper end of the box indicates the third quartile (75% percentile) and the lower one the first quartile (25% percentile), so that the box size indicates the AIR_j. In the central position is the AM_j, which divides the data into two set of the same size. In general, in a representation of this type, the longer the box and the whiskers are, the more scattered is the distribution of data. The line representing the median indicates symmetry.

Figure 2 represents the box and whiskers plot for the AM_j and AIR_j of the reviewers in this study. Almost all the distributions are quite asymmetrical, and the median coincides in many of the cases with the limits of the quartiles. This is because student assessments tend to concentrate on one value of the scale, so that one grade is much more prevalent than all others.

Other preliminary conclusions about biases can also be obtained. For instance, a high restriction of range, that is, a high tendency to rate every work with the same grade, may be discovered just analyzing the AIR_j and focusing on small boxes. This is the case of reviewers #2, #6 and #24 (with $AIR = 0$), but also reviewers #4, #9, #13, #14, #15, #17 and #23 (with $AIR_j < 0.5$). Moreover, the value of the AM_j can also give some interesting information: considering that 7.5 is the midpoint of the possible grades (formally, the midpoint is 5, since the range of possible values is [0,10], but actual values are always above 5 since no peer dares to rate low grades), reviewers #5, #6, #7, #9, #10, #11, #20 and #22 may be in risk of having some central tendency bias because of their AM_j having values between 7 and 8. This risk is particularly high in the case of reviewers #6, #9 and #10 because of their low AIR_j. The case of reviewers #2, #4 and #8 are also interesting because of their extreme AM_j.

The study of these metrics is interesting since they provide information about dispersion and symmetry but the conclusions that can be obtained are quite limited. However, a joint analysis of these metrics and the average distance to diagonal and its standard deviation could give us some clues that contribute to shed light to the possible biases. In the following sections a revision of this preliminary conclusions is presented.

4.3 Average distance to diagonal and standard deviation

The $\overline{DD}_j$ and the s_{DDj} are used to estimate possible leniency (tendency to give high ratings) and harshness (tendency to be severe) biases. In this case, the use of a mean and standard deviation graph can be interesting. For each reviewer, the average distance from his or her assessments to the canonical assessment (placed on the diagonal of the confusion

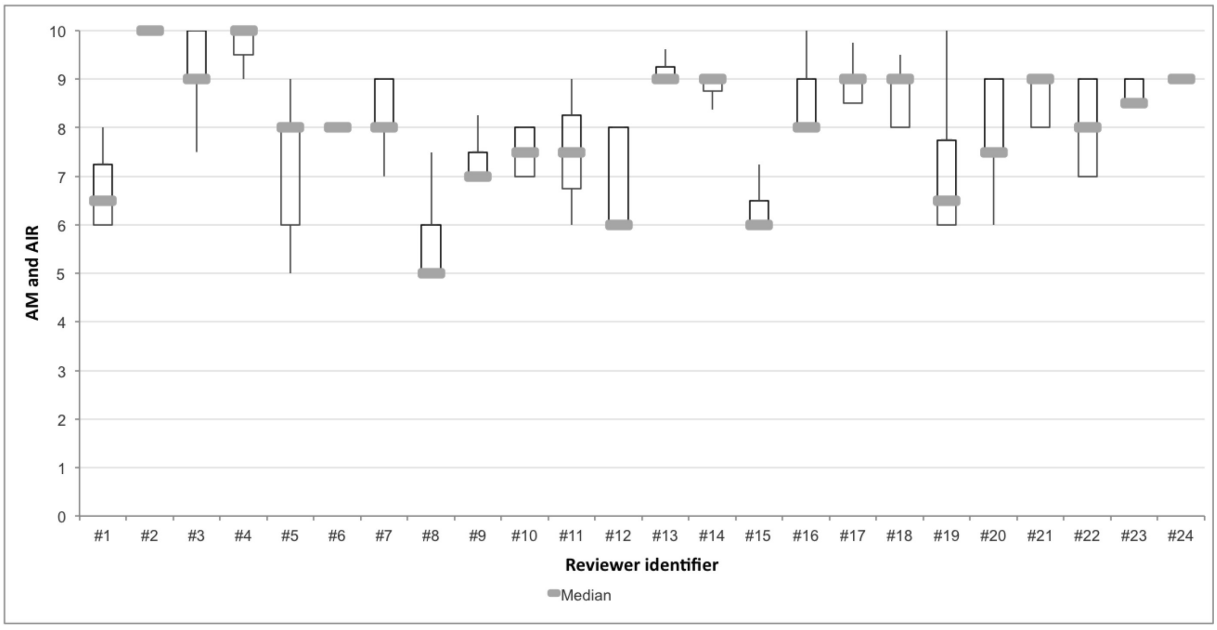

Fig. 2. Box and whiskers graph of assessment media and its interquartile range for each reviewer.

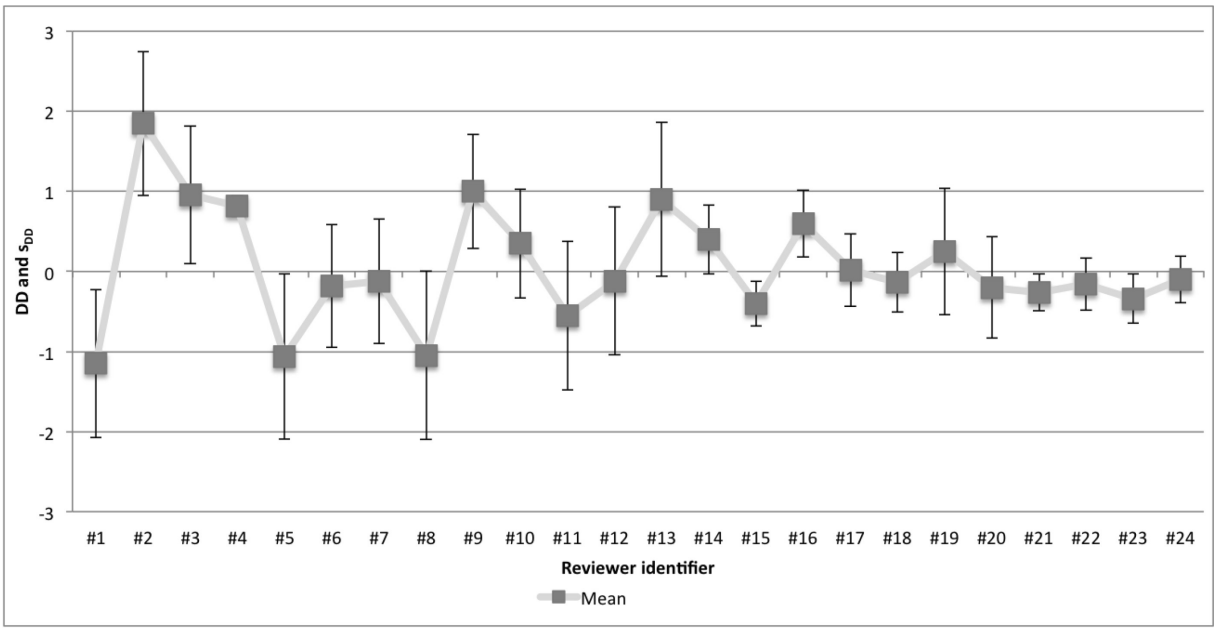

Fig. 3. Average distance to diagonal and standard deviation graph for each reviewer.

matrix) is represented, as well as the standard deviation as a measure of the dispersion of the assessments of this reviewer. The reviewers whose mean is displayed near the 0 value are those whose assessment are, in average, closer to the canonical assessment, indicating that there is not leniency of harshness bias. When $\overline{DD}_j$ has a positive value, there may be a leniency bias (greater as the value is higher), while a negative value indicates a possible harshness bias (greater as the value is lower). The s_{DDj} modulates how robust is the estimation of these possible biases.

Figure 3 displays the $\overline{DD}_j$ and the s_{DDj} of the reviewers participating in this study. Some interesting general conclusions can be obtained. It can, be seen at a glance that the reviewers on the right have a much steadier behavior than the one on the left. Because of the order established, the ones on the right are the ones with highest success rate (see Fig. 1). Reviewers from #17 to #24 have a $\overline{DD}_j$ value around 0, so there is neither important leniency nor harshness bias. Moreover, s_{DDj} is quite low in most cases, so the differences between their particular assessments and the canonical ones are low. However, when SR_j is low, the biases are, in many cases, evident.

The joint study of this graph with the one of Fig. 2 sheds light about the behavior of the reviewers and their possible bias. For instance, observing Fig. 2, reviewers #2, #6 and #24 were candidates to have a high restriction of range bias, since $AIR_j = 0$. Fig. 3 corroborates that reviewers #2 and #6 have a high restriction of range because they have a high s_{DDj}, so their assessments are usually very different to the canonical ones. However, reviewer #24 has $AIR_j = 0$ but a very low value of s_{DDj}, that is, his or her assessments were very similar among them but also very close to the canonical ones. In this case, it may mean that this is a very good reviewer (no important differences with the canonical assessment and high values of SR_j and AD_j) who has had to evaluate very similar works.

Another interesting case are those of reviewers #6, #9 and #10 because of their low AIR_j and midrange values of AM_j. In this case we should notice the value of s_{DDj}. They all have a high value of s_{DDj}, so there are important differences between their evaluations and the canonical ones, dismissing the possibility of having been assigned mid-quality works and confirming a probable central tendency bias.

The case of reviewers #2, #4 and #8 were also highlighted in the previous section because of their extreme AM_j. The very positive value of $\overline{DD}_j$ in the case of reviewers #2 and #4 indicates a clear leniency bias, while the very negative value of $\overline{DD}_j$ for reviewer #8 indicates an obvious harshness bias. Other cases of leniency are that of reviewers #3 and #13, and other of harshness are #1 and #5.

4.4 Special cases

In the following sections two representative cases are deeply studied. The first selected reviewer is #24, a clear example of individual with a high value for both SR_j and AD_j. The second one is reviewer #2, another paradigmatic case, in this case of those individuals with a low value of both SR_j and AD_j. For each reviewer the proposed metrics and the

confusion matrices are calculated (see Tables 2 and 3). Although each confusion matrix is of size 10 x 10 (ten possible classes corresponding to ten possible grades), only an extract is shown, since the other cells have a value of 0. Column 11 (*FP*) indicates the number of false positives per class; row 11 (*FN*) presents false negatives per class. Finally, column 12 and row 12 respectively indicate the precision and sensitivity per class. As previously observed, the sum of the cells in the main diagonal indicates the number of works in which the reviewer's assessment coincides with the final assessment.

4.4.1 Reviewer with a high value for both success rate and agreement degree

The representative reviewer is #24. This is the case of a reviewer who behaves very similarly to the canonical classification and, moreover, he or she almost always coincides with the other peers. Table 2 shows the proposed metrics and the confusion matrix for this reviewer.

This reviewer has the highest success rate in his or her assessments (a value of 1) and a high agreement degree with the other reviewers (a value of 0.89). The values of both metrics are fully related, and they seem to indicate that this is a reviewer with solid arguments and a great insight in his or her reviews. The assessment median and its interquartile range indicate that the central grade is 9 and his grades are all very close (AIR_j = 0), indicating a possible restriction of range bias. However, this bias is discarded because the very low value of s_{DDj} indicates that his or her assessments are very close to the canonical ones. Moreover, the value of $\overline{DD_j}$ is almost 0, so there are neither leniency nor harshness biases.

The confusion matrix shows that he or she has 5 evaluations, a precision per class of 0 and a sensitivity per class of 1, for those classes of which examples are provided. The high value of sensitivity indicates that the reviewer is able to properly distinguish between the different classes, that is, he or she is very careful with the small details that allow the correct classification. The low value of precision indicates that the dispersion of his or her assessments is very low, that is, these assessments are always very close to the canonical classification. In short, this is the case of a reviewer with solid arguments and a great insight in his or her assessments.

4.4.2 Reviewer with a low value for both success rate and agreement degree

Reviewer #2 is just the opposite case. Reviewers with low values of SR_j and AD_j do not agree either with the canonical assessments or with their peer's assessments. Table 3 shows the proposed metrics and the confusion matrix for this reviewer.

This is a reviewer with a very low success rate when evaluating (a value of 0) and a relatively low agreement degree with the other reviewers (a value of 0.62). These low values may correspond to a novel or negligent reviewer. The assessment median and its interquartile range are very signifi-

Table 2. Metrics (left) and confusion matrix (right) for reviewer #24

Metric	*Value*
SR	1
AD	0.89
AM	9
AIR	0
$\overline{\boldsymbol{DD}}$	-0.1
s_{DD}	0.29

Estimated grade \ *Actual grade*	*7*	*8*	*9*	*10*	*FP*	*Precision*
7	0	0	0	0	0	
8	0	1	0	0	0	1
9	0	0	4	0	0	1
10	0	0	0	0	0	
FN	0	0	0	0	5	
Sensitivity		1	1			

Table 3. Metrics (left) and confusion matrix (right) for reviewer #2

Metric	*Value*
SR	0
AD	0.62
AM	10
AIR	0
$\overline{\boldsymbol{DD}}$	1.85
s_{DD}	0.9

Estimated grade \ *Actual grade*	*7*	*8*	*9*	*10*	*FP*	*Precision*
7	0	0	0	0	0	
8	0	0	0	0	0	
9	0	0	0	0	0	
10	2	0	2	0	4	0
FN	2	0	2	0	0	
Sensitivity	0		0			

cant in this case: the central grade is 10 and his grades are all very close (AIR_j=0) indicating a possible restriction of range bias. In fact, the confusion matrix reflects that all his or her assessments have a value of 10. Moreover, the high value of s_{DDj} corroborates the restriction of range bias and the very positive value of $\overline{DD}_j$ indicates a clear leniency bias.

The confusion matrix shows that this reviewer has a precision per class of 0 and a sensitivity per class of 0, for those classes of which examples are available. The low value of sensitivity indicates that the reviewer is not able to identify the subtle differences between classes. In this case, precision is not significant, since no classification is correct. In short, he or she is a reviewer who pays little attention to the details that make the difference and behaves by giving all works the highest grade.

5. Conclusions

Peer review has become a very important element in the evaluation systems. In some cases, it complements other measurement methods, as in the case of evaluation among students who normally complements the assessment of the teacher. In other cases, however, it becomes the only element or at least the primary one of the evaluation process. Such is the case of peer reviews for publications or conferences, the process of reviewing research projects for obtaining grants or the assessment in massive learning platforms like MOOCs. The benefits of peer review have been highlighted in many areas, but in this type of evaluation remains a subjective component inherent in the processes with human intervention. This component can be interesting from several points of view, but it must be properly controlled. In short, it is important to assess the assessment process itself. This has led us to consider the key question proposed at the beginning of the paper: Is it possible to establish some criteria for evaluating the work of the reviewers in a peer evaluation system?

In this article we have tried to answer this question, drawing a parallel between the work of a reviewer in a process of peer assessment and the operation of an automatic classifier. This has allowed us to leverage the usual measures in evaluating the quality of automatic classifiers to establish the quality of peer assessment. In this way important work done in this area can be leveraged to open a new line of study on peer review systems.

To illustrate this proposal, the case of peer assessment in the activities of a subject belonging to a Master course has been analyzed. This is the case of a numerical grading in the interval [0,10] but it has finally been treated as a multiclass classification (10 classes, corresponding to the division of the interval into 10 grade ranges). Besides the confusion matrices, six new indicators have been defined: success rate (the proportion of assessment made by a reviewer that are the same as the canonical assessment, similar to the concept of validity); agreement degree (it measures the agreement degree of each reviewer with others, with a similar meaning to the concept of reliability); assessment median of a reviewer and its interquartile range (central value and the dispersion of the assessments of a reviewer that allow the estimation of central tendency and restriction of range biases); and average distance to diagonal and its standard deviation (it is the mean of the differences of the assessments made by the reviewer and the canonical assessment, so that they allow us to determine possible leniency and harshness bias). Once each reviewer is characterized, it corresponds to the responsible for the system (that is, the teacher) to determine what actions to perform. For example, eliminating the evaluations of these reviewers to consider introducing outliers could be determined, or analyzing the history of this reviewer's assessments because he or she could have an eccentric but interesting point of view. Anyway, the method provides indicators of the reviewer's task and the detection of different profiles.

This experience is very preliminary and there are many paths to study, but an important work line could be developed in the future. From this study, we aim to apply other common metrics in the area of automatic classifiers to the case of peer assessment, to define our own metrics, to conduct a study about the exact meaning of each indicator and to understand and improve the process of peer review. Another interesting development in the future is the implementation of the proposed metrics as plugins for some of the more popular learning management systems, as well as the inclusion of the metrics in MOOCs.

References

1. J. M. Campanario, The peer review system: many problems and few solutions, *Revista española de Documentación Científica*, **25**, 2002.
2. J. Kay, P. Reimann, E. Diebold and B. Kummerfeld, MOOCs: So Many Learners, So Much Potential..., *IEEE Intelligent Systems*, **28**, 2013, pp. 70–77.
3. W. Greller and H. Drachsler, Translating Learning into Numbers: A Generic Framework for Learning Analytics, *Educational Technology & Society*, **15**, 2012, pp. 42–57.
4. A. Mulligan, L. Hall and E. Raphael, Peer review in a changing world: An international study measuring the attitudes of researchers, *Journal of the American Society for Information Science and Technology*, **64**, 2013, pp. 132–161.
5. R. Grangel Seguer and C. Campos Sancho, Contratos de aprendizaje y evaluación entre iguales para responsabilizar al alumno de su aprendizaje, in *Actas de las XIX Jornadas sobre la Enseñanza Universitaria de la Informática (Jenui 2013)*, J.

M. Badia Contelles, S. Barrachina Mir, and M. M. Marqués Andrés, Eds., (Universitat Jaume I, 2013).
6. M. Marqués, J. M. Badía and E. Marínez-Martín, Una experiencia de autoevaluación y evaluación por compañeros, in *Actas de las XIX Jornadas sobre la Enseñanza Universitaria de la Informática (Jenui 2013)*, J. M. Badia Contelles, S. Barrachina Mir, and M. M. Marqués Andrés, Eds., (Universitat Jaume I, 2013).
7. P. Sánchez and C. Blanco, Una metodología para fomentar el aprendizaje mediante sistemas de evaluación entre pares, in *Actas de las XIX Jornadas sobre la Enseñanza Universitaria de la Informática (Jenui 2013)*, J. M. Badia Contelles, S. Barrachina Mir and M. M. Marqués Andrés, Eds., (Universitat Jaume I, 2013).
8. K. Topping, Peer Assessment Between Students in Colleges and Universities, *Review of Educational Research*, **68**, 1998, pp. 249–276.
9. M. Á. Conde, L. Sanchez-Gonzalez, V. Matellan-Olivera, R. Lera and F. Javier, Application of Peer Review Techniques in Engineering Education, *International Journal of Engineering Education*, 2017.
10. M. S. Ibarra Sáiz, G. Rodríguez Gómez and M. Á. Gómez Ruiz, La evaluación entre iguales: beneficios y estrategias para su práctica en la universidad, *Revista de Educación*, 2011.
11. M. Sokolova and G. Lapalme, A systematic analysis of performance measures for classification tasks, *Information Processing & Management*, **45**, 2009, 427–437.
12. P. M. Sadler and E. Good, The Impact of Self- and Peer-Grading on Student Learning, *Educational Assessment*, **11**, 2006, pp. 1–31.
13. R. L. Weaver and H. W. Cotrell, Peer evaluation: A case study, *Innovative Higher Education*, **11**, 1986, pp. 25–39.
14. A. M. Langan and C. P. Wheater, Can students assess students effectively? Some insights into peer-assessment, *Learning & Teaching in Action*, **2**, 2003.
15. R. Ferguson and S. B. Shum, *Social learning analytics: five approaches*, 2012, 23, ACM Press.
16. P.-L. Hsu and K.-H. Huang, Evaluating Online Peer Assessment as an Educational Tool for Promoting Self-Regulated Learning, in *Multidisciplinary Social Networks Research*, **540**, L. Wang, S. Uesugi, I.-H. Ting, K. Okuhara, and K. Wang, Eds. (Springer Berlin Heidelberg, Berlin, Heidelberg, 2015).
17. T. Staubitz, D. Petrick, M. Bauer, J. Renz and C. Meinel, *Improving the Peer Assessment Experience on MOOC Platforms*, 2016, pp. 389–398, ACM Press.
18. K. J. Thiry, *Factors That Affect Peer Rater Accuracy in Multirater Feedback Systems* (BiblioBazaar, 2011).
19. N. Falchikov and J. Goldfinch, Student Peer Assessment in Higher Education: A Meta-Analysis Comparing Peer and Teacher Marks, *Review of Educational Research*, **70**, 2000, pp. 287–322.
20. D. Magin, Reciprocity as a Source of Bias in Multiple Peer Assessment of Group Work, *Studies in Higher Education*, **26**, 2001, pp. 53–63.
21. I. AlFallay, The role of some selected psychological and personality traits of the rater in the accuracy of self- and peer-assessment, *System* **32**, 2004, pp. 407–425.
22. A. M. Langan, C. P. Wheater, E. M. Shaw, B. J. Haines, W. R. Cullen, J. C. Boyle, D. Penney, J. A. Oldekop, C. Ashcroft, L. Lockey and R. F. Preziosi, Peer assessment of oral presentations: effects of student gender, university affiliation and participation in the development of assessment criteria, *Assessment & Evaluation in Higher Education*, **30**, 2005, pp. 21–34.
23. H. Luo, A. Robinson and J.-Y. Park, Peer Grading in a MOOC: Reliability, Validity, and Perceived Effects, *Online Learning*, **18**, 2014.
24. M. Freeman, Peer Assessment by Groups of Group Work, *Assessment & Evaluation in Higher Education*, **20**, 1995, pp. 289–300.
25. W. Cheng and M. Warren, Peer and Teacher Assessment of the Oral and Written Tasks of a Group Project, *Assessment & Evaluation in Higher Education*, **24**, 1999, pp. 301–314.
26. D. J. Magin, A Novel Technique for Comparing the Reliability of Multiple Peer Assessments with that of Single Teacher Assessments of Group Process Work, *Assessment & Evaluation in Higher Education*, **26**, 2001, pp. 139–152.
27. A. Tversky and D. Kahneman, Judgment under Uncertainty: Heuristics and Biases, *Science*, **185**, 1974, pp. 1124–1131.
28. F. E. Saal, R. G. Downey and M. A. Lahey, Rating the ratings: Assessing the psychometric quality of rating data, *Psychological Bulletin*, **88**, 1980, pp. 413–428.
29. M. D. Back and D. A. Kenny, The Social Relations Model: How to Understand Dyadic Processes: The Social Relations Model, *Social and Personality Psychology Compass* **4**, 2010, pp. 855–870.
30. R. Thompson, Reliability, Validity, And Bias In Peer Evaluations Of Self Directed Interdependent Work Teams, 24 June 2001, 6.845.1–6.845.37.
31. R. Molina-Carmona, R. Satorre-Cuerda, P. Compañ-Rosique and F. Llorens-Largo, *Performance measures for peer assessment*, 2016, pp. 341–347, ACM Press.
32. Y. Yang, An Evaluation of Statistical Approaches to Text Categorization, *Information Retrieval*, **1**, 1999, pp. 69–90.
33. M. Bordons and M. Á. Zulueta, Evaluación de la actividad científica a través de indicadores bibliométricos, *Revista Española de Cardiología*, **52**, 1999, pp. 790–800.

Rafael Molina-Carmona received his BSc and MSc in Computer Science from the Polytechnic University of Valencia, Spain in 1994, and his PhD in Computer Science from the University of Alicante, Spain in 2002. He is a professor at the University of Alicante, and he belongs to the department of Computer Science and Artificial Intelligence. He is also a researcher at the Industrial Computing and Artificial Intelligence research group and his interests are mainly the applications of Artificial Intelligence to different fields: computer-aided design and manufacture, computer graphics, learning, gamification and information representation. His first works were focused on Artificial Intelligence applied to computer-aided design and manufacture, space reconstruction and grammatical models for virtual worlds generation. He has published more than 20 papers and he has also directed three Theses in this fields. Moreover, he is now participating in a powerful research line about technology-enhanced learning and creativity, including videogames, gamification, learning analytics and information representation. He has co-authored more than 10 papers and he has co-directed two Theses in this field. He is a member of AENUI (Association of University Teachers of Informatics) and of the *Cátedra Santander-UA de Transformación Digital* (Santander-UA Chair of Digital Transform).

Rosana Satorre-Cuerda has a PhD in Computer Science (University of Alicante, 2002). Her specialty includes programming, stereoscopic vision, educational games, engineering education, and teacher training in ICT. Since 1994, she works as a lecturer in the Department of Computer Science and Artificial Intelligence at the University of Alicante (Alicante, Spain), where she is Professor of University since 2008. She has held the position of Deputy Director of Department between 2000 and 2004, Director acting Department between 2004 and 2005, Deputy Director of the Studies of Informatics Polytechnic School between 2005 and 2009, and from 2009 until 2012 Secretary of the Polytechnic School. Her thesis was related to issues of stereoscopic vision, although since its inception the University has devoted many efforts

to education and teacher training in and through the computer. Assistant Principal in her period of Informatics degrees, she coordinated the development of new curricula Degree in Computer Engineering, implanted at this time at the University of Alicante. She has several papers about the use of AI techniques applied to several problems. She participates in educational innovation projects related to the EEES. She is a member of AENUI (Association of University Teachers of Informatics) and of the *Cátedra Santander-UA de Transformación Digital* (Santander-UA Chair of Digital Transform).

Patricia Compañ-Rosique has a PhD in Computer Science (University of Alicante, 2004). She has held positions of leadership and management since she joined the University of Alicante: Deputy Head of Computer Engineering of the Polytechnic School (2009–2012) and deputy director of the Polytechnic School (2012–2013). In these periods she was directly involved in the development of new curricula in the Computer Engineering Degree and Master's. She has taught various subjects throughout her teaching career, especially computer programming and artificial intelligence. Her research lines are within the application of AI techniques: evolutionary algorithms for solving mathematical problems and neural networks applied to coastal engineering problems. Furthermore, she works in game development and the application of digital technologies to education. She has had several papers published related to the use of stereoscopic vision for segmentation and recognition as well as the reconstruction of space from grammatical models. All these lines have a common denominator, which is the use of AI techniques to deal with a wide range of different problems. She participates in many educational innovation projects related to the EEES. She is a member of AENUI (Association of University Teachers of Informatics) and of the *Cátedra Santander-UA de Transformación Digital* (Santander-UA Chair of Digital Transform).

Faraón Llorens-Largo (http://blogs.ua.es/faraonllorens) obtained his BSc and MSc in Computer Science in 1993, and a PhD in Computer Science by the University of Alicante in 2001. He also has a BSc in Education since 1982. He has been Head of the Higher Polytechnic School of Alicante (2000–2005) and Pro-Vice-chancellor of Technology and Educative Innovation at the University of Alicante (2005–2012). He is now head of the *Cátedra Santander-UA de Transformación Digital* (Santander-UA Chair of Digital Transform) at the University of Alicante, devoted to exploring new trends in digital transformation. He has received many awards related to education, like the Professional Sapiens 2008 award, from the Official Association of Computer Scientists of Valencia, or the AENUI award to educative quality and innovation 2013. He is currently professor at the University of Alicante and his research interests are focused on IT governance and the uses of Artificial Intelligence, games and gamification to improve education. He is a member of AENUI (Association of University Teachers of Informatics).