Active Learning through Collaborative Knowledge Building using an Automatic Free-Text Scoring System in a B-learning environment

Diana Pérez-Marín, Raquel Hijón-Neira and Liliana Santacruz
Computer Science Faculty, Universidad Rey Juan Carlos, Madrid, Spain

Abstract

According to active learning, students should be responsible of their own learning. Automatic free-text scoring allows teachers to provide open-ended questions with their correct answers to a computer system, so when students answer the questions, they get immediate feedback. However, teachers are usually overloaded with many tasks, and they may not have time to create the questions with the correct answers. Therefore, in this paper, we provide teachers, for the first time, with a procedure that combines active and social pedagogic theories, free-text scoring technologies and blended learning, so that students create the questions and correct answers, and get more involved in courses that could be found boring as they are unrelated to the main topic of their degree. To test the procedure, we have asked a group of 124 Pre-Primary and Primary Education University students to follow it in a Computer Science course. Out of the 124 students, 41 fulfilled all the tasks requested. Our hypothesis was that those students would be able to increase their academic performance and levels of engagement compared to the rest of the students. The results gathered provide statistic evidence to support that hypothesis.

Keywords

Free-text scoring; Active Learning; Collaborative Knowledge Building; Blended Learning; Computer Assisted Education

1. Introduction

According to constructivism, knowledge can be seen as socially constructed (Vygotsky, 1978), **all learning is active in a certain sense** (Nunan, 1990; Simons, 1997; Brown, 2000), and students are able to construct their knowledge when they encounter problems that they have to solve in active situations (Good and Brophy, 1994; Lesgold, 2004).

Moreover, it has been studied that learning is most effective and productive if it is goal-oriented and self-regulated (Nicol & Macfarlane-Dick, 2006). Learning is also regarded as more fun when it occurs in interaction and collaboration with others based on constructive processes of knowledge and skill acquisition. Group activity increases discussion, experimentation, enthusiasm, and participation.

In the last decades, how technologies can be used for education has also been studied. Computers have been more and more used as support for several pedagogic theories. The new learning environments try to involve the students in activities, and change the role of the students from passive recipients of information (students only digest or memorize facts) to active participants in the construction of their knowledge (students are engaged in their learning experience). The role of the teachers has evolved from being the only owners of the knowledge to become facilitators in powerful learning environments (Gagne, 1985; Ashton-Hay, 2006).

However, not all students accept the proposed role change so easily. They may fear failure because of the new more active and collaborative approach to learning. In particular, it has been claimed that the main threat that active learning and collaborative knowledge building pedagogic theories face when trying to be applied in **powerful learning environments** is human. Some University students may prefer just to take notes during the lesson and later study on their own at home without having to meet or talk to other students. To memorize information and to avoid collaboration with others could be seen easier for them (Brown, 2000).

We have also seen that assessment in many powerful learning environments keeps being limited to tests, fill-in-the blank exercises or matching activities. In those cases, students do not have the opportunity to express themselves and use their own words. **Free-text scoring** allows students to answer open-ended questions. Teachers usually have to provide a set of questions to a computer system with their correct answers. The system usually compares the students' answers to the correct answers provided by the teachers with the core idea that the more similar the student's answer is to the correct answers, the better it is, and thus the higher it should be scored. The system can provide immediate feedback for each question as a numerical score, a comment or both (Mitchell et al. 2002; Attali & Burstein, 2006; Pérez-Marín et al. 2009).

In this case, teachers may have problems with this approach, because they are usually overloaded with many tasks, and they may not have time to create the open-ended questions with the correct answers to feed the free-text scoring system. Even, when they create the questions with the answers, the language that teachers use may be totally different to the language used by the students preventing the computer system to provide a good automatic evaluation of the questions.

To sum up, as shown in Figure 1, there are pedagogic theories such as active learning and collaborative knowledge building that seem to bring benefits to students, powerful learning environments that can apply these theories, and free-text scoring systems that automatically evaluate students' answers so that they can self-regulate their learning. In this paper, we answer the call for more research into how to combine the active and social pedagogic theories, with the new powerful learning environments and free-text scoring technologies as requested by researchers such as Hoang and Arch-Int (2013).

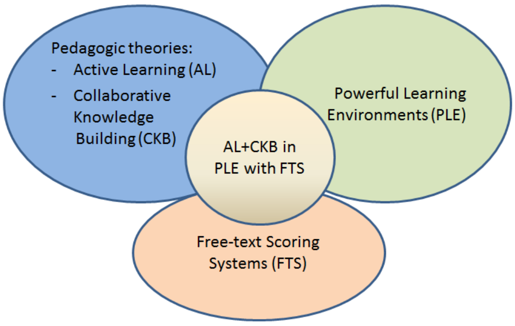

Pedagogic theories:
- Active Learning (AL)
- Collaborative Knowledge Building (CKB)
Powerful Learning Environments (PLE)
AL+CKB in PLE with FTS
Free-text Scoring Systems (FTS)

Figure 1. Overview of the fields related to the proposal of the paper to provide active learning with collaborative knowledge building in Powerful Learning Environments with free-text scoring

It is our proposal to use a **Blended Learning (b-learning) methodology**, i.e. to combine face-to-face instruction with the use of computers for education (Graham, 2005). We have asked students to use a free-text scoring system to create their own questions and answers for each lesson of the course following the principles of active learning and collaborative knowledge building as proposed by Good and Brophy (1994).

We have considered that the participation in the experience should be voluntary because of the human fear to change. It is our intention that students start to see the possibilities of new teaching strategies as opportunities as opposed to impositions. The results of the experience will not affect their score in the course (i.e. we will not give a percentage of the final score to those students). On the other hand, we also consider that those students should be rewarded somehow. For instance, by providing them with a certificate indicating their commitment to the experience.

Moreover, not all the questions and answers proposed by the groups were published. Students had to vote in their groups which questions and correct answers they like the most to be published in the free-text scoring system. Later, they have to answer and pass those questions. This is to prevent forgetting what they have learned, trying to cut the forgetting curve (Ebbinghaus, 1913).

It is our **hypothesis (H) that students more involved in the procedure would increase both their academic performance and levels of engagement**. To test that hypothesis, in the 2012/2013 academic year, we asked a test group of 124 Pre-Primary and Primary Education University students to become the creators of the questions and their correct answers in groups (usually 5 students per group) in the Willow free-text scoring system (Pérez-Marín et al. 2006).

From them, 41 students were involved in all the tasks (creating questions, correct answers, voting, answering and passing the questions) fulfilling all requirements during the course. We call those students GIS (group of involved students). 83 students were involved in some tasks of the procedure because they did not get so involved with the course. We call those students GAS (group of average students). Finally, 38 students were also evaluated as the control group. They did not know about the procedure, so we could compare their results with the results achieved by the test group (both GIS and GAS).

The results gathered provide significant evidence to support H. In particular, GIS increased their score in a post-test taken at the end of the course up to 8.5 (SD=1.25) from a 4.9 (SD=1.5) average score in a pre-test taken at the beginning of the course, which is extremely statistically significant according to an unpaired t-test with a two-tailed p value less than 0.0001. This improvement is statistically significant greater than the improvement of the control group from 5.24 (SD=1.55) up to 7.84 (SD=1.57), and it is also statistically significant greater than the improvement of GAS from 5.51 (SD=1.76) up to 8.30 (SD=1.37). On the other hand, the improvement of GAS is not statistically significant compared to the control group.

GIS were also able to increase their final score in the course with a 9.04 value (in 0-no knowledge up to 10-maximum knowledge) scale (SD=0.84), which according to an unpaired t-test is extremely statistically significant with a two-tailed p value equals to 0.23. The average final score of GAS was 7.1 (SD=3.3) and the average final score of the control group was 4.9 (SD=3.7). The improvement of the GIS average final score is statistically significant compared to GAS and the control group.

It has also been registered that 4 out of the 5 most voted groups had at least 2 GIS **increasing their level of engagement** in keeping posting questions and correct answers in the following lessons of the course. GIS also showed more interest at class, and higher levels of motivation towards a course that it is not usually their favourite given that, in general, Pre-Primary and Primary Education University students do not enjoy having a course of technology which is quite different from the rest of their courses.

This study pretends to serve as a foundation for teachers interested in how to combine active learning, collaborative knowledge building, and educational technology so that they can provide their students with the tools to improve their academic performance and levels of engagement in courses that students may find boring per se.

The novelty of the approach lies in the lack of research on combining active learning, collaborative knowledge building and free-text scoring. To our knowledge, **no papers have been written on how these methodologies can be used to provide students with tools to improve their behaviour** towards courses that they may find boring as they are unrelated to the main topic of their degree (i.e. in our case, Pre-Primary and Primary Education University students do not tend to like Computer Science courses).

The paper is organized as follows: Section 2 reviews the main pedagogic theories in which the proposal is based, and overviews free-text scoring; Section 3 describes our proposal; Section 4 details how the proposal was applied in an experimental study and provides the results gathered; and, finally the paper ends with a discussion of the main ideas of this work in Section 5.

2. State-of-the-art review

The state of the art is organized into two main sections: Section 2.1 reviews the main pedagogic theories in which the proposal is based (we present them separated as they are found in the literature), and Section 2.2 overviews the free-text scoring field. It is not our intention to provide a comprehensive review of the state of the art of free-text scoring, which can be found in Pérez-Marín (2009).

2.1. Pedagogic theories

According to a constructivism pedagogic view, Good and Brophy (1994) claimed that there are four aspects necessary for learning:

1) **Learners construct their own meaning.** Students are not passive receptacles. Students have tried to understand the information that comes to them, and manipulate it to insert the new knowledge into their own belief system.

2) **New learning builds on prior knowledge**. When students manipulate the information that comes to them to integrate into their own belief system they need to find connections with their previous information, and accept or discard old information.

3) **Learning is enhanced by social interaction.** Students in social settings learn better because they have the opportunity to compare and share their ideas with others. When trying to solve conflicts with the belief system of their colleagues, learning happens.

4) **Meaningful learning develops through "authentic" tasks.** Learning activities must be chosen so that they simulate as much as possible real life activities.

These four aspects are highlighted here because they are mentioned by many other authors such as Simons (1997), Brown (2000), Lesgold (2004) and Cooperstein & Kocevar-Weidinger (2004). All of them consider that learning is active, and that students are able to construct their knowledge when they encounter problems that they have to solve in active situations.

According to those authors, lessons should start with a problem or a question that students need to solve. That way, they are able to construct their knowledge and learn more efficiently, even having fun (Cooperstein & Kocevar-Weidinger, 2004). During this process, students can also be helped with a scaffolding or a supportive framework (Vygostky, 1978), which guides them through a series of small steps. The idea is that the instructor motivates the students to keep asking, without giving the answers, but supporting the students' search (Gagne, 1985). Moreover, the instructor should help students to prevent that they forget what they learn as soon as they leave the class (Ebbinghaus, 1913).

However, according to Cooperstein & Kocevar-Weidinger (2004), this learning method requires a great deal of time, and according to Ashton-Hay (2006), some students may have fear of change, and ask for traditional lesson in which the teacher talks and they just listen and take notes, without having to follow the active learning approach. Exams can also be a factor that inhibits active learning for these students.

Learning should not only be active, but it should also be social (Vygostky, 1978; Good & Brophy, 1994; Moskaliuk et al. 2012; Bloom et al. 2013). In particular, the study of Bloom et al. (2013) serves as a foundation to study social dimension along with knowledge and cognitive process. They found that enhancing community building supports learning in various knowledge levels and improve the students' cognitive processes.

2.2. Free-text scoring

Computer-Assisted Assessment (CAA) is the research field that studies how to use computers to automatically evaluate student work (Pérez-Marín et al. 2009). Nowadays, CAA has many possibilities of application, such as scoring the students' assignments (summative assessment), producing feedback to discover if the students have learned what the teacher intended (formative assessment), and evaluating assessment effectiveness (Blayney & Freeman, 2003).

According to most authors, the main goal of CAA is not to substitute teachers, but **to support them in their tutoring task** (Mason & Grove-Stephenson, 2002). Therefore, CAA is typically formative although it can also be used with summative purposes. Most of the initial work in CAA was devoted to designing closed questions, such as fill-in-the-blank or Multi-Choice Questions (MCQ). However, many authors agree that MCQs do not really measure the higher cognitive skills (Birenbaum et al. 1992; Foltz et al. 1999; Parsons et al. 2003; Mcgrath 2003; Mitchell et al. 2003). Although there has always been hard critics about the idea of a computer grading human essays, the advances in Natural Language Processing (NLP) and Machine Learning techniques, the popularization of e-learning environments, the lack of time to give students appropriate feedback (despite the general assumption of its importance) and the conviction that MCQs cannot be the only computer-based assessment method have promoted the **development of free-text scoring**.

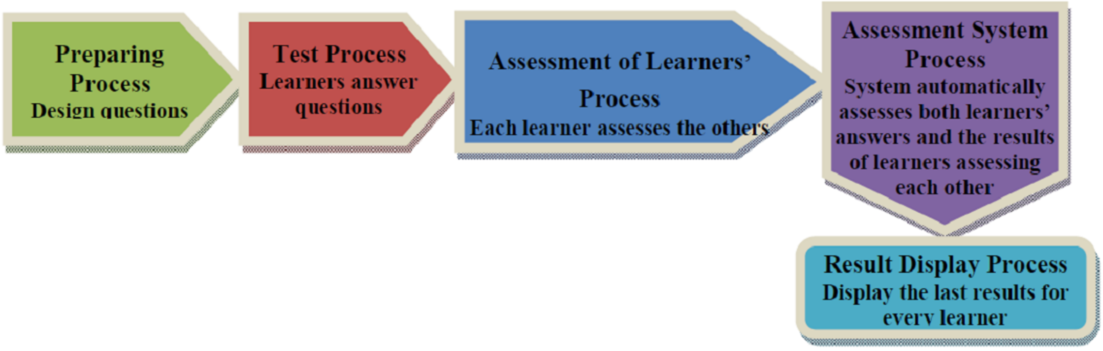
Automatic assessment of students' free-text answers can be seen as including two different sub-types: automatic assessment of short answers and automatic assessment of essays. Sometimes the same tool can evaluate both kinds but, in general, the boundaries between the two tasks are clear and most CAA tools only evaluate either essays or short answers. Some systems that will be considered out of the scope of this review are semi-automated computer-based essay marking systems (Marshall and Barron, 1987), systems that assess the student ability to summarize (Kintsch et al. 2000), and systems to improve the student writing skills (Wiemer-Hastings and Graesser, 2000).

There are several approaches to perform the automatic evaluation, most of which **compare the student's answer against some reference** (ideal answer) or template. In order to grade the technical writing quality, one traditional approach is to look for direct features in the text, such as word number or word lengths, and to use them to infer more abstract measures such as variety, fluency or quality (Page, 1966; Christie, 2003).

Some free-text scoring systems are AEA (Kakkonen et al. 2005), based on the use of variations of Latent Semantic Analysis; Apex Assessor (Dessus et al. 2000), based on the use of Latent Semantic Analysis; ATM (Callear et al. 2001), based on the use of Information Extraction; Automark (Mitchell et al. 2002), based on the use of Information Extraction; Auto-marking (Sukkarieh et al. 2003), based on the combined use of NLP and pattern matching; BETSY (Rudner and Liang, 2002), based on the use of statistical techniques; CarmelTC (Rosé et al. 2003) based on the use of Machine learning; EGAL (Datar et al. 2004) based on the use of Natural Language Processing; E-rater (Burstein et al. 1998) also based on the use of Natural Language Processing; IEA (Foltz et al. 1999) based on the use of Latent Semantic Analysis; IEMS (Ming et al. 2000), based on the combined used of pattern matching and clustering; Jess (Ishioka & Kameda, 2004) based on the use of pattern matching; and, Willow (Pérez-Marín et al. 2006) based on the combination of statistic techniques and shallow NLP.

Feedback is also an important factor in automatic free-text scoring as an instructional tool and as a motivational factor (Bruner et al., 1956; Good and Brophy, 1994). Free-text scoring systems usually provide feedback after each question evaluated, as a score, a comment, or both; and, in some cases, they even provide general feedback to have an overview of the evolution of the student in the course (Pérez-Marín, 2007).

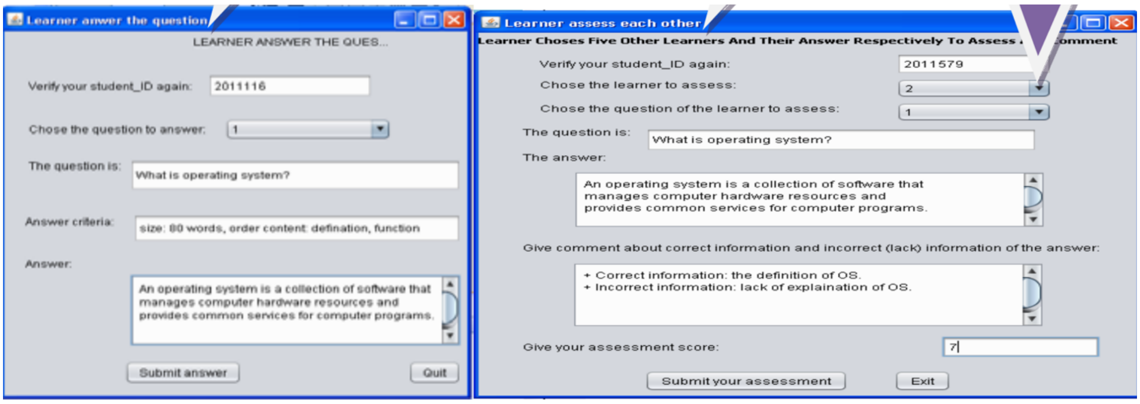
Recently, **Hoang and Arch-Int (2013) highlighted the social aspect of free-text scoring**. They proposed a new assessment method using open-ended questions with feedback enhancing collaboration and interaction of learners at the same time. They claimed that free-text scoring not only must evolve to keep improving the technical quality of the evaluation, but it should also take into account more human and social factors, evolving towards a multi-dimensional assessment that correlates with learning in present-day social networks. Figure 2 shows the multi-dimensional free-text scoring process proposed by Hoang and Arch-Int (2013) with active learning and collaborative knowledge building.



Preparing Process
Design questions
Test Process
Learners answer questions
Assessment of Learners' Process
Each learner assesses the others
Assessment System Process
System automatically assesses both learners' answers and the results of learners assessing each other
Result Display Process
Display the last results for every learner

Figure 2. Multi-dimensional free-text scoring process (source: Hoang and Arc-Int, 2013)

As can be seen, and as indicated by the pedagogic theories, the process starts with questions that students must answer. The difference with traditional free-text scoring relies on the third step, it is not just limited to the core idea of making a comparison between the student answer and the teachers' answers, but now each learner assesses other learner answers, and the system takes into account both learners' answers and the results of learners assessing each other, displaying the results for every learner as feedback.



Learner anwer the question
LEARNER ANSWER THE QUES...
Submit answer
Quit
Learner assess each other
Learner Choses Five Other Learners And Their Answer Respectively To Assess . omment
Submit your assessment
Exit

Verify your student_ID again: 2011116

Chose the question to answer: 1

The question is: What is operating system?

Answer criteria: size: 80 words, order content: defination, function

Answer:

An operating system is a collection of software that manages computer hardware resources and provides common services for computer programs.

Verify your student_ID again: 2011579

Chose the learner to assess: 2

Chose the question of the learner to assess: 1

The question is: What is operating system?

The answer:

An operating system is a collection of software that manages computer hardware resources and provides common services for computer programs.

Give comment about correct information and incorrect (lack) information of the answer:

+ Correct information: the definition of OS.
+ Incorrect information: lack of explaination of OS.

Give your assessment score: 7

Figure 3. Screenshot of a multi-dimensional free-text scoring system (source: Hoang and Arc-Int, 2013)

Figure 3 shows a sample screenshot of the system based on this process. As can be seen, on the left, each student has a text area in which s/he answers the question chosen. The evolution is on the right, in which the student chooses the question of another learner to assess it, and the system will take into account scores provided by other students and the automatic score to complete the assessment of each student's answer. This method has increased by 3.6% the accuracy of the free-text scoring, and has enhanced the interaction and collaboration of the students in the virtual learning environments. They are now aware of other students' answers as well as their own answers, and they have the opportunity of gaining more knowledge from the comments of other students.

3. Proposal

According to Good and Brophy (1994) constructivism pedagogic view, combining active learning and collaborative knowledge building pedagogic theories, we propose to **ask students to collaboratively create the questions** with their correct answers for the lesson of a course in a free-text scoring system in groups, **vote** the questions for each lesson that they consider the best, and **answer** in the free-text scoring system the four most voted questions of each lesson. Figure 4 shows a diagram of the procedure proposed.

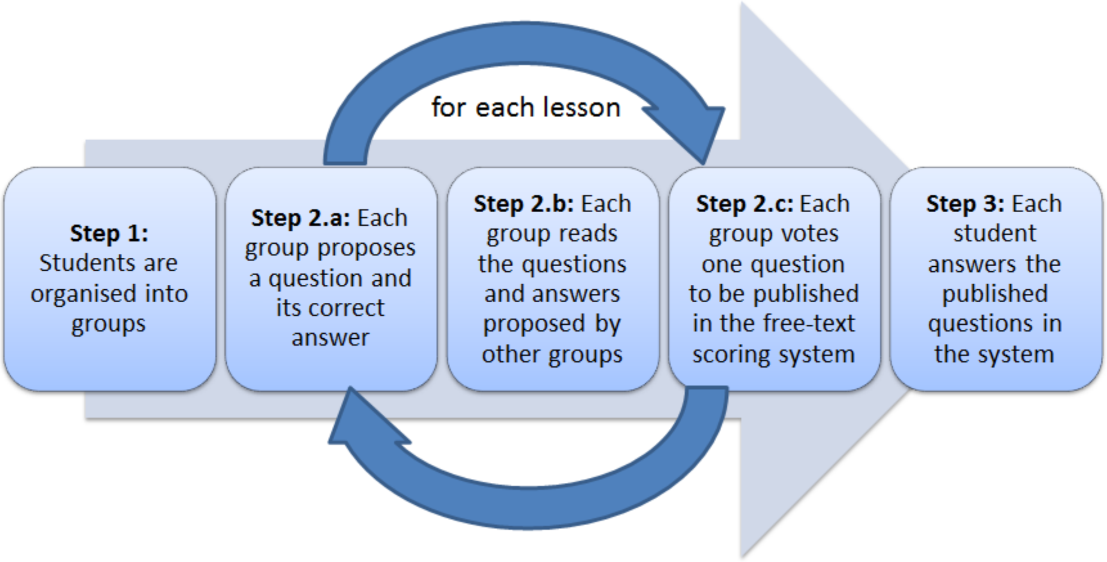

for each lesson
Step 1: Students are organised into groups
Step 2.a: Each group proposes a question and its correct answer
Step 2.b: Each group reads the questions and answers proposed by other groups
Step 2.c: Each group votes one question to be published in the free-text scoring system
Step 3: Each student answers the published questions in the system

Figure 4. Procedure proposed to combine active learning and collaborative knowledge building in a free-text scoring system and improve students' learning efficiency and engagement

The proposal follows the principle that **learners construct their own meaning.** As can be seen in *step 2.a* students have to create their own questions and answers. The proposal also follows the principle that **new learning builds on prior knowledge**. The questions that students have to create are associated to lessons of a course. Each lesson of the course is based on the previous lessons. It means that students need the knowledge of previous lessons to create new questions and answers for more advanced lessons.

The principle that **learning is enhanced by social interaction** is also taken into account**.** Students cannot create the questions alone; they belong to groups created in *step 1* and they have to create the questions in their groups. Moreover, students must read the questions and answers of other groups (*step 2.b*) and agree which questions vote for each lesson (*step 2.c*). Students also have to talk and agree the questions and answers that they would like to vote of other students. **It follows the multi-dimensional free-text scoring approach described by Hoang & Arch-Int** (2013). The main difference is that, in our proposal, students do not evaluate on their own the answers provided by other students, but they have to talk in their groups and vote the proposed questions and answers proposed by other groups.

Regarding the principle that **meaningful learning develops through "authentic" tasks,** the proposal does not limit the type of questions and answers provided by the students. They should be adapted according to the type of course, so that they are as much "authentic" as possible in each case.

The role of the teacher in this proposal is to serve as a guide. The reason is based on the need of some scaffolding or supportive framework highlighted by Vygostky (1978) and Gagne (1985). That way, students are helped and motivated to keep asking. Moreover, teachers are responsible to publish the most voted questions in the free-text scoring tool, and the procedure finishes with an individual study of these questions carried out by each student (*step 3*) to try reducing the forgetting curve as much as possible (Ebbinghaus, 1913).

It is also important to highlight that the procedure should be applied in a Blended Learning context (Graham, 2005) to avoid the time problems warned by Cooperstein & Kocevar-Weidinger (2004). Blended Learning (b-learning) combines face-to-face lessons with on-line lessons. Thus, it allows teachers to focus on the topics that need face-to-face interaction with students in class, and the rest of topics that can be reviewed before or after the class, students are assisted with some computer technology.

In particular, students cannot work on the free-text scoring system during the class. Given that the face-to-face time with students is usually limited (2-3 hours per week); it is proposed that face-to-face lessons are devoted to solve problems and doubts with teachers. On the other hand, students can collaborate with their groups before and/or after class assisted by the computer.

Due to the fear of change that some students may experience (Ashton-Hay, 2006), the procedure is proposed as voluntary. We do not recommend using a certain percentage of the score to evaluate the participation of the students in the procedure. On the other hand, students can be encouraged with some certificate or with extra credits to reward the extra effort of working outside the class. Students, who are not participating in the experience, are not penalized. They can answer the questions in class, listen to the teacher, take notes, and they do not need to use the free-text scoring system.

Finally, a continuous evaluation is proposed. We consider that the evaluation should be based on the work performed during the course rather than in a final exam, which according to Ashton-Hay (2006) can be a factor that inhibits active learning, and assessment should be related to the way that students have been taught.

4. Experimental study

4.1 Design

Figure 5 shows the experimental study design followed. As can be seen, the **target population was Pre-Primary and Primary Education students**. We chose that target population because we are Computer Science teachers, and we are used to teach students who enjoy learning about computers. Since the last three years, we are also responsible of teaching to Pre-Primary and Primary education students, who want to become teachers, and they are not so passionate about learning about computers.

From the target population, **we selected a sample of 162 students** (22% men, 78% women) who were the students of the groups in which we were teachers in the 2012/2013 academic year at our University. All the same, we did not want to make the procedure compulsory due to the problems mentioned of fear to change and being unable to follow a non-traditional teaching methodology (Ashton-Hay, 2006). Moreover, we did not want the procedure to directly affect their score. Therefore, we randomly assigned the 162 students into a **control group** of 38 students who did not know about the procedure, and a **test group** of 124 students who were given the possibility of following the procedure. Thus, they were given a talk in which we explained the procedure and the Willow free-text scoring system (Pérez-Marín et al. 2006) and an account to log into the system. They were also grouped to create and vote the questions.

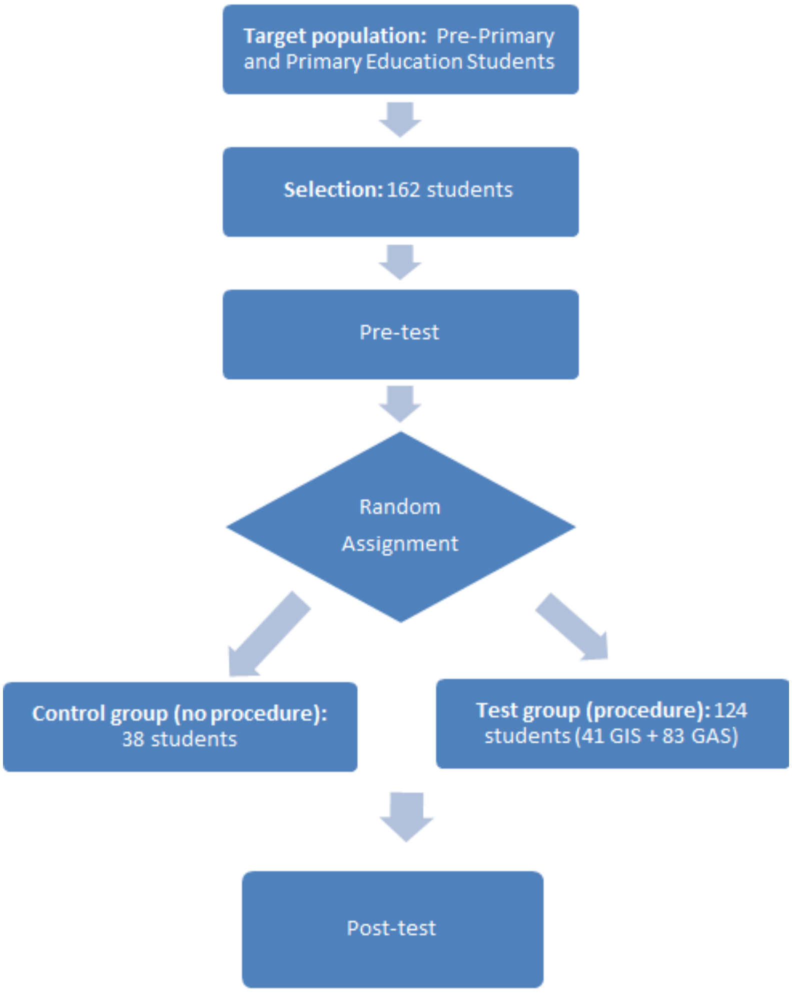

Target population: Pre-Primary and Primary Education Students
Selection: 162 students
Pre-test
Random Assignment
Control group (no procedure): 38 students
Test group (procedure): 124 students (41 GIS + 83 GAS)
Post-test

Figure 5. Experimental study design

All students took a 10 multiple-choice on-line pre-test individually at the beginning of the course. During the course, teachers supported the students of the test group to keep posting questions and their correct answers for the next lessons, and to vote the question they considered most adequate to be published in Willow for each lesson.

Two groups were distinguished in the test group according to their level of involvement with the procedure: the group of involved students (GIS) who fulfilled all requirements, and the group of average students (GAS) who stopped participating at some point during the course or missed some activity. Finally, all students took a 10 multiple-choice on-line post-test individually.

It is our **hypothesis (H) that students more involved in the procedure would increase both their academic performance and levels of engagement**. To test that hypothesis, two indicators were measured: the academic performance of the students and their engagement level.

1) The **academic performance of the students (A)** was measured according to Equation 1, where *n* is the number of students.

$$\mathrm{A} = \frac{\sum_{i=1}^{n} \mathrm{score(post\text{ - }test) - score\ (pre\text{ - }test)}}{n}$$

Equation 1. Formula to calculate the academic performance of the students

The statistically significance of the difference in A between the GIS (involved students who followed the procedure) and non-GIS (both GAS and control group students who did not meet all the requirements to complete the procedure) was measured, as well as the comparison between their final scores in the course. Provided that A is statistically greater in GIS, we consider that the procedure can increase the academic performance (even when no percentage of the score is provided because of taking part in the experience).

2) The **engagement level (E)** of the students according to Equation 2.

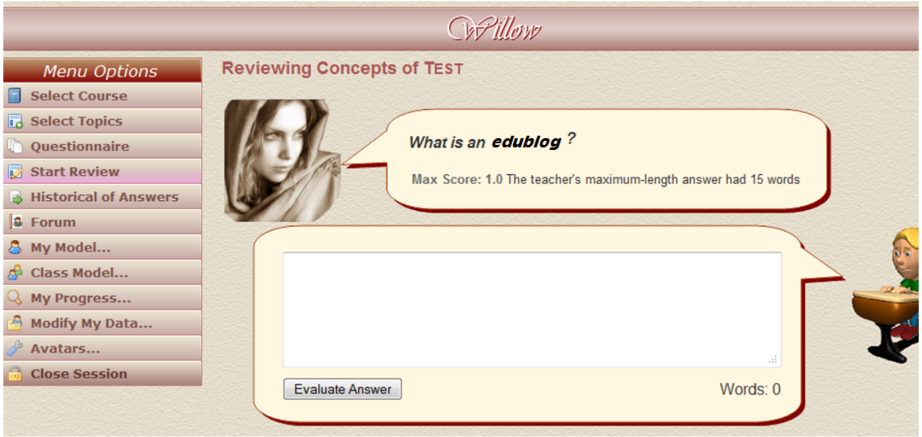
$$\mathrm{E} = \frac{\%(\text{votes received in GIS groups})}{\%(\text{votes received in non - GIS groups})}$$

Equation 2. Formula to calculate the engagement level

The perseverance in publishing questions and answers for all lessons, and the success in receiving votes are measured by Equation 2. If E is greater, equal or less than 1, it indicates that GIS have higher engagement levels than the rest of the students, and thus, that the procedure can increase the engagement level of the students.

4.2 Tools

The tools chosen to put into practise the proposal described in Section 3, according to the design described in Section 4.1 were a 10 multiple choice on-line test in Google Drive, and the free-text scoring tool Willow (Pérez-Marín et al. 2006). The reasons for choosing those tools were that either they have been created by us so we have full access to all their features, or they were already installed in our computers, and we have free access to all the data gathered together with automatically generated statistics. Figure 6 shows a snapshot of Willow.



Reviewing Concepts of TEST

What is an *edublog* ?

Max Score: 1.0 The teacher's maximum-length answer had 15 words

Words: 0

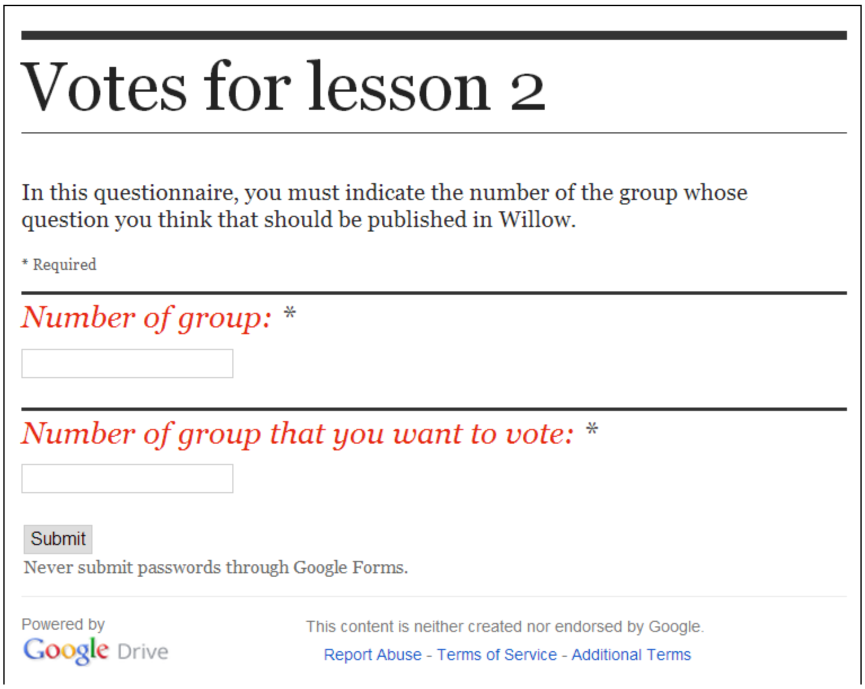
Figure 6. Snapshot of a question in the Willow free-text scoring tool

The conversation in Willow is driven by the agent (the woman on the left in Figure 6), which asks the questions introduced in the system (in Spanish or in English) and waits for the student's answer to be typed in the text area. After answering the question, Willow compares the student's answer to the correct answers provided by that question, with the core idea that the more similar they are, the higher the score. The feedback provided to the student can be just the score, the correct answers or a combination of both. Students have the possibility of changing the automatic evaluation in case that they consider that it is wrong.

The reason for allowing self-assessment is to prevent students thinking that they are taking an exam, or to feel that the score is going to be sent to the teacher. On the other hand, students are told that Willow is just a support system to help them study before and/or after class according to a Blended Learning methodology (Graham, 2005).

Students are also told that Willow cannot be used if they have not previously studied the lessons. The idea is not to replace the teacher, but to make students think about what they know about the course material, and if they need to keep studying new lessons. Willow has been used both in technical and non-technical domains by more than 500 students all over the world.

Figure 7 shows a snapshot of the voting questionnaire in Google Drive. It was created to be really simple so that students only needed to type the number of group that has published in the forum of Willow the question that they considered was the best candidate for each lesson.



Votes for lesson 2

In this questionnaire, you must indicate the number of the group whose question you think that should be published in Willow.

* Required

Number of group: *

Number of group that you want to vote: *

Submit
Never submit passwords through Google Forms.
Powered by
Google Drive
This content is neither created nor endorsed by Google.
Report Abuse - Terms of Service - Additional Terms

Figure 7. Snapshot of the questionnaire to vote for a question of a group in Google Drive

4.3 Development

Table 1 shows the development schedule of the experimental study. This schedule was published so that only the students of the test group have it available, and they knew when they had to publish a new question in the forum of Willow (always identified with their creators group number), or when they had to vote in to the Google Drive questionnaire after reading the questions and correct answers published in the forum of Willow for each lesson.

Table 1. Development schedule of the experimental study

Date	Task
Before February 7th	Wait for the mail with the link to the pre-test, the group number to participate in the experiment study, and the user-password account information to get access to Willow
February 7th – February 17th	Each group must publish in the forum of Willow one question and its correct answer for lesson 1
February 17th – February 24th	1) To vote in Google Drive the question of a group for lesson 1. 2) To publish in the forum of Willow one question and its correct answer for lesson 2.
February 24th – March 3rd	1) To vote in Google Drive the question of a group for lesson 2. 2) To publish in the forum of Willow one question and its correct answer for lesson 3.

March 3rd – March 10th	1) To vote in Google Drive the question of a group for lesson 3. 2) To publish in the forum of Willow one question and its correct answer for lesson 4.
March 10th – March 17th	1) To vote in Google Drive the question of a group for lesson 4. 2) To publish in the forum of Willow one question and its correct answer for lesson 5.
March 17th – March 24th	1) To vote in Google Drive the question of a group for lesson 5. 2) To publish in the forum of Willow one question and its correct answer for lesson 6.
March 24th – March 31st	1) To vote in Google Drive the question of a group for lesson 6. 2) To publish in the forum of Willow one question and its correct answer for lesson 7.
March 31st – April 7th	1) To vote in Google Drive the question of a group for lesson 7. 2) To publish in the forum of Willow one question and its correct answer for lesson 8.
April 7th – April 14th	1) To vote in Google Drive the question of a group for lesson 8. 2) Teacher publish the most voted questions in Willow
April 14th – May 5th	To answer the published questions in Willow and take the post-test (individually)

Finally, the teachers checked the groups that have followed the procedure for all lessons. Students who belonged to groups who have not published a question for a lesson, or who have not voted in Google Drive for all lessons were removed from the GIS and they were assigned to GAS. Moreover, the teachers checked which students have answered and passed all published questions in Willow, and completed both the pre and post test. Again, students who did not meet all requirements were removed from the GIS and passed to GAS. Finally, 41 GIS and 83 GAS were identified in the test group.

4.4 Results

At the end of the experimental study, the two indicators explained in Section 4.1 were measured. Regarding the **academic performance of the students (A)** measured according to Equation 1 (see Table 2), it was found that the 41 GIS increased their score in a post-test taken at the end of the course up to 8.5 (SD=1.25) from a 4.9 (SD=1.5) average score in the pre-test taken at the beginning of the course, which is extremely statistically significant according to an unpaired t-test with a two-tailed p value less than 0.0001. This improvement is also statistically significant compared to the improvement of the control group from 5.24 (SD=1.55) up to 7.84 (SD=1.57) with p=0.0145, and it is also statistically significant greater than the improvement of GAS from 5.51 (SD=1.76) up to 8.30 (SD=1.37) with p=0.0392. On the other hand, the improvement of GAS is not statistically significant compared to the control group (p=0.6164).

Table 2. Academic performance results (GIS=group of involved students who fulfilled all requirements of the procedure; GAS=group of average students who fulfilled some requirements; control group students who did not follow the procedure)

Test group						Control group		
GIS	**Pre-test**	**Post-test**	**GAS**	**Pre-test**	**Post-test**	**All**	**Pre-test**	**Post-test**
Mean	4.9	8.5	Mean	5.51	8.3	Mean	5.24	7.84
SD	1.5	1.25	SD	1.76	1.37	SD	1.55	1.57
n	41		n	83		n	38	

GIS were also able to achieve a 9.04 final score value (in 0-no knowledge up to 10-maximum knowledge) scale (SD=0.84). The improvement of these scores is statistically significant, according to an unpaired t-test with $p < 0.0001$ in comparison to the final score achieved by the control group (4.9, SD=3.7); and, it is also a statistically significant improvement to the final score achieved by GAS (7,1, SD=3.3) with $p = 0.0003$.

Table 3. Final scores (GIS=group of involved students who fulfilled all requirments of the procedure; GAS=group of average students who fulfilled some requirements; control group students who did not follow the procedure)

Test group				Control group	
GIS	**Score**	**GAS**	**Score**	**All**	**Score**
Mean	9.04	Mean	7.1	Mean	4.9
SD	0.84	SD	3.3	SD	3.7
n	41	n	83	n	38

Regarding the **engagement level (E)** of the students measured according to Equation 2, the data gathered was calculated by counting the number of votes received by questions in which all members were GIS, at least one member was GIS, and none of the members was GIS, and their percentages. Table 4 shows the results gathered.

Table 4. Number of votes received by questions of groups and their percentage

Type of group	Votes	Percentage
All members are GIS	28	11
At least one member is GIS	121	47
No members is GIS	108	42

The engagement level is calculated as:

E(all members are GIS) = %(votes received in GIS groups)/%(votes received in non-GIS groups) = 11/42 < 1

E(at least one member is GIS) = %(votes received in GIS groups)/%(votes received in non-GIS groups) = 47/42 > 1

As can be seen, it is not necessary that all members of the group are GIS to keep the engagement level. The minimum needed is two GIS to keep publishing questions and correct answers for the next lessons in the forum of Willow, and voting in Google Drive in the indicated dates. GIS students also showed more interest at class, and higher levels of motivation towards a course that it is not usually their favourite given that, in general, Pre-Primary and Primary Education students do not enjoy having a course of technology which is quite different from the rest of their courses. Nevertheless, some comments provided by GIS were the following (translated from Spanish to English):

- *"The procedure helps me to organize my time study".*

- *"Studying like we have done this year is not so boring, and the course material starts to make sense".*

- *"Following the procedure I can easily keep up to date the work for the course".*

- *"By creating the questions and finding the correct answers I can focus better than just answering questions provided by the teacher".*

4.5 Practical implications and limitations

The results gathered provide statistic evidence to support H. The main practical implication is that the combination of active learning and collaborative knowledge building using free-text scoring tools in Blended Learning (b-learning) environments can increase both the students academic performance and their levels of engagement.

In particular, this is interesting in courses that students may find boring because they are not so related with their degree, or they are different from the rest of courses of their degree (e.g. more or less theoretical or practical). However, it should also be highlighted that GIS are usually good students (i.e. they make an effort to study the course and they usually pay attention to lessons). It could be the case that even without applying the proposal GIS could have achieved higher scores and A, as it has happened in other courses in which no procedure was applied. Nevertheless, it should also be taken into account the comments of GIS indicating that the procedure has helped them to achieve their goals, so it could be thought that without the procedure the effort needed to reach the same academic performance would have been higher even for GIS.

Finally, another limitation that can be found in this study is the relevance of the teacher in the application of the procedure. It has been observed that different teachers provided different levels of support, and they are more or less demanding with their results. It could be the case that even applying the same procedure in two different courses with two different teachers the results may be different because of the teacher. If teacher A is really demanding, the results could be less significant than if teacher B is less demanding and the increase in academic performance can be higher. Nevertheless, it should also be taken into account the engagement indicator and the higher levels of motivation and involvement with the course registered by GIS irrespectively of the teacher.

5. Discussion

According to many authors, learning should be active, and collaboration and interaction improve the academic performance of the students. It has also been studied how students considered collaborative learning as fun, and they are more engaged to the tasks to perform. However, traditional learning keeps being usual in many educational institutions, and some students are still afraid of changing their role from a passive recipient of knowledge to become an active creator of their own knowledge.

In this paper, **we have answered the need for more research into how powerful learning environments can be extended with social free-text scoring tools in Blended Learning (b-learning) environments**. That way, learning can be more active and collaborative without diminishing face-to-face time in class. Moreover, free-text scoring can be extended so that it is not only taken into account the comparison between the student answer and some correct answer provided by a teacher, but other students' answers are also taken into account.

The novelty of the approach lies in the lack of research on combining active learning, collaborative knowledge building and free-text scoring. To our knowledge, **no papers have been written on how these methodologies can be used to provide students with tools to improve their behaviour** towards courses that they may find boring as they are unrelated to the main topic of their degree (i.e. in our case, Pre-Primary and Primary Education University students do not tend to like Computer Science courses).

Therefore, a procedure to combine active learning, collaborative knowledge building and social free-text scoring has been presented with three main steps: *(1)* to organize the students into groups; *(2)* to ask the students to create and vote questions for each lesson in their groups using a free-text scoring tool; *(3)* to answer the questions individually.

162 Pre-Primary and Primary Education University students were randomly assigned into a control group (no procedure) and a test group (procedure). These students have also the particularity that they want to become teachers, and they do not usually enjoy the Computer Science courses they have in the degree.

In particular, 38 students were assigned to the control group, and 124 students were assigned to the test group. From the test group, 41 students fulfilled all the requirements during the course (GIS), and 83 students fulfilled some requirments (GAS).

Our hypothesis was that GIS would be able to increase their academic performance and levels of engagement compared to the rest of the students in the study. **The results gathered provide statistic evidence to support that hypothesis**. Therefore, we would like to encourage teachers who want to increase the academic performance and levels of engagement of their students to try active learning, collaborative knowledge building and new social computer assisted tools.

We would also like to launch a call for more research into this combination of methodologies. It is because, from our experience, we have seen that they may have a good potential to improve the behaviour of University students in courses, which they may find boring, as they are unrelated to the main topic of their degree.

6. References

Ashton-Hay, S. (2006), 'Constructivism and Powerful Learning Environments: Create Your Own!', *in* QUT Digital Repository: http://eprints.qut.edu.au/, ed., '9th International English Language Teaching Convention "The Fusion of Theory and Practice"', Middle Eastern Technical University - Ankara, Turkey.

Attali, Y. & Burstein, J. (2006), 'Automated Essay Scoring With e-rater V. 2', *Journal of Technology, Learning, and Assessment* **4**(3).

Birenbaum, M.; Tatsuoka, K. & Gutvirtz, Y. (1992), 'Effects of response format on diagnostic assessment of scholastic achievement', *Applied psychological measurement* **16**(4), 353-363.

Blayney, P. & Freeman, M. (2003), Automated Marking of Individualised Spreadsheet Assignments: the impact of different formative self-assessment options, *in* 'Proceedings of the 7th Computer Assisted Assessment Conference'.

Bloom, M. J.; Kurian, J. C.; Chua, A. Y. K.; Goh, D. H. L. & Lien, N. H. (2013), 'Social question answering: Analyzing knowledge, cognitive processes and social dimensions of micro-collaborations', *Computers & Education* **69**, 109–120.

Brown, H. (2000), *Principles of Language Learning and Teaching.* Pearson Education.

Bruner, J.; Goodnow, J. & Austin, G. (1956), *A Study of Thinking*, Wiley, N.Y.

Bruner, J.; Goodnow, J. & Austin, G. (1956), *A study of thinking*, Wiley, N.Y.

Burstein, J.; Kukich, K.; Wolff, S.; Lu, C.; Chodorow, M.; Bradenharder, L. & Harris, M. D. (1998), Automated Scoring Using A Hybrid Feature Identification Technique, *in* 'Proceedings of the Annual Meeting of the Association of Computational Linguistics', pp. 206-210.

Callear, D.; Jerrams-Smith, J. & Soh, V. (2001), CAA of Short Non-MCQ Answers, *in* 'Proccedings of the 5th International Computer Assissted Assessment conference'.

Christie, J. (2003), Automated essay marking for content - does it work?, *in* 'Proceedings of the 7th International Computer Assisted Assessment Conference'.

Cooperstein, S. E. & Kocevar-Weidinger, E. (2004), 'Beyond active learning: a constructivist approach to learning', *Reference Services Review* **32**(2), 141-148.

Datar, A.; Doddapaneni, N.; Khanna, S.; Kodali, V. & Yadav, A. (2004), 'EGAL - Essay Grading and Analysis Logic', http://www.d.umn.edu/~tpederse/Courses/CS8761-FALL04/Project/Readme-Boca.html.

Dessus, P.; Lemaire, B. & Vernier, A. (2000), Free Text Assessment in a Virtual Campus, *in* 'Proceedings of the 3rd International Conference on Human System Learning', pp. 61-75.

Ebbinghaus, H. (1913), *Memory: A Contribution to Experimental Psychology*, Teachers College, Columbia University, New York, NY..

Foltz, P.; Laham, D. & Landauer, T. (1999), 'Automated Essay Scoring: Applications to Educational Technology', *proceedings of EdMedia* **99**.

Gagne, R. (1985), *The Conditions of Learning*, Holt, Rinehart & Winston, New York, NY.

Good, T. & Brophy, J. (1994), *Looking in Classrooms*, Harper Collins College Publishers, New York, NY.

Graham, C. R. (2005), Blended Learning Systems: Definition, Current Trends, and Future Directions, *in* 'Handbook of Blended Learning: Global Perspectives, local designs', Pfeiffer Publishing, , pp. 3-21.

Hoang, L. & Arch-Int, N. (2013), 'Assessment of Open-Ended Questions using a Multidimensional Approach for the Interaction and Collaboration of Learners in E-Learning Environments', *Journal of Universal Computer Science* **19**(7), 932-949.

Ishioka, T. & Kameda, M. (2004), 'Automated Japanese Essay Scoring System: JESS', *Proceedings of the 15th International Workshop on Database and Expert Systems Applications*, 4--8.

Kakkonen, T.; Myller, N.; Timonen, J. & Sutinen, E. (2005), Automatic Essay Grading with Probabilistic Latent Semantic Analysis, *in* 'Proceedings of the 2nd Workshop on

Building Educational Applications Using NLP, Association for Computational Linguistics', pp. 29--36.

Kintsch, E.; Steinhart, D.; Stahl, G. & the LSA Research Group (2000), 'Developing Summarization Skills through the Use of LSA-based Feedback', *Interactive Learning Environments*.

Lesgold, A. (2004), 'Contextual requirements for constructivist learning', *International Journal of Educational Research* **41**(6), 495-502..

Marshall, S. & Barron, C. (1987), 'MARC-Methodical Assessment of Reports by Computer', *System* **15**(2), 161-167.

Mason, O. & Grove-Stephenson, I. (2002), Automated free text marking with paperless school, *in* 'Proceedings of the 6th International Computer Assisted Assessment Conference'.

Mayer, R. (1987), *Educational Psychology: A Cognitive Approach*, Little, Brown, Boston, MA.

Mcgrath, P. (2003), Assessing Students: Computer Simulation vs MCQs, *in* 'Proceedings of the 7th Computer Assisted Assessment Conference', pp. 243-246.

Ming, Y.; Mikhailov, A. & Kuan, T. (2000), 'Intelligent Essay Marking System', *Learners Together, NgeeANN Polytechnic, Singapore*.

Mitchell, T.; Aldridge, N.; Williamson, W. & Broomhead, P. (2003), Computer Based Testing of Medial Knowledge, *in* 'Proceedings of the 7th Computer Assisted Assessment Conference', pp. 249-267.

Mitchell, T.; Russell, T.; Broomhead, P. & Aldridge, N. (2002), Towards Robust Computerised Marking of Free-Text Responses, *in* 'Proceedings of the 6th Computer Assisted Assessment Conference'.

Mitchell, T.; Russell, T.; Broomhead, P. & Aldridge, N. (2002), Towards Robust Computerised Marking of Free-Text Responses, *in* 'Proceedings of the 6th Computer Assisted Assessment Conference'.

Moskaliuk, J.; Kimmerle, J. & Cress, U. (2012), 'Collaborative knowledge building with wikis: The impact of redundancy and polarity', *Computers & Education* **58**, 1049–1057.

Nicol, D. & Macfarlane-Dick, D. (2006), 'Formative assessment and self-regulated learning: a movel and seven principles of good feedback practice', *Studied in Higher Education* **31**(2), 199-218.

Page, E. (1966), 'The Imminence of Grading Essays by Computer', *Phi Delta Kappan* **47**(1), 238-243.

Parsons, H.; Schofield, D. & Woodget, S. (2003), Piloting Summative Web Assessment in Secondary Education, *in* 'Proceedings of the 7th Computer Assisted Assessment Conference'.

Pérez-Marín, D. (2007), 'Adaptive Computer Assisted Assessment of free-text students' answers: an approach to automatically generate students' conceptual models', PhD thesis, Escuela Politecnica Superior, Universidad Autonoma de Madrid.

Pérez-Marín, D.; Alfonseca, E.; Rodríguez, P. & Pascual-Nieto, I. (2006), Willow: Automatic and adaptive assessment of students free-text answers, *in* 'Proceedings of the 22nd International Conference of the Spanish Society for the Natural Language Processing (SEPLN)'.

Pérez-Marín, D.; Pascual-Nieto, I. & Rodriguez, P. (2009), 'Computer-assisted assessment of free-text answers', *The Knowledge Engineering Review* **24**(4), 353–374.

Rosé, C.; Gaydos, A.; Hall, B.; Roque, A. & VanLehn, K. (2003), 'Overcoming the Knowledge Engineering Bottleneck for Understanding Student Language Input', *Proc. of the 11h International Conference on Artificial Intelligence in Education (AIED'03)*.

Rudner, L. & Liang, T. (2002), Automated Essay Scoring Using Bayes' Theorem, *in* 'Proceedings of the annual meeting of the National Council on Measurement in Education'.

Simons, P.Stern, D. & Huber, G., ed. (1997), *Active Learning for Students and Teachers*, Frankfurt: Lang., chapter Definitions and theories of active learning, pp. 159-173.

Sukkarieh, J.; Pulman, S. & Raikes, N. (2003), Auto-marking: using computational linguistics to score short, free text responses, *in* 'Proceedings of the 29th IAEA Conference, theme: Societies' Goals and Assessment'.

Vygotsky, L. (1978), *Mind in Society: The Development of Higher Psychological Processes*, Harvard University Press, Cambridge, MA.

Vygotsky, L. (1978), *Mind in Society: The Development of Higher Psychological Processes*, Harvard University Press, Cambridge, MA..

Wiemer-Hastings, P. & Graesser, A. (2000), 'Select-a-Kibitzer: A computer tool that gives meaningful feedback on student compositions', *Interactive Learning Environments*.