

POEM-COPA Collaborative Open Peer Assessment

Pierre Collet, Raaj Seereekissoon, Isaac Abotsi, Marie Michaud-Maret,
Anna Scius-Bertrand, Emma Tillich, and Pierre Parrend

1 Introduction

A long time ago, education was provided by personal tutors who were paid by rich families to take care of 3 or 4 children. This type of education was of very high quality, as the tutor could adapt his tuition to the capabilities and inclinations of each student, therefore providing personalized education. Unfortunately, this was also very expensive, meaning that very few people were educated. The advent of schools allowed for many more people to learn how to read, write and count but this was only possible through mass education, with classrooms of 30 pupils and national education programmes that are identical for all. This means that slow students are often left behind whilst bright students have to wait for the others. This is even more so with the advent of Massive Open Online Courses (MOOC) [7], that are currently invading the world of e-education. With MOOCs, a single course can be followed by thousands of students.

The aim of the Personalised Open Education for the Masses (POEM) platform is to use complex systems to create an intelligent Learning Management System that

P. Collet (✉) • R. Seereekissoon • I. Abotsi • E. Tillich • P. Parrend
CSTB Team, ICUBE Laboratory, Strasbourg University, Strasbourg, France

e-laboratory on Education, CS-DC UNESCO UniTwin, Paris, France
e-mail: pierre.collet@unistra.fr

M. Michaud-Maret
Institut de Formation des Métiers de Santé du Nord Franche-Comté, Belfort, France

A. Scius-Bertrand
e-laboratory on Education, CS-DC UNESCO UniTwin, Paris, France
École Pratique des Hautes Études, Paris, France

is able to educate hundreds of thousands of students along personalized trajectories, depending on their previous knowledge, skills and experience, as with the personal tutor of ancient times.

One could think that massive education and personalized education are antagonistic objectives but on the contrary, they are in synergy.

A long time ago, personal tutors would use their teaching experience to find the best series of exercises on topics adapted to each child they were in charge of. Then, if they detected a particular skill or interest in literature, science or the arts in their pupil, the tutors (who were often multifaceted) would adapt their teaching to nurture and develop this inclination for a more personalized education.

Their pedagogic experience increased with the number of children they taught, as trial and error improved their tuition skills. Having the opportunity to statistically study large numbers of educational trajectories, modern Intelligent Tutoring Systems can draw conclusions on previous successes and failures to improve their interactions with students online, and are in the position of predicting the best future for a student.

More accurate predictions require assimilating data of a massive number of such trajectories (which once more is what good professors do as their experience increases). For this reason, the participation of everyone to such an educational ecosystem is extremely desirable. Not only will it improve the system, but anyone wanting to resume their education will be quickly learning things they do not already know. Such synergy between massive and personalized education is only possible within a social intelligent ICT platform. The aim of POEM is therefore to implement an educational ecosystem responding to the objectives of 4P-education, i.e.:

- **Participative** (to collect data and reinforce the “experience” of the system),
- **Predictive** (to guide the student using the elaborated experience),
- **Preventive** (to avoid failure) and
- **Personalized**, thanks to multi-level Quality Measurements allowing for an experience tailored to each student.

Many functionalities must be developed in order to develop such a comprehensive platform, typically:

- **Constructing and visualizing dynamic Knowledge Maps** of domains, to help students determine their objectives.
- **Developing individual MOOC and curricula trajectories.** POEM conjectures that, given an individual profile, the best next incremental step is determined in probability by the distribution of the choices of previous learners with similar profiles. This conjecture is developed in the Personalized Educational Man-Hill Problem, because of the similarity with ants’ collective behaviour, which is known to quickly find optimal paths towards food sources [2–4, 6].
- **Providing inter-tutoring** between students, which is needed if direct Student-Teacher interaction is impossible due to the very large number of students. In such cases, POEM provides each student with a tutor, who is also a student but

more advanced in the same curriculum. The student can ask questions to his tutor. If the tutor cannot provide an answer to a difficult question, he himself can forward the question to his own tutor and so on, until there is no tutor anymore and the question reaches a professor.

- **Offering an automatic skill-level assessment system** depending on success/failure along the personalized trajectories of students. This is implemented through Elo-points as in chess or Tennis ranking.
- **Offering a high-quality assessment of open answers to open questions.** This is proposed through peer assessment, which is recognized to be of mutual benefit.
- **Provide crowdsourcing** by letting students bring new questions and new content as part of their evaluations. Good questions and content will find their way into participatively evolved trajectories, while poor content will eventually get discarded.

This paper will focus on the two last points, implemented in the COPA (Collaborative Open Peer Assessment) module of the POEM¹ educational ecosystem of the CS-DC (*Complex Systems Digital Campus*) UNESCO UniTwin.²

2 Peer Evaluation of Open Answers to Open Questions

Peer evaluation has been studied for a long time [16] but was only recently experimented extensively, in distance learning with platforms such as Spark (Self and Peer Assessment Resource Kit) or more recently in different MOOCs [9, 15]. It is quite frequently used for operational teaching, such as management or software development [10] but remains a top down approach, from the teacher to the learner, as topics are not dynamic and imposed to students. In the evaluation of these technical teachings, peer assessment shows a limited deviation that can be lower than 3 % [11], which means that it is quite accurate (at least in computer science, in which this experiment was published). It is also efficient for the assessment of complex tasks such as composition, typically when associated with appropriate coaching [12], which is more difficult to put into practice within MOOCs.

Peer assessment is often better accepted by students than an evaluation performed by the teachers. It enables them to improve the content of the course as well as their ability to evaluate others.

One of the current limitations of available solutions is the static character of the set of questions, which increases the risk of cheating. By including in the assessment the requirement to ask a question (that will be posed to other students), COPA provides a solution to this problem by turning learners into producers of new knowledge and analyses, that will feed the system to create a virtuous circle.

¹<http://poem.unistra.fr>.

²<http://cs-dc.org>.

COPA therefore inserts into xMOOCs the transmission of specific knowledge [13], some dynamic elements of connectivist cMOOCs [14] based on individual experience and interaction between learners.

3 Principles and Implementation of COPA

Implementation of a COPA evaluation needs several databases:

- A database containing the courses (videos, Powerpoint or PREZI presentations, PDF or Word documents, etc.),
- A database containing questions,
- A database containing answers,
- A database containing students and professors.

Then, COPA evaluation uses three stages: a participative stage where students must ask a question on the course, a more passive stage where they should answer three questions and then, a third stage where they should evaluate the answers given by other students.

Before the COPA evaluation takes place, students should have followed a course (face-to-face or video) on which they are to be evaluated. If digital contents are associated with the course, they can be stored in the courses database, so that students can access them for reference during all stages of the COPA evaluation.

Interestingly enough, because COPA allows open questions and answers, topics are not limited to hard sciences where one can expect exact answers to precise questions. COPA can also be used to evaluate knowledge in social sciences, skills, literature, arts, a.s.o.

3.1 COPA Phase 1

Write a question on the course and
Provide a model answer.

When opening phase one, students have access to the course but rather than being asked to answer some questions, they are asked to pose a question on the corresponding course. This activity is much more demanding than answering a question because the students must be creative in order to ask relevant and interesting questions. Indeed, the quality of their questions will be rated (by other students) against questions from the pedagogic team, so if the question they imagine is less interesting than the teacher's questions, they will not get a good grade on the exercise.

Then, after they have been asked to pose a question, an even more challenging task is asked from them: they must provide a model answer for the question they have imagined.

Here again, the quality of the provided model answer will be evaluated by other students.

4 COPA Phase 2

Answer (and evaluate the quality of) three questions.

This phase has several purposes:

1. assessing if the students have understood the content of the course,
2. evaluating the quality of the questions posed by students in Phase 1 and
3. improving the quality of the database of student questions.

The contents of this phase are inspired by CAPTCHAs [1] and Re-CAPTCHAs [7]. CAPTCHA is an acronym for “Completely Automated Public Turing test to tell Computers and Humans Apart”. Indeed, during 1950 Turing came out with a very famous test [5] to allow humans to determine if an unseen interlocutor is a human or a machine, a CAPTCHA can be seen as a reverse Turing test, created to allow computers to determine whether their interlocutor is another computer or a human.

Re-CAPTCHAs make use of the time and energy given by humans to pass the CAPTCHA test in a constructive way, i.e. to solve a problem for the computer (cf. Fig. 1.) Because POEM-COPA is run by a computer, the algorithms used in POEM-COPA are computer-oriented, not human-oriented, and therefore CAPTCHA-like. As in re-CAPTCHAs, COPA is not only asking the students to answer the questions, but also asking them to rate the relevance, originality and quality of the formulation of the questions they are asked.



Fig. 1 Re-CAPTCHA: the computer not only tests whether its interlocutor is human or not (by asking him to decipher the twisted text on the right) but also asks him what is the word on the left (that it could not recognize *via* Optical Character Recognition because it was badly printed)

However, because the aim of this phase is to evaluate how much of the course has been understood by the student, it is important that most of the questions come from the pedagogic team.

4.1 *Answering 3 Questions*

Re-CAPTCHAs are typically divided into two parts: a part to tell if the user is human or not, and another “crowdsourcing” part where the collaboration of users is sought.

In COPA, 2 questions come from a database of questions validated by the pedagogical team (providing for an approved evaluation), and one question comes from another student (crowd-sourcing part, where the participation of the user is sought) but of course, the user does not know which question among the 3 is by the student.

This 2/3–1/3 proportion means that students are mostly evaluated on validated questions (with model answers provided by the pedagogical team).

Each provided answer will be anonymously evaluated by 3 other students (peers), following the current practice in scientific journals/conferences where the quality of submitted research is also evaluated by peers.

4.2 *Constrained Evaluation of Questions*

In this collaborative part, students are asked to evaluate the relevance, originality and formulation of the 3 questions they are asked, with the aim of evaluating the quality of the student question that is posed along with 2 questions from the pedagogical team.

The risk of self-assessment between students is to observe some bias induced by the type of training undergone by the students: in competitive training (ending with a competitive exam), it is in the interest of the students to give bad marks to the others, in the hope of obtaining better relative marks. On the contrary, in courses where all students above a certain grade pass the exam, there is no competitive pressure so students may decide mark other students more generously than they otherwise would.

In order to reduce such bias, students are not asked to give grades to the questions but to rank them *via* a constrained rating, meaning that in effect, the question asked by a student will be compared to the questions asked by the pedagogical team.

Students must give 0–5 points to each of the 3 questions but they only have exactly 10 points that they must distribute entirely.

This method imposes that only the cases of Table 1 can be encountered. One can see in this table that:

Table 1 Rating combinations and marking

Teacher Q1	Teacher Q2	Student Q	Teacher average	Difference	Grade	Mark/5
0	5	5	2.5	2.5	7.5	5
1	4	5	2.5	2.5	7.5	5
2	3	5	2.5	2.5	7.5	5
3	2	5	2.5	2.5	7.5	5
4	1	5	2.5	2.5	7.5	5
5	0	5	2.5	2.5	7.5	5
1	5	4	3.0	1.0	6	4
2	4	4	3.0	1.0	6	4
3	3	4	3.0	1.0	6	4
4	2	4	3.0	1.0	6	4
5	1	4	3.0	1.0	6	4
2	5	3	3.5	-0.5	4.5	3
3	4	3	3.5	-0.5	4.5	3
4	3	3	3.5	-0.5	4.5	3
5	2	3	3.5	-0.5	4.5	3
3	5	2	4.0	-2.0	3	2
4	4	2	4.0	-2.0	3	2
5	3	2	4.0	-2.0	3	2
4	5	1	4.5	-3.5	1.5	1
5	4	1	4.5	-3.5	1.5	1
5	5	0	5.0	-5.0	0	0

Col 3 and 4: The quality of the student question can never be equal to the average quality of the 2 questions of the pedagogical team.

Col 5: In 11 cases, the student question gets better grades than the average of the teacher questions, vs 10 cases when it is deemed worse.

Col 6: If one adds 5 points so that the student gets grades between 0 and 7.5 and if

Col 7: one multiplies by $5/7.5$ in order to get the grades back into a $[1,5]$ interval, the student question gets a mark (Col 7) that is identical to the grade given to him by the student undergoing the COPA evaluation.

Students whose question have been estimated as being of slightly lower quality than the average of the teacher's questions will get $3/5$, which is fine as it can be considered difficult for students to only obtain points if their question is better than the questions of the teacher.

If the question of the student is used several times, it will get marked several times. COPA will use the average mark, weighed by the number of evaluations.

4.3 Improving the Quality of the Questions in the Database

Because there are obvious questions for all courses, there is a great risk that many students will ask the same questions. Then, some of the obvious questions may also be asked by the pedagogic team.

This could cause a problem as it allows for the possibility that the selected student question be semantically identical to the question provided by the pedagogic team, in which case the student undergoing the COPA evaluation will be faced with two identical questions (one from a student and another one from the pedagogic team).

If this is the case, the student has the possibility to indicate that 2 questions are identical. If the student clicks on the “similar questions” button, the student question will be replaced by a question from the pedagogical team, therefore guaranteeing that all 3 questions are different.

Then, the student question that was noted as being similar to one of the teacher’s questions will be flagged and not be selected along with the incriminated teacher question in the future.

Another button is available for students to signal if the contents of a question are inappropriate. Inappropriate questions are removed from the students question database and an email is sent to the pedagogic team containing the question and the email of the student who submitted it.

5 COPA Phase 3

Evaluate 9 answers from other students and
Evaluate the quality of model answers.

The current practice in science is that new research submitted to a conference or journal is evaluated by 3 experts on the domain. Because no “super-scientist” exists, the “experts” cannot be anyone else other than scientists (i.e. peers) working in the same field as the scientist who submitted the work, so basically, research is driven by anonymous peer reviewing.

We pose that (even if double blind peer reviewing has its pros and cons) what is currently good enough for research evaluation could also be used for student evaluation.

Each of the three answers given in phase 2 will be evaluated by 3 peers, to determine if the answer is correct or not. Such student peer review is very desirable because:

1. The online evaluation scheme is not limited to Multiple Choice Questions: students have brains that can be used to analyse the provided answers. Using students as evaluators allows COPA to use human brains to evaluate the quality of *open responses to open questions*.

2. Evaluating somebody else's answers is not a waste of time for the evaluators: all teachers know that studying someone else's point of view has many pedagogic virtues.
3. Students take their task seriously, as they know that the mark they give could have some influence. Asking them to be evaluators *involves* participating students.

Practically speaking, because in phase 2, students must answer 3 questions and because each of the three questions must be reviewed by 3 other students, phase 2 creates the need for 9 reviews (3×3).

5.1 Evaluation of 9 Answers

Phase 3 therefore consists of 9 evaluations of other student's answers. In order to help the evaluating student in his task, he/she is presented with:

1. the question that was posed,
2. the model answer that was proposed by the person who wrote the question (teacher or student) and
3. the answer to be evaluated.

Because the evaluating student is presented with the question and its model answer, this is an occasion for him to see 9 more questions and answers than those he had to answer in phase 2. Then, he has to fully understand the question and the proposed model answer in order to form a judgement on the quality of the student answer he must evaluate. Doing this is not a waste of time for the evaluating student: because he is involved (his accurate evaluation will have an influence on another student's mark) the time he spends on this task is of high pedagogic quality.

Here again, a constrained evaluation scheme is given to the evaluating student so that he/she is not tempted to underrate or overrate the other student's answers.

Because the added value of POEM is more to have students revise and understand their lessons (formative evaluation) rather to evaluate how the lessons have been understood (summative evaluation) or evaluate how well students compare with peers (normative evaluation), in the first versions of POEM, the students had to distribute *exactly* 30 points over the 9 answers to be evaluated. 30 was chosen so that the evaluating student had to give an average of 3.22 points to the 9 answers, in order to be encouraging for students.

However, even though this was meant as an encouragement to students to get involved in POEM as it made it difficult for them to participate and to get bad marks, this hard constraint of 30 points was considered as too strong by many students.

Indeed, Table 2 shows the probabilities for each grade to be given: if an evaluator decided to give 0 to someone, he would have 30 points to distribute over 8 questions only, meaning giving an average of 3.75, which is a lot. Then, if an evaluator gave 3×0 because he evaluated that 3 answers were plainly wrong, he/she had no other choice than giving 5 to all the other answers.

Table 2 Possible occurrences of the different grades and probabilities

Grade	0	1	2	3	4	5
Grade occurrences	16,808	25,488	36,688	50,288	65,808	82,384
Probabilities	6.06 %	9.19 %	13.22 %	18.12 %	23.72 %	29.69 %

Table 3 Possible occurrences of the different grades and probabilities

Grade	0	1	2	3	4	5
Grade occurrences	615,553	732,033	842,499	939,099	1,014,399	1,062,279
Probabilities	11.83 %	14.06 %	16.19 %	18.04 %	19.48 %	20.40 %

In the current version, more leeway was given to evaluators: they could distribute between 22 (average of 2.44/5) and 30 (average of 3.33/5) to distribute over the 9 answers. This increased the number of possible grades combinations from 277,464 to a whopping 5,205,862 and the number of possible combinations with no 0 or 5 jumped from 2598 to 159,436.

The probabilities for each grade to be given was more even (cf. Table 3). Evaluators could still not give bad grades to all answers (impossible to give a lower average than 2.44, which is pretty close to 2.5/5) while still being able to give greater than average marks to everyone (encouraging for participating students).

More importantly, this made it easier to distribute grades 0 and 5. Previously, giving 2×0 meant you had to give at least 2×5 and therefore 5×4 : $0 + 0 + 4 + 4 + 4 + 4 + 5 + 5 = 30$.

With the less constrained notation, evaluators had more choice to give 0s or 5s.

5.1.1 Grade Interpretation to Create Marks

Because evaluators have more leeway in the way they can distribute grades, it is considered that *grades from 0 to 5* are interpreted as *marks from 0 to 3* the following way:

Grades 0 and 5: are given by *choice* rather than by constraint. It is therefore desirable to respect the evaluator’s choices when they are extreme. *Grade 0* gives **mark 0** and *grade 5* gives **mark 3**.

Grades 1 and 2: if grades 0 and 5 are given by choice, more constraints apply on grades 1, 2, 3, 4 in order to stay within the 22/30 points limit. Therefore, *grades 1 and 2* give **mark 1**.

Grades 3 and 4: Similarly, *grades 3 and 4* give **mark 2**.

Thanks to this relaxation of constraints, what was previously ranks can now be considered as a summative evaluation. For an evaluated student, each of the 3 questions he answered on phase 2 is evaluated three times between 0 and 3. The marks are summed and then multiplied by 5/27 in order to obtain a global summative mark within [0, 5].

5.2 *Evaluation of Model Answers*

As said above, in order to help the evaluator in his task, he is presented with:

- the question that was posed,
- the model answer that was proposed by the writer of the question,
- the answer to be evaluated.

Because the writer of the question may be a student, it is possible that the student posed a good question, but has not thoroughly understood it and therefore, did not provide a good or complete enough model answer.

Therefore, whenever shown a model answer, the evaluator is also asked to evaluate the quality of the model answer proposed by the writer of the question, by giving it a grade between 0 and 5.

Because COPA is thought as a collaborative platform, if the given grade is 2 or less, this means that the evaluator clearly thinks that the model answer could be improved. A pop-up window therefore opens, asking him to provide a better model answer. In the future, both model answers will be shown to future evaluators, who will then have to evaluate two model answers.

Because this is a participative grade, no constraint is given on the grades that can be given. If the model answer is evaluated several times, the average mark is given weighted by the number of evaluations.

5.3 *Evaluation of the Quality of the Marking*

The conscientiousness of the evaluator can also be evaluated. Indeed, if all 3 evaluators do their job correctly, they should give equivalent marks to a same answer. One can then determine how different the marks of an evaluator are with reference to the marks given by his/her two colleagues, on the same question. To this effect, one can compute the difference it makes on the average mark to include his mark or not.

Supposing the evaluator gives the maximum grade (5, i.e. mark 3) on an answer where the other two gave 0. The average when including his mark is 1, while the average without his mark is 0. It is possible to sum all the differences made by this user (maximum 9) and evaluate the quality of the conscientiousness of evaluator e as:

$$C_e = 9 - \sum (\overline{m} - \overline{(m - m_e)})$$

with $\overline{(m - m_i)}$ the average of the marks of the other evaluators than e .

The mark is multiplied by 5/9 in order to bring it into the $[0, 5]$ interval.

6 Global Marking

For each course, POEM-COPA requires many interactions involving the student, from writing a question to writing a model answer, to answering 3 questions (while evaluating their quality) to evaluating 9 answers (and evaluating 9 model answers).

COPA provides the professor with two marks: one corresponding to a summative evaluation (within $[0, 5]$) and with participation marks coming from:

- the evaluation of the quality of the provided question ($[0, 5]$ weighted),
- the evaluation of the quality of the provided model answer ($[0, 5]$ weighted),
- the evaluation of the consciousness of the student as an evaluator ($[0, 5]$).

The POEM-COPA interface asks over how many points the student should be graded (usually 20 in France, but could be anything). Then, a selector is provided for the teacher to tell if he wants to combine (or not) the summative and participative evaluations, and offers two parameters to weigh the marks.

Finally, each of the three participative marks (PQ, PMA and C) can also be weighted depending on the importance the teacher wants to associate with each mark.

7 Conclusion

POEM-COPA has been mainly used along face-to-face classes, for more than 10 different courses up to now, at Strasbourg University (in computer science and English courses), French National Engineering School for Advanced Techniques (ENSTA), Institut de Formation des Métiers de Santé du Nord Franche-Comté, in a training course for nursing auxiliaries and others.

Students always have been quite enthusiastic about using the COPA collaborative peer to peer formative evaluation. In one of the feedback polls, 97 % of the students stated the need to open their courses and research content over the Internet in order to fulfill the evaluation. 72 % found that COPA allowed them to be better prepared for the final exam and 71 % said they would like COPA to be used in other courses.

8 Scientific Validation

This paper has been unanimously validated in a collaborative review mode with the following reviewers:

- Christophe Schnitzler, Strasbourg University, Sport Sciences Faculty
- Roeris Gonzalez Sivilla, from University of Camaguey “Ignacio Agramonte Loynaz”.

Acknowledgements We thank the Strasbourg University IDEX Investissements d’Avenir Program for funding the edX-POEM project.

References

1. Ahn LV, Blum M, Hopper NJ, Langford J (2003) Captcha: using hard ai problems for security. In: Proceedings of the 22nd international conference on theory and applications of cryptographic techniques. Springer, Berlin, pp 294–311
2. Johnson J, Shum SB, Willis A, Bishop S, Zamenopoulos T, Swithenby S, MacKay R, Merali Y, Lorincz A, Costea C et al (2012) The FuturICT education accelerator. *Eur Phys J Spec Top* 214(1):215–243
3. Louca J, Johnson J, Bourguine P, Portelli PCT, Scius-Bertrand A, Lenhard W, Escalona M, Taramasco C, Kohlhase M, Cointet J, Collet P (2013) Poem platform for massive personalized education. In: Proceedings of the European Conference on Complex Systems (2013)
4. Semet Y, Lutton E, Collet P (2003) Ant colony optimisation for e-learning: observing the emergence of pedagogic suggestions. In: 2003 IEEE swarm intelligence symposium, SIS 2003, Indianapolis, IN, 24–26 April, pp 46–52. <http://dx.doi.org/10.1109/SIS.2003.1202246>
5. Turing AM (1950) *Comput Mach Intell* 59:433–460
6. Valigiani G, Biojout R, Jamont Y, Lutton E, Republique CB, Collet P (2005) Experimenting with a real-size man-hill to optimize pedagogical paths. In: Proceedings of the 2005 ACM symposium on applied computing (SAC). ACM, New York, pp 4–8. <http://doi.acm.org/10.1145/1066677.1066683>
7. von Ahn L, Maurer B, McMillen C, Abraham D, Blum M (2008) Recaptcha: human-based character recognition via web security measures. *Science* 321(5895):1465–1468. <http://science.sciencemag.org/content/321/5895/1465>

First Complex Systems Digital Campus World
E-Conference 2015

Bourgine, P.; Collet, P.; Parrend, P. (Eds.)

2017, VIII, 424 p. 120 illus., 96 illus. in color., Hardcover

ISBN: 978-3-319-45900-4