Machine and social intelligent peer-assessment systems for assessing large student populations in massive open online education

Conference or Workshop Item

Machine and social intelligent peer-assessment systems for assessing large student populations in massive open online education

Cristian Jimenez-Romero[1], Jeffrey Johnson[1], Ricardo De Castro[2]
[1]The Open University, Milton Keynes, MK7 6AA, UK,
[2]Corporación Universitaria Reformada, Barranquilla, Colombia
cristian.jimenez-romero@my.open.ac.uk, jeff.johnson@open.ac.uk, rcastro@unireformada.edu.co

Abstract: The motivation of the European Etoile project is to create high quality free open education in complex systems science, including quality assured certification. Universities and colleges around the world are increasingly using online platforms to offer courses open to the public. *Massive Open Online Courses* or MOOCs give millions of people access to lectures delivered by prestigious universities. However, although some of these courses provide certification of attendance and completion, most do not provide any academic or professional recognition since this would imply a rigorous and complete evaluation of the student's achievements. Since the number of students enrolled may exceed tens of thousands, it is impractical for a lecturer (or group of lecturers) to evaluate all students using conventional hand marking. Thus in order to be *scalable*, assessment must be automated. The state-of-the-art in automated assessment includes various methods and computerised tools including multiple choice questions, and intelligent marking techniques (involving complex semantic analysis). However, none of these completely cover the requirements needed for the implementation of an assessment system able to cope with very large populations of students and also able to guarantee the quality of evaluation required for higher education. The goal of this research is to propose, implement and evaluate a computer mediated social interaction system which can be applied to massive online learning communities. This must be a scalable system able to assess fairly and accurately student coursework and examinations. We call this approach "machine and socially intelligent peer assessment". This paper describes our system and illustrates its application. Our approach combines the concepts of peer assessment and *reputation systems* to provide an independent computerised system which determines the degree and type of interaction between student peers based on a reputation score which emerges from the marking behaviour of each student and the interaction with other individuals of the community. A simulation experiment will be reported showing how reputation-based social structure can evolve in our peer marking system. A pilot experiment using a population of ninety 16-year old high school students in Colombia measured the marking accuracy of our system by comparing the statistical differences between the scores resulting from teacher marking (the 'gold standard'), peer assessment using average scores, and our intelligent reputation-based peer assessment. This addresses the research question: to what extent does the proposed approach improve peer marking in terms of marking accuracy and fairness? We report the first results of this experiment, summarise the lessons learned, and describe further work.

Keywords: MOOCs, automated marking, peer assessment, reputation systems, complex systems education, Etoile

1. Introduction

The motivation of the European Etoile project is to create high quality free open education in complex systems science, including quality assured certification (http://www.etoileplatform.net/index). In the first instance we are focused on providing postgraduate education for masters and doctoral students and other researchers. This means that, although we can assume that our students will be well motivated with good study skills, we share the same general challenges of open online learning.

Massive Open Online Education is gaining not only traction but legitimacy through the offer of open online courses given by prestigious academic institutions around the world. Academics from Harvard,

Stanford, the Santa Fe Institute, MIT and many others institutions are giving online lectures in several subjects to millions of students. The access to these lectures is in most cases without any academic restriction and without any fee.

The offer of massive Open Online Courses (MOOCs) can be mostly found in dedicated online education platforms (e.g: Coursera), which offer an extensive curriculum of courses in a variety of subjects including mathematics, computer science, natural sciences and social sciences.

Giving massive and ubiquitous access to high level education provides a valuable contribution to the expansion of knowledge in society, especially in poorer regions of the world where a large part of the population does not have access to university level education**.**

Thanks to their content, quality and accessibility, MOOCs allows students from everywhere to gain significant knowledge and skills for personal and professional development. However, the question is, to what extent is it possible to get academic and professional recognition from a MOOC? The answer concerning academic accreditation is very limited. Some courses offer a certificate of completion on completion of all lessons. Such certificates provide in some cases a record of the student performance and proof of participating in the course but do not provide any university credits that can be transferred to a higher education qualification.

Granting academic recognition in a MOOC would imply a rigorous and complete evaluation of the student's achievements. Taking into account the large number of students enrolled in a course which often exceeds tens of thousands, it is impractical to evaluate all students using conventional hand marking. Thus in order to be scalable, assessment must be automated.

The absence of academic accreditation of MOOCs represents an issue of accessibility to widespread of certified higher education. We propose in this paper an alternative to overcome this barrier through the design of a model capable of handling large student populations while producing high quality assessment. We implemented the algorithms described in our model in a computerised assessment system and evaluated this in an experimental study performed with ninety 16 year-old students in a secondary school and also in a simulated scenario with larger populations of students.

2. Automated Assessment

The state-of-the-art in automated assessment includes various methods and computerised tools including multiple choice questions, and intelligent marking techniques (involving complex semantic analysis). However, none satisfy all the requirements of an assessment system able to cope with very large populations of students also able to guarantee the quality of evaluation required for *certified* higher education. Exclusive use of any single format for assessment is not recommended (American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education, 1999).

Multiple choice questions

One of the most commonly used assessment formats is *multiple choice testing* (Haladyna, 1999; McDougall, 1997). This mature and well known assessment method is recognized as an efficient way to evaluate a multidisciplinary range of knowledge (Haladyna, 1999). Once a multiple choice questionnaire has been created, it is possible to use computational tools for assessment of larger student populations. However, there are significant drawbacks:

- multiple-choice assessment is not suitable to evaluate all types of knowledge.
- testing student understanding beyond the superficial can require great ingenuity
- multiple choice questions can be time-consuming to prepare.
- questions can be ambiguous (but statistical methods make this easy to detect and correct).
- students are not able to demonstrate partial knowledge of the subject in question.
- students may gain marks by chance, spuriously indicating knowledge and learning
- completing many multiple choice questions can be repetitive, boring and demotivating.

Multiple choice assessment is a useful evaluation method when combined with another assessment formats (e.g. open questions) but using it exclusively is not enough to guarantee complete and accurate knowledge evaluation.

Intelligent and Short answer marking

This method refers to open questions with free text entry where the student has to write an answer usually no longer than four sentences. Here the student *constructs* rather selects answers (www.jisc.ac.uk, 2013). The aim is to obtain in the shortest possible text a definition or a set of facts. Short answers can then be evaluated by a computerised interpreter which may perform syntactic and semantic analyses of the text. The ability and accuracy of these interpreters may vary according to the technical complexity of the engines. Some of them use algorithmic manipulation of keywords while others use complex computational linguistic techniques to perform spell checking, syntax normalization, morphologic analysis, pronoun resolution and many others levels of linguistic processing (Butcher *et al*, 2010). The limited amount of words permitted by the interpreters does not allow students to create larger essays which may be necessary for more complex questions. Moreover, marking engines may not perform well for questions with an unpredictable range of valid answers *e.g.* 'what is freedom?' (www.jisc.ac.uk/media/documents/projects/shorttext.pdf (April 2013).

There is specialized commercial software for essay evaluation e.g. ETS, which use model essays (pre-graded scripts) as templates to compare aspects such as style, vocabulary, length, etc. However, It is important to keep in mind that computerised tools are not really "aware" of the items or subjects they evaluate.

Short answer marking can be very effective and can provide students with hints or further remedial material if at first they get the answer wrong, *e.g.*, an incorrect response can generate automatic feedback telling the student to read a particular page of text, view a video, or follow a link directing them to material enabling them to answer the question. In this way the student can be diagnosed as initially not knowing something, but can be rewarded for their learning by being given credit at second or subsequent attempts (Johnson *et al*, 2012, page 231)

Peer Assessment:

Peer Assessment, or Peer Marking, is a method in which assessment is done by students who are also part of the same evaluated population or who study the same subject to be evaluated. This is considered an innovative assessment strategy (McDowell and Mowl, 1996) which has demonstrated to improve student learning process. Peer assessment also has a positive impact on teachers or tutors by allowing them to use their time more efficiently, and also to get the results of student assessments in a shorter time (Sadler *et al*, 2006). Because of these benefits, and taking into account that the size of the assessed population is equal to the number of assessors, we see in this method a potential solution to the gap between MOOCs and certificated higher education. However, it is necessary to consider reliability and validity before going for a pure peer assessment approach.

Reliability of Peer Assessment: one of the first questions that arise when applying peer assessment is how accurate and reliable it is. Studies have shown that in some scenarios peer assessors tend to overrate their peers giving higher marks than the teacher would do (Falchikov, 2002; Roach, 1999). The opposite case has been also observed, where the tendency is to underrate the assessed peers (Penny and Grover, 1996 in Heywood, 2000, 387).

Peer marking is widely used in schools, often with a single student peer-marking another student's work, even though more than one peer marker is recommended (Bostock, 2000). This has the problem that for any piece of work some students will be penalised (or advantaged) if their single peer-marker gives a low (or high) score compared to the mark that would be given by a teacher. This problem may be alleviated by an argument that these discrepancies will be evened out over a number of pieces of work.

Multiple peer marking addresses this problem since it gives a distribution of marks. If the markers are highly competent their marks will be similar. If the markers are less competence there will be greater variance in the marks. Thus a high variability in the scores indicates that at least one of the peer markers is giving erroneous marks. Taking the average of the marks reduces the error compared to the worst marker and increases the error compared to the best marker. In the absence of knowing which markers are good or bad, taking the average is a compromise that reduces the frequency of large errors made by individuals at the expense of accepting an average error for the group.

In principle it is assumed that for any assessment there is a 'gold standard' of marking provided by the teacher. In other words, all teachers would give the same mark for a given piece of work, and the quality of peer marking is defined relative to this standard.

In practice the marks of experts can vary considerably. At the Open University in the UK many courses have thousands of students and their work is assessed by teams of markers. Quality assurance procedures involve systematic sampled second marking. While most second marking confirms the original mark, sometimes there are variations outside a limit of 15% which identifies weak markers and triggers third marking. Such considerations are outside the scope of this paper. In our experiment all the students answers were marked by a single teacher and our results are relative to the quality and consistency of his marking.

Improving the quality of Peer assessment

Taking into account that the abilities and performance of students for peer assessment are not homogeneous, it is necessary to find mechanisms to reduce the error resulting from the use of a group of assessors. One aspect where more emphasis has been put is on achieving feedback which is as objectively as possible. Several mechanisms have been proposed for eliciting Honest Feedback (Prelec, 2004; N. Miller, P. Resnick, and R. Zeckhauser, 2005), among which the well known Peer prediction method (Miller et al, 2005) where honest reporting between peer assessors is sustained by a Nash equilibrium. Another mechanism aimed at eliciting honest feedback has been proposed by Hütter, C.; Kimmerle, T. & Böhm, K. (2012) where an incentive mechanism is used to reward assessors according to their feedback quality. Increasing honest feedback in a peer assessor population may have a significant impact on the quality of scores. However, truthful or honest marking is only one of the factors affecting accurate marking. Highly deviated marking can occur due to other reasons, e.g. lack of knowledge about the evaluated question-answer or misunderstanding of the assessment criteria. Thus, having different factors affecting marking behaviour does require a mechanism able to identify and quantify in general terms the marking accuracy of each assessor. Probabilistic frameworks (Witkowski, J. & Parkes, D. C., 2012; Carpenter, 2008) using bayesian inference have been proposed as a strategy to predict the gold standard and classify assessors according to their score quality, by assigning weights to them.

Similarly to Witkowski et al (2012), we also implemented a weighting mechanism according to the calculated performance of each assessor. However, our system uses a more simplified analysis with a shorter set of equations. As a complementary feature, our system automatically assigns assessors to the assessed population based on a probabilistic distribution so that each assessed individual has an equal chance to be evaluated by a group of poor and good assessors.

3. The hypothesis being tested

Our approach involves giving peer markers a *reputation* score that discriminates the better markers from the weaker markers. Weak marking can be systematic, *e.g.* always too high or always too low. This can be seen as a calibration problem and in principle can be detected by its systematic nature. In contrasts some weak markers give *inconsistent* marks with work of the same quality being given different marks.

Our approach is based on the underlying hypothesis that peer markers giving similar scores are more likely to be good markers than peer markers giving deviant scores. On this basis the reputation of each marker can be recalculated after each session, as explained in Section 5.

Hypothesis

The aggregate mark given to an answer by triples of peer markers on the basis of dynamically calculated reputation and marker selection will be closer to the gold standard mark than the average score of triple marking with no use of reputation information.

4. Description of Experiment

To evaluate our approach, we conducted a study with 90 students and one teacher for a duration of eight weeks. One control group and one experimental group were created. Each group has a population of 45 students selected at random. Each student was assigned a unique code which was used to identify them within this study.

The students had four hours of classes per week. The first hour on the first day was used for teaching. The next day half of the second hour was used for students to complete new tests, and half an hour used to peer mark the answers from a previous session. On the third day the hour was used for teaching, and on the fourth day the fourth hour was again used for answering questions (half an hour) and peer-marking answers from the previous session.This in four weeks, it was possible to get seven samples of peer marking.

For the control group, each copy is randomly distributed to the control population (after removal of the student whose exam is being distributed since self assessment is out of the scope of this study). Once an exam has been peer reviewed, the score is calculated based on the average from the score given by the 3 peer assessors.

For the experimental group, the identification code of each student which presented the examination is given to our system. This generates a list for each exam script with the code of its 3 peer assessors. Once an exam has been peer reviewed, the score given by each peer assessor is entered in the system. This, calculates the exam score based on the weighted sum of the peers reputation.

During the first month of study, students have been assessed and applied peer assessment twice a week for a total of 560 marked exams equivalent to 1680 peer reviews. To analyse the accuracy of peer assessment, the teacher marked all exams as the ‘gold standard’.

We are comparing the differences between the scores calculated in the control and the experimental group with the scores given by the teacher.

The experiment was done in (Spanish) *linguistics* lesson corresponding to what is called *English Language* in the UK, which includes elements of spelling, grammar, comprehension, forming shorter abstracts of text (precise), and general concepts of literature. An example question is

> *Explique como se crea un guion tecnico y los pasos a seguir*
>
> which translates into English as
>
> *Explain how to create a technical script (in a theatre) and the steps to follow*.

The students’ answers to the questions had two sheets of paper. The first sheet had their name and other information. The second sheet included the question and their answer.

The preparation for each lesson included making three photocopies of the question-answer sheets for the previous lesson with hand written codes on each identifying the student and one of the three selected peer markers for this students’ answer. The section algorithm usually assigned different peer markers to each student’s answers.

In the second thirty minutes the students were given three scripts to peer mark, these being answers given by their peers in the previous lesson.

Students were motivated to give good answers since the gold standard assessment of the teacher is used to grade the course. They were motivated to peer assess other students well by being told that their performance would also count to the overall assessment the course.

The data reported in this paper correspond to the first four weeks of the experiment. In the first week the process of answering questions and peer marking was explained to the students. The test college has a high reputation in its city and its students are highly motivated with relatively good academic achievements, and the students all complied with the procedures and tried to give good peer assessments.

5. The experimental peer-assessment system:

The system has three components:

Management of students' data:

The system keeps track of the students' information by using a reputation score on a scale of 1 to 100. There are two types of reputation: one for the role as assessor and one for the assessed role. The assessor reputation score indicates the performance of the individual as a peer evaluator. The assessed score measures the performance of the individual as a student and is built from the result of one or more assessments. Both types of reputation scores are subject specific, allowing independent measures in both assessor and student roles for each specific subject. For instance, an individual may have a higher reputation as assessor in mathematics, a lower score as assessor in history but a higher score as a student in the same subjects.

A student can be associated to several subjects depending on the courses on which they are enrolled. This is supported in our system by a many-to-many relationship between the student and subject entities.

Selection of Peer assessors for marking

On completion of an examination, the system generates for each given answer a pool of n (3 in our system) assessors. This pool is selected from the assessors population which has been previously associated to the subject of the question/answer.

In order to achieve a fairer assessment, the pool of assessors is created by using a probabilistic selection based on the reputation score of each assessor. The objective is for every student to have the same probability to be assessed by at least one higher-reputation marker, one lower-reputation marker and one random marker when the pool size is of three or larger. In order to achieve this, each assessor is associated with a variable known as selection preference Sp which is initialized with the reputation value of the corresponding individual.

Once, an individual has been selected to be an assessor for a pool, their Sp variable is assigned a value equal to the average of the reputation of the entire population associated with the subject of the evaluated answer. This reduces the probability of the same individual being selected immediately after as high/low-reputation assessor, thus preventing an overload of work and also giving another individuals the opportunity to become assessors in a pool.

Recovering preference: each time a new pool of assessors is selected, the Sp variable of each individual i tends to reach the value of their corresponding current reputation Rp. Thus, the probability of being selected as high/low marker grows overtime to a value proportional to the individual's current reputation. The growth over time of Sp is indicated by a recovery rate Rcr (1.0 in our system). The recovery process of $Sp_i = \begin{cases} Sp_i + Rcr, & if \quad Sp_i < Rp_i \\ Sp_i - Rcr, & if \quad Sp_i > Rp_i \end{cases}$

The mechanisms for selection of random and high/low reputation assessors are defined as follows:

High Reputation assessor: for the selection of this marker the "Fitness proportionate selection" algorithm is used. This is also known as "Roulette wheel selection" and is widely used in genetic algorithms. This mechanism allows the reputation of each potential peer-assessor to be associated with their probability of selection. If fi is the reputation of individual i in the assessors population (associated to the question), with $f_i = reputation_i$, its probability to be selected in the markers pool is $p_i = \frac{f_i}{\sum_{j=1}^{N} f_j}$, where N is the number of individuals in the pool population **(1)**

When an individual i is selected, this is locked by the system and cannot be selected again for the same assessors pool, thus avoiding the issue of repeated assessors for the same student answer.

Low Reputation assessor: This individual is also selected based on the Fitness proportionate selection mechanism. However, the fitness fi is defined as the difference between the maximum scale value(100 in our system) and the reputation of individual i in the assessors population. $f_i = Max.ScaleValue - reputation_i$, with a probability to be selected given by:

$$p_i = \frac{f_i}{\sum_{j=1}^{N} MaxScaleValue - f_j} \quad \textbf{(2)}$$

This low reputation assessor also remains locked until the pool population has been totally formed.

Random assessor: this individual is randomly selected from the non-locked assessors in the population associated to the subject of the question.

Weighted Score Calculation:

The score of each answer is calculated on a scale from 1 to 100 based on the weighted sum of the scores given by the peers in the pool.

The weight w of the score s of an assessor i depends on their reputation r and is defined as follows: $w_i = \frac{r_i}{\sum_{j=1}^{M} r_j}$, where M is the number of individuals in the assessors-pool population. **(3)**

Thus, the final score S_c of the answer is expressed as: $S_c = \sum_{i=1}^{M} w_i s_i$ **(4)**

The following example illustrates this process. Suppose the following pool of three peers has been selected to grade a student answer.

	Peer 1	Peer 2	Peer 3
Assessor Reputation	45	22	71
Given score	70	68	40

The weights of each peer are:

$$Peer1_w = \frac{45}{45+22+71} = 0.33, \quad Peer2_w = \frac{22}{45+22+71} = 0.16, \quad Peer3_w = \frac{71}{45+22+71} = 0.51$$

Applying the formula (4) with the above calculated weights the final answer score is:

$$S_c = 0.33 \times 70 + 0.16 \times 68 + 0.51 \times 40 = 54.38$$

Adjusting the reputation score:

Once the score of an answer has been calculated, the next step is to adjust the reputation scores of each pool assessor. In order to do this, the difference $\Delta sc_i = |sc_i - Psc|$, between the score sc given by each assessor i and the calculated pool score Psc is taken as the criteria to determine if the given score falls within a given acceptance range Th_r (15 in our system):

$$Wi = \begin{cases} W_+, if \quad Th_r \geq \Delta sc_i \\ \\ W_-, if \quad Th_i < \Delta sc_i \end{cases} \quad \textbf{(5)}$$

where W_i represents the weight to be subtracted (w_-) or added (w_+) to the current reputation score of the assessor i .

In case Δsc_i falls within the range Th_r the positive weight w_+ is calculated as follows:

$$w_{i+} = \frac{1}{(\Delta sc_i + Th_r \times 0.5)} \, x \, (\Delta rp_i + 1) \Big/ 5 \times learning_p \quad \textbf{(6)}$$

where $\Delta rp_i = |Rp_i - B_{Rp}|$ is the difference between the reputation Rp of assessor i and the highest reputation value B_{Rp} in the pool. $learning_p$ is a constant which determines the increase factor of the reputation (we used values between 1 and 10).

In case Δsc_i falls outside the range Th_r the negative weight w_- is calculated as follows:

$$w_{i-} = -\frac{(\Delta sc_i + Th_r \times 0.5)}{Max_{sc}} \, x \, (\Delta rp_i + 1) \Big/ 10 \times learning_n \quad \textbf{(7)}$$

where Max_{sc} is the maximum scale value (100). $learning_n$ is a constant which determines the decrease factor of the reputation.

Knowing Wi from formulas 5, 6 and 7 the new reputation Rp of assessor i is adjusted as follows:

$$Rp_i = Rp_i + Wi \quad \textbf{(8)}$$

6. Results

Table 1 shows the experimental peer marking results for sessions 5, 6 and 7. For example, in the leftmost column (Session 5) Student-1 was assessed at 62.75% by the three peer markers and 60% by the teacher (2.75% difference) , while Student-2 was assessed by the three peer markers as 61.75 compared to 40% by the teacher (20% difference). The columns marked Error show the differences between the combined peer marker scores and the teacher's score. At the bottom of these columns is the mean error, 12.43% for Session 5, 10,31% for Session 6 and 9.3 % for Session 7.

Curso 11-5	Q5_peers	Q5_teach	Error	Q6_peers	Q6_teach	Error	Q7_peers	Q7_teach	Error
Student									
1.	62.75	60.00	2.75	67.18	80.00	12.82	57.27	50.00	7.27
2.	61.57	40.00	21.57	78.95	60.00	18.95	40.60	30.00	10.60
3.	63.75	20.00	43.75	47.35	40.00	7.35	50.26	70.00	19.74
4.	57.48	60.00	2.52	58.73	40.00	18.73	48.98	50.00	1.02
5.	57.20	40.00	17.20	63.60	60.00	3.60	35.30	50.00	14.70
6.	76.56	60.00	16.56	58.62	60.00	1.38	66.04	80.00	13.96
7.	77.20	60.00	17.20	79.78	80.00	0.22			
8.	97.51	70.00	27.51				62.70	40.00	22.70
9.									
10.	64.34	50.00	14.34				71.96	60.00	11.96
11.	79.85	40.00	39.85	67.53	60.00	7.53			
12.	89.86	70.00	19.86				85.23	90.00	4.77
13.	56.19	40.00	16.19	67.34	60.00	7.34	66.70	50.00	16.70
14.	69.33	90.00	20.67	75.74	80.00	4.26	43.39	30.00	13.39
15.	79.80	70.00	9.80				58.12	50.00	8.12
16.	68.82	60.00	8.82	61.87	60.00	1.87	42.71	80.00	37.29
17.	34.61	40.00	5.39				61.81	60.00	1.81
18.	66.90	60.00	6.90	61.68	60.00	1.68	50.00	50.00	0.00
Cu									
19.									
20.	90.00	90.00	0.00	71.27	80.00	8.73	72.42	70.00	2.42
21.	60.00	60.00	0.00	93.64	60.00	33.64	51.33	30.00	21.33
22.	83.84	80.00	3.84	66.95	60.00	6.95	52.22	40.00	12.22
23.	62.30	60.00	2.30	76.26	80.00	3.74	74.18	60.00	14.18
24.	76.98	70.00	6.98	64.23	80.00	15.77	58.46	60.00	1.54
25.	63.64	80.00	16.36	83.41	60.00	23.41	39.61	40.00	0.39
26.									
27.	79.54	60.00	19.54	47.04	60.00	12.96	76.79	70.00	6.79
28.	53.42	80.00	26.58	44.86	60.00	15.14	74.12	60.00	14.12
29.	60.89	60.00	0.89	68.42	60.00	8.42	63.24	60.00	3.24
30.	63.04	80.00	16.96	52.20	60.00	7.80	55.48	80.00	24.52
31.	81.90	80.00	1.90	80.24	60.00	20.24	63.33	60.00	3.33
32.	55.75	50.00	5.75	46.58	40.00	6.58	78.25	70.00	8.25
33.	45.85	50.00	4.15	83.77	60.00	23.77	69.11	70.00	0.89
34.	75.46	80.00	4.54	60.08	60.00	0.08	79.60	80.00	0.40
35.	69.79	60.00	9.79				63.10	60.00	3.10
36.	50.00	50.00	0.00	64.23	60.00	4.23	71.51	70.00	1.51
37.	90.89	60.00	30.89	56.32	60.00	3.68	64.87	60.00	4.87
38.	74.08	80.00	5.92	56.11	60.00	3.89	52.42	50.00	2.42
39.									
40.	89.69	90.00	0.31	65.36	90.00	24.64	73.44	80.00	6.56
			12.43			10.31			9.30

Table 1. Experimental peer marking results for sessions 5, 6 and 7.

	Session 1	Session2	Session 3	Session 4	Session 5	Session 6	Session 7
Control	42%	32%	23%	14%	15%	11%	13%
Experimental	31%	29%	20%	13%	12%	10%	9%

Table 2. Mean errors for experimental and control cohorts.

Our findings in Table 2 reinforce previous research that shows peer marking improves as students become more experienced (Hanrahan, S.J. & Isaacs, G. (2001)). These data, displayed in Figure 1, show that the mean difference between the peer marking and the teacher decrease over time. The first two sessions are shown against a grey background to indicate that the values for the reputation-based experimental marking are arbitrary since the initial reputations are not formed or stable.

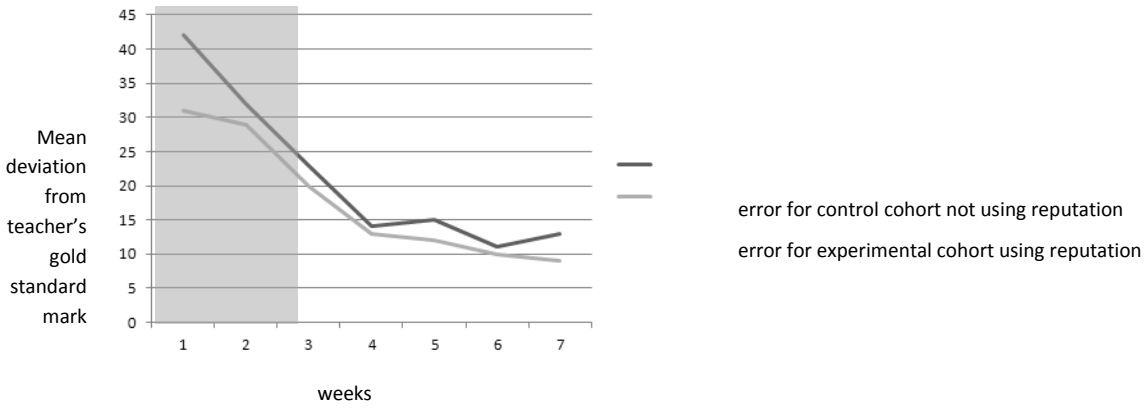

Figure 1. The mean errors for the reputation-based scores are lower than the average-based scores

Figure 1 gives a summary of our results and shows that the performance of the reputation-based system gives results about 2% - 5% closer to the gold standard than taking averages. Statistical tests (Table 3) suggest that this effect may not be significant for the first few sessions while the students are learning how to do peer marking, but by sessions 5, 6 and 7 they are unlikely to have occurred by chance. Thus the original hypothesis has been validated by this study, suggesting reputation-based systems may give a small benefit in performance compared to score-averaging systems.

Deskreptive Statistiken

		N	Mittelwert	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
						Lower Bound	Upper Bound		
Error	Control	102	13,43	9,63	,95	11,54	15,32	,00	60,00
	Experimental	100	10,73	9,67	,97	8,81	12,65	,00	43,75
	Total	202	12,09	9,72	,68	10,74	13,44	,00	60,00

ANOVA

		Sum of Squares	df	Mean Square	F	Significance
Error	Between Groups	367,22	1	367,22	3,94	,05
	Within Groups	18620,84	200	93,10		
	Total	18988,06	201			

Table 3. An analysis of variance suggests that the control and experimental mean errors are significantly difference for the last three sessions.

7. Discussion

These results suggest that our reputation-based peer marking can give aggregate results within 9% of the teachers' mark. In principle 9% difference is sufficiently close to conclude that peer marking is almost as good as the 'gold standard', especially since the teacher's marks are all multiples of ten. However care must be taken with such an interpretation.

Our goal is to create a scalable marking system that can be used as the basis of high quality certification. Whereas by the final session the mean difference reduced to 9%, inspection of the rightmost column in Table 1 shows a high proportion, 6 out of 34, of the differences were above 15%, a value considered unacceptable by the Open University that would trigger remarking. Thus as it stands our reputation-based peer marking system has not yet been demonstrated to deliver the high quality marking that we require.

Computer simulation experiments of our reputation-based system suggest that the errors continue to reduce after eight or more sessions. Since the experiment is continuing beyond the seven sessions reported here, we will soon be able to test this empirically.

Apart from these expected improvements we intend to make further developments to the system. For example, we could ask students to provide feedback on their assessment, saying whether or not they felt it was accurate. This will provide further data for the peer marker reputations. Also currently we do not use the time series of marker differences implicit in the data, and these may also help to discriminate strong and weak peer markers.

In summary we consider this experiment to give a strong indication that our reputation-based peer marking can be developed to give the robust high quality assessment required for accreditation.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. *Standards for educational and psychological testing*. Washington, DC: AERA, 1999

Bostock, S. (2000). *Computer-Assisted Assessments—Experiments in Three Courses.* From Learning Technology website, Keele University. (Last Accessed: 25 May 2013).

Boud, D. and Holmes, H. Self and peer marking in a large technical subject, 63-78 in Boud, D. Enhancing Learning through Self Assessment, London: Kogan Page, 1995.

Butcher, P. G., Jordan, S. E., 'A comparison of human and computer marking of short free text student responses', *Computers & Education*, Volume 55, Issue 2, Pages 489-499, September 2010

Carpenter. Multilevel bayesian models of categorical data annotation. Technical Report available at http://lingpipe-blog.com/lingpipe-white-papers/, 2008.

Falchikov, N. (2002). 'Unpacking' Peer Assessment'. In P. Schwartz & G. Webb (Eds.). *Assessment: Case Studies, Experience & Practice from Higher Education*. London: Kogan Page.

Haladyna, T. M. *Developing and validating multiple choice test items (2nd Ed.)*. Mahwah, NJ: Lawrence Erlbaum Associates, 1999.

Hanrahan, S.J. & Isaacs, G.. 'Assessing Self- and Peer-assessment: the Students' Views'. *Higher Education and Development,* Vol. 20, No. 1, pp. 53–70, 2001.

Heywood, J. 2000 Assessment in Higher Education, London: Jessica Kingsley Publishers

Hütter, C.; Kimmerle, T. & Böhm, K. (2012), Peer-Supervised Learning with Built-In Quality Control Based on Multiple-Choice Questions: A Case Study., *in* Carlo Giovannella; Demetrios G. Sampson & Ignacio Aedo, ed., 'ICALT' , IEEE, , pp. 453-457.

Johnson J., Buckingham Shum, S., Willis, A., Bishop, S., Zamenopoulos, T., Swithenby, S., MacKay, R., Merali, Y., Lorincz, A., Costea, C., Bourgine, P., Louçã, J., Kapenieks, A., Kelley, P., Caird, S., Bromley, J., R. Deakin Crick, R., Goldspink, C., Collet, P., Carbone, A., Helbing, D., 'The FuturICT Education Accelerator', *Eur. Phys. J. Special Topics* 214, 215–243, 2012.

McDougall, D. College faculty's use of objective tests: State-of-the-practice versus state-of-the-art. *Journal of Research and Development in Education*, 30(3), 183-193, 1997.

McDowell, L. and Mowl, G., Innovative assessment - its impact on students, 131-147 in Gibbs, G. (ed.) Improving student learning through assessment and evaluation, Oxford: The Oxford Centre for Staff Development, 1996.

Miller, N., P. Resnick, and R. Zeckhauser, "Eliciting Informative Feedback: The Peer-Prediction Method," *Management Science*, vol. 51, no. 9, pp. 1359–1373, 2005.

Prelec D., “A Bayesian Truth Serum for Subjective Data,” *Science*, vol. 306, no. 5695, p. 462, 2004.

Roach, P. (1999). ‘Using Peer Assessment and Self-Assessment for the First Time’. In *Assessment Matters in Higher Education*. Buckingham [England]; Philadelphia, PA: Society for Research into Higher Education & Open University Press. pp. 191–201.

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

Witkowski, J. & Parkes, D. C. (2012), Peer prediction without a common prior., *in* Boi Faltings; Kevin Leyton-Brown & Panos Ipeirotis, ed., 'ACM Conference on Electronic Commerce' , ACM, , pp. 964-981