Computers & Education 68 (2013) 440–457

Contents lists available at SciVerse ScienceDirect

Computers & Education

journal homepage: www.elsevier.com/locate/compedu

A web based collaborative testing environment

Ricardo Conejo [a], Beatriz Barros [a,*], Eduardo Guzmán [a], Juan-Ignacio Garcia-Viñas [b]

[a] E. T. S. Ingeniería Informática, Universidad de Malaga, 29071 Malaga, Spain
[b] E.T. Ing. Forestal, Universidad Politécnica de Madrid, 28071 Madrid, Spain

ARTICLE INFO

Article history:
Received 19 March 2013
Received in revised form
3 June 2013
Accepted 6 June 2013

Keywords:
Collaborative learning
Learning environments
Collaborative testing
Computerized testing
Improving classroom teaching

ABSTRACT

This paper presents a *computer supported collaborative testing* system built upon the Siette web-based assessment environment. The application poses the same set of questions to a group of students. Each student in the group should answer the same question twice. An initial response is given individually, without knowing the answers of others. Then the system provides some tools to show the other partners' responses, to support distance collaboration. Finally a second individual answer is requested. In this way assessment and collaboration activities are interlaced. At the end of a collaborative testing session, each student will have two scores: the *initial score* and the *final score*. Three sets of experiments have been carried out: (1) a set of experiments designed to evaluate and fine tune the application, improve usability, and to collect users' feelings and opinions about the system; (2) a second set of experiments to analyze the impact of collaboration in test results, comparing individual and group performance, and analyzing the factors that correlate to those results; and (3) a set of experiments designed to measure individual short-term learning directly related to the collaborative testing activity. We study whether the use of the system is associated with actual learning, and whether this learning is directly related to collaboration between students. Our studies confirm previous results and provide the following evidence (1) the performance increase is directly related to the access to other partners' answers; (2) a student tends to reach a common answer in most cases; and (3) the consensus is highly correlated with the correct response. Moreover, we have found evidence indicating that most of the students really do learn from *collaborative testing*. High-performing students improve by self-reflection, regardless the composition of the group, but low-performing students need to be in a group with higher-performing students in order to improve.

1. Introduction

Collaborative learning scenarios should engage learners in cognitive and metacognitive activities. They ought to promote a conscious cooperative development of shared knowledge in a common context and enrich learners' individual understanding of the world. But, as Nussbaum et al. (2009) point out, collaboration itself does not necessary yield learning. Effective collaborative scenarios should involve learners in situations that require a reflection on their own knowledge as well as their colleagues' in a grounding process that attempts to acquire more refined and mature knowledge. Measuring the effects of computer supported collaborative learning, during or after collaboration, is a challenge that can provide teachers and learners with tools "*to monitor and evaluate their solo and collaborative processes and products during actual engagement of learning*" (Gress, Fior, Handwin, & Winne, 2010).

It has been demonstrated that students' interaction and peer help are very useful for learning. Assessment can be also viewed as a learning activity. Black and William (2009) developed a theory for *formative assessment*, based on their ideas about the use of assessment by teachers for learning practices to improve student achievement. This theory is currently known as *assessment for learning* (Cooper & Cowie, 2010). The system we are describing in this paper is mainly used for formative assessment because it provides a score at the end of the test that indicates to the student if he is progressing adequately. The system is also a tool for learning, because during the collaborative testing students reflect and argue about their answers with their peers, they discuss about the questions posed, and give and receive feedback, and so they improve their understanding of the subject.

[*] Corresponding author. Tel.: +34 952 13356 (office).
E-mail address: bbarros@lcc.uma.es (B. Barros).

0360-1315/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.compedu.2013.06.001

Assessment can also be combined with collaborative learning practices. There are different ways of doing so. When a student evaluates himself, it is called *self-assessment*. If he evaluates his partners and is evaluated by them, it is commonly known as *peer assessment* (van Gennip, Segers, & Tillema, 2010; Kollar & Fischer, 2010). If two or more partners, including students and teachers, discuss and grade something, we call it *collaborative assessment* (Kwok & Ma, 1999). *Collaborative assessment* is nowadays extensively used in Massive Open Online Courses (MOOCs) like Coursera, although it is not called so (Waldrop, 2013). On the other hand, when two or more students take an assessment together (which is assessed by a teacher), previous research refers to it as *collaborative testing* (Rao, Collins, & DiCarlo, 2002; Sandahl, 2009). *Collaborative assessment* refers to the task "assessment" that is taken in common. For example, in a collaborative assessment environment the score of an assignment is determined as a result of a negotiation or agreement between two or mode evaluators. On the other hand *collaborative testing* refers to the task "testing" which is taken in common. That is, the same question is posed to a group of students that solve it in common. Then the answer is assessed by the teacher, or by an automatic grader, which is the case of the system described in this paper.

The main difference between *collaborative assessment* and *collaborative testing* is that in the first case the students participate in the elicitation of the assessment criteria. It usually applies in cases where the task to be assessed is complex, it may be an essay or an assignment. On the other hand, collaborative testing involves simpler tasks where there is no thoughtful discussion about the assessment criteria, like quizzes. However, *self-assessment*, *peer assessment*, *collaborative assessment* and *collaborative testing* share a common objective: to make students reflect on the answers given by others, explain and justify their own answers and foster reflection on their own learning (Ochoa, Guerrero, Pino, Collazos & Fuller, 2003; Swan, Shen, & Hiltz, 2006).

In this research study, we present a *computer supported collaborative testing* system designed for the Web. The system is domain independent, and can be used with different types of questions, including multiple-choice with single or multiple answers, and short open questions that can be corrected automatically by means of regular expressions. The application poses the same set of questions to a group of students. Each student in the group should answer the same question twice. An initial response is given individually, without knowing the answers of others. Then the system provides some tools to show other partners' responses to support distance collaboration. Finally a second individual answer is requested. In this way assessment and collaboration activities are interlaced. The application does not coerce the collaboration, neither the final consensus between the students in the group. All communications are asynchronous, so that a student can freely decide whether to collaborate with others or just take the test as a free rider. We will see that in most cases they did collaborate; which is a good indicator of the usefulness of the system.

According to Dillenbourg (1999), effective collaborative learning scenarios are those that increase the probability of interaction among people that will trigger learning mechanisms. He describes four ways to address that: (i) *To set up initial conditions*, like group composition, group size, etc. However, Dillenbourg acknowledge that it is difficult to set up initial conditions that guarantee effectiveness of collaborative learning. (ii) *To over-specify the collaboration contract with a scenario based on roles*; (iii) *To scaffold productive interactions*, providing semi-structured interfaces and trying that peers focus more on the task; and (iv) *To monitor and regulate the interaction*. To meet these requirements a collaborative framework has been implemented on the Siette testing environment (Conejo et al., 2004) The framework provides a high flexibility to select different initial conditions, like pre- or post- definition of groups, selection of group size, number of questions, question selection criteria, etc. (see section 3). The collaboration has been defined using a script, structured enough to be able to organize and observe the collaboration process, but not too rigid, so as to allow a rich, open and flexible interaction (Dillenbourg & Hong, 2008). Special care have been applied to help students to focus on the learning objectives, instead of being distracted by other aspects not related to those objectives, which is a common problem in collaborative learning, like off-topic conversations (Paulus, 2009) or procrastination (Michinov, Brunot, Le Bohec, Juhel, & Delaval, 2011). Finally, all the activities carried out by the students have been logged, and a set of indicators about the discussion process has been defined. This will allow us to gauge the quality and quantity of collaboration and will allow us to implement in the future some monitoring actions.

The research presented in this paper differs from previous publications in the *collaborative testing* literature. Robinson, Sweet, & Mayrath (2008) describe a computer-based system that delivers tests for a group of students who take the test together and are assessed together. Most of the other experiments are based on paper-and-pencil and multiple choice questions combined with some kind of face-to-face interaction in the classroom. They also separate the individual and group assessment in time (Bjornsdottit, 2012). In our case, the test is delivered and assessed by a computer application and does not require face-to-face collaboration. Our system allows distance collaboration using computer supported collaborative tools. The system organizes the collaboration in a way that focuses discussion on the current topic, avoiding distraction and providing immediate feedback if necessary. We have defined a sequence of individual and collaborative activities that are interlaced during the test. This allows the individual and group to be tested simultaneously, and facilitates the analysis of the role of each member and its correlation with individual and group performance. A side benefit of a computer-supported system is that the collaboration process can be recorded and measured. The system is thus a workbench to experiment with and analyze in depth the different factors that are involved in collaborative testing.

The paper describes 12 experiments (some of them repeated for cross validation) in which 463 students from 2 public universities in Spain were involved from 2004 to 2011. Different domains (from Computer Science to Botany), different cases (allowing students to access the system from home or in a controlled environment in classrooms) and different parameters (number of students in each group, number of questions posed, etc.) have been considered in this work in order to guarantee the generalization of the conclusions, and to study their potential influence on the results. Experiments are grouped into three sets: (1) a set of experiments designed to evaluate and tune up the application, improve usability, and to collect users' feelings and opinions about the system; (2) a second set of experiments to analyze the impact of collaboration on test results, comparing individual and group performance, and analyzing the factors that correlate to those results; and (3) a set of experiments designed to measure individual short-term learning directly related to the collaborative testing activity. We study whether the use of the system is associated with actual learning, and whether this learning is directly related to collaboration between students. As will be shown, the evidence suggests that the learning that occurs is clearly related to the collaboration activity and that it is relatively independent of the prior knowledge of students' and group composition. Although not all of the results were always conclusive at 95% statistical confidence, they are consistent with previous findings of other researchers in collaborative testing and provide new evidence and a deeper knowledge of the collaborative testing process.

The paper is structured as follows: Section 2 presents a summary of related work, the goal of which is to contextualize the contributions of this research. Section 3 first gives a brief description of Siette that allows readers to appreciate the technical aspect of the development of the collaborative learning environment and the description of the experiment conditions. Next, we describe the collaborative script and the collaborative testing environment that has been implemented, and the data we measure with it. Section 4 describes the experiments. Finally, Section 5 summarizes the conclusions that can be extracted from the work done.

The main contributions of this paper can be summarized in these five points: (1) A description of a new domain independent system to allow web-based computer supported collaborative testing. (Described in Section 3 and evaluated in Section 4.1.) (2) All students increase their performance on tests by mean of collaboration. (Supported by section 4.2, Tables 3 and 4.) (3) Students freely get to a consensus that usually correlates to the right answer (Supported by section 4.2, Tables 5–7.) (4) All students do learn in the short-term by using the collaborative system. (Supported by section 4.2, Table 9.) (5) To improve more, groups should be composed by at least a high-performance student. (Supported by section 4.3, Table 10.)

2. Related work

In this section we present a summary of previous work, and highlight the principal findings obtained. The section is divided into two subsections. In the first one we have collected together some research results on discussion groups and knowledge convergence during collaborative learning activities. The second subsection focuses specifically on previous work on *collaborative assessment* and *collaborative testing*.

2.1. Discussion groups and knowledge convergence

Schellens and Vackle (2005) argued that "*discussion groups with high discussion activity perform better*". This conclusion is partially backed up by the research studies described in this paper, although using a different environment and experiment. In the former case, the authors highlighted the importance of the task structure and the length of discussion activity. In our research context, however, we have considered the number of times that a user looks at other students' responses, the number of messages exchanged and the total length of messages, as indicators of the collaborative activity.

When activities are carried out in groups, it is interesting to explore how knowledge converges in a group after collaboration. Social interaction and the exchange of ideas are important aspects of learning that lead to *knowledge convergence*. The authors point out that individuals benefit from the collaborative learning process. Weinberger, Stegmann, and Fischer (2007) have conceptualized knowledge convergence and distinguished between *knowledge equivalence* and *shared knowledge*. The former refers to "*learners becoming more similar to their learning partners with regard to the extent of their individual knowledge*"; the latter considers "*that learners have knowledge of the very same concepts as their learning partners*". There are two main aspects of knowledge convergence, (i) the *process convergence* in which learners make their ideas explicit, contribute, compare and organize their opinions into a reasoning approach. At this point, convergence occurs when there is reciprocal influence between partners that leads to an increased similarity of the cognitive responses within the group (Fisher & Mandl, 2005) and (ii) *outcome convergence* refereed to the study of the results built individually by the learners and the relationship of these products with the group result. In this paper we explore how knowledge converges as a result of the collaborative process.

Laurillard (2010) reports that on-line interaction between students has a positive effect on students' learning. A text-based online conversation tool that allows on-line interaction could be implemented using a chat. In our research, a chat with sentence openers with free conversation is used to facilitate on-line communication. This semi-structured tool allows us to define a set of indicators to measure collaboration activity.

2.2. Collaborative assessment and collaborative testing

Two interesting experiments (Jermann & Dilemburg, 2003) were carried out in a web-based environment combining argument-based collaboration with questionnaires (about opinions) and sharing facilities for working and reflecting in groups. PECASSE (Guoli, Gogoulou, &

Table 1
Comparison of the experiment results.

	J1	J2	B
Evaluated prototype	2nd	3rd	3rd
Date	Dec. 06	Nov. 07	Jan. 08
Number of users	33	36	18
(1.1) My expertise with computers	4.12 ± 0.23	3.89 ± 0.26	3.17 ± 0.55
(1.2) I like computer-based testing	3.64 ± 0.36	3.83 ± 0.23	3.67 ± 0.64
(1.3) I use the chat frequently	3.65 ± 0.43	3.66 ± 0.40	3.22 ± 0.69
(2.1) I have enjoyed taking the collaborative test	4.33 ± 0.34	4.47 ± 0.21	3.41 ± 0.58
(2.2) I think I have learnt from my partner	4.00 ± 0.35	4.06 ± 0.32	3.12 ± 0.63
(2.3) The partner's answer helped me	3.65 ± 0.33	3.75 ± 0.32	3.18 ± 0.58
(2.4) The discussion with my partner helped me	4.15 ± 0.31	4.31 ± 0.27	3.65 ± 0.60
(3.1) The system works fine	3.62 ± 0.33	4.42 ± 0.31	3.17 ± 0.57
(3.2) I always knew where I was	3.99 ± 0.36	4.03 ± 0.36	4.44 ± 0.31
(3.3) I easily got/obtained my partner's answers	4.04 ± 0.38	4.53 ± 0.29	4.00 ± 0.48
(3.4) I could easily communicate with my partner	3.59 ± 0.38	4.51 ± 0.26	4.33 ± 0.48
(3.5) The chat annotation was useful	2.66 ± 0.31	3.17 ± 0.47	3.44 ± 0.57
(3.6) The structured chat was useful	2.70 ± 0.41	–	–
(3.7) It is better to use Siette with the collaborative frame	4.38 ± 0.37	4.58 ± 0.23	3.78 ± 0.47
(4.1) My overall rating of the system is	4.06 ± 0.35	4.44 ± 0.19	3.83 ± 0.35

Table 2
Summary of research questions about performance improvement and main conclusions.

Research question	Table	Summary of conclusions	Section
(Q1): Do student improve their final score taking a collaborative test?	Table 4	The average of final answer scores is always higher than the average of the initial answer score ($p << 0.05$)	4.2.1
(Q2): Does the use of the collaborative framework correlates with the score improvement?	Table 5	The students that have used the collaborative framework improve their performance more than those that have not used it. ($p < 0.05$)	4.2.2
(Q3): Do student reach to a consensus in their responses? *(Q4): Does the use of the collaborative framework correlates with higher consensus?* *(Q5): Does the reach to a consensus correlates with higher score improvement?*	Table 6	On the average, a consensus is commonly reached (between 50% and 90%) Those students who have used the tool were more likely to reach a consensus in their final answers. ($p < 0.05$, in 4 tests out of 6). Students that reach to a consensus increment their performance more than those that do not ($p < 0.05$ in 5 tests out of 6)	
(Q6): What features of the collaborative framework are most effective, allow viewing other students answers, allow communication using the chat tool or the combination of both?	Table 7 Table 8	Best results are obtained when students use both the chat tool and "view answer" button. Although the differences are not very significant	
(Q7): Does the length of discussion correlates with higher improvements	Table 7 Table 8 (right part)	No significant differences in the improvement obtained in questions where a short or long discussion took place	

Table 3
Description of the tests included in the second set of experiments.

Test ID	Description	Date	#Students	#Questions
1377	Compilers: Grammars and parsers	2007, Nov. 27	53	15
5225	Programming: JAVA	2006, Dec. 11	34	20
5769	Compilers: LEX	2007, Nov. 06	58	17
6107	Compilers: Attribute grammars.	2008, Mar. 12	17	15
6128	Botany: Vegetal anatomy	2008, Jan. 25	17	25
7147	Compilers: LEX	2008, Oct. 28	19	15

Grigoriadou, 2006) came up with an approach combining individual and collaborative learning where assessment is considered an important aspect, but which focuses more on elaboration, review and evaluation of activities rather than on *collaborative testing*. Valdivia and Nussbaum (2009) used multiple-choice questions in collaborative environments with small groups. They found that the use of small groups fosters an improved participation of each individual and that collaboration improves their performance. The QSIA system (Rafaeli, Barak, Dan-Gur, & Toch, 2004) defines an interactive environment for learning, assessment and knowledge sharing. In this case, questions are created in a learning community and students learn independently, using the generated database of knowledge systems. The results of the assessments are used for recommendation purposes in future interactions within the learning community. These systems are closer to *collaborative assessment*. Our research is different from them because: (1) questions are used for assessment; (2) our system interleaves individual and collaborative phases, and (3) the answers are automatically evaluated by the system.

Bjornsdottit (2012), in his thesis dissertation, surveys the current proposals that involve *collaborative testing* and summarizes most of the research topics and published findings. According to this survey, *collaborative testing* groups are mostly composed of 2–3 people, occasionally by 3–4, and rarely by more. The type of assessment is mostly based on multiple-choice and fill-in-the-blank questions, and less

Table 4
Description of the average results obtained in each test.

Test ID	Total number of questions answered	Average of initial answers score (%)	Average of final answers score (%)	Average diff between final and initial answers (%)	Average of initial agreement (%)	Average of final agreement (%)
1377	795	81.5 ± 1.9	88.0 ± 1.6	6.5 ± 1.5	45.4 ± 3.5	87.7 ± 2.3
5225	627	38.9 ± 4.5	46.2 ± 4.5	7.4 ± 2.9	52.8 ± 3.9	70.0 ± 3.6
5769	985	53.3 ± 3.2	70.5 ± 2.9	17.5 ± 2.8	35.8 ± 3.0	81.3 ± 2.4
6107	330	33.8 ± 6.9	68.3 ± 6.1	34.5 ± 6.7	10.0 ± 3.2	57.9 ± 5.3
6128	382	44.0 ± 5.2	61.7 ± 4.9	17.7 ± 4.2	39.0 ± 4.9	74.9 ± 4.4
7147	247	28.3 ± 6.8	49.1 ± 6.8	20.7 ± 5.9	43.3 ± 6.2	78.1 ± 5.2

Table 5
Score of students who used the collaborative framework with those who did not.

Test ID	The collaborative framework was not used			The collaborative framework was used		
	Number of questions	Average final score (%)	Average diff score (%)	Number of questions	Average final score (%)	Average diff score (%)
1377	299	86.8 ± 2.6	4.1 ± 1.9	407	88.7 ± 2.0	7.9 ± 2.0
5225	409	41.9 ± 5.6	1.5 ± 1.5	218	53.4 ± 7.4	18.5 ± 6.9
5769	131	72.1 ± 7.1	13.3 ± 5.6	854	70.2 ± 3.1	18.1 ± 2.9
6107	47	52.1 ± 18.4	25.5 ± 20.3	283	71.0 ± 6.4	36.0 ± 7.0
6128	39	58.5 ± 14.7	10.7 ± 10.7	343	62.1 ± 5.2	18.5 ± 4.5
7147	49	37.1 ± 17.0	12.1 ± 9.4	198	52.0 ± 7.3	22.8 ± 6.9

Table 6
Score of students that used the collaborative framework with those who did not.

Test ID	Average final agreement in questions where the student did not use the collab. frm. (%)	Average final agreement in questions where the student did use the collab. frm. (%)	Average diff. score in questions where the students did not reach an agreement (%)	Average diff. score in questions where the students reach an agreement (%)
1377	82.6 ± 4.3	90.7 ± 2.6	2.4 ± 3.2	7.1 ± 1.6
5225	63.6 ± 4.7	82.1 ± 5.1	0.9 ± 0.9	10.1 ± 3.5
5769	60.3 ± 8.4	84.5 ± 2.4	6.4 ± 4.6	20.0 ± 3.0
6107	57.4 ± 14.3	58.0 ± 5.8	27.8 ± 10.3	39.4 ± 8.7
6128	71.8 ± 14.3	75.2 ± 4.6	11.9 ± 7.8	19.6 ± 4.9
7147	65.3 ± 13.5	81.3 ± 5.4	8.1 ± 9.9	24.2 ± 6.9

Table 7
Use of the collaborative framework.

Test ID	Average diff. score using just the "View answer" button (%)	Average diff. score using just the "chat room" button (%)	Average diff. score using both (%)	Average diff. score where conversation was shorter than average (%)	Average diff. score where conversation was longer than average (%)
1377	5.8 ± 4.8	6.8 ± 9.7	8.4 ± 2.3	7.8 ± 2.7	8.1 ± 3.1
5225	(not available)	18.5 ± 6.9	(not available)	15.0 ± 16.0	19.5 ± 7.5
5769	6.4 ± 5.0	25.8 ± 7.7	17.5 ± 3.4	16.6 ± 3.7	20.2 ± 4.6
6107	48.6 ± 19.0	38.0 ± 14.7	32.4 ± 7.8	36.0 ± 9.2	36.0 ± 10.9
6128	29.0 ± 18.7	18.3 ± 7.8	17.4 ± 5.7	17.3 ± 5.9	20.0 ± 7.0
7147	13.3 ± 26.0	22.0 ± 9.7	23.6 ± 10.0	26.5 ± 9.6	19.0 ± 9.8

frequently by short questions and essays. Most assessment items in this research were written in terms of the "knowledge" or "comprehension" levels of Bloom's taxonomy. Meseke, Nefzinger, and Meseke (2010) indicate that "*It is the potential of collaborative testing to improve critical thinking that is most promising*." They propose further studies that could consider an analysis of assessment item depth. The third set of experiments described in this paper is a first step in this direction.

All of the previous studies have found that students have a more positive reaction toward *collaborative testing* compared to traditional testing. Many studies have demonstrated that collaborative testing reduces anxiety, improves satisfaction, engagement and motivation and is perceived as an effective learning activity (Pandey & Kapitanoff, 2011; Zimbardo, Butler, & Wolfe, 2003). In this research, psychological aspects have not been evaluated, but the responses to the surveys used to evaluate system performance and usability is congruent with these findings.

One of the major claims of *collaborative testing* research is that it improves test performance (Bjornsdottit, 2012). This increment is mostly defined as the difference between an individual pre-test and a group test taken afterward. Generally there is no dispute on this issue (Haberyan & Barnett, 2010; Sandahl, 2009; Simkin, 2005). Some authors claim that both high and low performing students do improve (Giuliodori, Lujan, & DiCarlo, 2008). Other authors have studied the influence of group composition (Jensen et al., 2011). In these studies the terms "low" and "high" performing refers to the comparative performance of all students in the class (that is dividing the whole class into two groups according to test results). Performance improvement is measured as the increment of scores between an initial individual test and an equal or equivalent collaborative test taken afterward. Of course it might be that low-performing students improve their scores more

Table 8
Use of the collaborative framework considering the final agreement as independent variable.

Test ID	Average agreement using just the "view answer" button (%)	Average agreement using just the "chat room" button (%)	Average agreement using both (%)	Average agreement where conversation was shorter than average (%)	Average agreement where conversation was longer than average (%)
1377	86.6 ± 8.2	90.9 ± 12.3	91.4 ± 2.7	90.9 ± 3.5	90.5 ± 3.8
5225	(not available)	82.1 ± 5.1	(not available)	73.6 ± 12.0	84.8 ± 5.5
5769	47.4 ± 11.2	84.0 ± 5.7	89.4 ± 2.4	81.3 ± 3.4	89.3 ± 3.3
6107	48.8 ± 15.5	53.1 ± 12.3	61.8 ± 7.2	60.0 ± 7.4	54.9 ± 9.2
6128	60.0 ± 19.6	72.1 ± 8.7	78.5 ± 5.5	71.1 ± 6.4	80.5 ± 6.4
7147	33.3 ± 65.3	79.3 ± 8.6	84.3 ± 6.9	82.2 ± 6.4	80.4 ± 7.9

Table 9
Summary of research questions for learning in the short-term and main conclusions.

Research question	Table	Summary of conclusions	Section
(Q8): Do students actually learn by using the Siette collaborative environment?	Table 11	Short-term learning is higher for those students that have used the collaborative framework ($p < 0.30$).	4.3.2 4.3.3
(Q9): Do low-level and/or high-level students improve their short-term knowledge by using the collaborative framework.	Table 12	Both groups increase their knowledge. Low-level students increase more than high-level students ($p < 0.59$).	4.3.2 4.3.3
(Q10): What is the composition of groups that makes collaboration more effective?	Table 12	The group composition is neutral for high-level students. Low-level students increments more in heterogeneous groups ($p < 0.10$).	4.3.2 4.3.3

Table 10
Description of the tests that were included in the third set of experiments.

Description	Date	#Groups	Group size	#Students	Total# of students
Experimental group A1	Nov 2008	9	2	18	78
Experimental group B1	Nov 2008	9	3	27	
Experimental group B2	Nov 2011	11	3	33	
Control group C1	Nov 2009	*(Individual test)*		52	88
Control group C2	Nov 2010	*(Individual test)*		36	

Table 11
Average scores obtained in the three sections of the test.

Description	#Students	Average initial score in the 1st section	Average initial score in the 2nd section	Average initial score in the 3rd section	Score increase from 1st to 2nd	Score increase from 1st to 3rd
Group A (Groups of 2)	18	40.28 ± 14.88	44.44 ± 13.90	51.85 ± 12.40	4.17 ± 20.36	11.57 ± 19.37
Group B (Groups of 3)	60	30.83 ± 8.50	40.21 ± 8.15	44.44 ± 7.46	9.38 ± 11.78	13.61 ± 11.32
Experimental group (A + B)	78	32.91 ± 8.51	41.19 ± 8.15	46.15 ± 6.41	9.38 ± 4.00	13.25 ± 9.79
Control group (classical test)	88	36.36 ± 7.11	35.71 ± 6.94	40.84 ± 6.41	−0.07 ± 9.94	4.47 ± 9.98

than high-performing students, simply because of the fact that they have “more room” for improvement. We also analyze the influence of the composition of the groups, with students of similar or different knowledge levels. As a general conclusion we have found that all students might increase their scores in *collaborative testing* just by knowing the partners’ responses.

There are some studies on the influence of collaborative testing in short-term learning or retention (mentioned in Sandahl, 2009). In some cases, short-term learning is confused with performance improvement (Rao et al., 2002). In our research, we distinguish between score improvement (previously called performance improvement) and short-term learning. Score improvement is clearly observed in all of our experiments (see section 4.2). Concerning learning in the short-term, we have found some evidence that indicates that most students benefit from *collaborative testing*. High-performing students improve by self-reflection, regardless the composition of the group, but low-performing students need to be in a group with higher-performing students in order to improve (see results of section 4.3, Table 10).

Even more controversial are the long-term learning effects of *collaborative testing*. Meseke, Bovec, & Gran, (2009) concludes that student performance is enhanced, but found no significant differences in a final exam. Some studies (Cortright, Collins, Rodenbaugh, & DiCarlo, 2003) observed that student retention of course content increases with the use of collaborative testing. This result is contradicted by Leight, Sanders, Clakins, and Withers (2012). Recently, Molsbee (2013) has reported a negative correlation between the use of collaborative testing marks and success in the subsequent courses of the formative program. In our opinion, there are many factors that contribute to the performance in a final exam and it is very difficult to isolate the influence of a given activity in long-term learning. In any case this issue is beyond of the scope of this paper and will require a new set of experiments.

Another topic of interest is whether or not *collaborative testing* requires students to reach a consensus in their answer, and which one is the best policy. Bjornsdottit (2012) concludes that there is no significant difference in the results obtained enabling or disabling this condition. In our research, consensus is not requested. However we have discovered that students tend to reach a common answer in most cases and that the consensus highly correlates with the correct response.

A common problem in *collaborative assessment* and *collaborative testing* is the assignment of grades to individuals that have been evaluated in a group. This problem leads to equality and fairness issues (Sharp, 2006). The system proposed in this paper mitigates this problem because there is no group answer. Two individual answers are provided for each question, before and after collaboration. Each student is responsible for his own answers. However, the score obtained from the first answers can be viewed as purely individual, but the second can additionally be viewed as a collaborative or group response. Although it is also an individual response, it has been influenced by the collaboration. The problem of combining both scores remains. A common solution to this problem is to obtain a weighted average of both (Zipp, 2007). This is the strategy that we follow in practice.

3. The collaborative environment of Siette

Siette is a web-based environment for managing and administering computerized tests. The system allows test construction, delivery and results analysis. Siette also includes a built-in framework where students can collaborate with their colleagues while answering test questions. For this purpose, a middleware layer has been placed over the basic interface of Siette, incorporating elements that facilitate the synchronization and collaboration among group members. Siette and its collaborative environment are domain independent. Thus, any test defined in this environment can be taken collaboratively or individually.

Briefly speaking, a collaborative learning is a type of assessment where two or more examinees answer the same questions, sharing information before sending their answers. The main idea behind a collaborative testing is to enhance the potential of assessment as a learning tool, originating an *assessment for learning* environment. However, there are many ways in which collaborative testing can be

Table 12
Analysis of groups composition.

	Score increase in S-groups	Score increase in D-groups	Total
L-students	2.95 ± 22.19	30.98 ± 25.71	15.52 ± 16.81
H-students	9.90 ± 20.75	8.65 ± 18.78	9.33 ± 14.81
TOTAL	6.63 ± 13.53	20.54 ± 13.40	13.25 ± 9.79

implemented. In the next subsections we are going to explain the features of the Siette environment, according to the natural stages that will have to be followed in order to use this environment collaboratively. Readers interested in more information about Siette can obtain it in previous publications (e.g. Conejo et al., 2004) and in the wiki pages of its documentation (http://www.siette.org).

3.1. Defining a collaborative testing

Siette includes an authoring tool for teachers, which is used to define questions, also called *items* in the assessment literature. This authoring tool allows defining different types of questions, such as multiple-choice, multiple responses, and even open answer questions automatically evaluated using regular expressions patterns and other techniques. Additionally, Siette allows the dynamic generation of isomorphic items from templates. That is, a teacher can define a question template that is instantiated dynamically when the question is posed to a student. Siette can be used as an “assessment for learning” environment (Guzmán, Conejo, & Pérez-de-la-Cruz, 2007a) since its questions could optionally include hints and feedback. Hints are explanations given before the question is answered, and feedbacks are messages that are presented after the answer is given. Regarding the collaboration issues of Siette, there is no limit as to the type of questions that can participate in a collaborative testing session, but some features should be limited, as will be shown later.

Questions are grouped in courses and, inside each course, they can be structured into a hierarchy of concepts. Each course has therefore its own question bank. After creating the question bank, the teacher should define a *test* or assessment session specification. A test definition includes among other characteristics:

1. The subset of questions from the question bank that could be administered to the students. Commonly, this is done by restricting issues such as the concepts involved in the test, the number of times the question has already been posed before, etc.
2. The question selection criterion, that is, the strategy to select dynamically the next question to be posed to a student during an assessment session. Siette allows predefined ordered tests, where all questions are presented in the same order to all of the students; randomized tests, where questions are selected randomly; and adaptive tests where questions are selected according to statistical criteria whose goal is to get the student knowledge estimation faster and more accurately.
3. The finalization criterion, e.g. the maximum number of questions to be posed, requested by the student himself/herself or, in adaptive testing mode, the accuracy threshold of the student knowledge estimation.
4. The assessment criterion. There are three main alternatives: the classical percentage of correct answers, the score obtained in the test (correct answers give positive scores and incorrect answers could also penalize), and a statistical criterion based on Item Response Theory. For more information, see (Guzmán, Conejo, & Pérez-de-la-Cruz, 2007b).
5. The display options, including the number of questions to be presented at the same time, and whether to show correct response after the student answer or at the end of the session; whether or not to allow the student to go back after answering a question; hints and feedback control, etc.
6. The access control; defining who can take the test and when it can be taken.
7. The assessment criteria. A teacher should select whether or not a test could be taken in a collaborative mode and, if so, he/she should define the criteria for group composition. Fig. 1 shows an image of the teacher interface used to define test criteria.

3.2. Creating groups

Assuming that a teacher has created and configured a test to be taken collaboratively, he/she should proceed with the creation of students' groups. This stage can be carried out in three different ways: (1) Manually, i.e. the teacher creates a set of groups and assigns the students to each group. (2) Randomly, that is, the teacher specifies the number of students that a group should contain. At run time, students are assigned to a group in the order in which they arrive. (3) The students define groups on their own at run time.

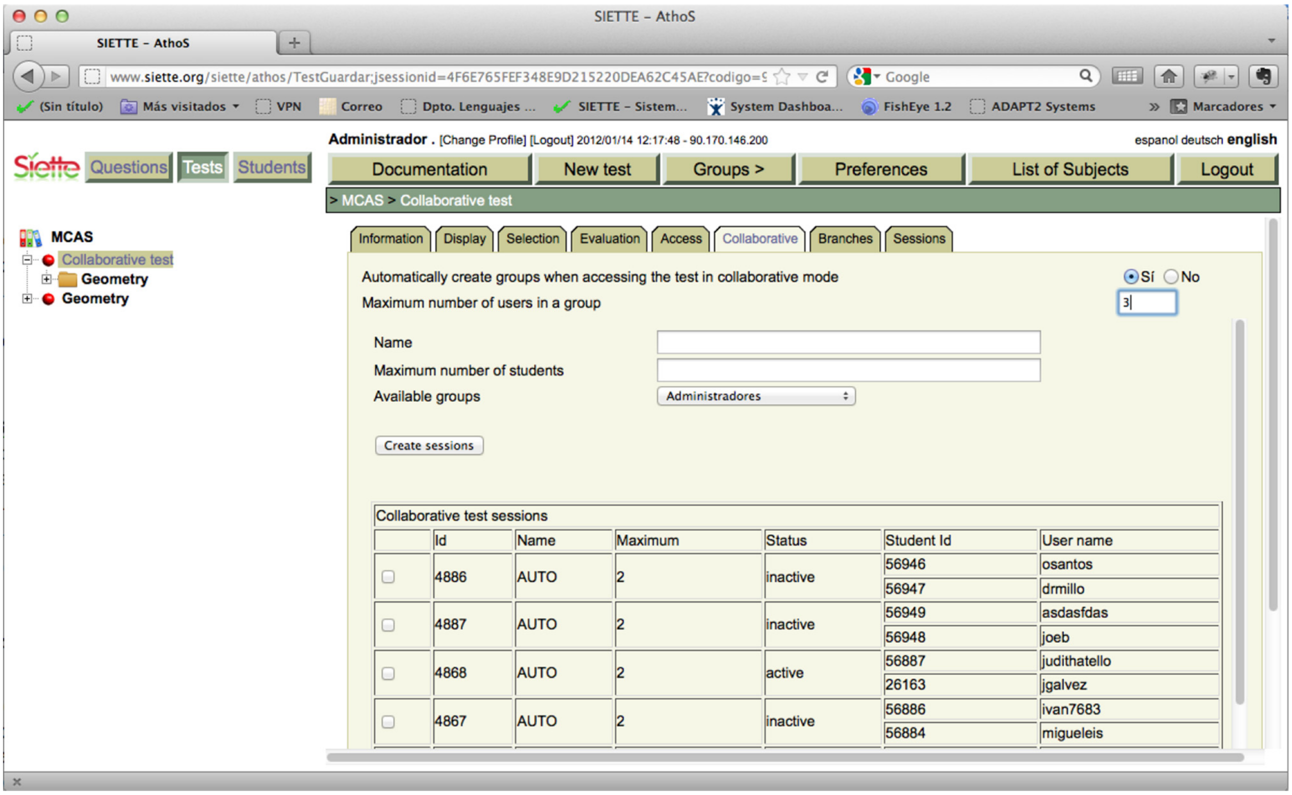
In order to take a collaborative test, a student has to log into Siette, choose the course and the test, and finally select the *collaborative mode* (if this mode is allowed by the teacher). After that, the student enters into the so-called *Collaborative Testing Hall* (also known as the *waiting room*), i.e. an initial *collaborative frame* that deals with group creation. The behavior of the system when the student enters the *Hall* differs depending on the group composition criteria:

Case 1: If the groups were predefined manually, when a student logs into the *Hall*, he/she will be directly assigned to the group that he/she belongs. Once all students in the group have entered, the test will start automatically.
Case 2: This is a simpler approach. The teacher has specified the number of students per group and thus, groups have been created automatically. In this case, students are assigned to groups in the order they entered into the *Hall*. This behavior produces a *pseudo-random* assignment.
Case 3: In this situation, either students or the teacher create the groups. The mechanism followed is similar to those used in web-based game platforms to meet players. Any user can initiate a group and define the maximum number of participants. Each participant should decide which group to join. When the group is completed, the collaborative test will start.

Furthermore, if the teacher enters into the *Hall*, he/she can create dynamically groups for all the students that were currently at the *Hall.* This option assigns students to groups randomly and is very convenient because it leverages the student for the cognitive overload of creating a group and selecting partners. Furthermore, this option guarantees random assignment.

3.3. Taking the collaborative test

Although Siette gives several question selection criteria, in a collaborative test the sequence of questions is always the same for all the members of a group, independently of the selection criterion of that test. The goal is to make the collaboration among the group members



Administrador . [Change Profile] [Logout] 2012/01/14 12:17:48 - 90.170.146.200

> MCAS > Collaborative test

Automatically create groups when accessing the test in collaborative mode ● Sí ○ No

Maximum number of users in a group 3

Name

Maximum number of students

Available groups Administradores

Collaborative test sessions

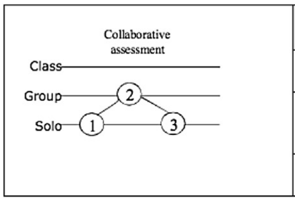
	Id	Name	Maximum	Status	Student Id	User name
□	4886	AUTO	2	inactive	56946 56947	osantos drmillo
□	4887	AUTO	2	inactive	56949 56948	asdasfdas joeb
□	4868	AUTO	2	active	56887 26163	judithatello jgalvez
□	4867	AUTO	2	inactive	56886 56884	ivan7683 migueleis

Fig. 1. Teacher interface used to define test criteria.

feasible. Once a question is posed to all the students of a group, a collaborative script in three phases is defined (see Fig. 2), following the representation proposal of Dillenbourg and Hong (2008):

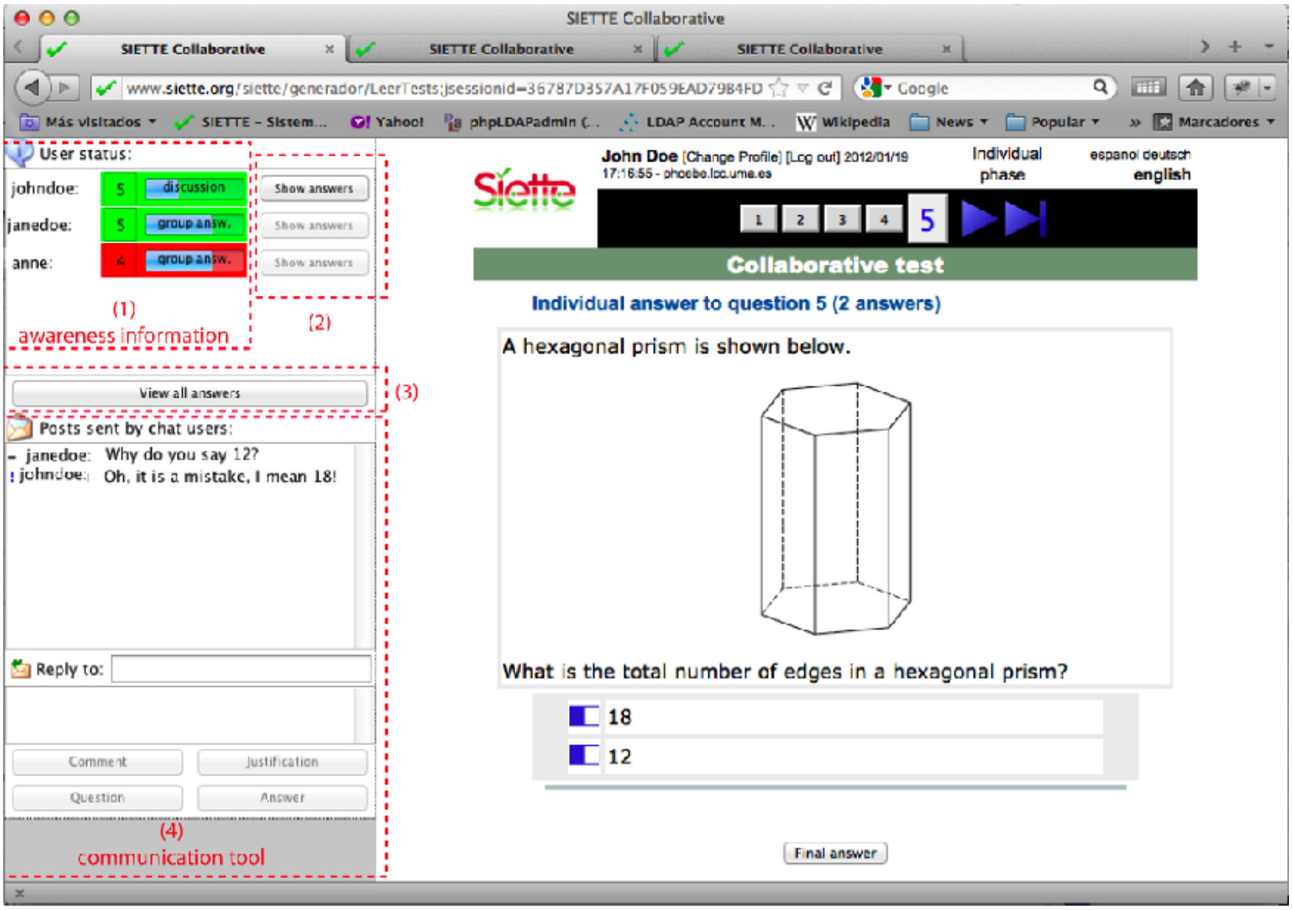
- *Initial response*: First of all, each student in a group gives their own answer to the current question. In this phase, test conditions are exactly the same as in a conventional test, since the collaboration frame is disabled and the students cannot interact with another.
- *Discussion phase*: The members of a group can exchange their answers and discuss a given question. For this purpose, the *collaborative frame* provides the user with various elements: (1) Awareness information, which indicates the status of other group members and also facilitates synchronization. This information is available throughout the entire answering process. (2) A "*View answer*" button located beside each member's nickname, showing his/her individual response when clicked. (3) A button "*View all answers*" showing the distribution of responses within the group. (4) A specially designed chat tool that allows students to argue with their colleagues by posting different types of messages. The content of this chat tool is reset before each question is posed and does not allow students to exchange messages if they are not at the same question. Messages can be additionally annotated with a *performative* (*Comment, Question, Answer, Justification*) in the sense of speech acts.
- *Final response*: After the discussion phase, the students are requested to give the final answer to the question. This final answer may be the same or different from their initial response. It is not compulsory to reach an agreement; each member of a group has to make their own decision. However, evidence shows that agreement is often reached.

As an example of the user interface, Fig. 3 shows a collaborative test on Geometry taken by three students. Two students (*johndoe* and *jannedoe*) are answering the 5th question, as can be seen in the awareness frame (in the upper left of Fig. 3), and have exchanged some messages using the communication tool (lower left of Fig. 3). The first student (*johndoe*) has clicked on "View all answers", and this information is shown in the right frame (right of Fig. 3). The second student (*janedoe*) has already decided to send the final answer, and the third one (*anne*) is still answering the 4th question.



PHASE	LEVEL	ACTIVITY
1	Individual	Individual initial response to question
2	Group	Show results and argumentative discussion about responses
3	Individual	Individual final response to same question

Fig. 2. The collaborative testing script.



janedoe: Why do you say 12?
johndoe: Oh, it is a mistake, I mean 18!

Individual answer to question 5 (2 answers)

A hexagonal prism is shown below.

What is the total number of edges in a hexagonal prism?

18

12

Fig. 3. The collaborative testing environment interface.

It is important to notice that the awareness information only indicates the progress status of each student of the group in the collaborative test. Each row of this information contains data from one member of the group. More concretely, it contains the question number and the collaborative script phase, changing from green to red or blue, if the other students are at the same question, before, or after the user's current question. Each member, at each question, can freely decide whether or not to collaborate with the other members of the group. That is, individuals can go ahead in the test without waiting for their colleagues to finish it. However, we have observed that most students do synchronize with each other.

Siette gathers all the information generated in the three script phases for posterior analysis. This includes the score obtained in the initial and final responses; the requests made by students to see their partners' answers, and all the messages exchanged through the communication tool. In order to analyze the collaboration activity within a group, some indicators can be obtained from that collected information. From an individual perspective, we have considered three indicators: (1) The number of times a student has requested to have a look at their colleagues' answer; (2) the number of messages a student has written; and (3) the total length of messages written. From the group perspective, similar indicators have been considered: (1) the total number of requests for other group members' answers; (2) the total number of messages in the forum; and (3) the total length of messages in the forum. Even though off-topic conversations during collaboration could take place and, consequently, they could partially disturb the experiments described in this paper, manual reviews of the chat messages concluded that almost all conversations were focused on the questions.

4. The experiments

In the previous section we have presented the system we have implemented. The initial motivation was to combine the Siette assessment environment with a new collaborative learning framework. According to previous research results, assessment can be used as a tool for learning. On the other hand, collaboration activities were also a promising way to improve learning. We hypothesize that a combination of both approaches would be beneficial for higher education students. In order to prove it we follow an underlying positivistic research philosophy and the deductive method. We have designed a set of experiments under controlled conditions to examine the validity of our hypothesis based on a quantitative approach. This section summarizes the mayor findings.

Our initial purpose was just exploratory, that is, to observe the students while they use the system in order to verify its performance, help us to improve it, and to develop new ideas. The main source of information for this task is the students' opinions collected by anonymous questionnaires that included close and open questions. One of the problems we faced was that the system was designed as a domain independent tool. Therefore the system target user population might have different expertise with computers, computer-based testing, or communication using chat; and that could affect the results. We addressed that problem selecting different sets of users, two of them composed by last year students of the School of Telecommunication Engineering at the University of Málaga, and one composed by first-year students at the Forestry School (the Polytechnic University of Madrid). The results are presented in section 4.1.

A second set of experiments was designed with a descriptive and explanatory purpose. Our aim was to find out the functional relationships between the quantitative variables involved in the collaborative testing. We were mainly interested in the relationship between score improvement and other measures like the length of communications, the degree of consensus in the students' answers, etc. In this case all data were collected automatically by the computer system. To increase motivation, students were told that the score obtained would be used as part of their grades. The system was originally designed for unsupervised use, that is, it is supposed that the students might use the system from their home. However, there were two problems with this approach: the environment conditions at home cannot be controlled (the students might communicate using other means, or use additional resources to solve the questions); and secondly, some students complaint that they have no computer connection from home. To avoid these problems all experiments in this section were carried out at the University computer laboratories. In order to verify the results we have repeated the same experiment in six different groups. The results are presented in section 4.2.

Finally, an interesting research question arose from previous experiments. *Do students really learn by using the system?* The experimental design in this case was more difficult to achieve. Long-term learning in a real environment is difficult to isolate. There are many variables that cannot be controlled, many different resources that contribute. As a result, we decided to study short-term learning. On the other hand, our students and we are always concerned with equity issues. If a measurement method is beneficial for some students it should be offered to the whole group. To solve that issue we have split the experiment in different courses and we offered the same evaluation option to all students in the same course. The whole experiment took four years to complete. The results are presented in section 4.3.

The analysis of data was carried out following a classical statistical approach. For all data in the following tables we have constructed the 95% confidence interval. To check whether the value is statistically significant at 95% confidence, applying the Student *t*-test, it is only needed to explore if the null hypothesis is excluded, that is if the interval excludes the zero value. In all tables data are presented in the form of sample means, plus/minus 95% confidence interval. Confidence intervals of the differences are not shown in tables but can easily be obtained from given values.

The results obtained with these experiments are bound to the sample conditions we have had. We are aware that the sample population has some bias intrinsically. We are only selected students of two engineering disciplines in two Spanish Universities. Can these results be extrapolated to the world higher education population of different disciplines? That question goes beyond the scope of this work. Nonetheless, the system is publicly available through the Web and we kindly invite other researchers to reproduce our results.

4.1. System evaluation

The main aim of this set of experiments was to improve the system performance and usability, and discover potential weakness or misbehaviors. The system was tested with different groups of students that took the collaborative test as a formative activity. The results of these experiments provided feedback from the users and some guidelines for improvement. The score obtained by students was not taken into account for their final evaluation. We have included this section in order to show what are the main difficulties we have faced, what we have learnt from our errors, and how we have solved them.

The first version of the Siette collaborative environment was released around March 2004. After that, another two new versions were released. During these years, the implementation has evolved by fixing some bugs, adding new features and improving the system's usability, according to the system evaluation process. However, the main features of the collaborative testing environment described in the previous section remain unchanged. Fig. 3 shows on the left the collaborative frame corresponding to the current version of the collaborative framework. Fig. 4 shows the two previous prototypes of this framework. The assessment framework (right part of Fig. 3) was almost the same in all prototypes.

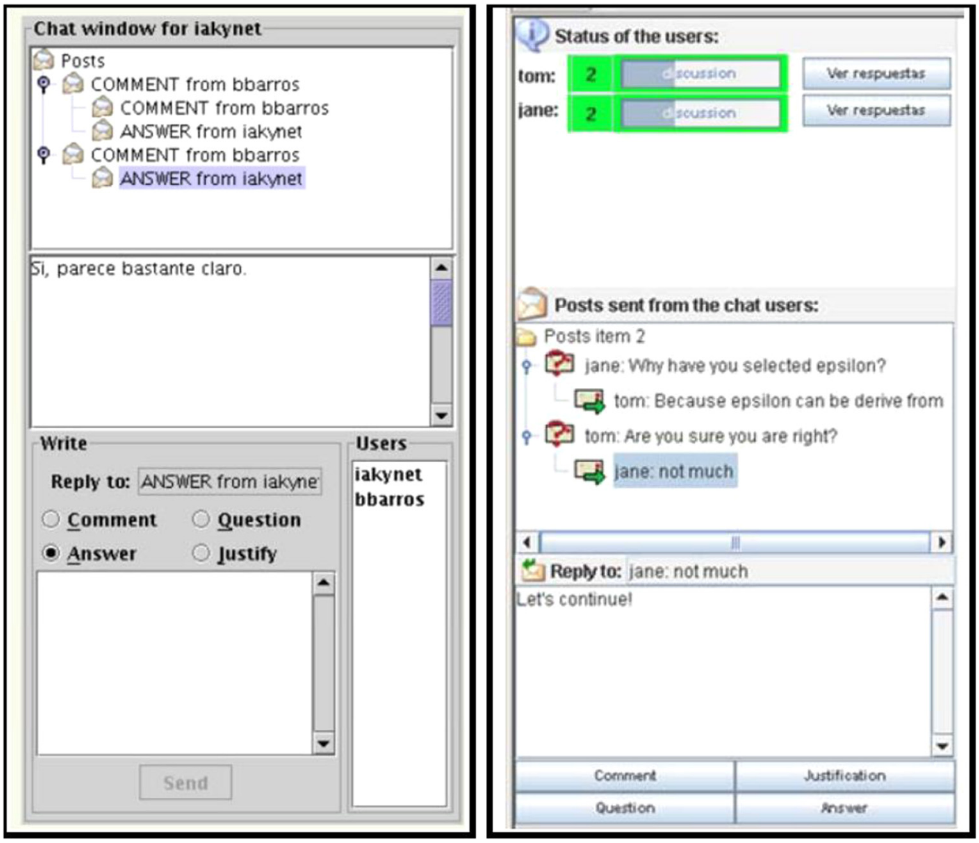
4.1.1. First prototype evaluation

In this first prototype (see Fig. 4 left), the chat tool was structured as a tree, but instead of displaying the message content directly, the tree simply displayed a message header. When the user clicked on the header, the message body was displayed in the panel below. In order to send a message, the user should have to select the previous message he wanted to answer (from the message tree), categorize his own message as "*Comment*", "*Question*", "*Answer*" or "*Justification*" (by using a radio button), write down the message body, and submit it. The student could also query other users' responses, by clicking on the nickname that appears on the bottom right panel (labeled with the text "*Users*"), but this prototype did have an awareness information panel yet.

To evaluate this prototype, we conducted an experiment with a small group of 12 M.Sc. and Ph.D. students from the Computer Science Schools at the University of Malaga and the Polytechnic of Madrid. The experiment involved a collaborative test about English grammar. Students were organized in groups of two people: one located at Malaga and the other at Madrid. This experiment was not part of any course and was mainly design for software testing. Even though the experience was not very fruitful for us, from the perspective of the collaboration, it gave us very useful data from the software testing point of view, since it addressed some non-qualitative information that we took into account carefully in the development of the following prototypes.

First of all, the students participating in that experiment found it difficult to synchronize with their partners. They used the chat room for that, but they were continuously asking one another about their current question number. The delay in the message exchange contributed to the overall mess and, once the first three or four questions were posed, the partners disengaged and finished the test on their own. This fact indicated the importance of awareness and establishing a desired goal: *Each user should know where he/she is and where his/her partners are*.

The second blocking problem found in the first prototype was related to the system performance. Siette had been previously tested under different user conditions, and proved to be stable with more than a hundred students taking tests simultaneously. However, the performance of the collaborative frame had not been tested before under real conditions. The system had a problem of scalability, and once the third or fourth group tried to begin a new session, the whole system crashed. We analyzed the problem and realized that the most time-consuming feature was the log of the users' actions. Logs were written in a separate XML file in the server side, and included a large set of



Chat window for iakynet

Posts
- COMMENT from bbarros
  - COMMENT from bbarros
  - ANSWER from iakynet
- COMMENT from bbarros
  - ANSWER from iakynet

Si, parece bastante claro.

Reply to: ANSWER from iakyne

○ Comment ○ Question
● Answer ○ Justify

Users: iakynet, bbarros

Status of the users:

tom: 2 discussion
jane: 2 discussion

Posts sent from the chat users:

Posts item 2
- jane: Why have you selected epsilon?
  - tom: Because epsilon can be derive from
- tom: Are you sure you are right?
  - jane: not much

Reply to: jane: not much

Let's continue!

First Prototype

Second Prototype

Fig. 4. Screenshots of the two previous collaborative frameworks.

information for further studies. To solve this problem, we adopted some realistic criteria: *Do not register unnecessary data, and replace XML files by a database as a persistence mechanism.*

The observation of the users' behavior during the test, revealed a third problem. The system assumed that the student had to provide an individual answer first. After that, using the chat room, he/she should discuss the solution to the current test question with the partner(s), and then provide the final answer. This cycle was repeated in the next questions. Tests in Siette can be configured to show the correct answer just after the student's response. A pair of students cheated the system by avoiding synchronization. That is, one student was answering question n, while his partner was answering question $n + 1$, and the second sent a message to the first through the chat room, giving the correct answer to question n. This problem generated a new system requirement: *Forbid users' communication when they are not answering the same question.*

4.1.2. Second and third prototype evaluation

Due to the many problems that arose during the experiment with the first prototype, we were not able to carry out a suitable formal evaluation. For this reason, we had to redesign and recode several parts of the software, taking into account the lesson learned during the evaluation of the first prototype. The second prototype was released in October 2006. It solved the main technical problems found in the previous prototype, and included the awareness features. It was the first fully operative prototype, but it still had some scalability problem that we will discuss later. The third prototype was released in March 2007, and included some minor changes compared to the second one, but it was significantly more stable and useful.

In order to complement the evaluation of both prototypes, we defined a questionnaire to be distributed to students after taking a collaborative test. The questionnaire was divided in four sections:

(1). **About the user.** The first section asked about the user's computer literacy, whether or not they had previous experience with computer-based testing and/or with chat tools, and if they liked them.

(2) **About the activity.** We asked if the activity was clear, if they enjoyed taking collaborative tests, and if they thought they had learnt from their peers. Also we wanted to know if they considered it useful to see their peers' responses, and whether they appreciated the communication with their partners.

(3) **About the system.** We asked about the system usability, its performance, layout and font sizes, etc. We also included four questions about the awareness, asking the users if they had easily found the chat room, the answer of their partners, their current situation (that is, the question of the test they are currently answering) and the situation of the others. We also asked about the use of the structured chat and the chat annotation buttons. Finally we asked if they preferred using Siette with or without the collaborative environment, and requested an overall rating of the system.

(4) **General comments.** We included some open questions about different components of the system, to encourage students' comments and proposals, and asked them to identify the advantages and disadvantages of collaborative testing with Siette.

All questions, but the general comments, were rated in a Likert scale from 1 to 5. The questionnaire was voluntary and anonymous. We gave the same questionnaire to three different student populations (J1, J2 and B) after they had taken the collaborative test. The first questionnaire was filled by the J1 group in December 2006 (second prototype). The second was filled by the J2 group in November 2007 (third prototype). In both cases, the sample came from undergraduate students of the last year of the School of Telecommunication Engineering at the University of Málaga. The test was about Java Programming Fundamentals. A third questionnaire (group B) was scheduled for January 2008, for first-year students at the Forestry School (the Polytechnic University of Madrid). The test topic was Botany. In all these three cases, the students had some previous experience with Siette, but it was the first time they used the collaborative frame. Table 1 shows the means (±confidence interval at 95%) of the most relevant results.

In all cases, the collaborative tests took between 1 and 2 h and consist in 20–30 questions of different formats including multiple choice, single and multiple answer and fill-in-the-blanks. However, the scores obtained were not taken into account for the summative evaluation of the students.

Regarding the open questions, in the group J1, most students (70% of the sample) pointed out that the main disadvantage of the system was its slowness. However, in group J2, nobody complained about this issue. Concerning the strengths, only 9% of group J1 considered the system as a useful tool for improving their learning. This percentage was increased in sample J2 (14%). The main strength remarked by the students of both samples was the collaboration. Students highly valued this feature (39% in J1 and 47% in J2) since, according to their words, it helps them to resolve their doubts, focusing on new issues not learned before, and learning new concepts, etc.

As can be seen in Table 1, the second prototype was evaluated very positively (question 4.1), with a mean of 4.06. Nonetheless, some students had problems with the system and results were not so good in terms of the system's performance (question 3.1). The students' comments about the difficulties they found with the second prototype highlighted a new problem. Let us imagine a test session in which three students are taking a collaborative test. Suppose that one of them has a computer problem or a network problem and is disconnected from the group. That student looses his/her session and has to re-enter into the system again. The other two continue the test, but the student is unable to connect with the group again. Therefore, in the third prototype we added a new requirement for future versions: *Keep track of student's actions and give him/her the chance of recovering the session if necessary*.

The information provided by the students in the questionnaire also revealed that they were not happy with the initial design of a threaded chat interface with additional annotations of the user's actions (questions 3.5 and 3.6). Most users commented that they preferred a flat chat, which they found easier and faster to use. Thus, for the third prototype we replaced the original tree-structured dialog to *simplify the user interface according to user preferences*.

We also noticed a problem with the system performance when all members of a group connected simultaneously. That caused some delays in the chat messages that were also commented on by some users. That was due to the architecture of the communication tool that was based on the server side. That decision was initially taken for two reasons: (1) we wanted to avoid firewalls using only communications through HTTP tunneling, and (2) we wanted to register the user's dialogs during the test for further analysis. We realized that the technical solution we adopted was not optimal, and requires a complete reimplementation of important parts of the system. We have postponed that task for future prototypes, but we have discovered what we should do: *Try to improve and measure the system performance*.

The results from the questionnaires of December 2006 and November 2007 were applied to the same population, and gave similar results i.e. statistically significant difference for four questions (questions 3.1, 3.3, 3.4 and 4.1) according to the classical hypothesis test with $p \leq 0.05$. Next, we are going to focus on these questions.

Question 3.1 measured indirectly the system performance and indicated a statistically significant improvement on it. Third prototype reduced the network traffic to a minimum, but still keeping the centralized architecture. Questions 3.3 and 3.4 measured the system usability that was also improved as a consequence of a better system performance. Finally, the overall rating of the activity (question 4.1) increased from 4.06 to 4.4.

The third experiment (see Table 1, column B) was conducted with a different population, and under different circumstances. The students had no background in computers and the test location was away from the server site. The system was evaluated positively (question 4.1), with a mean of 3.83, but lower than in the first population, and the system performance also was good. However, regarding the usability issues, i.e. questions 3.2, 3.3, 3.4 and 3.5, they were as satisfactory as the results of the J2 group, whose members also evaluated the third prototype. With respect to the open questions and comments of this group, 33% explicitly declared that the system did not need any improvement, and the rest made no comment. Asked about the advantages and disadvantages, 44% of the students in this sample listed collaboration as the main strength of the system. Some of them, 22%, explicitly said that it is good as a learning tool, and 16% pointed out that it is very time-consuming. Finally, this third session also brought about/pointed to a relevant fact. At the beginning, before starting the collaborative test, the students were instructed about the use of the system like in the former experiments. Nevertheless, the students' groups were not predefined by the teacher, as had been the case in the previous experiments, and the group initialization took longer than expected. For the next release, we plan to improve collaborative group creation and (optionally) provide the option of leaving the waiting room. In addition, at the beginning of the collaborative test, some students got confused when the chat tool was disabled in the *initial response* phase. This confusion was overcome in all cases after the second question. A lesson learned is that: *A training session with 2–3 questions would be useful for beginners.*

Finally, we have to mention that in the three samples, there was no clear correlation between any question of the questionnaire and the final rating (question 4.1). The highest is a positive, but not significant correlation (0.58) with system apparent performance (question 3.1).

4.2. Measuring the impact of collaboration in test results

A second set of experiments was designed to measure the effect of collaboration on the test results. With this set of experiments we would like to answer the questions listed in Table 2.

We have included in this study the data collected in different tests of three different subject matters taken from 2006 to 2008 (see Table 3). First column of Table 3 contains the internal ID of the test in the Siette system. Second one is the name of the test (which also includes the subject matter). Third column shows the date of the experiment. Fourth and fifth columns show, respectively, the size of the student sample and the number of questions involved in the test. Note also that we have excluded from this study the tests with less than 15 students and two others where the Siette log system failed. All tests were taken collaboratively by pairs of students, except for the one with ID 6107, which was taken by groups of three students. In all cases, students were randomly assigned to a group. The individuals were undergraduate students (M. Sc. in Computer Science and M. Sc. in Forestry Engineering) in the second, third or fourth year of their studies, depending on the degree. They performed the experiment in the labs of their university, but appropriately distributed (mainly in different rooms) in order to ensure that the communication was only computer-supported. Using the students' performance data, we carried out several studies, as will be explained in the next subsections.

In all cases, the collaborative tests took between 1 and 2 h and consist in 15–25 questions of different formats including multiple choice, single and multiple answer and open questions automatically corrected by means of regular expressions matching. See Siette description for the types of questions that can be used and the scoring procedure (Conejo et al., 2004). The collaborative test was included as a formative activity during the course. The scores obtained in the collaborative test were averaged with the scores obtained in other formative activities like programming assignments or individual tests. To sum up, the collaborative test activity represent between the 10% and 20% of the final score of the formative evaluation. The final grade of the students also includes an exam. So the actual effect of the collaborative test in the summative evaluation is still marginal.

It is important to notice that two individual test scores compose the final score of the collaborative test: the score obtained from the initial responses and the score obtained from the final (after collaboration) responses. There are different options to score the activity as a whole in order to include it as a component of the final grade. In general, we have adopted eclectically a 50%–50% scheme that in our opinion is a good compromise between the pure individual evaluation and the group evaluation. In some cases, as mentioned in the corresponding experiments, we have also used a scheme of 0%–100% assignment to emphasize the collaborative work. Zipp (2007) proposed a 25%–75% scheme in a similar situation. However, these issues are beyond our research scope at the moment.

4.2.1. Study about the impact of using the collaborative environment in the test final score

First of all, we studied the *impact of using the collaborative environment on the results obtained in the tests (Q1).* As explained before in Section 3.3, in collaborative tests each student has to provide two answers to the same question. The initial answer is given before the discussion and without knowing their partner's answer, and the final answer after the discussion process. Data are summarized in Table 4. Note that in Table 4 the total number of questions answered may be less than the product of the students' number multiply by the questions number (Table 4), since some students did not complete the test. Each question was scored with a maximum of 1 point; therefore the average of question scores is equivalent to the average scores obtained by all students.

As can be seen in Table 4, the average of final answer scores is always higher than the average of the initial answer score, and this result is statistically very significant ($p << 0.05$), independently of the difficulty of the test, which is proportional to the average of initial or final answers scores. This result is consistent with previous findings of *collaborative testing* research (Haberyan & Barnett, 2010; Sandahl, 2009; Simkin, 2005).

The percentage of initial agreement seems to be related to the difficulty of the tests, but the final agreement is similar in most cases. The lowest value corresponds to the tests taken by groups of three students, where a consensus is more difficult to be reached.

The results in Table 4 can be explained by different factors. At the beginning of the test, students were told that their actual grades would be based on their final answers, without taking into account their initial answer. Consequently, the increase in the result might be explained either by the process of self-reflection, or because they paid less attention to the first answer.

4.2.2. Study about the impact of using the collaborative framework

We were also interested in the effect of the collaborative framework to test whether or not the increase of the final score is due to the use of the collaborative framework (*Q2*). Accordingly, we analyzed separately those questions in which the student used this tool and those where the student did not.

Table 5 shows that most students used the collaboration framework in each question before giving their final answers (70.2% of all cases). Concerning the absolute scores obtained by students, data do not indicate a clear relationship between the final score obtained and the use of the collaborative framework (comparing values in columns "Average Final score"). In general, the average of the final score obtained in questions, where the communication tool was used, was higher than in those where this tool was not used. However, sometimes the opposite situation occurs. The explanation of this fact could be the following: Easier questions might have a higher score rate, and require less discussion, or perhaps students with higher scores might not be willing to spend their time using the chat tool. The most significant variable is the difference between the score obtained according to the final and to the initial answers (comparing values in columns "Average Diff score"). This variable is relatively independent of the student's knowledge level or the question difficulty, and indicates an improvement in the result obtained. Evidence suggests a positive difference in both cases, but in those situations where the students used the collaborative framework, the difference is higher (the results are statistically very significant with $p << 0.05$ in all cases).

A second dimension to consider in this study is the degree to which the final answers get to a consensus, (Q3) the correlation with the use of the collaborative framework (Q4), and the difference between the final and initial scores (Q5). Table 6 shows the results. For all cases, there was a clear relationship between the use of the collaborative framework and the final agreement, and a positive correlation between this and the difference between final and initial scores. That is to say, those students who have the tool were more likely to reach an agreement in their final answers and are, therefore, more likely to improve their results. These results are statistically significant in all cases.

These results do not contradict those obtained by Bjornsdottit (2012), that concludes that a compulsory consensus requirement is not an important feature in *collaborative testing*. In fact, according to our data, the consensus is reached in most cases, although in our system it is always optional.

We have also explored how some features of the collaboration framework might influence the difference between the final and initial scores (*Q6*). As mentioned before, the collaborative framework has two main facilities: The awareness information panel, and the chat room

(Fig. 3). We have studied the use of both facilities and the consequences in the measured variable. Results are displayed in Table 7, and should be compared with the corresponding columns of Table 5. According to these data, we can conclude that using just the "view answer" button (of the awareness information panel) or the chat tool does not always lead to a better score. The best results are obtained when using both facilities. However, these results are not fully statistically significant for all tests because there are not enough data in some cases (most students have used both facilities).

Additionally, we have *explored the influence number and the length of the messages in the chat for each question (Q7)* (see right part of Table 7). In this sense, we have found no significant differences in the improvement obtained in questions where a short or long discussion took place. These experiments suggest that the students use the chat tool as much as they needed. May be some questions require a longer discussion, whereas others may be answered just by looking at and reflecting upon the partners' answers. There is also no clear correlation between the initial agreement and the number and length of messages in the chat. This issue is probably related to the question difficulty, or to the agreement in the initial answer.

Table 8 shows similar data to Table 7 but considering the final agreement as the independent variable. The main conclusion is that the agreement is more often achieved when both tools are used, and using only the "view button" does not guarantee achieving a higher agreement, but using the chat does. The best results are clearly obtained when using both facilities. Regarding the length of the discussion, there is no clear correlation with the final agreement.

The manual analysis of the chat log files does not contribute to arise any new hypothesis about the influence of the discussion in the score improvement, nor on the degree of final agreement. However it reveals that almost all (above 99%) of the messages exchanged were about the current question that was posed. There were almost no off-topic conversations. This fact is considered to be very positive, because it is one of the main concerns in CSCL systems as a prerequisite for effective learning. In our opinion the assessment environment is responsible of this behavior. Students felt that they do not have time to loose while making an assessment and so they focus on the subject.

4.3. *Measuring learning in the short-term*

The results of the previous studies cannot conclusively prove actual learning gain for a student using the collaborative testing environment. Perhaps it is due to the fact that students have more time to think about their second answer, and the question review process itself could explain the increase in the final score. In other words, an alternative hypothesis could be that the longer the time to think about the answer, the higher the student's score. Therefore, some important research questions still remain unanswered (Table 9) and will be studied in this section.

4.3.1. *Hypothesis*

Our hypothesis is that actual learning occurs while taking the test, by means of collaboration with other colleagues. This hypothesis relies on previous findings in the fields of Computer Supported Collaborative Learning (CSCL) (Stahl, Koschmann, & Suthers, 2006), that are consistent with the constructivist theory of learning that argues that interactive activities where learners play active roles are more effective than activities where learners are passive. However, we have designed a set of experiment to evaluate the hypothesis in this case.

To perform the experiment described below, we have taken advantage of the fact that the tool we want to analyze is an assessment system and students are being continuously assessed during their interaction with the tool. The classical approach to system evaluation consists of the following sequence: pre-test, "treatment", and post-test. In order to measure an increment in the knowledge level, it should be guaranteed that the pre-test and post-test scores are comparable, i.e. both tests should have exactly the same difficulty level. This is difficult to achieve during test designing and construction, and thus an alternative approach is usually adopted: splitting students in two testing groups (plus a control group), by random selection, and alternating the use of the two tests as a pre-test or as a post-test.

4.3.2. *Study design*

For this experiment, we administrated the same test several times in two different ways, i.e. collaboratively and individually, for different student samples. The test contained 25 questions about the LEX tool, used to build a Lexical Analyzer in the context of Compiler Construction. In particular, we focused on an example of a LEX code where some mistakes were introduced and asked the students to debug the code by true/false questions (such as "*Code in line 7 is incorrect, it should be replaced by*…" or "*Swapping lines 11 and 12 has no effect*…", etc.). Questions were independent of each other and they were presented randomly in each test. For each question the score was $+1$, or -1, if the right or wrong option was selected respectively. Alternatively, students could leave the question unanswered, without being penalized.

The same set of questions was posed to undergraduate students of Computer Science Engineering as a practice activity in a controlled environment in different years, from 2008 to 2011. As mentioned before, each year we changed the test conditions, from a classical test without collaboration to collaborative test with groups of 2–3 members. Students' samples varied from year to year but the rest of the variables (the course, the time of the year, and the semester of the degree course, etc.) were the same.[1] Furthermore, students were asked to complete the test in November, when they had already finished studying the LEX program. There was no time limit to answer the test. Before analyzing the data, we discarded incomplete sessions and sessions of abnormal duration (less than 10 min), since the latter suggests that the student were not paying enough attention to the test but were just answering randomly. The 25 questions were posed in a random order to different students. Randomization of question position was also applied when the test was administered collaboratively, but taking into account that all the members of a given group always have to take the same questions in the same order.

Any student test session can be divided in three sections, i.e. G_1, (the first 8 questions) G_2 (the 9 questions from position 9–17) and G_3 (the last 8 questions), simulating the administration of three tests. The initial section can be considered as a *pre-test* and the last section as the *post-test*. According to the *Law of large numbers*, the random selection of questions guarantees that, on average, all the test sections have the

[1] Of course from a purely statistical point of view, the selection of the sample could have been improved by a random selection over the whole population which would have mean that some students would have been assigned to different test condition at each course. However, this would lead us to inequity issues inside a course. This is the reason why all students of the course of 2009 and 2010 took the test individually, and all the students of the course 2008 and 2011 took the test collaboratively.

same difficulty. That is, each of the 25 questions has equal probability of being part of the initial, middle or last part of the test. Given the null hypothesis that there is no learning, it implies that the average of the results obtained by the students according to the first section of the test would be comparable to the average obtained using the last section. With this experiment, we are able to reject the null hypothesis in the experimental group (collaborative testing) with the standard 95% confidence, but clearly not for the control group (individual testing).

The results corresponding to individual administration of the test were used as control groups. In the collaborative test administration, students answered the questions twice (following the procedure explained in section 3.2): once before and once after the collaboration phase. If we focus only on the initial responses of the students, the test can be considered equivalent to a classical test, and consequently, it can be assumed (as usual) that the test score depends only on the student's knowledge level. This is commonly known as θ_i (for the i-th student). In classical test theory, it is assumed that θ_i is constant and the test scores are simply statistical estimators. If the test scores increase from the first to the last section of the test, that will indicate that learning has happened. Notice that using the test initial responses, we are measuring not only better performance or score improvement, but actual learning. Nonetheless the opposite does not always hold, that is, it may be the case that actual learning occurs but the test scores might not reflect this. As a counterexample, let us consider a test on traffic signs. Let us imagine that the student learns every sign after answering each question. This improvement is not reflected in the test scores if the subsequent questions ask about different traffic signs (in fact that was the case in a failed experiment we carried out in the subject of Botany). This fact relates to the intrinsic meaning of *knowledge level*, which is not clearly defined in the test literature and is only addressed as a *latent trait*. If learning does occur but the subject corpus is very large, the increase of the average knowledge level θ can be too small to be measured by a test with just m questions. To avoid this situation, the experiment should be restricted to a reduced subset of a whole subject. That is, the test subject should be selected in such a way that students can achieve significant learning while taking the test (i.e. in 1 or 2 h). In this experiment test questions require not only knowledge and comprehension, but higher level of learning processes in the sense of Bloom's taxonomy like analysis and synthesis.

4.3.3. Results and discussion

Table 10 shows the information about all times the test was administered between 2008 and 2011. The first three rows correspond to the tests administered collaboratively (experimental groups) and the last two rows, to the tests administered individually (control groups). The experimental group comprised a total of 78 students. The collaborative tests were taken in a total of 9 groups of 2 or 3 people (depending on the session) and 20 groups of 3 people. Students were informed that the test scores (both the initial and the final score) would be considered as part of their course marks. As a result, students were motivated to take the test more seriously. The control group consisted of 88 students that took a classical test without the collaboration framework. The collaborative test sessions took an average of 21.8 min of answering time (this value is computed as the time slot between posing a question and the initial response), while the classical test sessions took 19.5 min. Unfortunately, the time of the discussion phase was not recorded.

With the performance data of the student samples of Table 11, we explored whether or not *learning was directly related to the use of the collaborative framework (Q8)*. Table 11 shows the results obtained. The third column represents the average of the 8 questions posed in the first section of the test, next column is the average of questions posed in positions 9–16, and the fifth column the average of the questions posed in positions 17–25. Remember that in each test, a given question might appear at any position, and therefore the average score of each section is comparable. All data are presented on a scale between 0 and 100, and the 95% confidence interval is displayed giving the standard error to be analyzed by a t-test.

The two columns on the right of Table 11 show the score increase from the first to the second and to the third section of the test. Score increase in the control group is the lowest of the three (just 4.47 points and not statistically significant), and appears from the second to the third section of the test. The average score increase (from the first to the third section) of the whole experimental group (group A and B) is 13.25 ± 9.79, which is statistically significant. The experimental groups increased more, and earlier. However, the difference between the increase in the experimental group and the control group (8.64) is not statistically significant at 95% confidence, but it is significant at a lower level of 70% confidence ($p < 0.30$).

We also studied the correlation between score increase (from the first to the third section) and the use of the collaborative framework, using the indicators described in the previous set of experiments. The result indicates a small but positive correlation between the score increase and the number of messages exchanged (0.12) and the average length of those messages (0.14). This positive correlation suggests that those students who have used the collaborative framework more have increased their knowledge level scores more.

Next, we explored the correlation between the score increase (from the first to the third section) and the estimated student knowledge level (*Q9*), that is, between score increase against the average score in the whole test. In the experimental groups, and also in the control group, this correlation was small but negative (−0.19 and −0.08), which suggests that low-level students increased their knowledge level more than high-level students. This fact is not surprising, since they had more "space" for increase. Notice that a high level student who solves all questions does not increase at all. For the experimental group, we have also studied the correlation between the average score increase (from the first to the third section) and the average knowledge level of each group, obtaining a positive result (0.41). This result does not go against the previous one, but suggests that in order to be effective, a group should be composed of both high and low level students.

Former result led us to another interesting question: *the group composition (Q10)*. In this experiment, students were randomly assigned to a group, thus the data reflected different types of group composition. In all the samples and groups we could identify the student with the highest and lowest score in the group. Let us call them the *H-student* and *L-student*, respectively. The average score increase (from the first to the third section) for the *H-student* was 9.33 ± 14.81, but for the *L-student* it was 15.52 ± 16.81. This result indicates that the lower level students were those who took more advantage of the *collaborative testing* compared to high-level students ($p < 0.59$). However, the most remarkable fact is that the collaborative framework benefits both low and high level students in short-term learning.

These results complement those obtained by Giuliodori et al., (2008). In that work they found that low and high performing students benefits for collaborative testing, but low performing students benefits more. It should be noticed that they refer to score improvement and we are measuring actual learning.

Finally, we studied the optimal composition of the groups, and discovered that score increase occurs mainly in those groups where there is a significant difference in the knowledge level of their components. We defined two conditions: groups where all members had relatively similar knowledge level, *S-groups*; and those where it was different, the *D-groups*. Each collaborative group was labeled as S or D group, according to the standard deviation of their components knowledge level. The average score increase in the *S-groups* was 6.63 ± 13.53, but in the *D-groups* the increase reached 20.54 ± 13.40. That is, heterogeneous groups seems to be more effective ($p < 0.15$). To sum up, combining both conditions, see Table 12, we have observed that *H-students* increase was constant and similar to the control group, but the highest increase took place for the *L-students* that belonged to the *D-group* (30.98 ± 25.71). That is, low-level students belonged to a group with a higher-level partner. The difference with the *L-students* that belonged to the S-group is 28.03 ± 33.96, which is significant at 90% ($p < 0.10$) Notice that we were always measuring the score according to the initial responses that is, before the communication with the partner, so the increase of score could only be explained by a better understanding of the problem, i.e. learning in the short term.

These results align with socially oriented theories of development and learning. For example, Vygotsky says that children could perform above their current level of development when collaborating with others of higher ability. On the other hand Piaget's equilibration theory predicts that students should experience greater reasoning gains and higher science achievement in homogeneous groups where equilibration (self-regulation) is more likely to take place. It is not our aim to prove or contradict any of these theories. The value of this experiment is limited to the population (higher education) and the collaborative testing system we are presenting.

5. Conclusion

In this paper, we have described a system that allows taking computer-based collaborative tests. The development of the system has been guided by a formative evaluation that has indicated some missing features, like the awareness facilities, and helped to improve others like the chat room. It has also pointed out the potential advantages, disadvantages and limitations. On the one hand, the interactive learning; on the other, the time required to take a test that is longer than in the individual mode.

From our point of view, the main strength of *collaborative testing* is that it combines the benefits of assessment practice and collaborative learning. There is a consensus in the community that assessment can be used not only as a tool to measure student knowledge, but also as a learning tool. An assessment environment can motivate learners, and the practice of answering questions could induce self-reflection and enrich students' instructional process with other ways of learning. On the other hand, collaborative learning has been demonstrated to improve learning and self-reflection. Explaining other people how to reach a conclusion or an answer to a certain question consolidates the acquired knowledge.

From the teacher's point of view, assessment is needed to determine the students' knowledge level in order to evaluate their competences or to adapt their forthcoming learning process. From the student's perspective, self-assessment contributes to increase metacognition. To sum up assessment is used in both cases as a measurement of knowledge.

The approach followed in this implementation has been to combine assessment and collaborative work in such a way that both objectives, i.e. measurement and learning, are met. A session of a collaborative testing in Siette is a sequence of questions answered twice: first in a standalone way and then after a discussion stage. Therefore, at the end of a collaborative testing session, each student will have two scores: the *initial score* and the *final score*. Focusing on the answers given in the first phase, the collaborative testing can be viewed as a classical individual assessment, since no communication is allowed before giving the initial answer. Nonetheless, the experiments described in this paper reveal, when the students' initial and final scores are compared, that learning does occur. Moreover, the experiments made during the formative evaluation of the system included questions asking students whether or not they considered they had learnt using the collaborative Siette. The questionnaire was administered to three different groups of students from two universities in three different subjects. The average response was around 3.5 in a Likert scale from 1 to 5. From the statistical point of view, in almost all cases tests scores increased from the initial to the final response when using the collaborative environment. That is, the average score of a student, according to the responses given before communicating with their colleagues, was almost always higher than the average score obtained considering the responses after the group communication. We can conclude, therefore, that the score improvement was linked to the use of the collaborative tools provided by the environment. In turn, there is also a good correlation between final response agreement and higher scores, which indicates that the collaboration process leads correct answers.

However, these results cannot conclusively prove that there is an actual learning gain for students who use collaborative testing environment. For instance, we could hypothesize that students simply had more time to think about their second answer, and the item review process itself could explain the increase in the final score. In other words, an alternative hypothesis might be that taking longer time to think about the answer could also increase the student's score. Therefore, an important question still remains unanswered: Do students actually learn using the Siette collaborative environment? It is hard to prove the hypothesis that actual learning occurs while taking the test in collaboration with fellow students. We do not claim that using collaborative testing will improve long-term learning better than other types of learning. We do believe that different types and methods are complementary and this is just a new and attractive technology-enhanced learning tool. On the other hand, short-term learning is difficult to measure. To perform the experiment we have taken advantage of the fact that the tool we want to analyze is an assessment system and students are continuously assessed during their interaction with the tool.

Certainly *collaborative testing* is not suitable in all cases. Personal academic certification could be the most extreme case. In our use of the system, we have never used collaborative testing in final exams. Accordingly, the collaborative facilities of Siette are currently used as a learning tool for self-assessment and for mid-term assessments, but not for formal grading. Almost all students who completed the questionnaire evaluated the system positively. They have appreciated the possibility of viewing the answers of their partners, but even more the communication between one another, as well as explanation and justification of answers.

Regarding the impact of using the collaborative tool, the results obtained in the tests show that the average of final answer scores is always higher than the average of initial answer scores, independently of the test difficulty. In addition, most of the students have used the collaboration tool in each question before giving their final answers. Concerning the absolute scores obtained by students, the data do not suggest a clear relationship between the final score obtained and the use of the collaborative tool.

With respect to the agreement in the final answers, the use of the collaborative tool and the difference between the final and initial scores, we have found that those students who have used the tool are most likely to come to an agreement in their final answers and are therefore more likely to improve their results.

Bearing in mind the features of the collaboration tool that could influence the difference between the final and initial scores, evidence suggests that using just the "view answer" button or the chat tool does not always lead to a better score. The average best results are obtained when using both facilities. Moreover, It seems that there are no significant differences when it comes to improvement in questions supported by short discussions, and those involving long ones. Finally, agreement is more often achieved when both tools are used; using just the "view button" does not guarantee obtaining a greater degree of agreement, while using the chat does.

We can clearly conclude that using the collaboration tool leads to a greater agreement in the student responses and an improvement in the score obtained in the test questions. Viewing the partner's answer alone does not always guarantee that the students in the group will reach an agreement or will improve their score. Communication between them (using the chat tool) is also required although the length of the communication does not seem to be a relevant factor in this case.

There is some strong evidence indicating that collaboration during assessment is beneficial to learning, and that the Siette collaborative environment provides sufficient support to students wishing to use it. The collaboration is beneficial both to low-level and high-level students, probably because explanation fosters self-reflection, which also leads to learning. There is weaker evidence that groups of three students perform better than groups of two.

Our collaborative testing has been used by undergraduate student, as a self-assessment and learning tool in different subjects at the University of Malaga and the Polytechnic of Madrid. Since the first release, around 3000 collaborative test sessions have taken place. Currently we are planning a new version of the framework that will use Javascript and Ajax instead of Java Applets. That will allow access from mobile devices and will open new possibilities for ubiquitous learning. However, the mayor findings of this paper about its desired functionality still remain relevant. Siette and its collaborative framework are available at http://www.siette.org.

Acknowledgments

Authors of this paper would like to thank to all the teachers and students who participated in the experiments. This work has been partially funded by the Consejeria de Economía Innovación y Ciencia of the Junta de Andalucía *(Spanish regional government)* through the research projects TIADA (P07-TIC-03243), PATIO (P08-TIC-4273) and DEDALO (P09-TIC-5105)

References

Bjornsdottit, A. (2012). *Evaluating the use of two different models of collaborative tests in an online introductory statistics course*. Ph.D. thesis dissertation presented at the Faculty of the graduate School of the University of Minnesota. Avilable on-line at. Retrieved June 2013 http://iase-web.org/documents/dissertations/12.Bjornsdottir.Dissertation.pdf.

Black, P., & William, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability, 21*(1), 5–31.

Conejo, R., Guzmán, E., Millán, E., Trella, M., Pérez-de-la-Cruz, J. L., & Ríos, A. (2004). SIETTE: a web-based tool for adaptive testing. *International Journal of Artificial Intelligence in Education, 14*, 29–61.

Cooper, B., & Cowie, B. (May 2010). Collaborative research for assessment for learning. *Teaching and Teacher Education, 26*(4), 979–986.

Cortright, R. N., Collins, H. L., Rodenbaugh, D. W., & DiCarlo, S. E. (2003). Student retention of course content is improved by collaborative-group testing. *Advances in Physiology Education, 27*, 102–108.

Dillenbourg, P. (1999). What do you mean by collaborative learning? In P. Dillenbourg (Ed.), *Collaborative-learning: Cognitive and computational approaches* (pp. 1–19) Oxford: Elsevier.

Dillenbourg, P., & Hong, F. (2008). The mechanics of CSCL macro scripts. *International Journal of Computer-Supported Collaborative Learning, 3*(1), 5–23.

Fischer, F., & Mandl, H. (2005). Knowledge convergence in computer–supported collaborative learning: the role of external representation tools. *Journal of the Learning Sciences, 14*(3), 405–441.

van Gennip, N. A. E., Segers, M. S. R., & Tillema, H. H. (August 2010). Peer assessment as a collaborative learning activity: the role of interpersonal variables and conceptions. *Learning and Instruction, 20*(4), 280–290.

Giuiliadori, M. J., Lujan, H. L., & DiCarlo, S. E. (2008). Collaborative group testing benefits high- and low-performing students. *Advances in Physiology Education, 32*, 274–278.

Gouli, E., Gogoulou, A., & Grigoriadou, M. (2006). Supporting self- peer- and collaborative-assessment through a web-based environment. In P. Kommers, & G. Richards (Eds.), *Proceedings of world conference on educational multimedia, hypermedia and telecommunications 2006* (pp. 2192–2199). Chesapeake, VA: AACE.

Gress, C., Fior, M., Handwin, A., & Winne, P. (September 2010). Measurement and assessment in computer-supported collaborative learning. *Computers in Human Behavior, 26*(5), 806–814.

Guzmán, E., Conejo, R., & Pérez-de-la-Cruz, J. L. (2007a). Improving student performance using self-assessment tests. *IEEE Intelligent Systems, 22*, 46–52.

Guzmán, E., Conejo, R., & Pérez-de-la-Cruz, J. L. (2007b). Adaptive testing for hierarchical student models. *User Modeling and User-Adapted Interaction, 17*(1–2), 119–157.

Haberyan, A., & Barnett, J. (2010). Collaborative testing and achievement: are two heads really better than one? *Journal of Instructional Psychology, 37*(1), 32–41.

Jermann, P., & Dillenbourg, P. (2003). Elaborating new arguments through a CSCL script. In P. Dillenbourg (Ed.), *Learning to argue*, Vol. 1 (pp. 205–226). Dordrecht: Kluwer.

Kollar, I., & Fischer, F. (August 2010). Peer assessment as collaborative learning: a cognitive perspective. *Learning and Instruction, 20*(4), 344–348.

Kwok, R., & Ma, J. (1999). Use of a group support system for collaborative assessment. *Computers & Education, 32*, 109–125.

Laurillard, D. (2010). Effective use of technology in teaching and learning in HE. In P. Peterson, E. Baker, & B. McGaw (Eds.), *International encyclopedia of education*, Vol 4 (pp. 419–426). Oxford: Elsevier.

Leight, H., Sanders, C., Clakins, R., & Withers, M. (2012). Collaborative testing improves performance but not content retention in a large-enrollment introductory biology class. *CBE–Life Sciences Education, 11*, 392–401.

Meseke, C. A., Bovec, M. L., & Gran, D. F. (May 2009). Impact of collaborative testing on student performance and satisfaction in a chiropractic science course. *Journal of Manipulative and Physiological Therapeutics, 32*(4), 309–314.

Meseke, C. A., Nefzinger, R., & Meseke, J. K. (2010). Student attitudes, satisfaction, and learning in a collaborative testing environment. *Journal of Chiropractic Education, 24*(1), 19–29.

Michinov, N., Brunot, S., Le Bohec, O., Juhel, J., & Delaval, M. (2011). Procrastination, participation, and performance in online learning environments. *Computers & Education, 56*(1), 243–252.

Molsbee, C. P. (2013). Collaborative testing and mixed results. *Teaching and Learning in Nursing, 8*(1), 22–25,.

Nussbaum, M., Alvarez, C., McFarlane, A., Gómez, F., Claro, S., & Radovic, D. (January 2009). Technology as small group face-to-face collaborative scaffolding. *Computers & Education, 52*(1), 147–153.

Ochoa, S., Guerrero, L. A., Pino, J. A., Collazos, C. A., & Fuller, D. (2003). Improving learning by collaborative testing. *Journal of Student-Centered Learning, 1*(3), 127–139.

Pandey, C., & Kapitanoff, S. (2011). The influence of anxiety and quality of interaction on collaborative test performance. *Active Learning in Higher Education, 12*(3), 163–174.

Paulus, T. M. (2009). Online but off-topic: negotiating common ground in small learning groups. *Instructional Science, 37*(3), 227–245.

Rafaeli, S., Barak, M., Dan-Gur, Y., & Toch, E. (November 2004). QSIA – a web-based environment for learning, assessing and knowledge sharing in communities. *Computers & Education, 43*(3), 273–289.

Rao, S. P., Collins, H. L., & DiCarlo, S. E. (2002). Collaborative testing enhances student learning. *Advances in Physiology Education, 26*, 37–41.

Robinson, D. H., Sweet, M., & Mayrath, M. (2008). A Computer-based, team-based testing system. In *Recent innovations in educational technology that facilitates student learning* (pp. 277–290). Information Age Publishing Inc.

Sandahl, S. S. (2009). Collaborative testing as a learning strategy in nursing education: a review of the literature. *Nursing Education Perspectives, 30*(3), 171–175.

Schellens, T., & Valcke, M. (November 2005). Collaborative learning in asynchronous discussion groups: what about the impact on cognitive processing? *Computers in Human Behavior, 21*(6), 957–975.

Sharp, S. (2006). Deriving individual student marks from tutor's assessment of group work. *Assessment & Evaluation in Higher Education, 31*(3), 329–343.

Simkin, M. G. (2005). An experimental study of the effectiveness of collaborative testing in an entry-level computer programming class. *Journal of Information Systems Education, 16*(Fall 2005, 16), 273–280.

Stahl, G., Koschmann, T., & Suthers, D. (2006). Computer-supported collaborative learning: an historical perspective. In R. K. Sawyer (Ed.), *Cambridge handbook of the learning sciences* (pp. 409–426). Cambridge, UK: Cambridge University Press.

Swan, K., Shen, J., & Hiltz, S. R. (2006). Assessment and collaboration in on-line learning. *Journal of Asynchronous Learning Networks, 10*(1), 45–62.

Valdivia, R., & Nussbaum, M. (March 2009). Using multiple choice questions as a pedagogic model for face-to-face CSCL. *Computer Applications in Engineering Education, 17*(1), 89–99.

Waldrop, M. (2013). Online learning: campus 2.0. *Nature News, 495*(7440).

Weinberger, A., Stegmann, K., & Fischer, F. (2007). Knowledge convergence in collaborative learning: concepts and assessment. *Learning and Instruction, 17*(4), 416–426.

Zimbardo, P. G., Butler, L. D., & Wolfe, V. A. (2003). Cooperative college examinations: more gain, less pain when students share information and grades. *The Journal of Experimental Education, 7*(2), 101–125.

Zipp, J. (2007). Learning by exams: the impact of two-stage cooperative tests. *Teaching Sociology, 35*, 62–76.