JOURNAL OF INTERACTIVE MEDIA IN EDUCATION

Santamaría Lancho, M, et al. 2018. Using Semantic Technologies for Formative Assessment and Scoring in Large Courses and MOOCs. *Journal of Interactive Media in Education*, 2018(1): 12, pp. 1–10, DOI: https://doi.org/10.5334/jime.468

ARTICLE

Using Semantic Technologies for Formative Assessment and Scoring in Large Courses and MOOCs

Miguel Santamaría Lancho, Mauro Hernández, Ángeles Sánchez-Elvira Paniagua, José María Luzón Encabo and Guillermo de Jorge-Botana

Formative assessment and personalised feedback are commonly recognised as key factors both for improving students' performance and increasing their motivation and engagement (Gibbs and Simpson, 2005). Currently, in large and massive open online courses (MOOCs), technological solutions to give feedback are often limited to quizzes of different kinds. At present, one of our challenges is to provide feedback for open-ended questions through semantic technologies in a sustainable way.

To face such a challenge, our academic team decided to use a test based on latent semantic analysis (LSA) and chose an automatic assessment tool named G-Rubric. G-Rubric was developed by researchers at the Developmental and Educational Psychology Department of UNED (Spanish national distance education university). By using G-Rubric, automated formative and iterative feedback was provided to students for different types of open-ended questions (70–800 words). This feedback allowed students to improve their answers and writing skills, thus contributing both to a better grasp of concepts and to the building of knowledge.

In this paper, we present the promising results of our first experiences with UNED business degree students along three academic courses (2014–15, 2015–16 and 2016–17). These experiences show to what extent assessment software such as G-Rubric is mature enough to be used with students. It offers them enriched and personalised feedback that proved entirely satisfactory. Furthermore, G-Rubric could help to deal with the problems related to manual grading, even though our final goal is not to replace tutors by semantic tools, but to give support to tutors who are grading assignments.

Keywords: formative assessment; latent semantic analysis; open-ended question; automatic feedback; automated essays assessment; MOOC

Introduction

In recent years, we have seen an increasing demand for higher education and life-long-learning programmes. At the same time, budgets of public universities have been cut, at least in Spain. To respond to this demand a growing supply of online courses and new modalities, such as MOOCs, have been put in place. A greater demand and fewer resources mean that quizzes tended to become the main assessment tool. This resulted in poorer feedback and a lack of personalisation of the learning process.

Furthermore, our students, as users of technologies, expect a quick and iterative feedback (Kiili, 2005; Oblinger, 2004). They love learning by trial and error. Nevertheless fast and iterative feedback in a high-population learning context can only be provided by the use of technology. However, feedback based on technology still offers limited solutions.

As economic history teachers, our learning outcomes include not only knowledge referred to this subject, but also soft skills like analysis, critical thinking, and so forth. We realise that quizzes, especially multiple-choice questions, have serious shortcomings for assessing learning outcomes. To assess critical thinking and so on we would need to use a mix of different kinds of assessment activities, such as multiple-choice questions, short open-ended questions about concepts or processes, and written comments about texts, maps, graphs or statistical data. Therefore, the challenge we faced was how to give quick and iterative feedback for open-ended questions in a sustainable way. That was the main reason for using semantic technologies.

The role of feedback on performance improvement and student engagement

According to different researchers (Black and William, 1998), feedback is the most powerful single factor to make a difference to a student's achievements (Hattie and Timperley, 2007).

The economies of scale provided by online courses pose a challenge for assessment. Furthermore, providing comments on assignments remains a significant component of a lecturer's workload. For this reason, the feedback to an individual student has declined significantly

UNED, ES

Corresponding author: Miguel Santamaría Lancho (msantamaria@cee.uned.es)

as the number of enrolled students increased (Gibbs and Simpson, 2005).

The effects of formative assessment are well known. The inclusion of continuous assessment in the teaching and learning process seems to produce clear and important effects on the quality and quantity of learning, as shown by several research papers. For example, Carrillo de la Peña and Pérez (2012) carried out a study on Spanish university students. They compared the academic results over three years of two groups of students in the subject of physiological psychology. They found that the students of an experimental group that received, in addition to a numerical grade, continuous and personalised qualitative information (feedback) on their performance (formative assessment):

- passed the course in a greater proportion;
- obtained higher number of distinctions for excellence;
- showed a greater degree of satisfaction, as compared to the students of the control group that was evaluated in a traditional way with a final evaluation with a numerical grade.

These results are in line with what is shown in other studies (see Larsen, Butler, and Roediger, 2008).

Feedback helps to reactivate prior knowledge, focus attention on the subject and encourage active learning. Moreover, it gives the student the opportunity to practise skills and consolidate learning. Feedback allows the student to monitor their progress and develop self-evaluation and critical thinking (Crooks, 1988; Gibbs and Simpson, 2005). The effects of formative assessment in promoting self-regulated learning are well known (Nicol and Macfarlane-Dick, 2006).

Giving personalised feedback on large online courses is a challenge

Computers that deliver automatic assessment are nowadays an essential part of a virtual learning environment (VLE). VLEs mostly display multiple-choice questions but they can include a wider range of assessment types, including true-false, fill-in-the-blank, matching, or numerical manipulation. Assessment can be embedded in multimedia materials. More sophisticated tools are conceived every day, but all share one common trait, which is to provide an objective assessment. Some are able to provide different styles of feedback, but to do so these tools have to deal with a limited, predictable range of possible answers.

The current crop of MOOCs has brought to the fore the need for a more nuanced and flexible means to assess complex exercises with thousands of participants on one course. Quick assessment is needed but qualified instructors cannot provide it (Sánchez-Vera and Prendes-Espinosa, 2015). The need for precise grading and rich feedback has been lately addressed with peer-assessment based on detailed, well-structured rubrics. These break assessment operations into smaller pieces that can be entrusted to the very students taking the course. However, although this may sound promising, there is still plenty of room for automated assessment devices based on semantic technologies.

Semantic technologies are helping with the challenge

Automated essays assessment (AEA) has a long history. The development of technologies such as word processing and the Internet encouraged the improvement of AEA systems. Also, the advances experienced since the 1990's in natural language processing facilitated the analysis of morphology (word structure), syntax (sentence structure) and semantics (meaning). The analysis of content was carried out through lists of keywords, synonyms and the frequency with which specific terms appeared (Shermis and Burstein, 2003).

In the last two decades three AEA products were developed. Two of them, MY Access and Criterion, provided numerical scores and some evaluative feedback that was comparable to that produced by humans. The scores were obtained by comparing the essays with equivalent human-scored essays. The third AEA, the Intelligent Essay Assessor, made use of latent semantic analysis, in which the semantic meaning of a given text was compared to a broader corpus of textual information (Landauer et al., 1997). This system focused on evaluating conceptual content and paid less attention to text style and grammar structure. This approach will require fewer human-scored essays because it relies on semantic analysis rather than statistical comparisons with previously scored essays (Warschauer and Ware, 2006). According to research, AEA scoring tends to be accurate. Some AEA systems have become embedded within automated writing evaluation systems that assign scores and give feedback on errors, and may include instructional scaffolding and learning management tools (Roscoe and McNamara, 2013).

Paradoxically, there has been not much research in distance education institutions, despite the fact that large numbers of students should have made these tools an obvious choice (Jorge-Botana et al., 2015). Concerns about plagiarism and identity-control issues have presumably hindered progress in this context, along with logistical matters related to access to computers at the examination place. At present, MOOCs represent, indeed, an open field for the implementation of this kind of application.

What we present here is a pilot test of an LSA-based automated free-text assessment system named G-Rubric. It was designed by a team of researchers at UNED's Department of Developmental and Educational Psychology and tested on a group of first-year college students of economic history at the same university. G-Rubric has proved able to provide fast and precise numeric assessment of free-text short answers (75–800 words). The system also gave enriched, personalised feedback that allowed students to improve their answers through a series of successive attempts. Our test has been limited to formative assessment. The reliability and student satisfaction seem promising enough to consider applying G-Rubric to the summative assessment (grading). The first steps towards this aim will be mentioned in this paper.

How G-Rubric works

Latent semantic analysis (LSA) is based on the concept of vector space models. This means using linear algebra for allocating lexical units in an n-dimensional vector space. LSA is a set of different procedures by which a textual cor-

pus, usually lemmatised and curated, is transformed into a semantic space. As a first phase, this corpus is expressed into an occurrence matrix, which usually has its terms as rows and paragraphs as columns. A second phase is applied to this matrix, which smooths the asymmetries in word frequencies. The third phase has made LSA famous, which is to apply to this matrix a dimension reduction technique using singular value decomposition (SVD). SVD provides a suitable space in which words and texts are represented in a few but relevant latent (with no meaning) dimensions. This space is handy for representing expert and student answers and calculating similarities between them. The more similarity among student-expert answers, the higher the score. Recently some authors have developed a promising procedure called inbuilt-rubric (Olmos et al., 2014), which transforms the k first latent dimensions of the original space into non-latent dimensions. The k first dimensions no longer reflect latent knowledge but rather conceptual axes spread from relevant words of the academic topics. This can allow conceptual feedback. The scores of the student answers in such k first dimensions indicate if the relevant concepts of the rubric are present in his answer. This technique has reached satisfactory results in real contexts (Olmos et al., 2016). We are describing just the procedure used by G-Rubric, the AEA used in our research.

For the economic history teachers involved in this study, the essential characteristic of G-Rubric is its ability to provide the student with three different kinds of feedback. First, a numeric grade for content, second an additional numerical grade for writing quality, and third, a detailed graphic feedback that plots the score in each conceptual axis of the rubric.

Teachers had to provide three different types of input to develop the task.

1. *Texts to build the corpus*. This is the raw material of the course – handbooks, reference texts and so forth – that will compound the corpus. For the development of our experience, we built a corpus on economic history using six different world economic history textbooks, all of them written in Spanish and published in the last 20 years.
2. *A specific semantic space*. To generate the space from the corpus, a specific program called Gallito Studio was used (Jorge-Botana et al., 2013). Then, the resultant space, including inbuilt-rubric space, is uploaded to a specific application programming interface called GallitoAPI (www.gallitoapi.net). The web interface for the assessment of free-text was named G-Rubric. We will usually refer to the whole system as G-Rubric, although it is important to remember that managing the multi-vector semantic space, which is the heart of the system, is done with GallitoAPI.
3. *Several learning activities*. These inputs are based on short, open-ended questions. For this task, we used Gallito Studio. To accompany each activity we prepared a canon answer, or 'golden text', with which students' answers would be compared. A series of conceptual axes (three to five per question) were prepared that were composed of a series of keywords that depict different regions of the semantic field the answer should cover. This golden text and the axes were tested with actual students' answers taken from past exams to check the accuracy of the numerical grade and the visual feedback drawn from conceptual axes.

Several iterations were needed to generate acceptable G-Rubric activities for a trial with students. This material allows the system to provide students with a numerical grade and graphical feedback. G-Rubric allowed users to select questions, submit answers, and receive feedback almost immediately.

To help to understand how G-Rubric works, we offer a sample of the activities our students did.

Once the student registers at the G-Rubric website and chooses the activity they can write down or paste an answer. We opted for a learning activity on the definition of the concept of mercantilism.

First attempt by the student:

> *"Mercantilism is a set of ideas and policies deployed in early modern Europe (16th, 17th and 18th centuries) aimed at strengthening the State through economic power, and specially focused on trade-balance surpluses and accumulation of precious metals (bullionism)."*

After submitting an answer the student receives the feedback seen on the left side of **Figure 1**. After examining this feedback, the student can review the prior answer

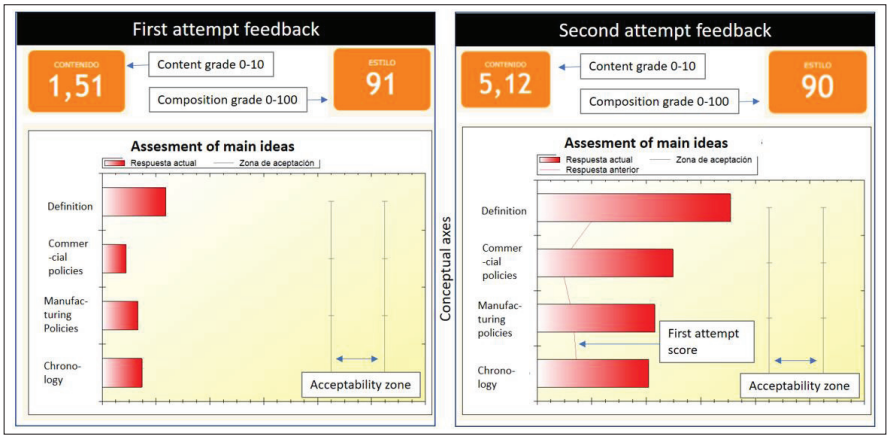

Figure 1: Screenshots of G-Rubric's feedback screen.

and make a new attempt adding, for instance, some new ideas about mercantilist policies (bold text in the second attempt).

Second attempt by the student:

> *"Mercantilism is a set of ideas and policies deployed in early modern Europe (16th, 17th and 18th centuries) aimed at strengthening the State through economic power, and specially focused on trade-balance surpluses and accumulation of precious metals (bullionism).* ***Amongst mercantilist policies, some outstanding, i.e., those focused on attaining surpluses in trade balance through tariff protection, prohibition of exports of gold, silver and raw materials, the creation of chartered trade companies, navigation acts, and commercial monopolies****".*

New feedback was produced, as seen on the right side of **Figure 1**. Then, the student could try again using the new feedback to improve the answer.

Experience of using G-Rubric to Give Formative Assessment (2015–2016)

It is important to point out that the trials carried out in 2015 and 2016 were focused on providing formative assessment. Our goal was to promote deep learning through iterative feedback, and not just provide grading scores. G-Rubric offers formative assessment because it allows as many attempts as lecturers desire and gives students immediate rich feedback. All trials have been conducted with first-year business administration degree students.

First experience with G-Rubric (May 2015)

With this first experience we had two goals. The first was to determine the efficacy of G-Rubric to promote learning, and the second was to establish its reliability for marking assignments. To develop this first trial, we asked for volunteers among our students and offered them a little reward (adding 0.25 point to their final mark of 10). We got 132 volunteers, we split them randomly into three groups and established different conditions for each group. Group 1 received rich feedback, both numerical and graphical, and had six attempts to answer. Group 2 received poorer feedback (only numerical), and had six attempts. Group 3 was the control group and received poorer feedback, and was only allowed one attempt to answer.

The students taking part in the trial would answer five short, open questions, similar to those they would find in their final exam. For each question, the student got a set of instructions referring to the number of words they were expected to write, how to use the tool to answer, and guidance for using the feedback they would get. Groups 1 and 2 could use their six attempts to improve their answers according to the received feedback. Each student could decide how many attempts they would make. The difference between the worst and the best mark achieved in each of the activities was used to measure the learning improvement of each student. Also, a questionnaire was used to measure student's agreement with the grades assigned by G-Rubric to their answers.

As can be seen in **Table 1**, in general, there was a learning improvement for group 1 as well as for group 2. Also, the difference between highest and lowest grades was higher for group 1, which received rich feedback. However, there was no significant difference in grades between the three groups in the final question, which was designed to measure learning derived from the use of G-Rubric.

Student agreement with the grades received was quite good, as seen in **Figure 2**.

Second experience with G-Rubric (April–May 2016)

The goal of the second trial was to improve the design of G-Rubric questions to foster learning and increase student satisfaction. To carry out this second trial, we increased from five to seven the number of objects (new questions)

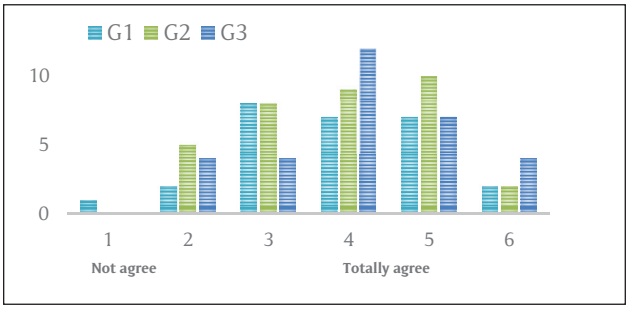

Figure 2: Student agreement with the grades received.

Table 1: Improved learning indicators.

		Average grade G-Rubric (/10)			Difference between max-min grade		
	Item	**G1**	**G2**	**G3**	**G1**	**G2**	**G3(1)**
1	Demographics regimes	6.9	6.5	6.4	0.52	0.69	0
2	Consequences of the neolithic revolution	6.5	5.9	5.6	1.06	0.95	0
3	Medieval European agrarian economies	6.2	7.4	5.5	1.10	0.78	0
4	Mercantilism	7.7	7.5	6.6	1.95	1.15	0
5	(Final) colonial commerce (2)	6.2	6.3	6.1	0	0	0

(1) G3 was the control group and had only one attempt per item, therefore no option to improve. (2) For Item 5 only one attempt was allowed.

Table 2: Student scores by item.

	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Average
Lowest grade	6.19	4.51	5.10	4.95	5.39	5.02	5.66	5.19
Highest grade	7.41	5.53	6.12	5.81	6.74	6.29	6.78	6.31
Difference in points	1.22	1.02	1.02	0.86	1.34	1.27	1.12	1.12
Difference %	19.7	22.6	19.9	17.4	24.9	25.4	19.8	21.7

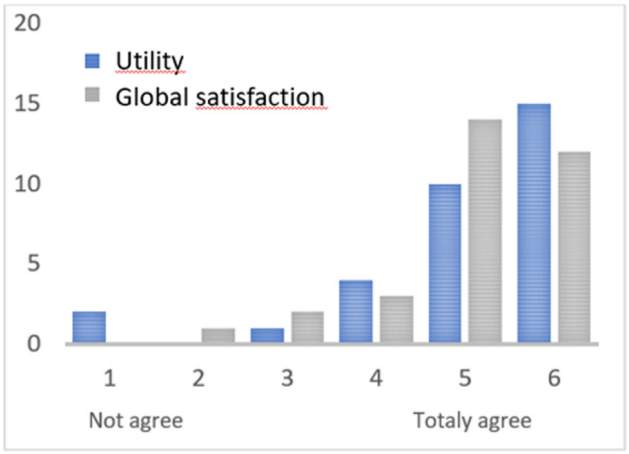

Figure 3: Utility and satisfaction with G-Rubric.

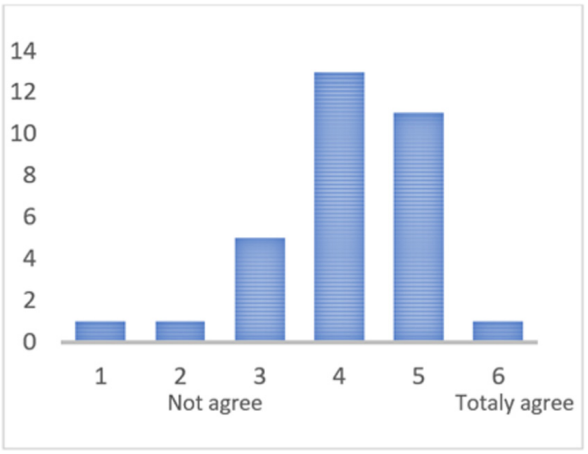

Figure 4: Student agreement with grade obtained.

offered to the students. To increase the number of volunteers the reward increased from 0.25 to 1 point. This reward was associated with the number of attempts made, rather than with the grades produced by G-Rubric because after the first experience we discovered that learning improved after several attempts at answering.

According to data in **Table 2**, the average grades obtained were satisfactory. It should be taken into account that we had recommended to the students that they should review the textbook before producing an answer. As we can see, after students accessed the feedback, they were able, on average, to improve their marks in the following attempts.

Notice that the best students were able to obtain high scores, close to those of the golden text produced by the lecturer and used by the system as a reference.

To analyse learning improvement (learning) we used the difference between the lowest and highest grade obtained by students. **Table 2** shows the difference by item, both in absolute terms and as a percentage. A 21.6% improvement average could be considered as remarkable, given that only three attempts were allowed. The different degree of improvement by item could be a consequence of various factors such as the quality of the question design and difficulty of the question.

We would like to point out some results of the satisfaction questionnaire that students completed after their experience (**Figures 3** and **4**).

According to **Figure 3**, students considered the experience useful and believed that they were better prepared for the final exam. Global satisfaction was also high.

Regarding the student's agreement with the grades received it could be said that it was quite satisfactory, shown in **Figure 4**.

To summarise, given the results of these trials, G-Rubric could be considered as a useful tool to provide accurate and formative feedback for short, open-ended questions (Santamaria Lancho et al., 2017). Our next goal was to analyse how this software could give support to the tutors.

Semantic Technologies can Help Tutors to Mark Assignments

After gaining experience with formative assessment, a new experiment was prepared to evaluate how G-Rubric could support tutor marking.

The tutor marking of open-ended questions presents two main problems, which has been described in the related literature. The problems are inter-examiner variability and intra-examiner reliability (Wakeford, 2003). In our opinion, a semantic tool such as G-Rubric could help to avoid both of them.

Are humans reliable when marking open-ended questions?

Open-ended questions are valid because they allow tutors to assess learning outcomes. In fact, higher-level outcomes such as analytical skill, construction of arguments and precise writing can be more efficiently assessed with open-ended questions. Because of this, many tutors have a preference for this kind of assessment, even if they are more time-intensive and harder to grade. The problems, however, arise when it comes to variability and reliability.

Whereas fairness is a qualitative measurement, reliability can be mathematically measured. For instance, we can establish the existence of inconsistencies across examiners (poor inter-examiner reliability) if there is one standard exam, and the assessment is more or less randomly assigned. The extent of theses inconsistencies can be even

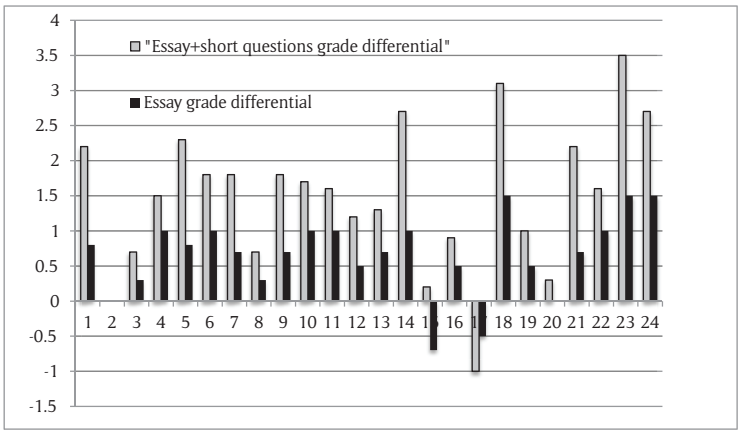

Figure 5: Differential in grades for doubly-graded exams (June 2012) (data from economic history final exams from Barcelona-CUXAM regional centre, June 2012).

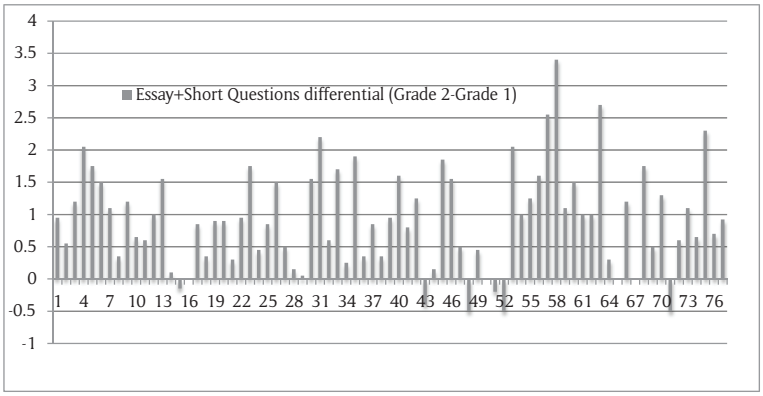

Figure 6: Differential in grades for doubly-assessed exams (data from economic history final exams from Valencia-Alzira regional centre, June 2013).

more precisely established if different graders independently grade the same exams. An analogous procedure could be followed to determine intra-examiner reliability, by carrying out two successive assessments of the same exams at two different moments in time. Statistical analysis of the marks awarded to various questions can also determine their adequacy if they show robust and consistent deviation from averages. All these measurements, however, require time and resources, which most teachers prefer to allocate to other tasks. In addition results might prove what many students suspect and most of the teachers refuse to consider, which is that a severe reliability problem exists in existing procedures.

Closely related to reliability is the issue of fairness. Students often view grading and assessment of free-text questions as subjective, and they think it is unfair or unreliable (Valenti et al., 2003). Students suspect a high degree of variability depending on who, when or how the essay is assessed, with the possibility of personal inclination or arbitrariness of the tutor. In contrast, automated test assessment is perceived as more objective, providing a high degree of consistency over time and space and excluding any bias derived from human intervention.

One of the advantages of having numerous students to grade and different tutors grading them, is that we can compare trends. That is the case of UNED, especially concerning thousands of first-year students. As mentioned before, economic history is a subject corresponding to junior courses in both economics and business degrees. Usually, high numbers of exams require several teachers grading the same questions, and we can thus measure inter-examiner variability.

A double-grading happened accidentally in two different years, when two teachers on the same academic team independently and unknowingly graded the same exams. Thanks to this event, we could evaluate our inter-examiner assessment. The results indicated significant differences in grading (**Figure 5**). The differential was in an average of 1.5 points over 8 (the total grade for the free-text questions of the exam), and 0.65 points over 3 in the case of essay questions only (a text or graph commentary). Furthermore, we could observe that there was a visible pattern, with one generous examiner systematically assigning higher grades than his colleague (with two exceptions) and only one instance of coincidence of marks (a 0 mark for a very poor answer). This difference would make the student's final grade differ substantially, meaning in 9 of 24 exams (37.5%) that the student would or would not obtain a passing grade.

In June 2013, there was another allocation mistake that led to double-grading of another 76 exams from Valencia UNED centre (**Figure 6**). Again, differentials in marks showed up, even though more limited (0.9 points on average over a total of 8 points at stake), and again with a clear upward bias in the case of instructor 2 (lenient) as opposed to stricter instructor 1. There were six

occurrences of a higher grade awarded by instructor 1, but always with differences under 0.5 points (**Figure 6**). Even if we could consider grading differentials below a 0.5 point threshold as acceptable, there would still be 49 out of 75 exams (almost two thirds) with substantial differences in grades depending on the instructor, going up in some cases as far as 3.4 points. Again, in many of these instances (16) the differential would affect the passing or not passing the exam.

In both cases, this variability happened despite efforts made to promote homogeneous grading with a shared and agreed rubric, including correct answers and grading criteria (although not entirely disaggregated). Despite differences in grades assigned by the examiners, high correlations were found between the marks corresponding to both the global score and the short questions. A lower correlation was found in the text comment scores (**Table 3**). This was probably due to the higher complexity of scoring a text commentary over concise questions.

These differences, even when detected in such a small number of instances, appear to justify claims of subjectivity or unfairness held by students, mainly when this evidence arises not from systematic testing, but from post facto analysis of accidental occurrences.

Claims of lack of reliability of short-answer, free-text questions provide a solid motive for the development of automated tools. They could be used either alone or with a human assessment to produce more reliable evaluation.

The use of G-Rubric could cope simultaneously with both problems, namely inter-examiner variability and intra-examiner reliability. We hypothesised that semantic tools such as G-Rubric would help to deal with the problems. Our goal was not to replace tutors with semantic tools, but instead to give support to tutors in grading student's assignments.

Table 3: Correlation between grades assigned by two examiners in 2012 and 2013.

	2012	2013
n	20	76
global grade	0.82	0.88
sort questions	0.85	0.87
text commentary	0.70	0.67

Reliability and validity of human and LSA-based assessment of essays

According to previous research, LSA-based assessment agrees with human graders' scores as much as different human graders' scores agree among them. Human and computer scores correlate around 0.80 to 0.85, with 40–60% perfect agreement and 90–100% adjacent agreement (human and computer scores within one point). See the summaries in Cohen, Ben-Simon & Hovav, 2003. This agreement does not depend on whether scoring was based on one golden answer or a sample of previously scored assignments. Even more, LSA-based evaluations of student assignments predicted results in a final exam (Seifried et al., 2012).

First attempts at giving support to the tutor with grading (October 2017–January 2018)

Taking previous evidence into account, the present study aimed at evaluating G-Rubric's capabilities to provide support to the tutors in grading. In our first approach to this goal, the objective was to compare the marks provided by tutors and G-Rubric to a certain number of TMAs. Economic history students had to write two TMAs per semester. They had to comment on a text, graph or statistical table. The resulting TMAs had an average of 800 words. Using G-Rubrics to mark this kind of assignments was a new challenge because previously it had only been used to mark short, simple questions with a more delimited answer.

To carry out this experience, the teaching team in charge of the economic history first-year subject established the following arrangements:

- a fragment of Adam Smith's *The Wealth of Nations* was selected and students were asked to comment;
- a rubric was built to minimise inter-examiner variability for as many as 37 tutors;
- a G-Rubric object, similar to those described above, was designed and its axes were aligned with the rubric used by tutors to mark the assignments.

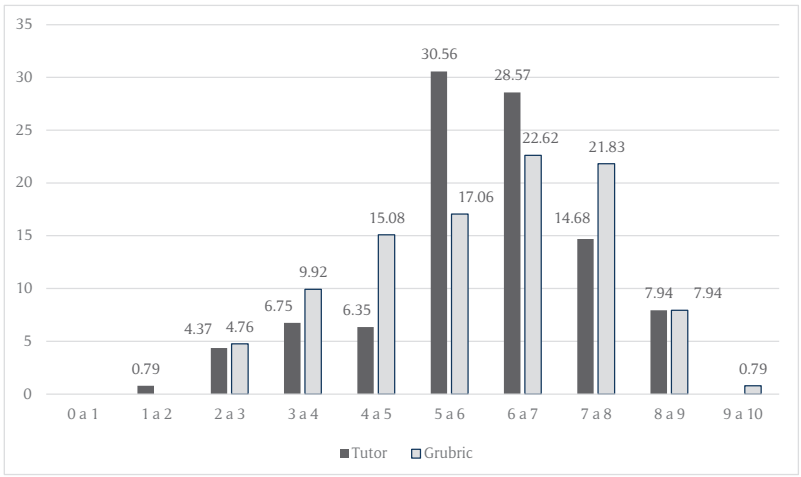

Figure 7: Percentage of TMAs graded by tutors and G-Rubric by grade range.

Once everything was ready, the students sent their TMA in a digital format through the Moodle platform as usual. Then the tutors graded these assignments using the rubric. Next, the teaching team used G-Rubric to grade the students' TMA again. A total of 252 TMAs were marked by both a tutor and G-Rubric. The 252 TMAs were graded by 37 tutors, although only eight of them marked more than 10 TMAs.

Results

The following is a summary of our results.

Overall, essay mean scores rated by tutors 5,95 (SD = 1.45) is very similar to those rated by G-Rubric 5,92 (SD = 1,61) (**Table 4**).

Grades distributions: analysis of frequencies

According to the marks, grade distribution was structured in five different meaning ranges. **Figure 7** displays the percentages for both TMAs graded by tutors and G-Rubric. Both distributions showed some differences. G-Rubric marks were more homogeneously distributed in comparison with the higher concentration of the tutors' marks in the ranges between five and seven.

Correlational analyses between tutors and G-Rubric marks and mean differences analysis

Pearson correlations between G-Rubric and tutor marks (all tutors globally considered) yielded a large effect size (.549**[1]). An independent sample *t* test yielded no significant differences between the means of tutors and G-Rubric marks, $t(251)$, $p = .720$, ns.

Additionally, a new variable was created (*mark difference*) subtracting the tutor's mark from G-Rubric one. The mean difference between both marks was –0.03 (SD = 1.46, Min = –3.79, Max = 4.23) for the total number of students.

Analysis of the homogeneity of G-Rubric and tutors' marks

The previous analyses were conducted without taking into account that 37 different tutors had marked TMAs, therefore introducing a potential source of variability among tutoring groups. For a closer analysis of the inter-group homogeneity of G-Rubric and tutor marks, Kruskal-Wallis non-parametric analyses were conducted, being *tutoring group* the independent variable and *marks* (tutors and G-Rubric gradings) and *mark difference,* the dependent variables of the study. The number of students per tutoring group varied between 1 and 48.

The results yielded by this analysis are shown in **Table 5**. As can be appreciated, *tutor mark* presented a significant inter-group variability, as well as *mark difference*. On the contrary, *G-Rubric marks* did not differ significantly between these same tutorial groups, proving, thus, its higher level of homogeneity.

This same analysis was conducted again, taking only into account the eight tutoring groups with 10 or more students. Results confirmed the previous ones, being even clearer in **Table 6**.

Finally, a *t*-test for dependent samples was conducted to analyse mean differences between tutors and G-Rubric marks for each one of the eight tutoring groups with more than 10 students. Only two groups yielded significant differences between tutor and G-Rubric marks (group 15, $t(32) = 2.69$, $p = .011$, and group 32, $t(12) = 2.19$, $p = .051$) in the direction of higher tutor average marks (**Figure 8**).

Table 4: Main descriptives of tutors and G-Rubric marks (N = 252).

	M	SD	Min	Max
Tutor's marks	5.95	1.45	1.55	8.54
G-Rubric marks	5.92	1.61	2.13	9.20

Table 5: Kruskal-Wallis analyses for the evaluation of marks homogeneity between the 37 tutoring groups.

	Tutor mark	G-Rubric mark	Mark difference
Chi-cuadrado	69.14	47.21	74.49
gl	36	36	36
p	.001	.100	.000

Conclusion – What Next?

Some conclusions can be drawn from our experience.

- Automated assessment software such as G-Rubric is currently mature enough and gave satisfactory results regarding accuracy. Results regarding students' satisfaction are also encouraging.
- The costs and complications of designing objects (questions) for G-Rubric are completely affordable for even small teams of teachers, with moderate learning costs concerning familiarisation with the system.
- Learning to work with G-Rubric was also easy for students. However, mastering the system and understanding visual feedback could take them a little longer than expected.
- The trial's results seem to show that interacting with G-Rubric can improve learning by giving detailed feedback in some ways:
  - encourages devoting more time to the task
  - increases 'earnings' in the quality of answers
  - increases motivation to work on activities
  - helps students to achieve better final answers – therefore it may soon become a viable tool for formative assessment.
- Comparing tutors' marks with G-Rubric grades, a

Table 6: Kruskal-Wallis analyses for the evaluation of marks homogeneity between the eight tutoring groups with 10 or more students.

	Tutor mark	G-Rubric mark	Mark difference
Chi-cuadrado	27.671	5.248	12.506
gl	7	7	7
p	.000	.630	.085

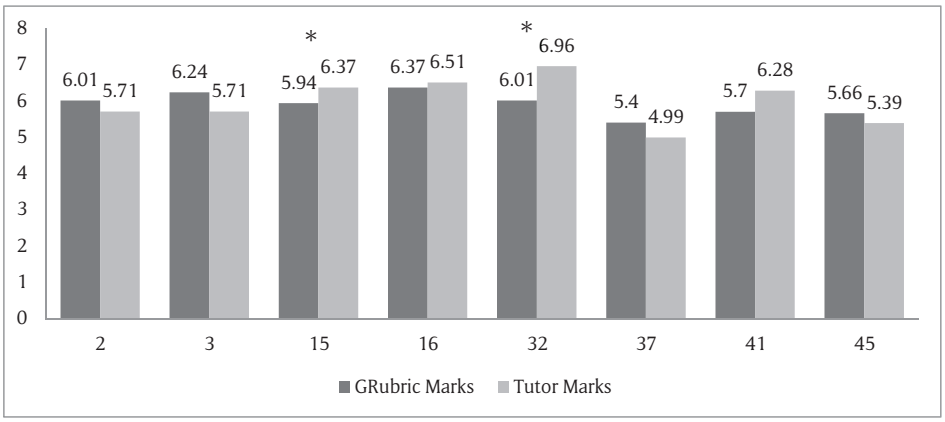

Figure 8: Means of tutors and G-Rubric marks and results of *t*-tests between marks for the eight tutorial groups with more than 10 students.
* $p < .05$.

remarkable correlation and no significant differences between the means has been found. Additionally, tutor scores presented a significant inter-group variability. On the contrary, G-Rubric marks did not differ significantly between these same tutorial groups, proving its higher levels of homogeneity. For these reasons, we think that G-Rubric could also be a useful tool to deal with the problems that characterise summative assessment such as inter-examiner variability and inter-examiner reliability. To do that, we propose that essays will be graded first using G-Rubric to allow the tutor to mark again, either validating or modifying the G-Rubric grades.

- MOOCs and large courses present a need for assessment that can provide immediate, rich and personalised feedback. To assess beyond multiple-choice questions or peer-assessed excercises, automated tools offer great promise that can be easily percieved when using G-Rubric.
- The main advantage of G-Rubric has to do with the feedback provided and the speed, precision and stability of assessment. Using open-ended questions as a part of formative assessment offers opportunities to promote learning through a series of iterations of writing-feedback-rewriting that enriches learning, both content and soft skills.
- The experience with G-Rubric indicates that the tool is able to assess and provide feedback to short-answer questions. But in addition it can handle long essays, such as those explained in our 2017–2018 experience.

Note

[1] ** The correlation is significant at the 0.01 level (bilateral).

Competing Interests

The authors have no competing interests to declare.

References

Black, P and **Wiliam, D.** 1998. 'Assessment and classroom learning'. *Assessment in Education: principles, policy & practice*, 5(1): 7–74 [online]. Available at: http://www.tandfonline.com/doi/abs/10.1080/0969595980050102 (Accessed 15 September 2017).

Carrillo-de-la-Peña, MT and **Perez, J.** 2012. 'Continuous assessment improved academic achievement and satisfaction of psychology students in Spain'. *Teaching of Psychology*, 39(1): 45–47. DOI: https://doi.org/10.1177/0098628311430312

Cohen, Y, Ben-Simon, A and **Hovav, M.** 2003. 'The Effect of Specific Language Features on the Complexity of Systems for Automated Essay Scoring' [online]. Available at: http://www.academia.edu/download/43641495/The_effect_of_specific_language_features20160311-18840-vbuncx.pdf (Accessed 30 August 2017).

Crooks, TJ. 1988. 'The impact of classroom evaluation practices on students'. *Review of educational research*, 58(4): 438–481 [online]. Available at: http://journals.sagepub.com/doi/abs/10.3102/00346543058004438 (Accessed 15 September 2017).

Gibbs, G and **Simpson, C.** 2005. 'Conditions under which assessment supports students' learning'. *Learning and teaching in higher education*, 1: 3–31 [online]. Available at: http://eprints.glos.ac.uk/3609/ (Accessed 14 September 2017).

Hattie, J and **Timperley, H.** 2007. 'The power of feedback'. *Review of Educational Research*, 77(1): 81–112. March 2007. DOI: https://doi.org/10.3102/003465430298487

Jorge-Botana, G, Luzón, JM, Gómez-Veiga, I and **Martín-Cordero, JI.** 2015. 'Automated LSA assessment of summaries in distance education: some variables to be considered'. *Journal of Educational Computing Research,* 52: 341–364. Available at: https://www.researchgate.net/profile/Guillermo_Jorge-Botana/publication/274252139_Automated_LSA_Assessment_of_Summaries_in_Distance_Education/links/551996f80cf244e9a458484e/Automated-LSA-Assessment-of-Summaries-in-Distance-Education.pdf. http://miau.gau.hu/miau/225/Annual_2017_Jonkoping_Proceedings.pdf (Accessed 15 December 2017).

Jorge-Botana, G, Olmos, R and **Barroso, A.** 2013. 'Gallito 2.0: A natural language processing tool to support research on discourse'. In: *Proceedings of the 13th Annual Meeting of the Society for Text and*

Discourse [online]. Available at: http://elsemantico.es/Documentos/Gallito2_Valencia_new.pdf (Accessed 4 November 2016).

Kiili, K. 2005. 'Digital game-based learning: Towards an experiential gaming model'. *The Internet and higher education*, 8(1): 13–24. DOI: https://doi.org/10.1016/j.iheduc.2004.12.001

Landauer, TK and **Dumais, ST.** 1997. 'A solution to Plato's problem: the latent semantic analysis theory of acquisition, induction, and representation of knowledge'. *Psychological review*, 104(2): 211 [online]. Available at: http://psycnet.apa.org/journals/rev/104/2/211/ (Accessed 15 September 2017).

Larsen, DP, Butler, AC and **Roediger, HL.** 2008. 'Test enhanced learning in medical education'. *Medical Education*, 42: 959–966. DOI: https://doi.org/10.1111/j.1365-2923.2008.03124.x

Nicol, DJ and **Macfarlane-Dick, D.** 2006. 'Formative assessment and self-regulated learning: a model and seven principles of good feedback practice'. *Studies in Higher Education*, 31(2): 199–218. DOI: https://doi.org/10.1080/03075070600572090

Oblinger, D. 2004. 'The next generation of educational engagement'. *Journal of Interactive Media in Education*, 1 [online]. (Accessed 4 November 2016). DOI: https://doi.org/10.5334/2004-8-oblinger

Olmos, R, et al. 2014. 'Transforming selected concepts into dimensions in latent semantic analysis'. *Discourse Processes*, 51(5–6): 494–510 [online]. Available at: http://www.tandfonline.com/doi/abs/10.1080/0163853X.2014.913416 (Accessed 4 November 2016).

Olmos, R, et al. 2016. 'Transforming LSA space dimensions into a rubric for an automatic assessment and feedback system'. *Information Processing & Management*, 52: 359–373 [online]. Available at: http://elsemantico.es/Documentos/transforming2.pdf (Accessed 15 December 2017).

Roscoe, RD and **McNamara, DS.** 2013. 'Writing pal: feasibility of an intelligent writing strategy tutor in the high school classroom'. *Journal of Educational Psychology*, 105: 1010. DOI: https://doi.org/10.1037/a0032340

Sánchez-Vera, MM and **Prendes-Espinosa, MP.** 2015. 'Beyond objective testing and peer assessment: alternative ways of assessment in MOOCs'. *RUSC Universities and Knowledge Society Journal*, 12(1): 119–130. (Accessed 15 December 2017). DOI: https://doi.org/10.7238/rusc.v12i1.2262

Santamaria Lancho, M, Hernández, M, Luzón Encabo, JM and **Jorge-Botana, G.** 2017. 'Writing to learn with automated feedback through (LSA) latent semantic analysis: experiences dealing with diversity in large online courses'. In: Volungeviciene, A and Szücs, A (eds.), *Diversity Matters!*, 331–339. Jönköping, 13–16 June 2017 [online]. Available at: http://miau.gau.hu/miau/225/Annual_2017_Jonkoping_Proceedings.pdf (Accessed 15 December 2017).

Seifried, E, et al. 2012. 'On the reliability and validity of human and LSA-based evaluations of complex student-authored texts'. *Journal of Educational Computing Research*, 47(1): 67–92 [online]. Available at: http://jec.sagepub.com/content/47/1/67.short (Accessed 17 January 2017).

Shermis, MD and **Burstein, J.** (eds.) 2003. *Automated Essay Scoring: A Cross-disciplinary Perspective.* Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Valenti, S, Neri, F and **Cucchiarelli, A.** 2003. 'An overview of current research on automated essay grading'. *Journal of Information Technology Education Research*, 2: 319–330. DOI: https://doi.org/10.28945/331

Wakeford, R. 2003. 'Principles of student assessment'. In: Fry, H, Ketteridge, S and Marshall, S (eds.), *A handbook for teaching & learning in higher education*, 42–61. Second edition, Kogan-Page: Sterling, VA.

Warschauer, M and **Ware, P.** 2006. 'Automated writing evaluation: Defining the classroom research agenda'. *Language teaching research*, 10(2): 157–180. Available at: http://ltr.sagepub.com/content/10/2/157.short (Accessed 5 November 2016).

How to cite this article: Santamaría Lancho, M, Hernández, M, Sánchez-Elvira Paniagua, Á, Luzón Encabo, JM and de Jorge-Botana, G. 2018. Using Semantic Technologies for Formative Assessment and Scoring in Large Courses and MOOCs. *Journal of Interactive Media in Education*, 2018(1): 12, pp. 1–10, DOI: https://doi.org/10.5334/jime.468

Submitted: 19 December 2017 **Accepted:** 10 May 2018 **Published:** 15 August 2018