OPEN ACCESS
EURASIA Journal of Mathematics Science and Technology Education
ISSN 1305-8223 (online) 1305-8215 (print)
2017 13(4):1085-1106
DOI 10.12973/eurasia.2017.00660a

Supporting Mediated Peer-Evaluation to Grade Answers to Open-Ended Questions

Maria De Marsico
Sapienza University, Rome, ITALY
Filippo Sciarrone
RomaTre University, Rome, ITALY
Andrea Sterbini
Sapienza University, Rome, ITALY
Marco Temperini
Sapienza University, Rome, ITALY

Received 1 March 2016 ▪ Revised 6 October 2016 ▪ Accepted 8 October 2016

ABSTRACT
We show an approach to semi-automatic grading of answers given by students to open ended questions (open answers). We use both peer-evaluation and teacher evaluation. A learner is modeled by her Knowledge and her assessments quality (Judgment). The data generated by the peer- and teacher- evaluations, and by the learner models is represented by a Bayesian Network, in which the grades of the answers, and the elements of the learner models, are variables, with values in a probability distribution. The initial state of the network is determined by the peer-assessment data. Then, each teacher's grading of an answer triggers evidence propagation in the network. The framework is implemented in a web-based system. We present also an experimental activity, set to verify the effectiveness of the approach, in terms of correctness of system grading, amount of required teacher's work, and correlation of system outputs with teacher's grades and student's final exam grade.
Keywords: open ended questions, assessment, peer-evaluation, grade prediction

INTRODUCTION AND MOTIVATIONS

The use of open-ended questions is widely considered as an important tool in educational settings: it allows to evaluate effectively the learner's skills and the achieved cognitive level (Bloom et al., 1956; Anderson & Krathwohl, 2000), in particular when the higher cognitive abilities are concerned (Birenbaum et al., 1992; Palmer & Richardson, 2003). On the other hand, it requires a good deal of teacher work. So, making the grading task less cumbersome, yet preserving its reliability, would both be beneficial for the teacher, and encourage wide use of open ended questionnaires.

Correspondence: *Marco Temperini, Dept. of Computer, Control, and Management Engineering, Sapienza University, Rome, Italy.*
marte@dis.uniroma1.it

State of the literature

- The power of open ended questions, as evaluation tool is well known; they allow for more accurate analysis of the student's knowledge, than multiple choice questionnaire, and also provide better opportunities for self-reflection.
- The pedagogical benefit of peer-assessment is also recognized in literature.
- There are approaches to the automatic or semiautomatic classification of open ended questions, with various aims, including the educational one. In the educational field, it seems that the diligent intervention of the teacher is much needed and not as a light task (this paper doesn't promise a light task for the teacher, though).

Contribution of this paper to the literature

- To our best knowledge the integration of peer-assessment into an educational environment for the management of open ended questions is novel.
- We have shown that the performance of our framework is good, in terms of grading accuracy, although not yet such good to be carelessly applied in the classroom.
- We have also shown that the framework already provides a very interesting predictive power of the final outcome of the student; this is conducive to a direct application in the classroom, allowing to devise remedial activities when a learner ('s model) is predicted to have insufficient results.

In this paper, we show an approach to the management of open-ended questions through peer-evaluation. In turn, peer-evaluation is another crucial means in education, allowing developing metacognitive skills, while exercising one's knowledge about a subject matter. As discussed in Metcalfe & Shimamura (1994), metacognitive abilities imply both "knowing, and knowing about knowing". Planning strategies and schedules to carry out a learning task, checking on one's own understanding about a subject, applying new concepts, and evaluating one's own progress in a task, are all abilities that help successful learning.

We have devised a framework to allow (semi-)automated grading of open answers. In our approach, we use peer-evaluation, although not exclusively. Each student is requested to grade some (e.g. 3) of her peers' answers. However, the peer-evaluation is strengthened by requiring that a "relevant" selection of answers is also graded by the teacher. Different strategies for "relevance" are detailed in the paper. We represent the network of data, associated to peers and teacher assessments, by a Bayesian Network (BN), where the students are modeled by their Knowledge level about the question topic (K) and by the effectiveness of their evaluations, denoted as Judgment (J: a variable depending on K). In the BN, the answers of a single student have an estimated Correctness (C: a variable depending on K as well). K, J, C are stochastic variables represented by some unknown probability distributions, updated by evidence propagation in the network by means of Bayesian propagation. When a student marks a peer's answer, a corresponding Grade (G) is added into the network. G is postulated to show a dependency on both the J of the student giving

the mark, and the C of the marked peer answer. In conclusion, the BN structure is composed by a set of students, each one represented by six nodes (i.e. six stochastic related variables) in a directed and acyclic graph.

The framework is implemented by the "OpenAnswer" web-based system, embedded in a more general Learning Management System (LMS) called sLMS. OpenAnswer supports peer assessment on open-ended questions, enriched by the intervention of the teacher. Such intervention takes place after the peer-evaluation process, and goes through the following iterative process:

1) The teacher is requested to grade an answer, selected by a stated selection strategy; different strategies have been devised, each one measuring the "relevance" of an answer as a source for additional information (how the teacher's grading of an answer adds information in the BN is sketched in the next step 2).
2) The teacher's grade replaces the estimated C for that answer; this causes a propagated update on the J and K variables of both the student that gave the graded answer and the peers that evaluated it, as well as on the C of the answers of such peers; evidence eventually propagates on the whole network.
3) Steps 1) and 2) are repeated until a given termination condition is met. The choice of possible termination conditions is detailed in Sec.3.
4) Once the termination condition is met, the teacher stops grading: grades for the answers not yet graded by the teacher are "inferred" by the system, basing on the present state of their C, and this completes the grading session.

The OpenAnswer Front-End provides suitable interfaces to administer questionnaires and perform grading sessions (for both students and teachers). Its Back-End allows configuring and using the modules managing the BN as explained in detail in Sec. 4.

Before considering the system "in production", an intense experimentation is needed, in order to determine the effects, and the relative effects, of the possible choices/combinations of selection strategies (cfr. point 1) and termination conditions.

A first set of experiments gave encouraging results (Sterbini & Temperini, 2012b; De Marsico, Sterbini & Temperini, 2014; De Marsico, Sterbini & Temperini, 2015). This paper reports on the results from a second experimental phase. We analyzed strategies and termination options, to verify the effectiveness of the overall approach, in terms of correctness of system grading, amount of required teacher's work, and correlation of system outputs with teacher's grades and student's final exam grade.

Currently, we are reporting on a two steps experimental process. In a first phase, we have been gathering data, by letting students answer questions and perform peer-evaluation, and asking the teacher to grade ALL the answers, so to have a reliable ground truth. In the second phase, we have been using the datasets created in the first phase: for each dataset, we have been simulating a session of use of the system with a different choice of both the selection strategy and termination condition. In these simulations, only part of the teacher

grades are used, as long as some of them are required to perform the steps 1) and 2) in the above-sketched process. In this way, we can test different combinations of strategies and termination conditions, and evaluate them. Section 5 reports on the ways of such evaluation.

Our main research question addresses the possibility to consider the system as a reliable substitute of the teacher in a relevant part of the grading work, allowing both to ease teacher's burden on each single questionnaire and to support a wider usage of assessment based on open-ended questions. This research question translates, in our opinion, into the support of the OpenAnswer framework to a twofold kind of modeling (one explicit and another implicit):

a) Student modeling - it is managed as an explicit representation of skills, also relevant for the system to compute the final grade of peer-assessed answers; basically K and J provide such a model in the BN.
b) Teacher modeling – even if not explicitly represented in the BN, it can be witnessed by the influence that teacher's grades and assessment criteria have on the overall behavior of the BN.

We are also interested in a second research question: the possibility to use the student model maintained in the system, for an evolving prediction of the learner's performance, in a course where open-ended questionnaires are regularly used on its various topics.

The measures adopted to test our research questions are discussed in Sec. 5, where we also show the results of an experimental activity and discuss them with respect to the research questions. Previously, the next session provides an analysis of related work, while Sec. 3 describes the logical framework of our teacher-mediated peer-assessment of open answers, and Sec. 4 shows the OpenAnswer system, that has been used to collect and examine data.

RELATED WORK

Peer-assessment (Kane & Lawler, 1978) is an activity in which a student (or a group of students) is allowed to evaluate other students' assignments (and possibly self-evaluate own assignments). It can be organized in different ways, yet a basic aspect is that it can be considered as one of the activities in which social interactions and collaborations among students can be triggered. It can also serve as a way to verify how the teacher can communicate to the students her own quality requirements with respect to the learning topics: if this happens, assessments from peers and from teacher agree better (Sadler & Good, 2006; Falchikov & Goldfinch, 2000). Moreover, teachers may save grading time, or they could put it to better profit by analyzing how students assess and self-assess: this provides more information about the students' ability to judge, and in turn, about their knowledge. Somervell (1993) stresses that peer assessment is not only a pure grading procedure, yet rather a crucial part of a learning process through which skills are developed, while, in turn, it is part of the self-assessment process.

(Kane & Lawler, 1978) points out various kinds of peer assessment:

- ranking: each group member ranks all of the others from best to worst according to one or more factors;
- nomination: each member of the group nominates the member who is perceived to be the highest in the group according to a particular factor or performance;
- rating: each group member rates each other group member on a given set of performance, e.g., by assigning grades.

(Douchy et al., 1999) provides a valid, if slightly dated, review on peer- (and self-, and co-) assessment. Here, self-assessment refers to the involvement of learners in making judgements about their own learning. It increases the role of students as active participants in their own learning by fostering reflection on one's own learning. Six main factors are discussed that can influence the quality of self-assessment: the influence of different students' abilities on the accuracy of self-assessment, the time effect, the accuracy of self-assessment in relation to teacher assessment, the effect of self-assessment, methods for self-assessment and the content of the self-assessment. The analysis is concentrated on aspects of validity, accuracy, fairness, effectivity. Self and peer assessment are combined when students are assessing peers, while being included in the group to assess. This combination fosters deeper reflection on the one's own learning compared to that of the other members in the group. Co-assessment implies that the teacher plays a significant role in the process: the participation of students and staff in the assessment process allows students to assess themselves but also allows the teacher to maintain the necessary control over the final results. A recent proposal (Põldoja et al., 2014) applies self- and peer-assessment to competency modeling and profiling for teachers. Teachers' digital competencies are analyzed, by a framework aiming to evaluate personal educational technology competencies.

In some works, peer and self-assessment marks are compared with teacher marks in order to assess their accuracy. Other studies aim to evaluating the inaccuracy implied by the fact of relying, all or partially, on this kind of assessment for students' performance evaluation. Falchikov & Goldfinch, (2000) provide a meta-analysis on 48 quantitative studies comparing peer and teacher marks. Peer assessments were found to be closer to teacher assessments when global judgements (marks) are required after a good understanding of assessment criteria. Another outcome of this meta-analysis was that peer assessments better resemble faculty assessments when rating academic products and processes, rather than professional ones. Finally, studies with high design quality appeared to be associated with more valid peer assessments than those having poor experimental design.

In (Cheng & Ku, 2009) the effects of reciprocal peer tutoring on student achievement, motivation, and attitudes are analyzed. Most interviews with students after the experimentation suggested designing cooperative projects, allowing students to pick own groups, and facilitating group cooperation as keys for a successful experience. In (Sterbini &

Temperini, 2009) we discussed how the integration of a reputation system into the learning environment, can have effect on peers' motivations, and increase the learning results".

With respect to peer-assessment, it has to be cited also the case of large scale peer-assessment, like in Massive Open Online Courses (MOOCs). It is known to achieve good marking accuracy results (Piech et al., 2013). However, it is worth noticing that this relies on the very large amount of information available, so the matter could be different (and less satisfying) in a normal class, where the number of students is not so massive.

Moreover, notwithstanding the advantages, peer-assessment may also suffer from biases, originated by several kinds of circumstances, such as friendship and reputation. Therefore, a mediated approach is sometimes preferred (like in the case of the approach presented in this paper). This implies the direct handling of a subset of the assignments by the teacher, to preserve good accuracy and reliability of results even with traditional classes, where less information is available than is the case of possibly huge online groups.

In regard to the automatic analysis of open answers several approaches have been proposed, not only for education. Methodologies of Computer Science, have been used in such an endeavor, ranging from data mining, through natural language processing and concept mapping, to semantic web techniques. These techniques also differ with respect to the different degrees of human intervention they require.

Applications from data mining tackle the problem of summarizing opinions out of marketing-commercial-oriented questionnaires. The extracted opinions, in these cases, are used to define and monitor the reputation of a product (Yamanishi & Li, 2002; Morinaga et al., 2002).

The approach in (Jackson & Trochim, 2002) has similar aims: the idea is to use concept mapping to develop a scheme of coding (basically semantic labels that tag the answers or parts of them) and then evaluating and re-examining such schemes. The process goes through five steps, and is good especially for applications on a "free list in context" type of answers. The codes are associated to parts of the answer, in a semi-automatic way. The human intervention of "coders" and "classifiers" is much needed. The labels allow to classify the chunks of answers and, consequently, to collect and classify the answers content. The just mentioned approaches are not deemed directly to education, yet such techniques show applications in e-learning, as reported in (Romero & Ventura, 2010).

An approach to (a partially semi-) automatic assessment of open-answers, through ontologies and semantic web technologies is in (Castellano-Nieves et al., 2011). There, the domain of knowledge of the questions, plus some other aspects of the educational process, are defined through ontologies. "Semantic annotations" are defined to label the questions by the ontological elements of the correct answer. The student's answer is also labeled in terms of the ontology; then the answer can be analyzed, by evaluating the similarity of its ontological elements against the question's ones; grading follows. Teacher's intervention in

the process is high at the beginning, when course ontology and questions' semantic annotations have to be defined, while it seems fairer later on, when the answers' semantic annotations (stated by the system) have to be checked, corrected and integrated.

(El-Kechaï et al., 2011) presents a system working as a cognitive diagnosing tool in an algebra-bound intelligent tutoring system. Therein open answers are analyzed to determine the implicit conceptions of the students, and to treat the uncovered mis-conceptions. The approach is based also on the use of a symbolic computation system, allowing for the practical manipulation of formulae. The system has proven worthy when applied to answers constituted by purely algebraic expressions, without parts comprising natural language. On the other hand, the free introduction of text in the answer would be much beneficial, making more explicit the reasoning applied by the learner while developing the answer.

Sterbini & Temperini (2012a) shows an approach to open answers grading, based on Constraint Logic Programming (CLP) and peer assessment, where students are modeled as triples of finite-domain variables. The CLP Prolog module supported the generation of hypotheses of correctness for answers (grounded on students' peer-evaluation), and the assessment of such hypotheses (also based on the answers already graded by the teacher).

The use of a Bayesian Network in an educational setting is of course not novel. An approach to student modelling is in (Conati et al., 2002). The authors propose the use of BNs to model the learner in an Intelligent Tutoring System (ITS). The aim is to support the relevant activities of knowledge assessment, plan recognition and prediction in problem solving tasks. The last two activities aim to uncover the intentions behind a student's decision, and to infer the next ones.

In OpenAnswer (Sterbini & Temperini, 2013) the peer is presented with a set of assessing criteria, to be used while marking; the criteria are defined by the teacher, who also is supposed to use them in her grading job. In our experience, too many criteria might result cumbersome for the peers. We have not investigated, though, on this aspect. In the literature, the specificity of "scoring criteria" has been identified as an important factor against the problem of having assessors that limit the range of their marks to a subset (typically in the high end) of the scale; in this case the problem is twofold: involving both peers leniency and shrinking of the marking scale (Miller, 2003).

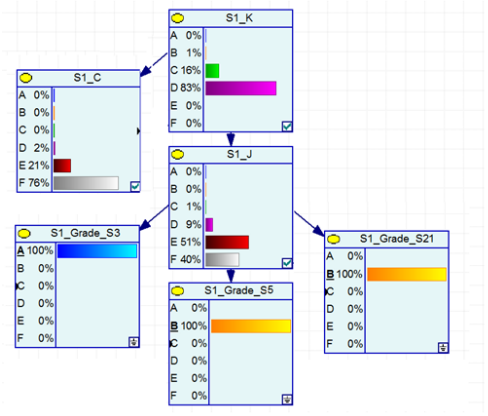
An aspect of the research in peer-assessment regards the number of peer-evaluations that a same job should undergo during the peer-evaluation process. In OpenAnswer this is configurable, with the default set to 3. In the literature, it is found that more feedbacks on the same job make the peer performing more complex revisions on her product, ending up with a better result (Cho & MacArthur, 2010).

THE ASSESSMENT FRAMEWORK

As mentioned, the OpenAnswer framework supports the use of peer-evaluation, exploited in the first phase (*marking*, by students), mediated by a second phase (*grading*, by teacher, that takes place after the abovementioned marking phase).

In the marking phase, the students are required to answer a question. In our present experimentations, we submitted one question at a time related to the course topics, in order to better highlight the peer-assessment dynamics. Each student also provides an assessment to the answers coming from a subset of the peers (usually 3). The questions built by the teacher, are annotated by assessment criteria that are supposed to be adhered to, both by the peers and (successively) by the teacher.

A student is modeled by a set of stochastic variables, plus three fixed values, all managed by a BN. Namely, for each student, the first stochastic variable *K* represents the learner's state of Knowledge; the second stochastic variable, *J* represents the learner's ability to Judge the answers of her peers. These two variables provide a kernel for an individual portion of the whole network (see **Figure 1**). The other variables in such individual part of the BN are:



P(C|K)	K						
		A	B	C	D	E	F
C	A	20%	9%	1%	1%	4%	1%
	B	40%	20%	9%	7%	6%	1%
	C	20%	40%	20%	12%	10%	1%
	D	12%	20%	40%	20%	15%	7%
	E	7%	9%	20%	40%	25%	40%
	F	1%	2%	10%	20%	40%	50%

Figure 1. (left) A sample fragment of the BN in OpenAnswer, showing the individual sub-network of student S1. K and J are stochastic variables computed during the network maintenance. C, the correctness of the answer given by the student S1, is a stochastic variable as well. On the contrary, the (three) evaluations given by this student (to the answers of peers S3, S5, S21) are just fixed: their values, together with those given by the other peers to the same answers, will concur in composing the distribution of probability for the related C values. (**right**) A sample CPT. It is the P(C|K) conditional probability table used in the experiments. It was devised based on experience and is fixed in the system, as well as for the other CPTs we used in this study. We discuss in the conclusions section about the possibility to learn the CPTs from previous experience with the classroom.

- The Correctness (C) of the learner's answer (inferred by the system, and possibly later on stated by the teacher);
- The Grades (G) the student gave to peers' answers (usually three in our work).

Some of these variables are *conditionally dependent* on others, so the values of certain variables influence the values of their dependent ones, making a process of evidence propagation to happen in the BN. The dependence of a variable from another is stated in the BN through the application of a Conditional Probability Table (CPT), see **Figure 1**.

In particular, the variables K, C and J, are computed according to the following observations:

- The ability to assess, represented by the J variable, is dependent on the knowledge about the topic, represented by the K variable. In fact, the skill measured by J is at a higher cognitive level in Bloom's taxonomy (Bloom et al., 1956), and this may well imply a dependency from K. Such dependency is defined in our framework by the P(J|K) CPT;
- Since the answer is an open one, we assume that the variable C has to be dependent on the variable K as well, by the P(C|K) CPT;
- For the values of the variables C and G we adopt the well-known grading range [A,…,F]. Since in our experiments we used a more traditional (as used in our country) decimal grading, we also provided and made use of a mapping as follows: F comprises marks in [0, 5.5), E is [5.5, 6.5), D is [6.5, 7.5), C is [7.5, 8.5), B is [8.5, 9.5), and A is [9.5, 10].

In this setting, C and the Gs control the evidence propagation in the BN. In the phase of peer-evaluation, each peer sets three Gs that are in turn associated to the assessed answers, so to determine the current value (distribution of values) of each learner's K and, in turn, of each answer's C.

In the subsequent phase of teacher grading, the framework supports the teacher by:

1) Suggesting an answer to grade that, according to one of the selection criteria detailed below, will propagate significant information into the overall BN;
2) Propagating into the network the added information provided by the teacher's grade. In fact, when the teacher grades the answer, the C variable that was represented by an unknown distribution of probability becomes a fixed value for the student. Once C is stated, there follow propagation effects:
    a. On the J variables of the peers that assessed that answer (since, now that C is fixed, something can be said about their assessment skill);
    b. On the K of the peer-evaluators,
    c. On the K of the student giving that answer,
    d. Eventually on the whole BN;

3) Iterating the previous steps until a termination condition is met. The termination condition states that new information, coming from additional teacher's grades, would be less decisive, so the teacher might as well stop grading;
4) Releasing the answer final grades: both those directly given by the teacher, and those inferred based on the current probability distribution of the associated C for the others.

In an OpenAnswer session, a set of variables is instantiated for each student (K, J, C, and a G variable for each peer-assessment given by the student). K and J are in the interval [A, F] as well. Once the network is created with the initial evidence, the Lauritzen belief propagation algorithm computes the initial probabilities (Huang and Darwiche, 1996), i.e., all the CPTs are instantiated.

We have devised two basic selection strategies, plus a random-based one, for suggesting the "next best answer to grade" to the teacher during the grading phase:

- max_wrong: selects the answer most probably mapping onto F: the rationale for suggesting to the teacher to grade such an answer first is in that a student would more easily accept F if coming directly by the teacher (rather than by the system);
- max_entropy: selects the answer showing highest entropy on C; the system knows less about it, since the associated G variables (peer grading) hold very different values, and its grading would pour more information into the system;
- random: selects randomly the next answer; this strategy is mostly used for testing purposes, to evaluate the difference with respect to a motivated strategy.

Two families of termination criteria for the teacher's grading are applicable:

- if max_wrong is the strategy to select the next answer to grade, we may apply one of the following three termination conditions:
    - no_wrong: the process stops when no answers would be automatically graded F by the system;
    - no_wrong2: as above but with p(grade=F) <= ½;
    - no_wrong3: as above, but with p(grade=F) <=1/3
- if max_entropy is the strategy to select the next answer to grade, we may apply one condition out of three again:
    - no_flip(N): the inferred grades corresponding to C variables have not changed in the last N grading steps (where N=1, 2, 3)
- if the strategy is the random one, any of the above termination criteria can be applied.

In **Figure 2** an UML activity diagram is shown, representing the operations described above. The three main actors, i.e., the OA System, the Teacher, and the Student are represented in the three vertical swimlanes.

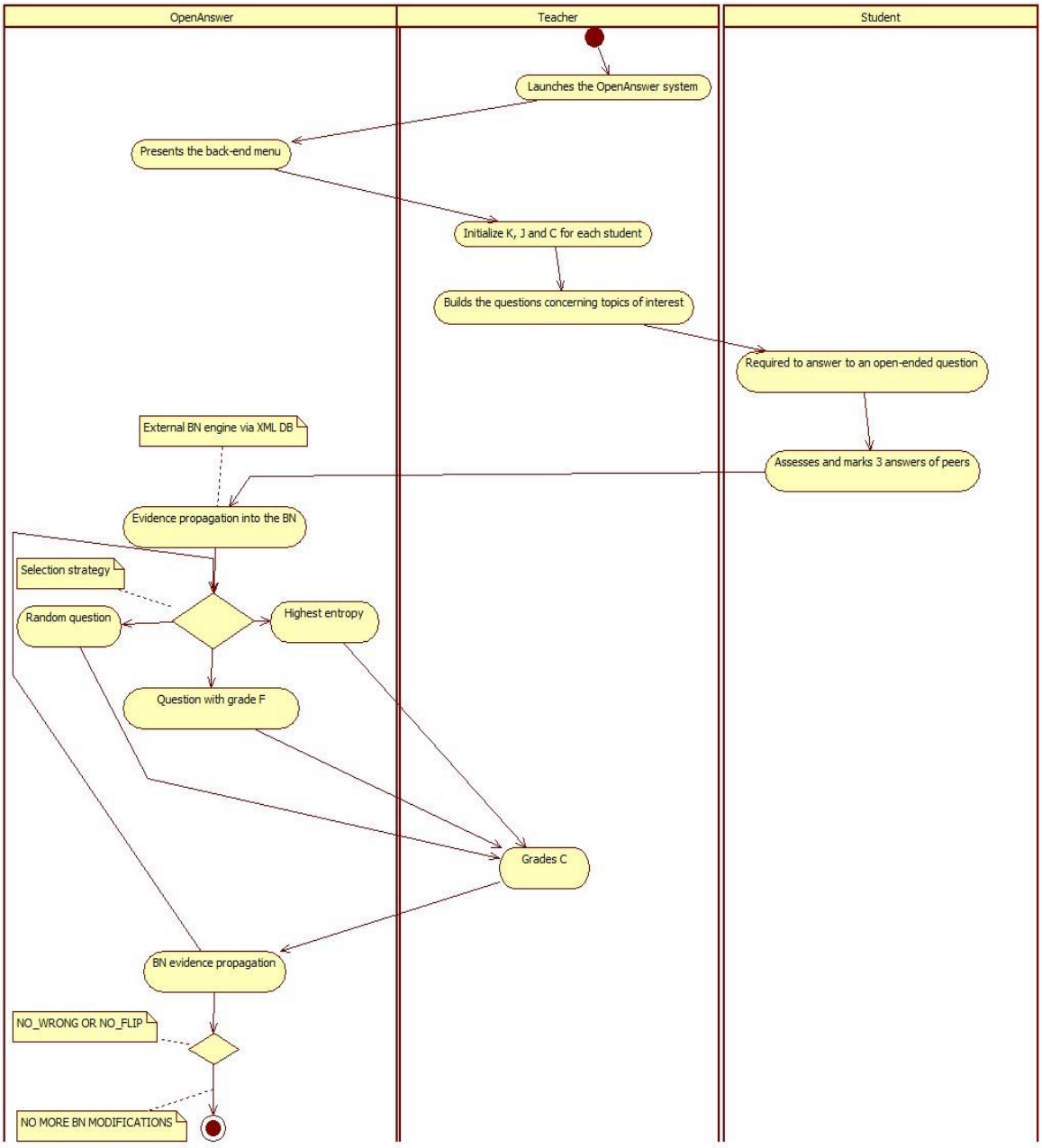


Figure 2. The Activity Diagram of the assessment process.

THE OPENANSWER SYSTEM

In this Section, we show the functional architecture of the OpenAnswer system with its embedding LMS. OpenAnswer is a web-based educational module, designed for Teachers and Students, capable to allow peer-assessment, with the aim of inferring a reliable grading of open-ended questions. It is embedded into the sLMS learning platform, a web-based 3-tier application, developed in php language and with data stored and managed through a XML and a MySQL database.

The Overall Architecture

From a management point of view, the sLMS system can be thought as divided in two parts: a *Front-End* and a *Back-End*. The Front-End provides functionalities depending on the class of user. It allows Teachers to define questions and questionnaires, to administer questionnaires, and to proceed to assessing the answers, possibly (yet not necessarily) with the support of the system. The same Front-End allows students to work on questionnaires (by answering and peer-assessing answers) and check the results of such a work, in terms of grades obtained in the questionnaires (the C values for the various answers) and of the present state of their model (K, J). The Back-End of the system allows managing the computational and representational aspects related to the BN and to the answers grading. This BN module, in the current version of the system, is an external module, in order to allow trying different BN models, as well as, in future, different machine learning approaches. Moreover, it implements the available selection strategies and termination conditions explained above, to support the teacher's evaluation work. The overall functional architecture of the sLMS system is shown in **Figure 3**, where the OpenAnswer module is highlighted with a different filling color and the BN Module is represented as an external module, connected to the XML database from which it receives the data gathered out of the peer-assessment sessions.

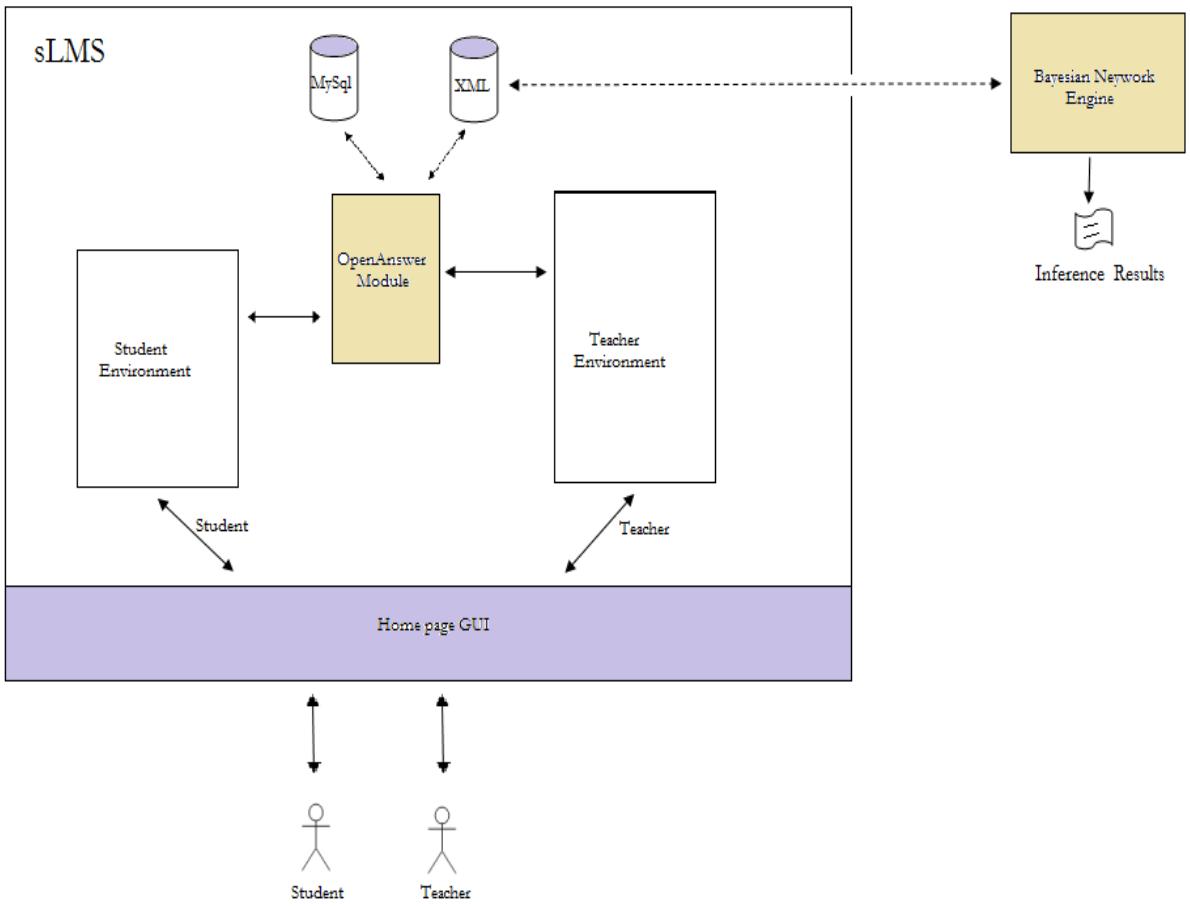


Figure 3. The functional architecture of the system: students and teachers can interact, respectively, with the student and the teacher environments through the front-end management modules while the Bayesian Network Module represents the back-end module.

The OpenAnswer Module

Here we provide a description of the phases of use of the OpenAnswer web-based software module. The module itself is launched from within eLMS. In particular, then, the teacher sees the peer-assessment environment as shown in **Figure 4**. Here the teacher can define questions and assessment criteria related to some topics of interest (*Questions Management* in figure).

Questions and criteria are defined separately. A new questionnaire can be defined (*Questionnaires Management* in figure) by selecting among the available questions, and contextually associating to each one of such questions some of the defined assessment criteria. An experiment in this module (*Experiments Management* in figure) is basically composed of a questionnaire and a class of students. An experiment can be held in different contexts and with different options of peer assessment, so there are different "sessions" for the same experiment (*Sessions Management* in figure).

The other functionalities in figure give access to the teacher's assessment task on the answers given by the students in a session, and to a visualization of the results of such session (peer's marking, and teacher's grading).

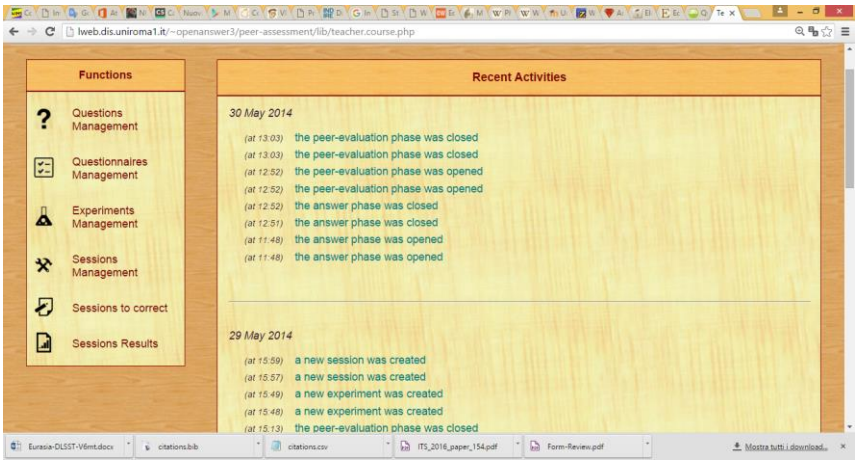


Figure 4. The peer assessment environment launched by the teacher. The logs shown in the figure regard the two groups of students in which we partitioned the class, to perform the experiment, so they appear to be repeated.

The questions are defined through an editor that allows defining them as formatted text enriched with images and other multimedia resources (**Figure 5**).

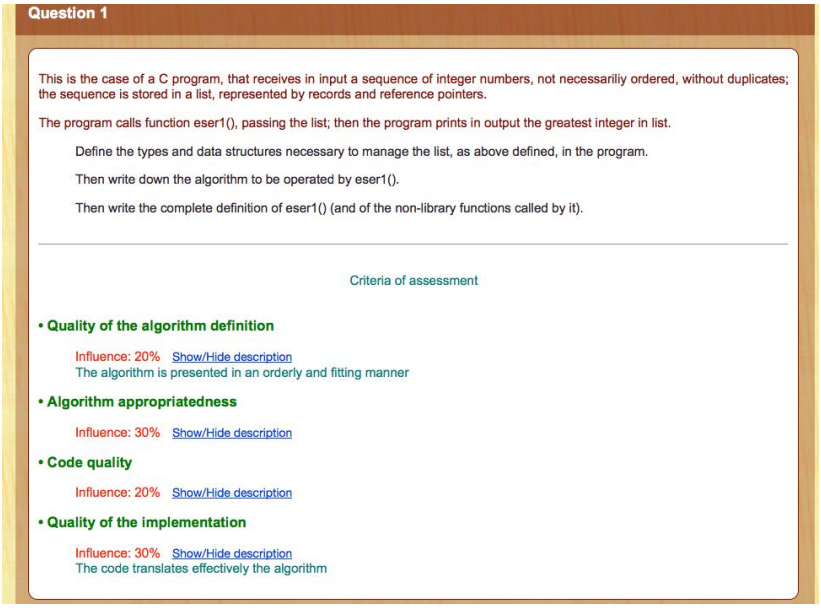


Figure 5. The OpenAnswer Front End: a question associated to its assessment criteria.

A session configuration (**Figure 6**) describes the use of a questionnaire by specifying i) how many questions do constitute the questionnaire; ii) how many answers will be evaluated by each peer; iii) whether the peer will know the name of the peer who submitted the answers she assesses; iv) whether self-evaluation (peer assessing peer's own answer) is possible.

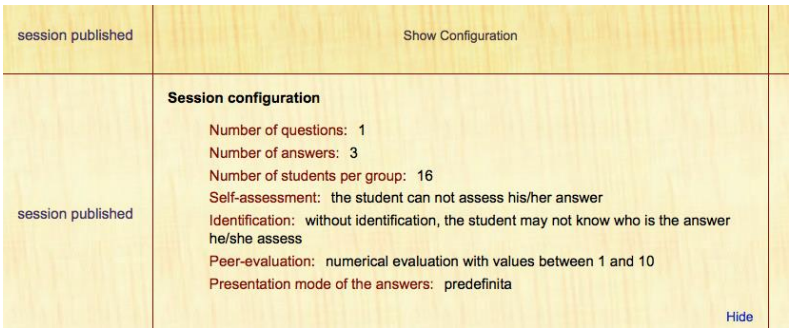


Figure 6. A session configuration (How the questionnaire will be administered).

During students' activity, the *peer assessment* module stores their answers and their peer assessments, working as a data gathering module. At the end of this activity, the assessment phase performed by the teacher can start. It develops as an iteration of the steps detailed earlier in Sec. 1. Firstly, the system suggests the next answer to grade (it appears on top of the list of answers available to the teacher). Secondly, the teacher selects an answer to grade (the teacher is free to choose a different one than the first in the list, although the system would follow a more stable process if the teacher accepted its suggestions). Thirdly, the teacher grades the selected answer.

EVALUATION AND TUNING OF THE SYSTEM COMPONENTS

As mentioned in the introduction, the main aim of our experimentation was to evaluate the behavior of our framework with respect to different settings/configurations. In this paper, the interest is focused only on the choice of a selection strategy and of a termination condition. Namely, out of the several selection strategies and termination conditions defined in Sec. 3, different combinations are possible, and from each one different overall performance of the framework may derive. Presently we are still in the experimental stage, so we actually ask the teacher to use the system while getting no advantage yet, in terms of saving grading time (see below the definition of the data gathering phase). On the contrary, when OpenAnswer will be "in production", and so actually used by teachers to manage questionnaires by grading only part of them, the merits of each possible combination of the supported selection strategies and termination conditions will have to be clear (and probably several of the possible combinations we analyze here will not be supported at all). At the same time, we also wanted to validate our framework of mediated peer-assessment, according to the research questions stated in the introduction. We are reporting on tests performed according to a two phases method:

1) *Data Gathering phase*: A session of use of OpenAnswer with a classroom is planned by the teacher, by the:
    a. definition of the questionnaire (usually with a singleton question in our experiments);
    b. configuration related to a number of peer assessments per student, and anonymity (we always required three markings of anonymous peer answers by a student);
    c. partition of the classroom in groups (due to the computational demands of the BN management we can have groups of peers not above 15/20 students).

    Consequently, the session takes place, by giving the students a time period for answering and a subsequent time period for assessing the (usually 3) peer answers.
    Then the teacher is required to grade all the answers, in order to provide the next phase with reliable ground-truth.
2) *Simulation phase*: we simulated several sessions of use of the full-fledged system. In each simulated session:
    a. First a dataset from the previous phase (D), a selection strategy (S), and a termination condition (T) are stated;
    b. Second a process of teacher grading of the questionnaire is simulated:
        i. an answer from D is selected for teacher's grading, using S;
        ii. the grading is simulated by extracting the real-teacher grade for that answer from D;
        iii. T is checked, to decide to continue the process (cycling again from point i.) or just to stop it.

As mentioned earlier, the sessions were simulated using the ground-truth data provided by the Data Gathering phase, that entailed three separate and sequential stages, involving three questions on computer programming (one question per each stage), at undergraduate university level (first year). During the first stage, two groups of 13 students were involved; two groups of 11 students participated in the second stage; and two groups, of 9 and 11, were active in the third stage. The students were always the same (except for some of them leaving the experimentation between the stages). The students were all from the same course (Programming Techniques – *Tecniche della Programmazione*). Between the stages there was an interval of 3-5 days.

In order to meet the first research question we measured the following variables in each simulation:

- *Length* (L) of the teacher's grading session, i.e. the number of teacher's (real) grades used during the simulated grading process. This, in fact, represents what the actual teacher's grading job would be;
- *Framework's accuracy*, i.e. how close the grade inferred by the framework is to the true one (i.e. to the teacher's grade coming from the ground-truth, which is "exact" by definition). In particular, we measured the number of answers whose grade has been

inferred exactly (OK), the number of grades that were inferred within a distance of one mark from the exact one (so it is either exact or wrong by one mark only); we call IN1 the latter measure. (We will also use IN2: the number of grades that were inferred within a distance of two marks from the exact one);

- *The correlation* of the grade inferred by the system Vs. the exact one.

The second research question is tackled by analyzing the correlation (Pearson) of the student model (K variable) with respect to the grade taken by the learner in the final examination of the whole course. The final exam is common to all the students in the course; it requires a written class-work (consisting in having to code functions to solve a stated problem), together with a discussion (of both the written work and of other topics in the course). The topics met in the questions answered during the experimentation are, of course, a subset of the course syllabus.

Analysis and discussion about the first research question

Table 1 shows some data deemed to help performance evaluation: the percentages are computed as the average of the related percentages in the various simulations. In particular, L (the Length of the grading phase) is expressed as the average percentage of answers for which the real teacher's grades was used in the simulation (i.e. it was extracted from the ground-truth of the dataset, in order to grade the answer selected for grading – cfr. step i. above). This average is computed on all the simulations performed. Accordingly, OK is the average percentage of grades that were inferred exactly: each grade is computed from the final state of the variable C for that answer, after that the termination condition was reached; this inference is exact if it is equal to the related teacher's grade stored in the dataset. In summary, (OK+L) is the overall exact grading produced by the framework (where OK is due to the system, and L to the teacher). Then, IN1 is the average percentage of grades inferred by the system with a difference of not more than one mark from the exact grade, and IN2 is the average percentage of grades inferred with a difference of not more than two marks from the exact grade. Actually IN2 does not show a particularly rich significance, especially in a marking system based on 6 elements: here we include it just to give a sense of closure to the picture. INFERRED is the total number of grades inferred by the system, so the ratio OK/INFERRED (and, to some extent, also the ratio IN1/INFERRED) might be considered as a representation of the overall performance of the grade inference. In the table we compare various combinations of strategies to select the next answer for the teacher to grade (including random selection) and related termination conditions. We also consider an additional strategy using only peer-evaluation ("none", i.e. no teacher grading): this corresponds to using (i.e. feeding our frameworks with), directly the raw data coming from the peer-evaluation contained in the dataset. In this case the inferred grade is computed basing on the distribution of probability assigned to C (the correctness of an answer), untouched by the effects (propagation) of teacher's grading. So, in the table, the column "none" relates to the "pure" peer-assessment case. Here L has no meaning (no teacher's grades were used), yet, to allow a comparison, we associated the percentage of grades correctly inferred from the overall peer

assessment to (OK+L). Accordingly, the cell with label (IN1+L) represents the percentage of grades inferred at a distance less_than or equal_to one from the exact grade. And similarly for the others.

Table 1. Performance comparisons on strategies and termination conditions: ME= max_entropy strategy and MW=max_wrong are placed on same rows, yet their values lie on separate columns and can be distinguished, as ME is applying only no_flip termination conditions, while MW uses only nowrong* termination conditions. For instance, ME with noflip(3) has a value of L equal to 54%, while, on the same row, MW with nowrong3 has L=26%. Moreover, "none" means "no teacher grading", and corresponds to pure peer-assessment in OpenAnswer. Finally, RANDOM means random selection of next answer to grade (using all our termination conditions).

	noFlip(1)	noFlip(2)	noFlip(3)	none	nowrong	nowrong2	nowrong3	
L	24%	41%	54%	(0%)	26%	12%	26%	**ME-MW**
	26%	50%	67%	---	65%	41%	46%	**RANDOM**
OK+L	46%	60%	67%	30%	44%	36%	44%	**ME-MW**
	50%	63%	79%	---	75%	56%	62%	**RANDOM**
IN1+L	78%	82%	86%	67%	79%	71%	79%	**ME-MW**
	76%	83%	85%	---	91%	81%	85%	**RANDOM**
IN2+L	96%	96%	97%	95%	98%	96%	98%	**ME-MW**
	94%	96%	99%	---	100%	97%	100%	**RANDOM**
OK/INFERRED	31%	36%	36%	30%	26%	29%	26%	**ME-MW**
	34%	25%	19%	---	41%	32%	36%	**RANDOM**
IN1/INFERRED	72%	72%	76%	67%	70%	67%	70%	**ME-MW**
	69%	65%	30%	---	74%	75%	77%	**RANDOM**
IN2/INFERRED	95%	94%	95%	95%	97%	96%	97%	**ME-MW**
	92%	92%	86%	---	100%	93%	100%	**RANDOM**

In **Table 1**, darker cells represent good results (that we considered reasonable to set as L not greater than 50%, OK+L at least 60%, IN1+L at least 80%, OK/INFERRED greater than 30%). Lighter grey cells point out good results yet associated to bad values for L (e.g. OK/INFERRED=36% is good, but it comes from a vexing 67% of teacher's grades, so this combination is not so good after all). Sadly, it does not seem so far, that our selection strategies improve over the random one to such a large extent. A random choice does a good job, although it has to be noted that it does it while necessarily exploiting our termination conditions. On the other hand, it is the max_entropy (ME) selection strategy that gets the best result: OK/INFERRED=36%, and L=41% (with no_flip(2), i.e. with the termination condition triggered when C didn't change after the propagation of the last 2 real teacher's answer grades). This allows us to see the merits of the various selection strategies, and also to conclude that the influence of the termination conditions appears to be greater than that of the selection strategies, which is somewhat expectable. To reach a conclusion regarding the first research question we also checked the Pearson correlation between the grade inferred by the system for an answer and the exact (teacher's) one.

Table 2 shows such correlation: the best, and more stable, correlation is still for the max_entropy (ME) strategy; the "pure" peer assessment conducted in the framework ("none") has also a good correlation. The grading inferred by the OpenAnswer system seems to be quite well correlated with the teacher's grading, confirming an effective influence of the teacher in the system.

Table 2. Correlation between inferred grade and teacher's grade, on all answers.

none	0.76
ME + no_flips	0.81 – 0.84
MW + nowrongs	0.59 – 0.69
RANDOM + noflips	0.62 – 0.73
RANDOM + nowrongs	0.47 – 0.56
MARK – GRADE	0.25

The table also reports on the direct correlation between the "raw" marks provided by the peers (not operated by the framework) and the teacher's grade (MARK-GRADE row). This is computed without the aid of the OpenAnswer and its BN: In this case we "inferred" the grades coming from the peer-evaluation just as the average of the peer marks. The fact that this correlation is really low (0.25) lets us conclude that the "quality" of the unprocessed assessment was low as well, and it couldn't be realistically used for grade prediction. On the other hand, the OpenAnswer framework, basing on the same data (in the case "none") reaches a much higher correlation with the exact grade (0.76): this should allow to conclude that the approach through BN implemented by the framework, is reasonable and fruitful. In conclusion, when we use the whole framework, and allow it to exploit the needed real teacher's grades, the correlation increases to quite good values between 0.81 and 0.84 (max_entropy again).

The data discussed here can't allow to say that OpenAnswer is ready to be deployed in real life classrooms: the amount of work still required to the teacher in the better cases, and the overall success rate of the inferred grading (OK) are still not good enough in our opinion. As for the latter, we would aim at achieving an OK/INFERRED of 75% at least, with a consequent IN1/INFERRED of 90% at least. On the other hand the data analysis, the comparison with "none" and "RANDOM", and ultimately the observation that a kind of influence of the teacher in the system does exist, allow to draw a confirming answer to the first research question.

Analysis and discussion about the second research question

Table 3 reports on the correlation between the value K of each individual student, after the end of the sessions simulations, and the grade obtained by the student in the final exam. The best results are in darker background, and generally max_entropy (ME) performs better. We remind that the simulations regarded a three stages process, involving the same students

and three successive questions; at each following stage the starting value for the student's K was the one computed by the end of the previous stage (in other words we used the K obtained through the answer to the previous question, as an initial value for the K to be computed during the grading session for the next answer). In the table, the data refers to the last stage results, so it can be considered as representative of the "final" student model, after the three stages (and the three answers). As such, even after only three peer-assessment sessions, the model shows a good correlation with the final grade in the course exam. The (three) questions used are related to just a part of the topics treated in the course; however, such parts are prominent ones in the set of course topics. The first question asked to define and implement an algorithm to determine the greatest data in a collection, stored as an array; the other questions were related to counting occurrences and finding the two greatest elements in a collection, represented as a dynamic data structure (linked list). We deemed the topics of the questions sufficiently representative of the topics met during the final course exam, so to make the student model sufficiently representative in turn. From the data we can conclude that the evolving model managed by OpenAnswer is promising as a prediction means for student performance, able to suggest, if needed, the timely administration of remedial activities.

Table 3. Correlation between K and the course final exam grade of the students.

noFlip(1)	noFlip(2)	noFlip(3)	none	nowrong	nowrong2	nowrong3	
0.71	0.69	0.72	0.53	0.60	0.54	0.60	**ME - MW**
0.66	0.57	0.63	---	0.46	0.59	0.74	**RANDOM**

CONCLUSIONS

Open ended questions are a powerful means to help checking the learner's state of knowledge. They are time consuming for the teacher, though, and if their use is to be fostered, then supporting the grading job becomes an important aim. Answering open ended questions is more challenging than filling in multiple choice questionnaires. It would be reasonable to expect some pedagogical advantages from such challenge in itself. However, it is sure that bringing peer-assessment into the picture may allow to enjoy for its many pedagogical advantages and, at the same time, help supporting the grading job of the teacher. In this paper we have shown an approach to the integration of peer-assessment into an environment for the management of open ended questions. We have shown that the performance of our framework is good, in terms of grading accuracy, although not yet such good to be carelessly applied in the classroom. We have also shown that the framework already provides a very interesting predictive power of the final outcome of the student; this is conducive to a direct application in the classroom, especially if a longer (than three) series of OpenAnswer sessions is programmed, during the semester, allowing to devise remedial activities when a learner ('s model) is predicted to have insufficient results. In terms of future work, there is space for improvement on both the BN model and the selection strategies used to determine the next answer for the teacher to grade. Regarding the BN model, the CPTs we used could be

improved at least by parametrizing them, and then optimizing with respect to the inference output quality. Moreover learning them from examples, through Machine Learning techniques, might improve the model. This line of study will be followed with further applications of the system (and so with further and bigger datasets available). Regarding the strategies, the max_entropy selection strategy shows comparatively good performances. A better effectiveness could be won through a selection strategy that does a deeper analysis of the information gain obtained by grading it. Finally, we have seen evidences of an influence of the teacher in the assessment framework, that can prelude to a more explicit definition of a teacher model related to peer-assessment. This could be done by defining for the teacher an instance of the same subnetwork template used for student, connected through a G variable for each graded student. This would also allow modeling the participation of more teachers.

REFERENCES

Anderson, L. W. & Krathwohl, D. R. (eds.) (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*, Allyn & Bacon, Boston, MA (USA)

Birenbaum, M., Tatsuoka, K. & Gutvirtz, Y. (1992). Effects of Response Format on Diagnostic Assessment of Scholastic Achievement, *Applied Psych. Measurement*, 16(4), 353-363 (1992)

Bloom, B.S., Engelhart, M.D., Furst,. E.J., Hill, W.H. & Krathwohl, D.R. (1956). *Taxonomy of educational objectives: The classification of educational goals. Handbook I*. McKay, New York

Castellanos-Nieves, D., Fernández-Breis, J., Valencia-García, R., Martínez-Béjar, R. & Iniesta-Moreno, M. (2011). Semantic Web Technologies for supporting learning assessment, *Inf. Sciences*, 181(9)

Cho, K. & MacArthur, C. (2010). Student Revision with Peer and Expert Reviewing. *Learning and Instruction*, 20(4)

Cheng, Y. & Ku, H. (2009). An investigation on the effects of reciprocal peer tutoring. *Computers in Human Behavior*, 25.

Conati, C., Gartner, A. & Vanlehn, K. (2002). Using Bayesian Networks to Manage Uncertainty in Student Modeling. *User Modeling and User-Adapted Interaction*, 12, 371-417

De Marsico, M., Sterbini, A. & Temperini, M. (2014). Adding time and propedeuticity dependencies to the OpenAnswer Bayesian Model of Peer-Assessment. In Proc. 13th IEEE Int. Conf. on Information Technology Based Higher Education and Training, ITHET

De Marsico, M., Sterbini, A. & Temperini, M. (2015). Towards a quantitative evaluation of the relationship between the domain knowledge and the ability to assess peer work. In Proc. IEEE ITHET 2015, 1-6

Douchy, F., Segers, M. & Sluijsmans, D. (1999). The use of self- peer and coassessment in higher education. A review. *Studies in Higher Education*, 24(3), 331-350

El-Kechaï, N., Delozanne, É., Prévit, D., Grugeon, B. & Chenevotot, F. (2011). Evaluating the Performance of a Diagnosis System in School Algebra, In Proc. ICWL, LNCS 7048, Springer

Falchikov, N. & Goldfinch, J. (2000). Student peer assessment in in higher education. A meta-analysis comparing peer and teacher marks. *Review of Educational Research*, 70(3), 287-322

Jackson, K. & Trochim, W. (2002). Concept mapping as an alternative approach for the analysis of open-ended survey responses. *Organizational Research Methods*, 5, Sage.

Huang C. & Darwiche, A. (1996). Inference in Belief Networks: A Procedural Guide. *International Journal of Approximate Reasoning*, 15, 225-263

Kane, L. S. & Lawler, E. E. (1978). Methods of peer assessment. *Psych. Bull.*, 85, 555-586

Metcalfe, J. & Shimamura, A. P. (1994). *Metacognition: knowing about knowing*. MIT Press, Cambridge

Miller, P. (2003). The Effect of Scoring Criteria Specificity on Peer and Self-assessment. *Assessment & Evaluation in Higher Education*, 28(4).

Morinaga, S., Yamanishi, K., Tateishi, K. & Fukushima, T. (2002). Mining product reputations on the Web. In Proc. 8th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data mining, KDD'02, 341-349

Palmer, K. & Richardson, P. (2003). On-line assessment and free-response input-a pedagogic and technical model for squaring the circle. In Proc. 7th CAA Conference, 289-300

Piech, C., Huang, J., Chen, Z., Do, C., Ng, A. & Koller, D. (2013). Tuned Models of Peer Assessment in MOOCs. In Proc. 6th Int. Conf. on Educational Data Mining, Memphis, USA

Põldoja, H., Väljataga, T., Laanpere, M. & Tammets, K. (2014). Web-based self- and peer-assessment of teachers' digital competencies. *World Wide Web,* 17, 255–269

Romero, C. & Ventura, S. (2010). Educational Data Mining: A Review of the State of the Art. *IEEE Trans. on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, 40(6), 601-618

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

Somervell, H. (1993). Issues in assessment, enterprise and higher education: the case for self-, peer and collaborative assessment. *Assessment and Evaluation in Higher Ed.*, 18, 221-233

Sterbini, A. & Temperini, M. (2009). Collaborative Projects and Self Evaluation within a Social Reputation-Based Exercise-Sharing System. In Proc. IEEE/WIC/ACM WI-IAT'09, Vol. 3, Workshop SPEL, 243-246.

Sterbini, A. & Temperini, M. (2012a). Dealing with open-answer questions in a peer-assessment environment. Proc. ICWL 2012. LNCS, vol. 7558, 240–248. Springer, Heidelberg.

Sterbini, A. & Temperini, M. (2012b) Supporting Assessment of Open Answers in a Didactic Setting. In Proc. IEEE 12th Int. Conf. on Advanced Learning Technologies (ICALT, Workshop SPeL), pp.678--679.

Sterbini, A. & Temperini, M. (2013). OpenAnswer, a framework to support teacher's management of open answers through peer assessment. In Proc. 43rd Frontiers in Education (IEEE FIE), pp. 164--170.

Yamanishi, K., & Li, H. (2002). Mining Open Answers in Questionnaire Data, *IEEE Int. Systems*, Sept-Oct, 58-63.