Performance variations of the Bayesian model of peer-assessment implemented in OpenAnswer

Response to modifications of the number of peers assessed and of the quality of the class

Maria De Marsico, Luca Moschella, Andrea Sterbini
Department of Computer Science
Sapienza University of Rome
Rome - Italy
{demarsico, sterbini}@di.uniroma1.it
moschella.1594551@studenti.uniroma1.it

Marco Temperini
Department of Computer, Control, and Management Engineering Antonio Ruberti
Sapienza University of Rome
Rome - Italy
marte@dis.uniroma1.it

Abstract— The paper presents a study of the performance variations of the Bayesian model of peer-assessment implemented in OpenAnswer, in terms of the grades prediction accuracy. OpenAnswer (OA) models a peer assessment session as a Bayesian network. For each student, a sub-network contains variables describing relevant aspects of both the individual cognitive state and the state of the current assessment session. Sub-networks are interconnected to each other to obtain the final one. Evidence propagated through the global network is represented by all the grades given by students to their peers, together with a subset of the teacher's corrections. Among the possible affecting factors, the paper reports about the investigation of the dependence of grades prediction performance on the quality of the class, i.e., the average level of proficiency of its students, and on the number of peers assessed by each student. The results show that both factors affect the accuracy of the inferred marks produced by the Bayesian network, when compared with the available ground-truth produced by teachers.

Keywords—peer assessment; Bayesian inference; student model; dependencies among cognitive abilities

I. INTRODUCTION

It is a common intuitive understanding that one realizes to have achieved good mastery of a topic/skill when able to explain it and to correct own peers. As a matter of fact, Bloom's taxonomy of educational objectives in the cognitive domain [3] formalizes such understanding. Comprehension, application, analysis, evaluation, and, finally synthesis of learning items, are deemed to require wider and firmer mastery than pure knowledge, intended as just remembering details about a topic and being able to report them. Even the revised version of the taxonomy in [1], that was conceived to further qualify knowledge abilities in relation to specific subjects, puts *remember*, *understand* and *apply* at increasing levels, while *analyze*, *evaluate* and *create* lay at the same top level. They are universally considered as higher metacognitive skills, that go well beyond the basic proficiency in a topic. Metacognitive activities require knowing about knowing [15], since they refer to higher order thinking processes. The ability to exercise an active control over the cognitive processes underlying learning, entails the ability to plan strategies and conceive working schedules, and to monitor comprehension and progress towards completion. In addition, the awareness of how to apply new concepts and rules, and the ability to evaluate activity outcomes even in relation to peers, is undoubtedly superior to passive acquisition of notions. For these reasons, peer-assessment is widely considered an important exercise not only to test, challenge and improve students' understanding of a topic, but also to achieve, exercise, and enforce higher metacognitive abilities. Students can get further advantage from teacher's assessment of the peer assessment. In fact, they can learn from smarter peers how to improve one's results [18] and grasp the rationale of how the achieved results underlie the grading process, by matching the grades they assigned with teacher's/peers' evaluation. Of course, as in teacher's assessment, the significance of the results also depends on the articulation of the exercising pattern. The more possibilities are left to the students to express their thought and knowledge in a not fortuitous ways (and possibly to make mistakes), the higher the reliability of the assessment. Open answers to specific questions allow proposing a variety of challenges to students, including exercises, short essays, free text answers to questions, etc. Therefore, they are much more informative than multiple-choice questionnaires, and better allow to evaluate both knowledge and evaluation abilities of students [17]. The other side of the coin is that they entail much more teacher's work, especially with a large group of students.

The OpenAnswer framework (OA) [20][21][22] aims at killing two birds with one stone. It allows (semi-)automated grading of open answers through peer assessment, with less grading burden as possible to the teacher. The starting point is a model of both the knowledge level and the judgement ability of single students and of the overall class.

OA assessment is organized into sessions. A session set-up entails to assign to each student a number (e.g., 3) of peers' answers to grade, possibly following some assignment criterion. The system provides the possibility to either carry out "pure" peer-assessment, or to involve the teacher in grading a subset of answers. This subset is dynamically identified step by step according to some relevant select-next-or-stop strategy. Both select and stop are each driven by specific rules.

OA framework is based on the theory of evidence propagation in Bayesian networks. Each student is represented through his/her cognitive/metacognitive level, which is modeled by discrete variables making up a fragment of a global Bayesian Network (BN). These discrete variables provide an estimate of the student's knowledge level on the topic (K), and of the student's ability to evaluate (J). In the same way, the answers of the student have an estimated correctness value (C). All variables, and in particular the latter

978-1-5386-3968-9/17/$31.00 ©2017 IEEE

one, are updated by evidence propagation. The global network and its topology are built by interconnecting the described single sub-networks. The interconnections are determined by the answers graded by each student, while the assessments provided by peers, and possibly by the teacher, are fed and propagated within the BN as evidence values. The final values of C variables represent the estimated correctness of the answers. Providing each student also with his/her final K and J values also spurs metacognitive awareness.

The model is quite rich, so that many investigations are possible regarding its parameters and rules. The final goal is to identify the best setting to achieve a sufficient accuracy with possibly a low engagement for the teacher in the pure grading process. Earlier works [8][20][21][22] analyzed several factors affecting the network. The research questions addressed here are:

RQ1) is there a "best" value for the number of peers each student should assess? In other words, is it possible to find out an “optimal” compromise between the number of peers assessed by each student, and the final grades accuracy? On one side one expects to obtain better predictions from more assessments, on the other side students hardly tolerate to assess too many peers.

RQ2) how much the prediction ability of the OA model changes when the class is composed by students with medium-high knowledge level, with respect to the case where their knowledge is medium-low? In this case it is reasonable to expect the model to show better results in case of a class of smarter students than in the case of weaker ones.

As the two expectations above could seem somewhat obvious, we could say that they can be considered a strong (meta) requirement for any good model of peer assessment.

In order to investigate the above ideas (and check our meta-requirement), we run experiments that simulate the teacher correcting a fixed 30% of the students (in general, this percentage is rather variable and driven by a specific stopping criterion), and results compare the grades predicted for the remaining students with the true teacher's grades (available as ground truth).

In order to compare networks with a varying number N of peers assessed, we have obtained many similar peer-assessment networks by using a dataset where 60 students assessed 20 peers each, and by selecting a subset of given size N of peers to be corrected by each student.

The same dataset can be used to create classes with the preferred distribution of students with medium-high/medium-low knowledge (as estimated from the ground truth), by selecting the subgraph induced by the desired groups of students.

The simulations with a varying number of N peers per student (RQ1) show that the best results are obtained for the maximum number of peers (4) we can assign with this dataset. Regarding the class composition, the simulation results show that OA predicts better from assessments made by medium-high students, with respect to the case of medium-low ones (RQ2). This result further confirms the strict relation between cognitive and metacognitive abilities. The paper continues with a summary of related work, presented in Section II. Section III presents the necessary details of OA framework, to comprehend the assessment setting and the possible research lines stemming from the devised architecture. Section IV presents the experimental setup and discusses the results. Section V draws conclusions and sketches future research.

II. RELATED WORK

The automatic extraction of relevant information from large amounts of (free-text) data typically relies on techniques of data mining and natural language processing. One of the contexts where these techniques are used is the design of marketing applications, when there is need to infer customer opinions and synthesize products reputation [26]. Concept mapping and coding schemes are an alternative used in [13]. As discussed in the introduction, the automatic analysis of open answers is among the most powerful learning assessment tools (knowledge tracing) [2], and relies on similar processing [11]. (Semi-)automatic assessment of open-answers in [4] exploits ontologies and semantic web technologies.

Assessment of peers’ work is a higher cognitive level activity [3]. Its educational use pursues different goals [24]: to allow the learner to appreciate the personal cognitive state and progress, and to evaluate and possibly amend gaps in learning achievements. According to [14], there is a relationship between the quality of the peers feedback, on a learner's job, and the quality of the final project submitted by the learner. The work in [6] presents a comprehensive study of peer assessment in a prototype educational setting. Of course, the real educational value of peer assessment has been the object of many other investigations. Without going too much backwards in time, an interesting study published in 1994 [19] presents the evaluation, in terms of correlations between sets of marks, of a collaborative peer, self and tutor assessment scheme. Among the issues addressed, there is the reliability of student-derived marks, and the learning benefits achieved by students participating in peer and self-assessment procedures. In particular, the study investigates the perceived tendencies of high achieving students to underestimate their performance, and of low achieving students to overestimate their ones. The reported results indicate that students have a realistic perception of their own abilities and can make rational judgements on the achievements of their peers. In addition, the positive implications of introducing peer and self-assessment schemes into undergraduate courses are discussed. The difference in evaluations carried out by high achieving students and low achieving students is especially related with RQ2 addressed here, and is also underlined by results in [23].

OA system is designed such that the evaluation of open answers through peer-assessment is carried out by modeling students by individual Bayesian Networks, which are interconnected through assessment. The work in [7] proposes a different machine learning approach to student modeling. In this case, Bayesian Network techniques are used to support learner's modeling in an Intelligent Tutoring System (ITS). Before the OA assessment session starts, the students are presented with a set of teacher-defined assessing criteria, to refer to while marking. This is an important step, since works in literature identify the specificity of "scoring criteria" as an important factor to avoid a range of peer marks possibly limited to the high end of the scale [16]. The results of the meta-analysis in [12] show that peer assessments were found to resemble more closely teacher assessments when using global judgements based on well understood criteria rather than when marking involves assessing several individual dimensions. An interesting aspect of research in peer-assessment regards the number of peer-evaluations that a same job should undergo during the process. In OA this is

configurable (default is 3) and is related with the RQ1 for this paper. Literature reports that more feedback on the same job makes the peer perform more thorough revisions, ending up with a better result [5]. In particular, the experiments in [23] address this aspect in relation to teenagers, and also investigate possible differences between the evaluation of individual versus group work. Reported results show that the acceptable accuracy is achieved with 3 or 4 raters for the assessments of individual performance, but 14–17 raters are needed for assessments of group work. Furthermore, the discrepancy between the ratings of students and experts was higher in group-work assessments then in individual-work assessments. This opens interesting research lines related to the evaluation of group-work.

III. OPENANSWER BAYESIAN APPROACH

The OA system models peer-assessment as a Bayesian Network (BN). This global network is composed by interconnecting individual sub-networks representing single students. Each such subnetwork is made of three discrete nodes/variables, making up the student model, and representing respectively:

- K: student's knowledge about the topic
- C: the correctness of student's answer
- J: student's ability to judge/assess a peer's answer
- plus one variable G for each grade given to a peer (G variables represent the interconnections among individual subnetworks).

Each Bayesian variable above is in general defined over a 6-valued discrete domain, ranging from A (best) to F (fail). In the case of the dataset used for these experiments [25] the discrete domain ranges over 5 values (ABCDF) because the initial grades were expressed on a 5-valued scale.

A Grade value (for the corresponding G variable) is injected into the network when a student marks a peer's answer, and propagates its effects depending on both the current value of J of the grading student, and on the current estimation of C of the answer corrected.

Variables C and J are assumed to be related to K, in particular they are represented by a conditional distribution of values conditioned by K (C | K and J | K). The motivation for the C dependency is because writing an essay cannot be carried out by a random guess (as it would be possible in the case of multiple-choice quizzes). As for J, the inspiring principle stems from Bloom's taxonomy of cognitive levels [3], where judging a peer's answer is considered as a more difficult task than knowing the topic and answering it. The distribution of values for G is conditioned by J and C (G | J, C). For each conditioned distribution it is necessary to model a Conditional Probability Table (CPT), describing the corresponding hypothesized probabilistic dependence.

An interesting issue is the initial probability distribution for K. A first alternative is to assume it flat (constant probability=1/6), i.e., no preliminary knowledge about the class. A second one is TgradeDist, i.e., the same probability distribution derived from the teacher grades of that assessment. Of course, in a real situation such information is only available ex-post, but it is useful in experiments to show what would happen if the system had some initial global information on the K distribution of the class, though no information on single students.

When teacher correction activity is included in the process, the systems supports the correction by suggesting the next answer to grade, according to some pre-selected rule, and by notifying the teacher when no further correction is needed, again according to a pre-defined stop condition. As for the stop condition, possible investigated alternatives can be found in [8]. However, in order to focus the comparison on other elements, in this work we stop the correction at 30% of answers. As for the choice of the next answer to grade, this is done trying to maximize the information gain achieved by its actual correction. Possible criteria are discussed in [8] and [9]. Among those, only the one achieving the best average results in past experiments is used here for comparison with the best topological criterion in [9]. Namely, the first rule is maxEntropy: the next answer to grade is the one presenting the highest entropy, in practice the one the system knows less about. The topological criterion is minAvgDistLInferred: it selects the next student in order to minimize the average distance among the inferred students and their nearest corrected peer (i.e. to reduce the average distance the information should traverse).

IV. EXPERIMENTAL RESULTS

A. Experiments setup

The benchmark used for experiments is a dataset of 60 students, each of which graded 20 peers [25]. This allows both building networks with different interconnection (answers graded by each peer) and to significantly divide the class according to student homogeneous knowledge level.

1) Number of corrections per student

To study the most suitable number of answers to grade per student (number of peers) the experiments were set up as follows:

- the dataset is preliminarily divided into groups, from 1 to 5 (i.e. from 60 to 12 students per group).; at the end of the computations the performance measures are averaged over such groups; this was mainly required because of the computational limits tied to the Bayesian network computation, which made very difficult to simulate bigger groups of students with more than 4 peers each;

- the number of peer assessments assigned to each student for a test session ranges from 1 to 4.

To ensure that in the generated essessments each answer received an almost uniform number of assessments from peers, a circular rule was followed: the peers to assess assigned to each student were picked from the available ones with ID greater than the assessing student (modulo the number of students in the group), so to obtain a ring-shaped network structure. This kind of network has a slightly higher diameter than a completely random one, and this is the reason why, regarding the selection of the next answer to correct by the teacher, we tested both the best information-based selection strategy (maxEntropy), and the above mentioned topology-based strategy (minAvgDistLInferred), which tries to keep minimal the average distance between the still uncorrected students and their nearest corrected peer in the network. Elsewhere [9] we have shown that a topology-oriented strategy performs better than our earlier best strategy (maxEntropy) when the network has high diameter.

To compare results among different strategies we use a fixed termination condition where we stop as soon as 30% of answers have been corrected by the teacher.

Respect to the initial probability distribution for K, we tested both flat and TgradeDist.

2) Quality of the class

The ranges that were separately considered for class knowledge level are the following: ALL (students achieving any ground-truth mark in ABCDF), MIDHI (marks in ABC), and MIDLO (marks in CDF)

For each range, the following set up was executed:

- as above, the dataset was divided into N groups (1 with 60 students, 2 with 30 students,…, 5 with 12 students) over which averaging of performance results was carried out;

- for each test session, only students with a mark in the given range were included in the simulation;

- for each student, the first 3 marks that that student had provided during the session were used in the simulation.

In this case, also, in order to obtain a symmetric distribution of marks, peers were rotated in the dataset in order to get a ring-shaped network structure. Because the ring-shaped network has a slightly higher diameter than one built by choosing peers completely at random, even in this group of simulations we have applied also the above mentioned topology-based selection strategy (minAvgDistLInferred).

To evaluate the quality of OA predictions we show the percentage of inferred grades predicted exactly (OK/INFERRED) and predicted within 1 grade from the correct one (IN1/INFERRED). As the available dataset contains the corrections made by 4 different teachers, we simulate the correction separately for each teacher. The average between all four teachers is summarized in the *Overall* column. It is worth underlining that the students made only one correction for each essay, while there are 4 different corrections from the teachers. Of course, only one introduction to grading criteria was presented to the class, which could have been more or less consistent with each of the teachers' grading styles. It is reasonable therefore to expect a different level of agreement between the students and the different teachers, and therefore different OA performances depending on the teachers (this issue is also mentioned in the related work section).

To get an idea of which teacher was more/less in agreement with the students' assessments, below we show the average absolute difference between student grades and teacher's (normalized) grades, (see TABLE 1). We notice that teacher with ID=107 shows the best agreement with the student's grades (1.3 grades on a 10 point scale), while the others show a higher disagreement (1.7 grades on a 10 point scale).

TABLE 1 AVERAGE PEER GRADE DIFFERENCE RESPECT TO TEACHERS' GRADES

TEACHER ID	5	107	887	1033
Average of DeltaPeerGrade	0.16	0.13	0.17	0.17

B. Experimental results

1) Number of peers corrected per student

Respect to the number of peers, in TABLE 2 we show the OA performances respect to an increasing number of peers corrected by each students, ranging from 1 to 4. In this case it is possible to initialize P(K) either as a *flat* distribution or as teacher grades distribution (*TgradeDist*).

The algorithm sketched above is used to select the set of peers to be corrected by each student, which produces a ring-shaped network. We test both the information-based maxEntropy strategy and the topologic-based strategy minAvgDistLInferred.

The table is colored to highlight the best results (in green) and the worst ones (in red).

TABLE 2 QUALITY OF PREDICTIONS DEPENDING ON AVERAGE NUMBER OF PEERS

	P(K) init. STRATEGY	flat	TgradeDist	flat	TgradeDist
		maxEntropy		minAvgDistLInferred	
Data	Num. Peers				
Average of OK/INFERRED	1	38%	40%	30%	39%
	2	41%	46%	41%	43%
	3	45%	48%	42%	43%
	4	45%	50%	44%	47%
Average of IN1/INFERRED	1	88%	91%	81%	89%
	2	90%	90%	88%	91%
	3	89%	90%	87%	88%
	4	92%	92%	91%	91%

From the table we notice that the best results are obtained for higher numbers of peers (4). This confirms our initial intuition that when we ask the students to correct more peers, OA gets more information and can predict better.

Moreover, as expected, when something is known about the overall abilities of the class (P(K)=TgradeDist) OA performs better, respect to the case where nothing is known (P(K)=flat).

Finally, probably because the averaged sub-groups are small (5 groups of 12 students) and thus the network diameters are low, the topology-oriented selection strategy (minDistLInferred) shows worse results than the information-based one (maxEntropy).

This provides a positive answer to RQ1: the number of peers assessed by each student influences the accuracy of the final result, the higher the better.

2) Quality of the class

In TABLE 3 we show the percentage of grades inferred by OA exactly or within one grade from the ground truth, both when the class is of mid-hi quality (RANGE=midhi, grades ABC) or when it is of mid-lo quality (RANGE=midlo, grades CDF). We further compare the results with the case where no selection has been made on the class (RANGE=all).

In this case we use only the P(K) initialization of type "flat" (i.e. no info on the class) because this information already underlies the session division of the class by selecting a subset

of students with a given set of grades (ABC or CDF). In this case TgradeDist, which would constrain the K values only on the corresponding 3 values, would give to the OA model too much information.

The table is colored to highlight the best results (in green) and the worst ones (in red).

TABLE 3 QUALITY OF PREDICTIONS DEPENDING ON QUALITY OF CLASS AND TEACHER

			TEACHER ID				
Data	RANGE	STRATEGY	5	107	887	1033	Overall
Average of OK/INFERRED	midhi	maxEntropy	54%	49%	52%	35%	47%
		minAvgDistLInferred	55%	51%	52%	45%	51%
	midlo	maxEntropy	26%	57%	8%	14%	26%
		minAvgDistLInferred	18%	34%	11%	28%	23%
	all	maxEntropy	45%	57%	43%	26%	43%
		minAvgDistLInferred	50%	50%	36%	31%	42%
Average of IN1/INFERRED	midhi	maxEntropy	95%	97%	94%	92%	95%
		minAvgDistLInferred	87%	100%	97%	98%	95%
	midlo	maxEntropy	82%	85%	53%	61%	70%
		minAvgDistLInferred	76%	90%	64%	80%	77%
	all	maxEntropy	90%	98%	90%	83%	90%
		minAvgDistLInferred	81%	100%	81%	83%	86%

From TABLE 3, column 'Overall', we see that with a midhi class OA predicts grades way better than in the case of a midlo class. This testifies that the adopted model satisfies the (meta) requirement that any peer assessment should predict better when students knows better the topic. At the same time we expect a generic class (line 'all') to perform in an intermediate way respect to 'midhi' and 'midlo', which the table shows.

From the table we notice that the performances of OA when the class is all made of midlo students decreases, since the propagation of less reliable information (poor grading) makes very hard for the Bayesian network to predict exactly (OK/INFERRED) or within 1 grade from the ground truth (IN1/INFERRED). This is clearly evident from the Overall column, which shows the average results over the 4 teachers. With this we answer RQ2, showing that a better class injects better information into the model, obtaining more precise predictions. This is somewhat expected, given the definition of the P(G|J,C) CPT, which on the C, D, F grades produces very shallow Gaussians (lower ability to deduce something from the grades given to peers).

Respect to the teachers, the best results are obtained on teacher 107, which is the one most agreeing with the student's grades. For the other teachers the OA performances are in general reasonably good, except for teacher 1033 where OA predicts poorly on exact grades and reasonably good within 1 grade. Elsewhere [10] we have shown that the Conditional Probability Tables (CPTs) describing the probabilistic dependences in the Bayesian network could be specifically tailored to the teacher-class pair. This can be done when a sufficient set of peer assessment session results are available. In this case we have used a unique "one size fits all" set of CPTs for all teachers, thus it's not unexpected to see worse performances respect to one of the teachers.

V. CONCLUSIONS

The presented experimental results show that OA performs better when each student assesses more peers, and on classes with better students.. More formally:

- they answer to RQ1 by showing that a higher number of peers (4 in this case) produces better predictions;
- they answer to RQ2 by confirming that a OA predicts better in the case of a better class.

While these outcomes may seem obvious, actually they confirm that the devised OA model behaves correctly with respect to the meta-requirements.

Because of computational constraints and of the dataset available, we were able to test only generated peer assessments with a maximum number of 4 peers. In a near future, it will be interesting to find a way to lift, at least partially, these computational limits and simulate cases with 5 or more peers per student.

References

[1] L.W. Anderson, D.R. Krathwohl (eds.), "A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives, Complete Edition," Pearson, 2000.

[2] J.R. Anderson, A.T Corbett, K.R Koedinger, R. Pelletier, Cognitive Tutors: Lessons Learned. The Journal of the Learning Sciences, 4(2), 167-207, 1995

[3] B.S. Bloom, M.D. Engelhart, E.J. Furst,. W.H. Hill, D.R. Krathwohl, "Taxonomy of educational objectives: The classification of educational goals. Handbook I," McKay, 1956.

[4] D. Castellanos-Nieves, J. Fernández-Breis, R. Valencia-García, R. Martínez-Béjar, M. Iniesta-Moreno, "Semantic Web Technologies for supporting learning assessment, Inf. Sciences, 181:9, 2011.

[5] K. Cho, C. MacArthur, "Student Revision with Peer and Expert Reviewing," Learning and Instruction 20(4), 328-338, 2010.

[6] H. Chung, S. Graf, K. Robert Lai, Kinshuk, "Enrichment of Peer Assessment with Agent Negotiation". IEEE Ttransactions on Learning Technologies, 4(1), pp.35-46, 2011.

[7] C. Conati, A. Gartner, K. Vanlehn, "Using Bayesian Networks to Manage Uncertainty in Student Modeling". User Modeling and User-Adapted Interaction, *12*(4), pages 371-417, 2002.

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

[9] M. De Marsico, A. Sterbini, M. Temperini, "Effects of network topology on the OpenAnswer's Bayesian model of peer assessment" Proc. Twelfth European Conference On Technology Enhanced Learning, (EC-TEL 2017), Tallinn 12-15 Sept. 2017, (submitted).

[10] M. De Marsico, A. Sterbini, M. Temperini, "Leveraging CPTs in a Bayesian approach to grade open ended answers", Proc. ICALT 2017, 3-Timisoara, Romania, 7 July 2017, 2017.

[11] N. El-Kechaï, É. Delozanne, D. Prévit, B. Grugeon, F. Chenevotot, "Evaluating the Performance of a Diagnosis System in School Algebra," International Conference on Web-Based Learning, LNCS 7048, (pp. 263-272). Springer Berlin Heidelberg. 2011.

[12] N Falchikov, J. Goldfinch, "Student peer assessment in higher education: A meta-analysis comparing peer and teacher marks," Review of educational research, 70(3), 287-322, 2000.

[13] K. Jackson, W. Trochim, "Concept mapping as an alternative approach for the analysis of open-ended survey responses," Organizational Research Methods, *5*(4), 307-336, Sage, 2002.

[14] L.X. Li, X. Liu, A.L. Steckelberg, "Assessor or Assessee: How Student Learning Improves by Giving and Receiving Peer Feedback." Br. J. of Ed. Tech. 41 (3), pages 525–536, 2010.

[15] J. Metcalfe, A.P. Shimamura, "Metacognition: knowing about knowing," Cambridge, MA: MIT Press, 1994.

[16] P. Miller, "The Effect of Scoring Criteria Specificity on Peer and Self-assessment," Assessment & Evaluation in Higher Education, 28(4), 383-394, 2003.

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

[18] P.M. Sadler, E. Good, "The Impact of Self- and Peer-Grading on Student Learning" Educational assessment, 11(1), 1-31, 2006.

[19] L. A. Stefani, "Peer, self and tutor assessment: Relative reliabilities," Studies in Higher Education, 19(1), 69-75, 1994.

[20] A. Sterbini, M. Temperini, "Dealing with open-answer questions in a peer-assessment environment," Proc. ICWL 2012. LNCS, vol. 7558, pp. 240–248. Springer, Heidelberg, 2012.

[21] A. Sterbini, M. Temperini, "OpenAnswer, a framework to support teacher's management of open answers through peer assessment," Proc. 43th Frontiers in Education (FIE 2013), 2013.

[22] A. Sterbini, M. Temperini, "Analysis of OpenAnswers via mediated peer-assessment," Proc. 17th IEEE Int Conf. on System Theory, Control and Compu-ting (ICSTCC 2013), 2013.

[23] Y. T. Sung, K. E. Chang, T. H. Chang, W. Yu,. "How many heads are better than one? The reliability and validity of teenagers' self-and peer assessments," Journal of Adolescence, 33(1), 135-145, 2010.

[24] K. Topping, "Peer assessment between students in colleges and universities," Rev. of Ed. Research, 68, pp. 249–276, 1998.

[25] A. Vozniuk, A. Holzer, D. Gillet, "Peer assessment based on ratings in a social media course," Proceedings of the Fourth International Conference on Learning Analytics And Knowledge (LAK '14). ACM, New York, NY, USA, 133-137, 2014.

[26] K. Yamanishi, H. Li, "Mining Open Answers in Questionnaire Data," IEEE Int. Systems, Sept-Oct, pp 58-63, 2002.