Towards a quantitative evaluation of the relationship between the domain knowledge and the ability to assess peer work

Maria De Marsico, Andrea Sterbini
Department of Computer Science
Sapienza University of Rome
Rome, Italy
{demarsico, sterbini}@di.uniroma1.it

Marco Temperini
Department of Computer, Control, and Management Eng.
Sapienza University of Rome
Rome, Italy
marte@dis.uniroma1.it

Abstract— **In this work we present the preliminary results provided by the statistical modeling of the cognitive relationship between the knowledge about a topic a the ability to assess peer achievements on the same topic. Our starting point is Bloom's taxonomy of educational objectives in the cognitive domain, and our outcomes confirm the hypothesized ranking. A further consideration that can be derived is that meta-cognitive abilities (e.g., assessment) require deeper domain knowledge.**

Keywords—Bloom's taxonomy; ranking of educational objectives; meta-cognitive abilities; Bayesian network;, peer-assessment

I. INTRODUCTION

In educational sciences, the distinction between *cognitive* and *metacognitive* activities sets the demarcation line between *knowing* and *knowing about knowing* [1]. The accepted definition of metacognition refers to higher order thinking which involves active control over the cognitive processes that are engaged in learning. This definition encompasses activities such as planning how to approach a given learning task, monitoring comprehension and ability to apply it, and evaluating progress toward the completion of a task. It is quite clear that metacognition plays a critical role in successful learning. As a consequence, it is important to investigate metacognitive activities and development to determine how to teach students to apply their cognitive resources in the best possible way through metacognitive control. This supports students with higher metacognitive abilities to also achieve better cognitive results. It is important to stress that, in their strict definition, is little or no distinction between domain-general and domain-specific metacognitive skills. The latter are domain-general in nature, and there are no specific meta-cognitive skills for certain subject areas. In other words, the metacognitive skills that are used to review an essay are the same as those that are used to verify an answer to a math question [2]. Notice that here we are setting a subtle yet important conceptual distinction between knowing a domain concept, and being able to assess one's knowledge, which is a metacognitive skill, even if it somehow entails domain knowledge as one of its premises. In particular, the ability to assess, and particularly to self-assess, is deemed to be fundamental for better learning. However, the ability to peer-asses is often classified as metacognitive. As a matter of fact, through self- and peer-assessment students should learn to see mistakes in their thinking and be able to correct any problems in future assignments. By grading peer work, students can learn to understand how the grading process should work. Moreover, by grading assignments, students may learn from smarter peers how to complete assignments more accurately and how to improve their test results [3].

Perhaps the first and best-known taxonomy of educational objectives in the cognitive domain is due to Bloom and his group [4]. According to them, there is an increase in learner's abilities going from pure knowledge (ability to remember, the lower level), to comprehension, application, analysis, evaluation and finally synthesis. The revised version by Anderson [5] includes remember, understand and apply at increasing levels, and finally analyze, evaluate and create at the same top level. In practice, in this work, even if we rely in principle on Bloom's levels, we conceptually compact them, and consider Knowledge of a topic related to a learning domain (K), and ability to judge peers' work (J), which are put in relation through the correctness of the given answers as assessed by the teacher (C), and the grades assigned to peers (G). In this work we present preliminary results provided by statistical modeling of the cognitive relationship between knowledge about a topic and the ability to assess peer achievements on the same topic. We are interested in determining the relationship between these two macro-abilities, which we summarize in knowledge and judgment. We compare Judgment and Knowledge by modeling peer assessment and optimizing the model parameters, which depend on their relative distance, to get the best match with the observed grades. The optimal parameters show that Judgment appears to be in general more difficult than Knowledge. To the best of our knowledge this is the first quantitative evidence in literature of the validity of Bloom's ranking of cognitive abilities.

II. PEER-ASSESSMENT

Pedagogical research along the line of socio-constructivism explores two main lines for improving students' achievements.

978-1-4799-1756-3/15/$31.00 © 2015 IEEE

First, personalization of learning processes is crucial in both classroom-based and distance learning. Second, social-collaborative e-learning is a winning strategy, but it requires special design too. As an example, in [6], the pitfalls for social interaction when collaborative learning is framed in a computer-supported environments, are identified in “taking for granted that participants will socially interact simply because the environment makes it possible and neglecting the social (psychological) dimension of the desired social interaction.” In [7], the authors discuss how learning from peers is facilitated by motivating students through reputation systems. In [8] 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. Peer-assessment can be considered as one of the activities that qualify collaboration among students. It is a process whereby single students as well as groups grade their own assignments and/or peers’ ones. This exercise may or may not entail previous discussion or agreement over criteria. The practice can also be employed to save teachers’ time. However, one concern is that students may give better grades than teachers, and this would make assessment unreliable. To this respect, the cited study by Sadler and Good [3] has shown that if students can understand the teacher's quality requirements, then there is a high level of agreement between grades assigned by teachers and students. This also means that the students are using a kind of implicit knowledge, of what must be evaluated and how, together with assignment-related abilities. Furthermore, if teachers look at how students grade themselves and their peers, they can derive more information for accurate grading. However, grade inflation may occur and other factors (e.g., friendship, reputation) may also affect peer grading especially if it is not blind. The alternative is to use a semi-unsupervised system that is able to infer the correctness of assignments not only from peer grades, but also from a sample of grades that the teacher generally provides and from which the remaining ones are inferred. Actually, peer-assessment has a significant potential to improve both students' understanding of assignment topics and their metacognitive skills. The integration of assessment and instruction, makes the student an active stakeholder who shares responsibility, reflects on own and peers’ achievements, collaborates in some sense with both peers and with the teacher. In particular, Somervell [9] underlines that peer assessment is not only a pure grading procedure, but it is rather a crucial part of a learning process through which skills are developed, and a part of the self-assessment process itself. It is possible to distinguish different forms of peer assessment [10]:

- 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.

A quite dated yet still valid review about self- and peer-assessment can be found in [11]. This paper provides a clear definition of self-, peer, and co-assessment. 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. As already mentioned, peer assessment is the process through which groups of individuals rate their peers. The studies on peer assessment reported in the survey focus on the aspects of validity, fairness, accuracy and effects. Self- and peer assessment are combined when students are assessing peers but they are also 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 assessments. Along such distinction, in some works peer and self-assessment marks are compared with teacher marks in order to assess the accuracy of peer or self-assessment. Other studies aim at evaluating the inaccuracy implied by completely or partially relying on this kind of assessment for students’ performance evaluation. It is also interesting to mention the results of the meta-analysis presented in [12] and taking into account 48 quantitative studies about peer-assessment aiming at 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 which have poor experimental design.

III. THE PROPOSED FRAMEWORK

The OpenAnswer system was designed to support semi-automatic grading of answers to open-ended questions (*open answers*). Evaluating open answers is widely considered as a more powerful tool to assess knowledge than closed answer tests (quizzes) [13]. However, it is much more demanding, for both the student and the teacher. OpenAnswer exploits peer assessment. From one side, it allows to measure the student’s ability to correctly evaluate the answers of their peers. From the other side, it may be the base for possible strategies to limit the amount of teacher’s grading. Many proposed systems especially focus on the first goal, while completely automatic grading techniques based on peer grading are still not reliable enough to be extensively used in a real educational context. The strategy adopted by OpenAnswer is a mixed one, where the teacher is required to assess some part of answers too. In this way, we aim at increasing the reliability of the final grades, while reducing at the same time the grading workload. The system suggests to the teacher the order of answers to check, according to a selection strategy chosen in advance. Manual correction can stop when some pre-defined condition (termination criterion) is met. As for the remaining answers, the system automatically completes the grading task using the knowledge collected so far and the results of the peer-assessment. In [14 - 17] we presented the OpenAnswer approach relying on a simple model of Bayesian networks. For each student, the system maintains an individual model that is built as a single Bayesian network. The variables of the model are an up-to-date evaluation of the learner's state of knowledge on the question topic (*K*) and of her ability to judge (*J*) answers given by peers on the same topic. These variables make up a kind of student profile useful for the dynamics of the system. During a peer assessment session, the individual networks of students are interconnected. For each question, we consider the correctness (*C*) of the answer given by the learner, and the grades (*G*) she assigned to answers from peers. *C* may come from the teacher as well. The latter two variables also control evidence propagation according to the measured correspondences between teacher's and peer assessments. The resulting compound network is continuously updated, and, as the teacher stops grading after a sufficient number of answers, allows to automatically evaluate those answers that were not directly graded by the teacher.

The Bayesian network underlying OpenAnswer model consists of the following set of finite-domain variables, all taking values from A to F:

- *K*: *Knowledge*; independent variable;
- *J*: *Judgment*; the ability to judge is considered at a higher cognitive-level ([4]); this suggests a dependency between *K* and *J*, represented by the *P(J|K)* conditional probability;
- *C (G)*: *Correctness*; we assume a dependence between *K and C*, which we represent in the network by *P(C|K).*

During each session, each student has to assess some (e.g. 3) peer answers, by assigning them a grade, or choosing the best one or choosing the worst one, and the variables modeling her are connected to those of the corresponding peers. In present experiments, we use the grade assignment.

For each OpenAnswer simulation a template Bayesian network is instantiated. It consists of *K, J, C* variables for each student; depending on the chosen peer-assessment settings, from one to three Bayesian variables (connecting the student's J and her peers' *C1, C2, C3*) are added. If *method = grade*, as in this paper, a *GradeX* variable for each assessed peer *X* is required. In particular, we adopt the range of grades [A – F] for both *C* and *G*, and K and J, with a simplification of the conventional mapping towards decimal marks (A reaches 10, B for 9.4-8.5, C for 8.4-7.5, D for 7.4-6.5, E for 6.4-5.5, F corresponds to all marks from 5.4 downwards). The values of the peer-assessment variables corresponding to each student are set to her specific assessment choice. Once the network is created with the initial evidence, the initial probabilities are computed by the Junction Tree belief propagation algorithm.

As for the teacher, the possible selection strategies to suggest the next answer to grade are:

- *max_wrong*: the next answer to grade is the most probably incorrect one;
- max*_entropy*: the next answer to grade is the one presenting the highest entropy (the one the system *knows* less about);
- *random*: mostly used for testing purposes, and not used in the present work; in practice, the next answer to grade is chosen at random.

As for the termination criterion, we have again a list of choices:

- *no_wrong*: no more answers exist which are automatically graded as wrong;
- *no_flip(N)*: the automatically computed grades remained stable in the last N correction steps;

In general, *max_wrong* is best associated with *no_wrong* and *max_entropy* with *no_flip*, while *random* can be associated with both the termination criteria.

IV. OPENANSWER SYSTEM TO VALIDATE BLOOM'S TAXONOMY

We are interested in determining the relationship between the two macro-abilities summarized from Bloom's taxonomy, namely knowledge (*K*), i.e., competence in a certain topic, and judgment (*J*), i.e. ability to correctly assess a peer about an assignment related to that topic. More in detail, we aim at modeling the conditional relations between: 1) the ability to judge the answers by a peer on a certain topic (*J*), and the correctness of one's answer to an assignment on the same topic as evaluated by the teacher (*C*), and 2) between the correctness of scores assigned to peers' answer to the same assignment (*G*), and the ability to judge *J* joined with *C*. In modeling such relations, we also consider *K*, i.e., the student's knowledge of the topic, intended in its wider sense of skills, abilities, etc., which is stored in the student's model. We model such relations as conditional probabilities in a Bayesian network, namely *P(J|C)*, derived through *P(J|K)* and *P(C|K)*, and *P(G|JC)* which are in turn modeled as Gaussian distributions. It is worth noticing that we do not connect *J* and *C* directly because they express two different macro-abilities (assessment and knowledge), and even at two different levels of generalization. *J* is general for the topic, while *C* is associated to the specific assignment, though related to the topic. On the other hand, *J* and *K* are both topic-based, while *C* directly depends on *K*, since they pertain to the same educational objective.

We investigate the distribution parameters, i.e., the parameter μ (the mean or expectation of the distribution), and the parameter σ (its standard deviation), separately for each relevant distribution. One of our main research questions is which is the better localization estimate of the value μ of *P(J|C)* with respect to the value for Correctness (*C*) of the student's answers. The other one is how much the conditional distribution *P(J|C)* is spread around the average value, i.e., how much variation can be expected in the dependence between *J* and *C*. The results show that the best localization of μ for *P(J|C)* is below the value for *C*, meaning that on average the ability to judge is lower than ability to solve the exercise, which in turn demonstrated that judgment is a "harder" task, i.e., it is at a higher cognitive level. We give more details on the experimental results in the next section.

V. Experiments

To test our hypotheses we have modeled each student as one Bayesian sub-network (see Figure 1), interconnected to her peer's corresponding sub-networks, as sketched above.

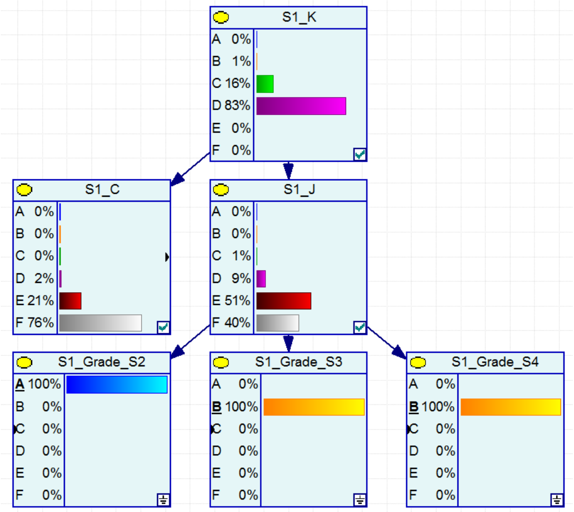

Figure 1: The student model of student S1 with her assessment of answers from peers S2, S3, S4

The subnetworks are interconnected through the probabilistic dependence of the *Grade* variables, associated to the assessment performed by a student, which depend both from her *Judgment* and, most importantly, from the *Correctness* of the assessed peer answers..

To understand how the *Judgment* ability of a student is placed respect to her *Knowledge* and to *Correctness* of her answers, we have modeled the probabilistic dependencies of *J*, *C* and *G*, as Gaussian distributions parametrized respect to their μ and σ values. In particular:

$$P(J \mid K) = Gauss(K + \delta_J, \sigma_J)$$

$$P(C \mid K) = Gauss(K + \delta_C, \sigma_C)$$

$$P(G \mid J, C_G) = Gauss(C_G + \delta_G, \sigma_G * J)$$

The parameters $\delta_J, \delta_C, \delta_G$, position the center μ of the respective Gaussian distributions relative to the *K* or C_G variables, while the corresponding σ_J, σ_C and σ_G describe their standard deviation. Notice that we use here C_G to denote the correctness of the answer that the student is grading rather than the correctness of her own answer, which is denoted by *C*. The δ_J, δ_G, δ_G, parameters could be also described as:

- δ_J added difficulty of doing peer assessment compared to knowing the topic
- δ_C added difficulty of doing the specific exercise compared to knowing the topic

Then, the difference $\delta_J - \delta_C$ denotes the added ability to judge a peer's answer compared to solving the exercise

- δ_G bias of the peer assessment grade respect to the correct one

Since they mostly represent the displacement of the values for other abilities with respect to the value for *K*, their most accurate values as returned from the experiments can provide a kind of ranking of the associated abilities. Notice that we model the different ability to judge a peer's answer by making the standard deviation of the $P(G \mid J, Ci)$ distribution linearly dependent on the *J* value (with A=1 and F=6), i.e.:

- A very good judge (*J=A*) has very high probability to judge her peer correctly (narrow Gaussian distribution)
- A poor judge can grade her peer further away from the correct C_G value (wide Gaussian distribution)

To find the correct values of the parameters we have used an objective function which evaluates how much the model is able to correctly deduce the remaining students' grades when the system is fed only a partial set of the teacher grades. In this respect, this work is also a further step in our earlier investigations aiming at making OpenAnswer a valuable support tool for the teacher correction. For each run, with a given set of values for $\delta_J, \delta_C, \delta_G, \sigma_J, \sigma_C, \sigma_G$ parameters, we simulate a partial correction where the teacher chooses the next answer to correct with one of the selection strategies, which implement the ones listed above:

- *maxEntropy*: the student with the present $P(C|K)$ with highest entropy (lower certainty) is chosen,
- *maxWrong*: the student with maximum $P(C=F|K)$ is chosen

We also remind the possible termination criteria:

- *noFlip(N)*: the deduced grades on the remaining students have been stable in the last N steps (with the maxEntropy strategy).
- *noWrong(W)*: there is no remaining answer with $P(C=F)$ bigger than W (with the maxWrong strategy)

When a sufficient set of answers has been corrected, as detected by the chosen termination criteria, the correction is stopped.

The datasets used in our correction simulations come from different experimental settings (in class tests/exercises):

dataset	level	topic	groups	students
A/170-171	University	multi-level cache systems	2	15 and 14
M/3-4	University	C programming with array	2	13 each
M/6-7	University	C progr, with linked lists	2	11 each
M/8-9	University	C progr. on searching in linked lists	2	9 and 11
I/3-4	H. School	Physics	2	14 and 12

	Delta_G																										
	MIN of AvgTotDV									AVG of AvgTotDV									MAX of AvgTotDV								
DJ-DC	2,0	1,5	1,0	0,5	0,0	-0,5	-1,0	-1,5	-2,0	2,0	1,5	1,0	0,5	0,0	-0,5	-1,0	-1,5	-2,0	2,0	1,5	1,0	0,5	0,0	-0,5	-1,0	-1,5	-2,0
4,0	2,72	2,08	1,80	0,94	0,85	0,65	0,51	0,48	0,67	3,72	3,05	2,57	2,08	1,89	1,76	2,27	2,39	2,55	4,82	4,39	3,96	3,40	3,06	3,08	4,52	4,92	5,03
3,5	2,34	1,98	1,37	0,94	0,77	0,62	0,55	0,45	0,62	3,55	2,94	2,48	2,02	1,78	1,64	1,88	2,01	2,24	4,76	4,31	3,93	3,36	3,06	3,06	4,52	4,87	4,98
3,0	2,27	1,76	1,32	0,96	0,76	0,60	0,54	0,46	0,60	3,39	2,81	2,38	1,94	1,70	1,57	1,71	1,82	2,07	4,67	4,22	3,85	3,27	3,07	3,06	4,43	4,75	4,86
2,5	1,71	1,46	0,92	0,86	0,76	0,53	0,44	0,50	0,57	3,21	2,68	2,26	1,85	1,61	1,49	1,52	1,63	1,85	4,49	4,09	3,70	3,17	3,07	3,06	3,76	4,44	4,69
2,0	1,55	1,20	0,86	0,84	0,69	0,51	0,38	0,51	0,57	2,97	2,54	2,13	1,75	1,53	1,40	1,40	1,47	1,64	4,36	3,90	3,53	3,12	3,07	3,07	3,06	4,10	4,36
1,5	1,28	1,01	0,85	0,80	0,56	0,53	0,38	0,51	0,54	2,74	2,35	1,98	1,65	1,45	1,33	1,32	1,38	1,47	4,20	3,71	3,39	3,07	3,07	3,07	3,07	3,66	4,00
1,0	0,34	0,31	0,31	0,43	0,49	0,51	0,37	0,50	0,51	2,45	2,12	1,80	1,54	1,38	1,29	1,28	1,32	1,39	4,05	3,59	3,33	3,07	3,07	3,07	3,07	3,07	3,51
0,5	0,30	0,31	0,37	0,51	0,44	0,47	0,41	0,47	0,49	2,20	1,94	1,68	1,46	1,33	1,25	1,25	1,29	1,35	3,92	3,46	3,22	3,07	3,07	3,07	3,07	3,07	3,07
0,0	0,32	0,36	0,47	0,48	0,39	0,43	0,43	0,47	0,46	2,02	1,80	1,59	1,42	1,30	1,25	1,25	1,29	1,35	3,81	3,38	3,11	3,07	3,07	3,07	3,07	3,07	3,07
-0,5	0,37	0,48	0,56	0,44	0,36	0,38	0,43	0,47	0,45	1,81	1,64	1,45	1,32	1,22	1,18	1,19	1,22	1,28	3,43	3,12	2,97	2,99	2,98	2,96	2,95	2,94	2,93
-1,0	0,45	0,55	0,50	0,41	0,33	0,34	0,40	0,42	0,44	1,62	1,48	1,33	1,22	1,16	1,13	1,16	1,20	1,26	3,21	2,93	2,88	2,88	2,85	2,82	2,80	2,77	2,75
-1,5	0,55	0,59	0,47	0,38	0,32	0,32	0,36	0,38	0,42	1,44	1,33	1,22	1,14	1,10	1,10	1,14	1,19	1,26	2,96	2,71	2,39	2,16	2,27	2,34	2,46	2,42	2,52
-2,0	0,62	0,56	0,45	0,37	0,31	0,31	0,33	0,37	0,41	1,28	1,20	1,13	1,09	**1,08**	1,10	1,15	1,21	1,28	2,74	2,46	2,16	2,20	2,28	2,30	2,35	2,47	2,50
-2,5	0,65	0,54	0,44	0,37	0,31	**0,30**	0,32	0,37	0,40	1,19	1,14	1,10	1,09	1,11	1,15	1,21	1,28	1,36	2,38	2,18	2,12	2,22	2,31	2,33	2,45	2,50	2,63
-3,0	0,65	0,68	0,62	0,49	0,50	0,56	0,60	0,64	0,71	1,14	1,13	1,13	1,15	1,19	1,23	1,29	1,36	1,46	**1,99**	2,03	2,18	2,24	2,38	2,45	2,45	2,53	2,63
-3,5	0,78	0,70	0,51	0,50	0,51	0,57	0,64	0,68	0,76	1,17	1,18	1,19	1,23	1,27	1,32	1,38	1,46	1,55	2,10	2,15	2,20	2,29	2,38	2,41	2,42	2,47	2,65
-4,0	0,72	0,72	0,71	0,65	0,68	0,73	0,80	0,86	0,97	1,27	1,28	1,32	1,35	1,40	1,46	1,52	1,58	1,66	2,12	2,17	2,24	2,31	2,42	2,43	2,44	2,54	2,69

Figure 2: MIN (left), AVG (center) and MAX (right) of AvgTotDV (green=better, red=worse) respect to different values of δ_J-δ_C and δ_G over the 3 datasets A,I,M

The letters in the labels identify the different groups of exercises, and the numbers identify the peer assessment IDs. For each student the difference between her inferred grade and her correct grade is computed and the overall average difference **AvgTotDV** over the whole peer assessment is computed and reported among the experimental results.

In the simulated correction process, the computation of the *noFlip* termination criteria requires mapping the estimated probability distribution $P(C_G)$ onto a discrete vote in the range [A-F]. To obtain this, we compute a continuous numeric grade from 0 to 10 as the weighted sum of the probabilities of the different intervals. The weight of each interval is the value of its center (A = 9.75, B = 9, C = 8, D = 7, E = 6, F = 2.75). Afterwards, we discretize the grade resulting from the weighted sum so as to take it back to the corresponding interval. This is done because if we just computed the difference between discretized grades and the teacher's grade, the latter being discretized as well, we should obtain only integer values. Given that both the teacher's grade and deduced ones are in a continuous range, we can increase the precision of the computed difference by considering the difference between continuous values instead of discretized ones. The results reported here are obtained by this procedure.

Getting a minimum **AvgTotDV** value is the objective function of our optimization.

After the first experiments we have noticed that the objective function appears to be continuous, and thus we have been able to reduce the parameter space explored by our optimization runs by separately optimizing the δ_J, δ_C, δ_G, and the σ_J, σ_C, σ_G, parameters as follows:

- First we have fixed σ_J=1.0, σ_C=1.0, σ_G=0.5 with an educated guess and explored the parameter space of the remaining parameters over the set of values:

  δ_J: from -2 to +2 in 0.5 increments

  δ_C: from -2 to +2 in 0.5 increments

  δ_G: from -2 to +2 in 0.5 increments

  This allowed us to find a good set of values for the following σ_J, σ_C, σ_G, optimization. Thus we have fixed δ_J=1.5, δ_C=-1.5, δ_G=0.5 and searched for best σ_J, σ_C, σ_G, which we found at:

  σ_J=1.0, $\boldsymbol{\sigma_C}$**=0.5**, σ_G=0.5

  which showed that our initial educated guess was almost correct.

- Finally we have re-optimized δ_J, δ_C, δ_G relative to the new fixed sigma parameters.

The resulting δ_J, δ_C,and δ_G values computed are described in Figure 2, where we show both MIN, AVG and MAX of the AvgTotDV objective function depending on the δ_J-δ_C and δ_G parameters.

Our simulations assume no previous known information on the student's *K* except the *P(K)* distribution of the current assessment,, which is obtained from the teacher's grades.

All units in the table are in grades, i.e. 1 = distance between two consecutive A,B,C,D,E,F values.

From the table we see that the best average (AVG) values (green color) reside in the bottom part, with negative δ_J-δ_C which shows that *J* is lower than the corresponding *C* (the average minimum 1.08 is obtained at δ_J-δ_C= -2). There seems not to be a significant *Grade* displacement as the average minimum 1.08 is obtained for δ_G=0.

By looking at the MIN (leftmost) set of columns we see that OpenAnswers could infer grades with very low error (0.3 grades) over some combination of parameters, selection strategy, termination criteria and data-set . More work is due to find the best combination applicable to all data-sets.

The MAX (rightmost) columns instead show that we are able to ensure an upper bound on the average induced grade error of less than 2 grades (1.99 at δ_G=2 and δ_J-δ_C=-3), which is promising. More work is needed to improve the strategies and termination criteria and to rule out particularly bad combinations (e.g. maxEntropy/noWrong).

The most interesting result is that in all cases the best AvgTotDV values are obtained in correspondence to negative δ_J-δ_C . Such negative values mean that *J* is lower than the corresponding *C,* both in relation to *K.* In other words, the students get a lower grade for *J* than for *C*, meaning that the ability required to provide a correct judgement is “harder” than that required to give the correct answer. This confirms the validity of Bloom’s ranking, that suggests that judgement lies at a higher cognitive level.

VI. Conclusions and future works

The affectivity of e-learning activities depends on several factors, such as, as a limited set of examples, the apt selection of learning material [18 - 21], the social and collaborative fruition of it [6, 7], and the assessment of the formative results, that has been the main topic in this paper. We have shown how, by looking for the best parameters for the OpenAnswer Bayesian network model of peer-assessment, experimental evidence that peer assessment (Judgment) is harder than solving the exercise (Correctness). To the best of our knowledge this is the first experimental demonstration of part of the Bloom’s taxonomy of cognitive levels. This finding points out possible issues that are related to peer assessment, e.g., its quality and reliability.

Our current simulations assume no previous known information on K pertaining to the student except the current P(K) distribution from the assessment, which is obtained from the teacher’s grades (ground truth). In a following experiment we will feed information on K from the assessment of earlier exercises on the same or similar topics to explore how much the inferred grades improve in presence of added info. Furthermore, we will also weight students’ grades according to their profiles, in order to identify the impact of student’s profile on the results. We will also investigate if the relationship between knowledge and judgment can be influenced by lack of anonymity of peer work. This may take a twofold aspect; the grading may be influenced by personal relationships among peers, as well as by the reputation of graded peers.

We are aware of the long way still to go. In our previous works we have just explored the factors that can influence peer-assessment, and also aimed at investigating the extent at which peer assessment can relieve the teacher from the burden of a full correction. This entails building a suitable model for propagating and merging information coming from both teacher’s and students’ grading. In this work we have experimentally used this same model to verify the greater knowledge required to judge than just to carry out a task. However, a very important point still to address is how to exploit the mixed strategy of OpenAnswer (a kind of co-assessment) to improve students’ judging abilities. Of course, a comparison between the grades assigned by different peers and by the teacher to a same school work can be beneficial to both improve the students’ grading competence and to point out possible flaws in learning. However, as for the present setting, this is out of the scope of our work, since it entails framing OpenAnswer within a consistent educational strategy, where the teacher can plan different usages of peer-assessment results.

References

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

[2] A.F. Gourgey, "Metacognition in basic skills instruction," Instructional science, Vol. 26, pp. 81–96, 1998.

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

[4] 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: Cognitive domain. New York: David McKay Company, 1956.

[5] L. W. Anderson, D. R. Krathwohl (eds.), A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives. Allyn and Bacon, 2000.

[6] K. Kreijns, P.A. Kirschner, and W. Jochems. Identifying the pitfalls for social interaction in computer supported collaborative learning environments: a review of the research. Computers in Human Behavior, 19, pp. 335-353, 2003.

[7] A. Sterbini and M. Temperini. Learning from Peers: Motivating Students through Reputation Systems. Proc. IEEE/IPSJ Int. Symp. on Applications and the Internet, Workshop SPEL, pp.305-308, 2008, doi.ieeecomputersociety.org/10.1109/SAINT.2008.107

[8] Y. Cheng and H. Ku. An investigation of the effects of reciprocal peer tutoring. Computers in Human Behavior, 25, 2009.

[9] H. Somervell, “Issues in assessment, enterprise and higher education: the case for self-, peer and collaborative assessment,” Assessment and Evaluation in Higher Education, 18, pp. 221-233, 1993.

[10] L. S. Kane, E. E. Lawler, “Methods of peer assessment,” Psychological Bulletin, 85, pp. 555-586, 1978.

[11] F. Dochy , M. Segers, D. Sluijsmans, “The use of self-, peer and coassessment in higher education: A review,” Studies in Higher Education, Vol. 24(3), pp. 331-350, 1999.

[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] Palmer, Richardson, On-line assessment and free-response input – a pedagogic and technical model for squaring the circle, Proc. 7th CAA Conference, Loughborough University, 2003

[14] A. Sterbini and M. Temperini, Correcting open-answer questionnaires through a Bayesian-network model of peer-based assessment. Proc. Int. Conf. on Information Technology Based Higher Education and Training, ITHET 2012, pp. 1-6 (2012). DOI:10.1109/ITHET.2012.6246059

[15] 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). pp. 164-170 (2013)

[16] A. Sterbini, M. Temperini, Analysis of OpenAnswers via mediated peer-assessment. Proc. 17th IEEE Int Conf. on System Theory, Control and Computing (ICSTCC 2013), Workshop SPEL, 2013.

[17] M. De Marsico, A Sterbini, M. Temperini, Adding propedeuticy dependencies to the OpenAnswer Bayesian model of peer-assessment, Proc. Int. Conf. on Information Technology Based Higher Education and Training, ITHET 2014

[18] F. Gasparetti, C. Limongelli, F Sciarrone. Exploiting Wikipedia for Discovering Prerequisite Relationships Among Learning Objects. In Proc. ITHET 2015, 11-13 June, 2015, Lisbon, Portugal.

[19] C. Limongelli, F. Gasparetti, F. Sciarrone. Wiki Course Builder: a System for Retrieving and Sequencing Didactic Materials from Wikipedia. In Proc. ITHET 2015, 11-13 June, 2015, Lisbon, Portugal.

[20] F.Gasparetti, C. Limongelli, F. Sciarrone A Content-based Approach for Supporting Teachers in Discovering Dependency Relationships Between Instructional Units in Distance Learning Environments. In Proc HCI Int. 2015, 2-7 August 2015, Los Angeles, CA, USA

[21] C. Limongelli, F. Sciarrone, M. Temperini. A social network-based teacher model to support course construction. Computers in Human Behavior, 2015, doi:10.1016/j.chb.2015.03.038