
A·L·I·C·E

Adaptive Learning via Intuitive/Interactive
Collaborative and Emotional systems

Project Number: **257639**
Project Title: ALICE: ADAPTIVE LEARNING VIA INTUITIVE/INTERACTIVE,
COLLABORATIVE AND EMOTIONAL SYSTEMS

Instrument: Specific Targeted Research Projects
Thematic Priority: ICT-2009.4.2:Technology-Enhanced Learning

Project Start Date: June 1st, 2010
Duration of Project: 24 Months

Deliverable: **D8.1.1: Initial Experimentation and Evaluation Results**
Revision: 1.1
Workpackage: WP8: Experimentation and Validation
Dissemination Level: Public

Due date: 10/31/2011
Submission Date: 10/31/2011
Responsible: UOC
Contributors: UOC, TUG, MOMA

This work is licensed under the Creative Commons Attribution-NonCommercial-NoDerivs 3.0 License. To view a copy of this license, visit <http://creativecommons.org/licenses/by-nc-nd/3.0/legalcode> or send a letter to Creative Commons, 171 Second Street, Suite 300, San Francisco, California, 94105, USA.

PROJECT CO-FUNDED BY THE EUROPEAN COMMISSION WITHIN THE
SEVENTH FRAMEWORK PROGRAMME (2007-2013)



Version History			
Version	Date	Changes	Contributors
0.1	28/04/2011	Creation of index and addition IWT-UOC integration information	Carlos Ors (UOC)
0.2	12/05/2011	Creation of an Annex for the Integration of IWT	Carlos Ors (UOC)
0.3	08/09/2011	Input of first experiments by TUG	Margit Höfler, Mohammad AL-Smadi, Christian Gütl (TUG)
0.4	01/10/2011	New document structure	UOC, TUG
0.5	10/10/2011	Revision of the UOC Annex	Carlos Ors (UOC)
0.6	25/10/2011	Input of all experiments	TUG, MOMA, UOC
1.0	31/10/2011	Final version	UOC
1.1	20/01/2012	Revision from reviewers' comments	UOC, MOMA

Table of Contents

1.1	Purpose	8
1.2	Methodology	9
1.2.1	Experimentation at UOC site	11
1.2.2	Experimentation at TUG site	12
1.2.3	Experimentation at MOMA site.....	12
2	R1. Upper Level Learning Goals.....	14
2.1	Research goals and hypotheses	14
2.2	Method	15
2.2.1	Participants.....	15
2.2.2	Apparatus and Stimuli	15
2.3	Evaluation Results.....	20
2.3.1	Activity levels in the IWT	20
2.3.2	Usability of the IWT	22
2.3.3	Emotional aspects.....	25
2.3.4	Questionnaire evaluation	26
2.4	Validation Results	14
2.4.1	The IWT as a valuable resource.....	28
2.4.2	Motivational aspects	33
2.4.3	Tutor assessment and knowledge acquisition	33
2.5	Conclusion	35
3	R2. Knowledge model contextualization: Experimenting the Knowledge model contextualization.....	14
3.1	R2-1. Knowledge model contextualization: Experimenting the Knowledge model contextualization from the instructor’s viewpoint (TUG).....	36
3.1.1	Research goals and hypotheses.....	36
3.1.2	Method	37
3.1.3	Evaluation Results.....	41
3.1.4	Conclusion	45
3.2	R2-2. Knowledge model contextualization: Experimenting the Knowledge model contextualization from the instructor’s viewpoint (UOC).....	45
3.2.1	Research goals and hypotheses.....	45
3.2.2	Method	47
3.2.3	Evaluation Results.....	48
3.2.4	Conclusion	55
4	R3. Semantic Connections Between Learning Resources.....	56
4.1	Research goals and hypotheses	56

4.2	Method	56
4.2.1	Participants.....	56
4.2.2	Apparatus and Stimuli	57
4.3	Evaluation Results.....	61
4.3.1	Activity levels in the CLR.....	61
4.3.2	Usability of the IWT	63
4.3.3	Emotional aspects.....	63
4.3.4	Questionnaire evaluation	67
4.4	Validation Results	68
4.4.1	The CLR as a valuable resource	68
4.4.2	Motivational aspects	71
4.4.3	Tutor assessment and knowledge acquisition	71
4.5	Conclusion	72
5	R4. Live and Virtualized Collaboration.....	56
5.1	Research goals and hypotheses	74
5.2	Method	75
5.2.1	Participants.....	75
5.2.2	Apparatus and Stimuli	76
5.2.3	Procedure	78
5.3	Evaluation Results.....	78
5.3.2	Usability of the VCS	79
5.3.3	Evaluation of the questionnaire	84
5.4	Validation Results	85
5.4.1	The VCS as a valuable resource	85
5.4.2	Motivational aspects	87
5.4.3	Tutor assessment and knowledge acquisition	88
5.5	Conclusion	89
6	R5. Storytelling.....	90
6.1	Research goals and hypotheses	90
6.2	Method	91
6.2.1	Participants.....	91
6.2.2	Apparatus and Stimuli	91
6.2.3	Procedure	93
6.3	Evaluation Results.....	93
6.3.1	Storytelling Learning Object Activity	94
6.3.2	Usability of the tool	95
6.3.3	Emotional aspects.....	97
6.4	Validation Results	99
6.4.1	First School “E.Striano”	101
6.4.2	Secondary School “Pitagora”	104

6.5	Conclusion	106
7	R6. A Serious Game for Civil Defence Training in School	108
7.1	Research goals and hypotheses	108
7.2	Method	109
7.2.1	Participants	109
7.2.2	Apparatus and Stimuli	109
7.2.3	Procedure	110
7.3	Evaluation Results.....	111
7.3.1	First School “E.Striano”	113
7.3.2	Second School “Pitagora”	115
7.4	Conclusion	116
8	R7. Affective and Emotional Approaches.....	118
8.1	Research goals and hypotheses	118
8.2	Method	118
8.2.1	Participants	118
8.2.2	Apparatus and Stimuli	119
8.2.3	Procedure	120
8.3	Evaluation Results.....	121
8.3.1	Affective/emotional Interface	121
8.3.2	Usability of the tool	122
8.3.3	Emotional aspects.....	125
8.4	Validation Results	127
8.4.1	First School “E.Striano”	128
8.4.2	Second School “Pitagora”	131
8.5	Conclusion	133
9	R8. Enhanced Wiki-Test and Peer-review for writing assignments	135
9.1	Research goals and hypotheses	135
9.2	Method	136
9.2.1	Participants	136
9.2.2	Apparatus and Stimuli	136
9.2.3	Procedure	138
9.3	Evaluation Results.....	139
9.3.1	Usability of the WIKI-tool	139
9.3.2	Task Awareness	140
9.4	Validation Results	141
9.4.1	Attitudes and experiences concerning peer-assessment.....	142
9.4.2	Group-Assessment.....	145
9.4.3	Tutor’s assessment	145
9.4.4	Emotional Aspects	141
9.5	Conclusion	145

10	R9. Assessment in Self-Regulated Learning	147
10.1	Research goals and hypotheses.....	147
10.2	Pre-study R9-0a: Evaluation of the automatically created questions	149
10.2.1	Method	149
10.2.2	Evaluation Results.....	152
10.3	Pre-study R9-0b: Evaluation of the automatically created questions	157
10.3.1	Method	157
10.3.2	Evaluation Results.....	158
10.3.3	Extracting concepts	158
10.3.4	Analysing the pedagogical quality of the automatically created questions regarding Bloom's taxonomy	159
10.3.5	Evaluation of concepts	159
10.3.6	Evaluation of questions generated by the AQC.....	160
10.4	R9-1: Investigating the AQC and the co-writing wiki in a complex learning environment	162
10.4.1	Method	162
10.4.2	Evaluation Results.....	166
10.5	Validation Results.....	170
10.5.1	Motivational Aspects.....	171
10.5.2	Group-Assessment.....	172
10.6	Conclusion.....	173
	References.....	175
	Annex A – Integration of IWT tools with real context of learning	177
	A1 Integration at UOC site.....	177
	A1.1 Introduction	177
	A1.2 Survey of tools for interoperability.....	178
	A1.3 Open knowledge Initiative (OKI).....	178
	A1.4 IMS Basic Learning Tools Interoperability (BLTI)	179
	A1.5 Adoption of BLTI for the integration with ALICE.....	180
	A2 Integration at MOMA site	182
	A3 Integration at TUG site	184

1 Introduction

This report describes the results of the initial experimentation, evaluation and validation activities of Work package 8.

The aim of ALICE is to build an adaptive and innovative environment for e-learning. To this end, personalization, collaboration, and simulation aspects are combined and also affective and emotional aspects are considered. In particular, two specific contexts will be considered in ALICE: science teaching at university and training about emergency and civil defence. Three different pilot sites are involved in the experimentation and validation: UOC, TUG and MOMA.

This report presents the results of the execution of the experimentation and validation plan of the research and technology developed in ALICE reported in [4]. To this end, a practical method oriented to the experimentation of the tools developed and organized as prototype scenarios and its validation in real situations in different educational fields is followed.

It is worth clarifying at this initial point that the experimentation, evaluation and validation activities reported here are not intended to report a technical testing plan of each of the individual developments of ALICE nor their integration process into IWT. A technical testing was instead conducted in last stages of the whole ALICE development by all participating parties that developed stand-alone prototypes as a result of their participated research tasks. These tasks tested and validated the beta prototypes with the intent of finding software bugs and first feedback from a small set of testers in a very controlled situations (see [11] for WP3 particular case).

Therefore, this document reports the results of the experimentation, evaluation and validation of ALICE considering all individual developments have been tested and integrated into the referenced platform IWT performing the role of the e-learning system (i.e., ALICE System). To this end, Annex A of this document reports the integration activities performed in each pilot site.

ALICE includes the following 6 work packages, which investigate the main aspects of the project and were involved in the experimentation and validation activities reported here:

- WP2 Affective and Emotional Approaches
- WP3 Live and Virtualized Collaboration
- WP4 Simulation and Serious Games
- WP5 New Forms of Assessment
- WP6 Storytelling
- WP7 Adaptive Technologies for e-Learning Systems

These base their research goals on [10] and [3]. The latter reports all ALICE requirements forming the starting point of the research activities and thus it is the main reference of this report.

1.1 Purpose

WP8 of ALICE has the objective of experimenting developed tools (delivered as independent working packages) and resources in order to provide feedback to theoretical and technological activities. It includes, as well, the evaluation and validation of the impacts of the innovative features offered by ALICE inside the selected learning and training environments. There are three different training sites where each tool, as a prototype will be experimented:

- UOC
- TUG
- MOMA schools network

The purpose of this report is to collect information about the experience of performing the different tasks where the experimentation and validation are based on in the different sites mentioned above.

The objectives and research goals to be achieved by experimentation and validation are to provide evidence, through extended episodes of trials by real learners and teachers, that the developed technological solution of ALICE is effective towards covering the identified user requirements and implementing the developed scenarios of use, as well as towards enhancing the learning experiences of the various users by contributing to more effective and efficient learning activities, more motivation and inspiration for learners and teachers in various formal and informal learning circumstances.

In particular, the following quality criteria are defined to evaluate and perform a follow-up of the realisation of the trials and how these allow for validating the artefacts and investigations developed in ALICE:

- C1. Simple and clear-cut of precise form, so that they can evaluate without ambiguities.
- C2. Objective, avoiding the subjectivity in its quantification.
- C3. Easily to obtain, with a reasonable effort.
- C4. Valid. They have to measure what it is attempted to measure.
- C5. Reliable. They have to offer the same result for different evaluators.

With the aim to identify these general criteria, it was considered the following evaluation objectives:

- O1. Completeness. The clear-cut criteria have to allow for evaluating each and every of the potentialities of ALICE.
- O2. Exploitation. To evaluate the possibilities of exploitation of the solution developed in ALICE.
- O3. Transfer. To evaluate ALICE applicability, and how the solution proposed is adapted and transferred to the consortium partners and at large at their countries'

educational and research environments. In addition, to evaluate aspects that influence to improve its transfer, such as the use and/or promotion of standards.

- O4. Research and technological innovation. To evaluate the degree of real innovation proposed in ALICE. Commitment solutions have to be planned in case that this objective goes into conflict with O2 and O3.
- O5. Impact. To determine the impact that has ALICE, translated into potentials beneficiaries of the solution.

For the purpose of this report, only objective O1 is considered which addresses the functional features and technological advances of ALICE.

1.2 Methodology

A comprehensive experimentation study is developed in this section for ALICE describing all activities that have been undertaken during the experimentation, evaluation and validation.

The study includes, for each requirement scenario, details on the goals and hypotheses, the method (including number and type of participants, apparatus and stimuli, and procedure), and the evaluation and validation results. This is the standard structure to report empirical results following APA guidelines (see [9] and Table 1)

Step	Description/Questions to be considered
1. Research Goals and Hypotheses	What is the purpose/are the goals of the planned study? Which hypotheses can be derived from the goals?
2. Method	
2.1 Participants	Selection/Description of the participants. <ul style="list-style-type: none"> • How many subjects are necessary/available? • More detailed description (age, gender,...) • Are there any constraints? (e.g., only undergraduates, gender, age ...) • Selection criteria (e.g., volunteers, participation for course credit,...). • Are they informed about the goal of the study?
2.2 Apparatus and Stimuli	How is the problem investigated in detail (with respect to the hypotheses)? What is measured? (e.g., students knowledge of Topic XY) How is the outcome measured/quantified? (e.g., questionnaire, frequencies of log-ins, ...)
2.3 Procedure	Description of the procedure of the planned study <ul style="list-style-type: none"> • Short summary of the main design, assignment of the subjects to the groups, ... • What is - in detail - the course of events during the study? (e.g., subject is assigned to the group X, has to fill out a questionnaire (pre-test); learning tool is introduced to the subject; subject is allowed to learn XY minutes; gets a further questionnaire (post-test),...)

Step	Description/Questions to be considered
3. Evaluation Results	What about the usability/functionality of the tool? (e.g., Was the system easy to use?) What did the students like/not like regarding the tool? Were the students aware of the functions (contribution graphs, actions) of the tool? What can be improved regarding the tool?
4. Validation Results	Results from the pedagogical and psychological perspective <ul style="list-style-type: none"> • Were the students motivated regarding the experiment? • Did the tool support their learning process?
5. Conclusion	What are the most important results with respect to the predefined goals?

Table 1: Reporting a study (APA style) [9]

The experimentation study has been localised to better address the local circumstances pertaining in each experimentation site of user group. Implementation parameters have been determined, such as necessary adjustments to the agenda and needs of the different user groups, technical and organisational preparations, additional technological tools development, selection of the best technical configuration for the specific purposes, etc.

This methodology takes as inputs the user scenarios from D1.1 of Work Package 1 [3] and performs the definition, integration and experimentation tasks of the resulting software components.

To pursue these goals, communities of user groups (in general, students and teachers/lecturers) were organised in each pilot site, which are educational environments with full or relatively limited e-learning quotes (e.g. full virtual education and blended learning), and in which the extended computational capabilities of ALICE enabled the exploitation by teachers and students of existing advanced educational technologies. For each scenario of use a devoted user group was developed drawing from two different contexts, namely Science Teaching at University and Civil Defense and Emergency.

The deployed system and scenarios of its use were exposed, through demonstration activities, to numbers of real users in real settings, with the aim to validate the findings of the pilots with feedback from, and observations of, random (and not anymore deliberately selected) users in various educational contexts. In each validation site, several experiments with numerous users performing authentic technology-enhanced learning tasks were performed.

Both in this iteration, and gradually, in next iterations of the experiments the size of user groups will be extended by dynamically involving more groups from other subjects and programs. Therefore, a main issue of the experiments is the organization and the management of the user-centred activities in the participating pilot sites. The exact way of implementation as well as the necessary parameters was determined. The timetable of the proposed activities was designed in order to be discussed with the teachers involved.

Next, all 9 scenarios experimented located in the 3 pilot sites are summarized.

1.2.1 Experimentation at UOC site

The following four scenarios (see [4]) were experimented at UOC:

R1. Upper Level Learning Goals

This scenario is purposed to provide a high level access to the learning offer in order to simplify the learning courses building process. The generation of a learning experience starts from the explicit or implicit request made by a learner in terms of needs to be satisfied (expressed in natural language).

R2. Knowledge model contextualization

The aim of this scenario is to build an ontological description of a teaching domain that is able to automatically adapt to a context. The resulting ontology can be used to build a personalized course with a different learning path, tailored on the needs of the learner.

R3. Semantic connection between learning resources

This scenario provides a set of semantic connections between learning resources and algorithms to automatically activate and deactivate such connections according to teaching and learning preferences as well as to context information.

R4. Live and virtualized collaboration

The goal of this scenario is to virtualize live sessions of collaborative learning to produce storyboard learning objects embedded in a learning resource (VCS) to be experienced and played by learners. During the resource execution, learners observe how avatars discuss and collaborate, how discussion threads grow, and how knowledge is constructed, refined and consolidated. This scenario was previously tested (see [2]).

We experimented with a combination of the four scenarios (IWT-ALICE classroom involving R1, R2, R3 and R4 scenarios) in one study and one main experiment with R4 scenario (see Table 2). These studies are described in the following chapters.

Study	Description	Schedule
Study R4	Experimenting with the Live and Virtualized Collaboration at UOC	October 2011
Study R1	Experimenting with the IWT-ALICE classroom on Upper Level Learning Goals	October 2011
Study R2-2	Experimenting the Knowledge model contextualization from the instructor's viewpoint	October 2011
Study R3	Experimenting with the IWT-ALICE classroom on Semantic connection between learning resources	October 2011

Table 2: Overview about the studies at UOC

1.2.2 Experimentation at TUG site

The following three scenarios (see [2]) were experimented at TUG:

R.2 Knowledge model contextualization

The aim of this scenario is to build an ontological description of a teaching domain that is able to automatically adapt to a context. The resulting ontology can be used to build a personalized course with a different learning path, tailored on the needs of the learner.

R.8 Enhanced WIKI-test and peer-review for writing assignments

In this scenario the performance of the learners is assessed by the peers during a (collaborative) WIKI activity. In addition, the learner him-/herself also self-assess his/her contribution. For the assessment of the group members' behaviour and their interactions, the instructor has to create rubric(s) that contain(s) the properties of the possible behaviours and interactions during the collaborative learning activity.

R.9 Assessment in self-regulated learning

The goal of this scenario is to provide a new form of assessment in which automatic question generation is used in order to create assessments in a self-regulated learning setting. The questions are created based on the selected content materials. In addition, they cover the required concepts of the learning content.

We tested the three scenarios in one pre-study (for scenario R9) and four main studies (see Table 3). These studies are described in the following chapters.

Study	Description	Schedule
Study R2-1	Experimenting the Knowledge model contextualization from the instructor's viewpoint	August 2011
Study R8-1	Experimenting the co-writing WIKI at TUG Graz	May to June 2011
Pre-study R9-0a & R9-0b	Evaluation of automatically created questions	December 2010 to February 2011
Study R9-1	Experimenting the automatic question creator and the co-writing WIKI in Self-regulated learning	July to August 2011

Table 3: Overview about the studies at TUG

1.2.3 Experimentation at MOMA site

The following three scenarios (see Deliverable D1.3) were experimented at MOMA:

R.5 Storytelling

The goal of this scenario is to allow an efficient learning about knowledge and behaviour to be adopted in civil emergency situation (like seismic event in Amusement Park) through the guided learning *narrative based*. The use of Storytelling as complex learning resource that combine guided, objectives oriented and adaptive process could contribute to improve

learning of the students that have a predisposition to the experiential learning and to demonstrate how such a didactic method, revised in a proper way according to an innovative architecture, is best suitable to the transmission of lesson learned.

R.6 A Serious Game for Civil Defence Training in School

The goal of this scenario is to allow an efficient learning about the risk managements through the delivery of a Serious Game (SG) in a personalized learning courses. The use of this kind of resource could contribute to improve the motivation and learning of the students that have a predisposition to the experiential learning.

R.7 Affective and Emotional Approaches

The goal of this scenario is to provide a new system able to recognize and evaluate the affective/emotional state of a learner for supporting and improving the learning. The questions are created based on the selected content materials.

We experimented the three scenarios in a real context by involving two secondary Italian schools belonging to the network schools that adopt the IWT platform (see Table 4).

These studies are described in the following chapters.

Study	Description	Schedule
Study R5	Experimenting the Storytelling Learning Object within an IWT-ALICE classroom on procedure to be performed in case of emergency	October 2011
Study R6	Experimenting the Serious Game within an IWT-ALICE classroom on procedure to be performed in case of emergency	October 2011
Study R7	Experimenting the Emotional tool within an IWT-ALICE classroom on procedure to be performed in case of emergency	October 2011

Table 4: Overview about the studies at MOMA

2 R1. Upper Level Learning Goals

The aim of this scenario is to provide a high level access to the learning offer in order to simplify the learning courses building process. The generation of a learning experience starts from the explicit or implicit request made by a learner in terms of needs to be satisfied expressed in natural language (see [5]). As a result, the Course Generation System (CGS) provides suitable learning resources that meet the learners' needs.

2.1 Research goals and hypotheses

To experiment with the upper level learning goals, we focused on the following goals and hypotheses as described in [4]:

Goals

G1.1: to develop a Course Generation System (CGS) able to generate a set of feasible courses starting from a need expressed in natural language by the learner.

G1.2: to ensure that generated courses cover the expressed needs and the (optionally) selected skills and contexts (taking into account the available learning material).

G1.3: to provide a user friendly interface for needs expression, courses generation, courses preview and course selection.

G1.4: to ensure that generated courses allow the effective learning of scientific concepts in selected domains.

G1.5: to identify possible ways of improving further the utility of the CGS.

Hypotheses

H1.1: a set of feasible courses can be effectively and efficiently created (in an easy and friendly way for the non-expert users) starting from a need expressed in natural language and, optionally, a skill and a context.

H1.2: the use of the CGS contributes to improve students' motivation.

H1.3: the use of the CGS contributes to improve students' understanding of domain concepts.

H1.4: the use of CGS contributes to increase students' activity levels.

H1.5: the use of the CGS contributes to reduce the time between the emerging of a new learning need and its fulfillment.

H1.6: generated courses are considered as a worthy resource by both instructors and students.

2.2 Method

2.2.1 Participants

In order to evaluate this scenario and analyze its effects in the learning process, 170 students enrolled in the course Software Engineering from the Computer Science and Multimedia degrees in the Fall term of 2011 at the UOC participated in the experience. Most of them (154) were from the Computer Science degree and a small group (16) was from the Multimedia degree. Both degrees share the same course “Software Engineering” in its curricula.

The students were equally distributed into 2 classrooms in the UOC virtual campus. Hence, each UOC classroom had 85 students, 77 from the Computer Science degree and 8 from Multimedia degree.

68 out of 170 students (40%) participated actively in the experience. We considered active participation the submission of an evaluation form at the end of the experience. Since the experiment was optional for all students, 60% of them chose not to send the evaluation form and thus they were excluded from the analysis.

41 out of 170 students (25%) also participated in the IWT experience. We considered active participation in IWT the use of the IWT prototypes and the submission of the evaluation form specific to IWT. Hence those 41 students belonged to the group of 68, which left a group of 27 who participated by submitting the form but did not use the IWT prototypes.

From the 68 participants we formed 2 groups for the experiment. One experimental group with 41 students who use IWT (60%) and one control group with 27 students who did not use IWT at all (40%). All of them submitted an evaluation form at the end of the experience.

Therefore, the sample of the experiment was formed by 68 students. For the sake of the experiment, we were only interested in the conglomerate of the experimental group. From this group we formed two sub-groups, 38 from the Computer Science degree (95%) and 3 from the Multimedia degree (5%). 33 students were male (83%) and 7 students were female (17%). The 27 students forming the control group studied at UOC only and did not enter IWT. Hence, whenever referring to IWT we mean the experimental group.

All students of the sample were supervised by one experimented tutor during the experiment.

2.2.2 Apparatus and Stimuli

All students had access to the IWT classroom (where the ALICE prototypes for R1 scenario were installed) from the UOC classroom (see Figure 1 below and Annex A1 for technical details of the integration).

The screenshot shows the UOC classroom interface. On the left, there is a sidebar with 'Salas Estudiantes' and a list of courses. The main content area is divided into several sections: 'Comunicació', 'Planificació', 'Activitats', 'Recursos', and 'Avaluació'. The 'Comunicació' section includes 'Tauler general [1]', 'Tauler', 'Fòrum [43]', 'Wiki', and 'Aula IWT' (circled in red). The 'Planificació' section shows two calendar views for 2011: 'octubre' and 'novembre'. The 'Activitats' section includes 'Calendari semestral' and 'Pla docent'. The 'Recursos' section includes 'Materials i fonts'. The 'Avaluació' section is currently empty. On the right, there are two tables listing activities with columns for 'Data', 'Títol', and 'Esdeveniment'.

Data	Títol	Esdeveniment
2	Estudi Mòdul 1	Liurament
3	Estudi Mòdul 2	Inici
16	Estudi Mòdul 2	Liurament
17	Estudi Mòdul 3	Inici
18	PAC 1	Liurament
19	Debat 1 Wiki	Inici
20	PAC 1	Solució
20	Pràctica 1	Inici
28	PAC 1	Qualificació

Data	Títol	Esdeveniment
2	Debat 1 Wiki	Liurament
6	Estudi Mòdul 3	Liurament
7	Estudi Mòdul 4	Inici
8	Pràctica 1	Liurament
10	PAC 2	Inici
10	Pràctica 1	Solució
16	Debat 2 Wiki	Inici
18	Pràctica 1	Qualificació
29	PAC 2	Liurament

Figure 1: UOC classroom with the access to IWT classroom

Once in the IWT classroom, students had access to the R1 scenario (see Figure 2 and [1])

The screenshot shows the IWT classroom interface. At the top, there is a breadcrumb trail: 'You are here: Home > Classrooms > Enginyeria del Programari'. Below this, there are three tabs: 'Courses', 'Formative Objectives', and 'Forum'. The 'Courses' tab is active, displaying a list of options: 'Express your formative needs', 'Navigate in the formative objectives', 'View recommended formative objectives', and 'My formative needs management'. A blue question mark icon is visible in the top right corner of the content area.

Figure 2: IWT classroom with a list of option to personalize and manage courses

In this scenario there are different functionality provided by the prototype (see [5] for a full description):

Express your formative needs: it allows the learner to indicate in natural language the learning goals he/she wants to achieve and to verify what are the most suitable (see Figure 3 and Figure 4).

Navigate in the formative objectives: allows the learner to view the complete collection of the ULLGs created by teachers.

View recommended formative objects: allows the learner to view a set of ULLGs the system suggests for him thanks to the recommender system integrated within ALICE.

My formative needs management: allows the learner to view their personal Formative Needs.

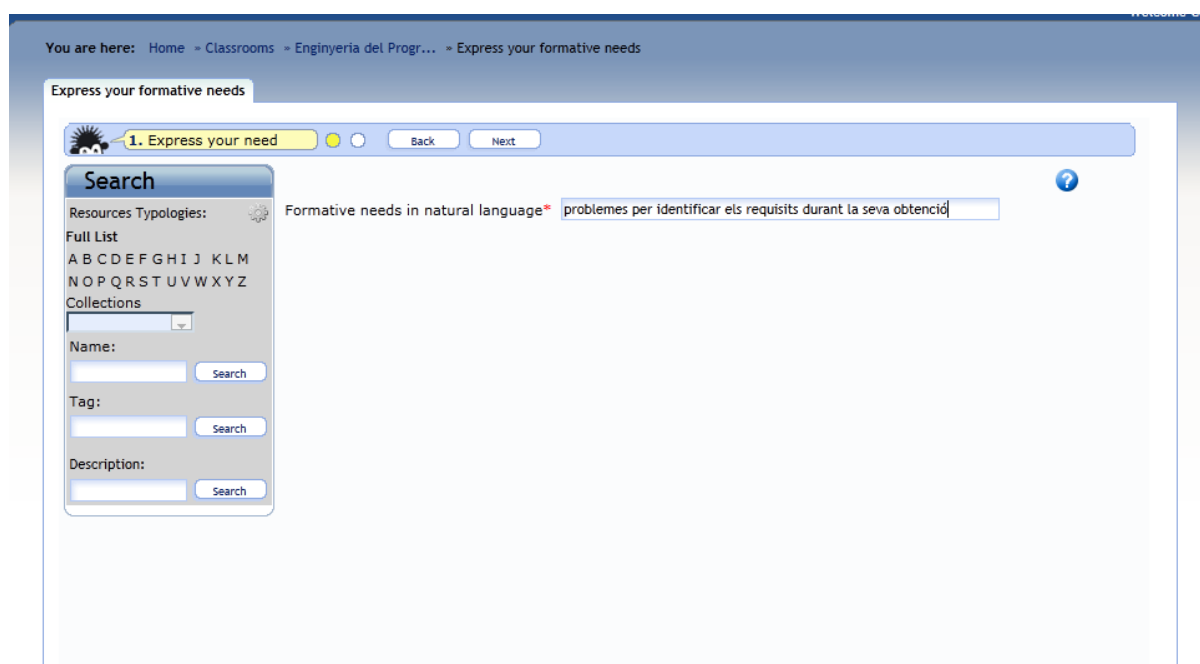


Figure 3: Express your formative needs. it allows the learner to indicate in natural language the learning goals he/she wants to achieve

You are here: Home > Classrooms > Enginyeria del Progr... > Express your formative needs

Express your formative needs

2. Add Ofal

Back Next

Search


Full List
A B C D E F G H I J K L M
N O P Q R S T U V W X Y Z

Name:
 Search

Description:
 Search


High level formative objective

Obtencio dels requisits

 Obtencio dels requisits La primera activitat relacionada a..


Author: Capuano Nicola
Relevance:100%

Problemàtiques de la Identificació..

 Problemàtiques de la Identificació de requisits En aquest ..


Author: Capuano Nicola
Relevance:61%

Casos d'ús

 Els casos d'ús són una tècnica de documentació de requisit..


Author: Capuano Nicola
Relevance:41%

Documentació dels requisits

 Anomenem especificació l'acte, típicament un document..

Author: Capuano Nicola
Relevance:32%

Gestio de requisits

 Gestio de requisits Gestio de requisits Estimació de requi..

Author: Capuano Nicola
Relevance:31%

pag. 1 of 1 go at page items shown in the page

Figure 4: List of the resulting learning resources

We used the SUS (System Usability Scale [6]) in order to investigate the usability of the IWT. The SUS is a simple, ten-item attitude Likert scale giving a global view of subjective assessments of usability. It is generally used after the respondent had an opportunity to use the system being evaluated. SUS scores have a range of 0 to 100 with an average score of 68, obtained from 500 studies. A Score above a 68 would be considered above average and anything below 68 is below average. A score above an 80.3 is considered an A (the top 10% of scores). Scoring at the mean score of 68 gets you a C and anything below a 51 is an F (putting you in the bottom 15%).

After the assignment, students of the experimental group were required to fill out a questionnaire that included the following 7 sections: (i) identification data (names and program they were enrolled); (ii) evaluation questions about the knowledge acquired with the course "Requisites" (Requirements); (iii) open questions evaluation on the IWT classroom supporting the course; (iv) test-based evaluation of the personalized learning system; (v) test-based evaluation on usability of IWT; (vi) test-based evaluation on the emotional state when using IWT; and (vii) a test-based evaluation of the questionnaire. Students submitting this questionnaire had the chance to increase their final grade of the course up to 20%. If the questionnaire was not submitted or with wrong responses the final grade would not decrease whatsoever.

For those students of the control group (i.e., they did not enter IWT during the experience), a different questionnaire was sent with only sections (i) and (ii) which had had to be filled. Students submitting this questionnaire had the chance to increase their final grade of the course up to 10%. If the questionnaire was not submitted or with wrong responses the final grade will not decrease whatsoever.

For qualitative statistical analysis, we summarized the open answers in the surveys. For the quantitative statistical analysis we employed basic statistics, such as Mean (M), Standard Deviation (SD) and median (Md).

For the section (v), as mentioned previously, we used the System Usability Scale (SUS) developed by [6] which contains 10 items and a 5 point Likert scale to state the level of agreement or disagreement. SUS is generally used after the respondent had an opportunity to use the system being evaluated.

Finally, to investigate in which emotional state the students were when they used the IWT platform, section (vi) concerned about the “emotional state” of students when using IWT which included 12 items of the Computer Emotion Scale (CES) [7]. CES scale is used to measure emotions related to learning new computer software. Research showed that the 12 items are describing four emotions:

- Happiness (“When I used the tool, I felt satisfied/excited/curious.”)
- Sadness (“When I used the tool, I felt disheartened/dispirited.”)
- Anxiety (“When I used the tool, I felt anxious/insecure/helpless/nervous.”)
- Anger (“When I used the tool, I felt irritable/frustrated/angry”)

The answer categories in this section are “None of the time”, “Some of the time”, “Most of the time” or “All of the time”.

The data from this experience was collected by means of the web-based forums supporting the discussions in each classroom. Moreover, quantitative and qualitative data were collected from questionnaires containing quantitative and qualitative questions, the answer categories varied between rating scales, multiple choice or open answers. Regarding the rating scales, for the majority of the quantitative questions we used the 5 point Likert scale, so that students could state their level of agreement or disagreement. The rating scale ranged from “I strongly disagree” (1), “I disagree” (2), “neither/nor” (3) to “I agree” (4), “I strongly agree” (5). Finally, quantitative data was also collected from IWT, VCS (see Section 5) and UOC Virtual Campus databases and log files.

2.2.2.1 Procedure

The in-class collaborative formal assignment in both groups lasted three weeks during the second third of the Fall term (October 2011) and consisted of studying part of the course “Software Engineering”. The part of the course corresponded with the topic “Requirements” which forms an essential goal of the course.

Students had two options: they either could study the topic “Requirements” only from UOC virtual classroom or, moreover, from the IWT virtual classroom. Hence, all students had to follow the teaching plan at UOC classroom and learn the mandatory material and perform the learning activities planned. In addition, any student who optionally wanted to complement the study of this topic at UOC with the study of the same topic at IWT could do so. The only requirement was to submit the questionnaire at the end of the experience to acknowledge participation in the experiment. Any student did not have access to IWT classroom before the experience while the access remained open after the end of the IWT course though with no support from the teaching staff.

Previous the experience, the topic “Requirements” had been modeled in IWT by using an ontology and concepts. Then it was contextualized into 2 contexts: GEI and GM, and specific contents for each context were then uploaded. Finally a personalized course called “Requirements” was created (see Section 3.1). The aim was to provide students with specific learning material in line with the specific needs expressed the CGS of IWT and the context they belonged to.

After the end of the experience, students received a questionnaire to be filled in order to evaluate the experience with IWT from the viewpoint of the CGS. Whether they belong to the experimental or the control group they received a specific questionnaire. Part of the evaluation consisted in identifying the knowledge acquired on the topic they have studied (in UOC classroom or, also, in IWT classroom).

2.3 Evaluation Results

Following the methodology described in Section 1.3, in this section we focus on the activity, usability and emotional aspects of the IWT tool (H1.1 and H1.4). We also include in this section the evaluation of the questionnaire. On the other hand, the analyses of the tool’s overall impact on student’s learning process are reported in Section 2.2.4 (Validation Results).

2.3.1 Activity levels in the IWT

In order to give a feedback about the students’ levels activity we should make a correlation between resource efficacy and levels of competency acquired (H1.4).

5 ULLG has been created in order to satisfy the learning needs. They have been associated to the corresponding ontology concepts.

Figures *Figure 5* and *Figure 6* show the quantitative analysis in terms of Media (M), Standard Deviation (SD) and Median (Md) relative to the competence acquired with respect to specific concepts of the topic course of Requirements of 2 contexts (GEI and GM)

- Students from GEI:

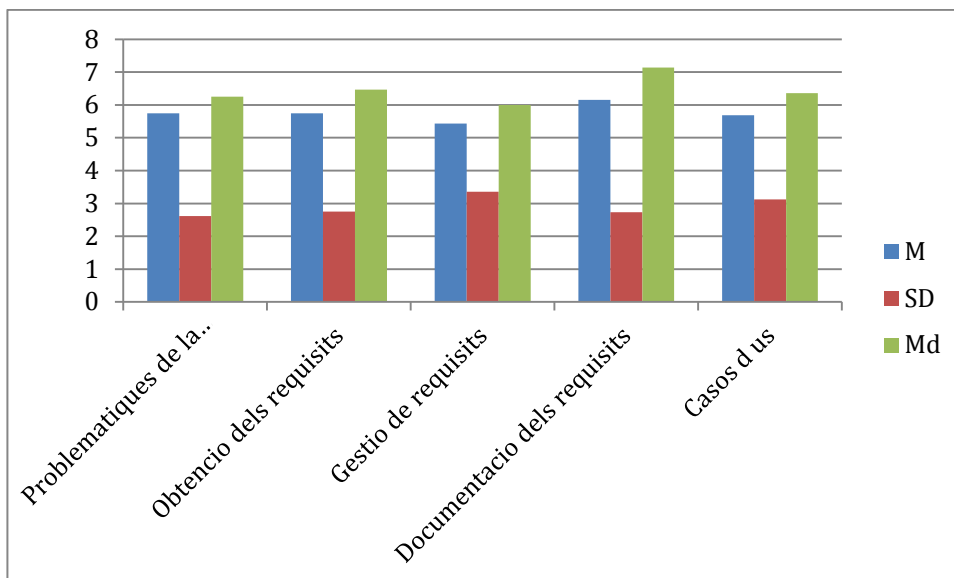


Figure 5: Comparison between ULLG and competence level for GEI

- Students from GM:

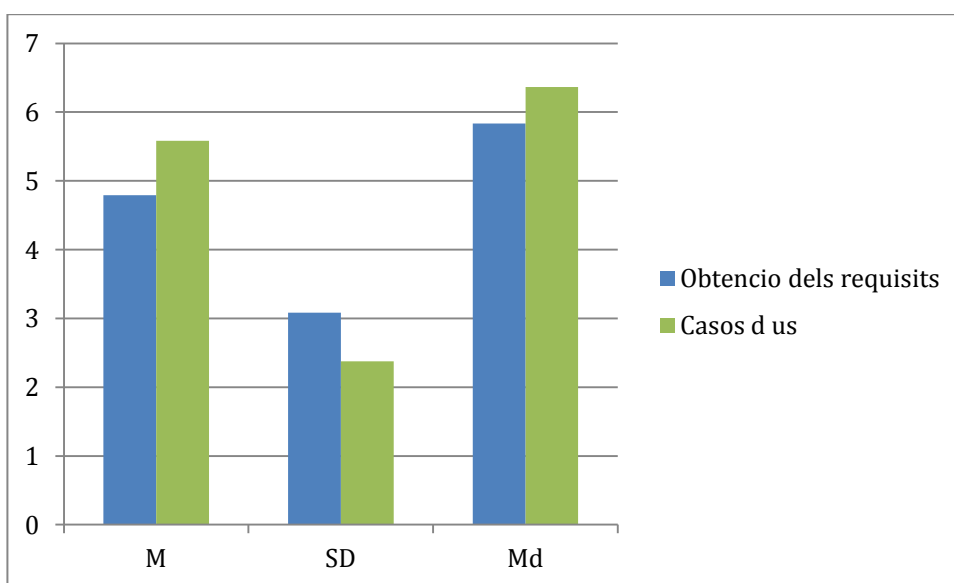


Figure 6: Comparison between ULLG and competence level for GM

In general, we denote good levels of the acquired competences that contributed to increase the students' activity levels. In particular the students from Computer Science degree have recorded a Media value better than the Students from the Multimedia degree. Indeed for the GM students the Figure 7 shows a high value of the standard deviation with respect to the two delivered ULLG. That it is due to a very low competences acquisition equivalent to zero for some students. That could be related to the more predisposition of the Computer Science' students with respect to the considered topics.

2.3.2 Usability of the IWT

To evaluate student's satisfaction with the tool regarding an efficient and user-friendly management (H1.1), we collected from students' ratings and open comments on the usability/functionality/integration of the tool.

To investigate the overall usability of the IWT system, we used the SUS (see Section 2.2) included in section (v) of the questionnaire. As mentioned, the answers were given on the 5-point Likert scale, so that students could state their level of agreement or disagreement. The rating scale ranged from "I strongly disagree" (1), "I disagree" (2), "neither/nor" (3) to "I agree" (4), "I strongly agree" (5).

SUS scores have a range of 0 to 100 with an average score of 68, obtained from 500 studies. A Score above a 68 would be considered above average and anything below 68 is below average. A score above an 80.3 is considered an A (the top 10% of scores). Scoring at the mean score of 68 gets you a C and anything below a 51 is an F (putting you in the bottom 15%).

After calculating the SUS score for each student, we got an average for 41 SUS scores of 60.78 thus below the SUS mean but nearby, which is a good score considering the first development iteration of the CGS and its integration in IWT. Next, we present the most relevant results of the SUS score by providing several statistics: Mean (M), Standard Deviation (SD) and Median (Md).

Students found the IWT particularly easy to use ($M = 3.42$, $SD = 1.00$, $Md = 3.5$) (See Figure 8). Students did not find much inconsistency with the IWT interface ($M = 3.3$, $SD = 1.1$, $Md = 3$) (See Figure 9). In addition, students stated that they did not need the support of a technical person to be able to use the IWT ($M = 1.92$, $SD = 0.88$, $Md = 2$) and they thought that most people would learn to use IWT very quickly ($M = 3.18$, $SD = 1.20$, $Md = 3$) (See Figure 11).

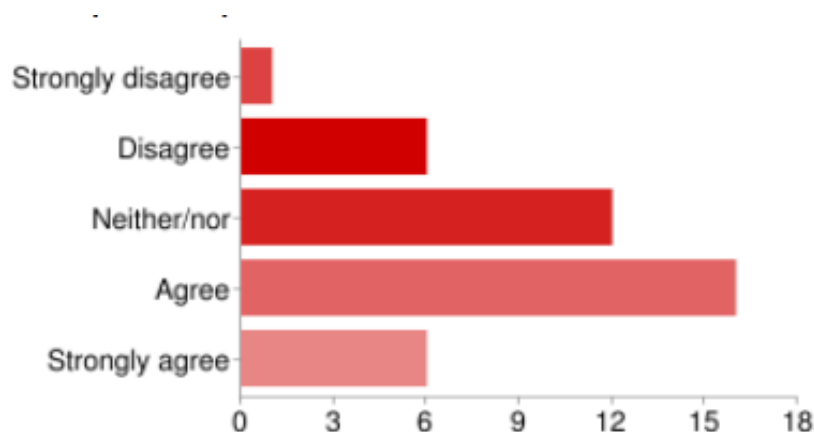


Figure 8: Results on the SUS item "I thought the system was easy to use".

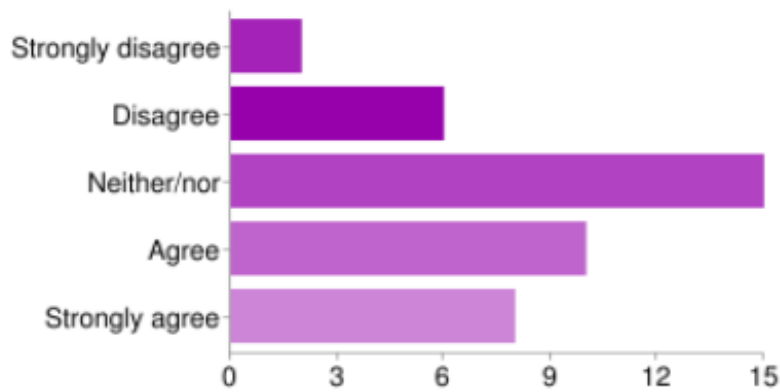


Figure 9: Results on the SUS item “I thought there was too much inconsistency in the IWT”.

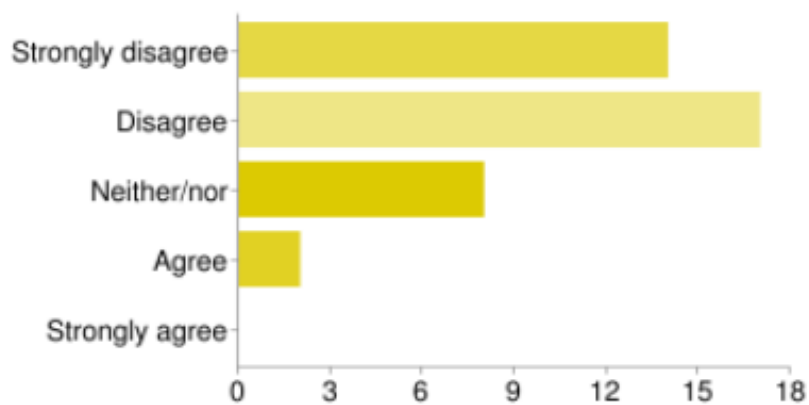


Figure 10: Results on the SUS item “I think that I would need the support of a technical person to be able to use the IWT”

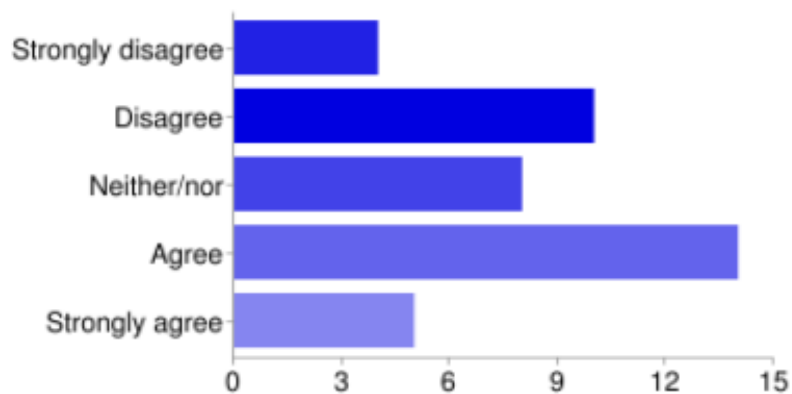


Figure 11: Results on the SUS item “I would imagine that most people would learn to use the IWT system very quickly”.

Many students (about 40%) reported that the study area of IWT was too small and very uncomfortable to read and move in the learning material (either PDF or Web format). They also missed the possibility to take notes on the own material. Some students indicated that the page navigation was unclear and the graphical interface not very pleasant while others found IWT comfortable and accessible. Finally, students found the system sometimes little responsive and performing slow.

In accordance with these results, students indicated in a balanced way they would and would not use the IWT system frequently ($M = 3.26$, $SD = 1.15$, $Md = 3.5$) in line with the overall SUS score of 60.78 and in Figure 12.

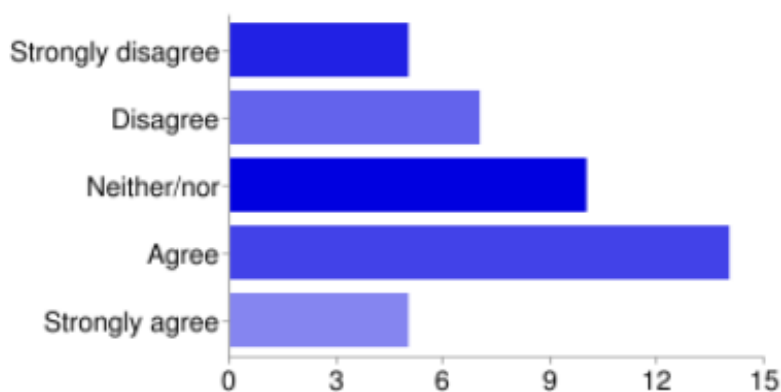


Figure 12: Results on the SUS item “I think that I would like to use this system frequently”.

Finally, students stated that the IWT was not very well integrated in the UOC classroom ($M = 2.65$, $SD = 0.96$, $Md = 3$) (see Figure 13). In particular, the access to IWT from UOC classroom was available only from within the communication area of the classroom though many students tried to find shortcuts and failed. The access to IWT was later on extended and accessible directly from multiple locations around the classroom. On the other hand, students appreciated to be able to accede to the IWT directly with neither re-authentication nor further navigation to the targeted web space.

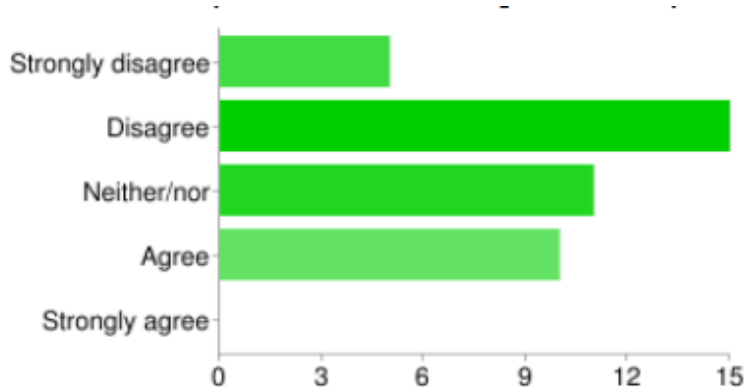


Figure 13: Results on the SUS item “I found the various functions in the IWT were well integrated”.

In overall, this is a very good result and very promising to face the second iteration of the project with a more advanced interface and having fixed the usability problems found in this first iteration.

2.3.3 Emotional aspects

Regarding the students' emotions during the work with the IWT tool (H1.1), we used the CES scale to analyze the emotional aspects. The answer categories and the scores to compute them are "None of the time" (0), "Some of the time" (1), "Most of the time" (2) and "All of the time" (3). The results from a 4-point rating scale (n=41) were as follows:

- Happiness (M=1.39, SD=0.73, Md=1) (Figure 14)

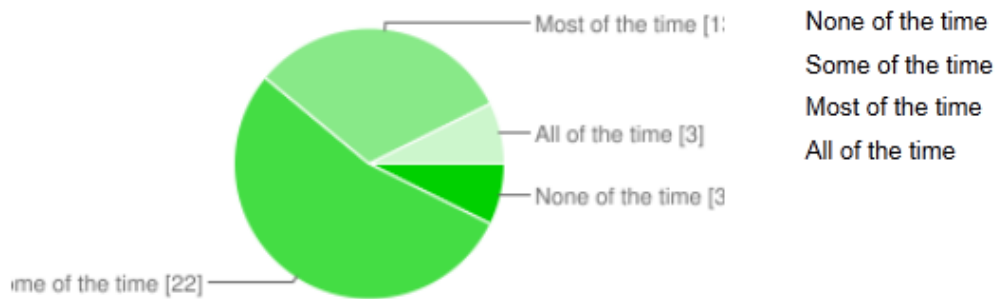


Figure 14: Results on the Happiness emotion

- Sadness (M=0.70, SD=0.49, Md=0) (Figure 15)

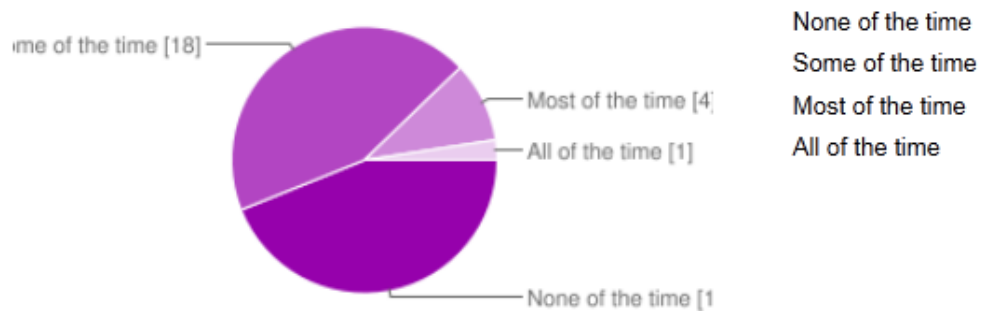


Figure 15: Results on the Sadness emotion

- Anxiety (M=0.60, SD=0.73, Md=0) (Figure 16)

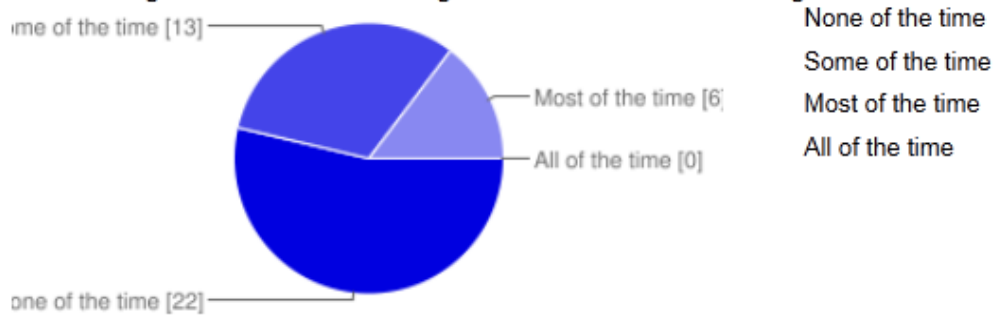


Figure 16: Results on the Anxiety emotion

- Anger (M=0.41, SD=0.74, Md=0) (Figure 17)

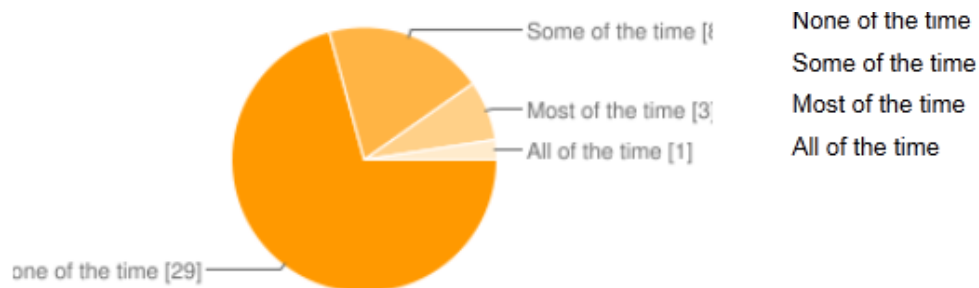


Figure 17: Results on the Anger emotion

Happiness emotion is nearby the mean and students felt significantly much more often happiness than sadness, anxiety or anger when learning by means of the IWT tool. In addition, students felt less anger and anxiety than sadness, being these two emotions low or very low (with standard deviation higher than the calculated mean, which emphasizes the very low mean values of both anger and anxiety). These results are in line with the results presented above concerning the evaluation of usability of the IWT about the SUS mean (see Section 2.3.2). As already discussed in the usability aspects, no appreciable signs of anger, anxiety and sadness emotions were reported by the students and the level of satisfaction (ie., happiness emotion) was.

In overall, this is a very good result and very promising to face the second iteration of the project, where the system will be improved and hence students will feel even better on the emotional scale..

2.3.4 Questionnaire evaluation

The questionnaire was designed not to be very intrusive in the students' responses by avoiding exceeding the length and/or time employed to fill it. Evaluation results of the suitability of the questionnaire design confirmed the expectations resulting in most of

students (73%) filling and submitting the questionnaire in less than 30 minutes (Figure 18) and 76% of them found it appropriate to evaluate the experience (Figure 19).

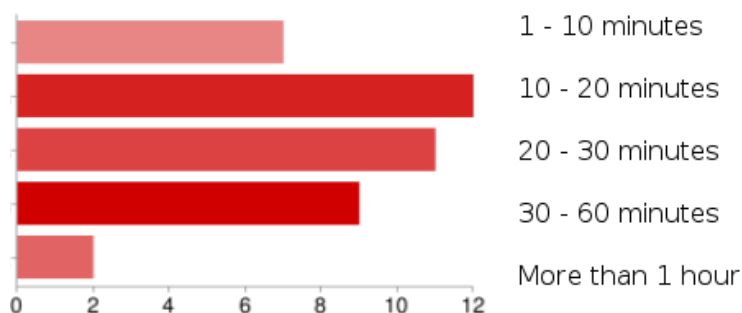


Figure 18: Time employed to fill the questionnaire

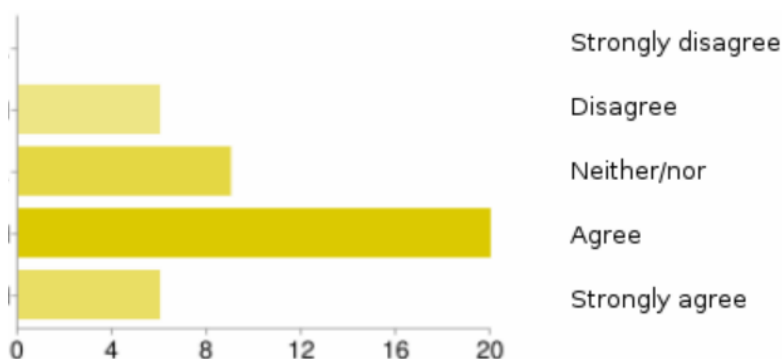


Figure 19: Appropriateness to evaluate the experience with the questionnaire

2.4 Validation Results

In this section we show the validation methodology that includes the following validation criteria and metric extrapolated by [4].

Validation criteria

- C1.1: To evaluate the level of fulfilment of the tool features.
- C1.2: To evaluate the level of satisfaction of the students that use the CGS.
- C1.3: To evaluate the increase in students' motivation caused by the use of the CGS.
- C1.4: To evaluate the increase in students' understanding of key concepts and students' results caused by the use of CGS.
- C1.5: To evaluate the increase in students' activity levels due to the use of the CGS.
- C1.6: To evaluate the level of satisfaction of the instructors with the inclusion of the CGS as a learning resource in their courses.
- C1.7: To evaluate the potential reduction of the time between the emerging of a new learning need and its fulfillment thanks to the CGS.

Validation metrics

- M1.1: Number of courses created with the CGS.
- M1.2: Time employed in creating each course with the CGS.
- M1.3: Number of students using the CGS.
- M1.4: Number of visits of learning objects alternative to those included in courses generated by the CGS.
- M1.5: Students passing the final test and/or with high marks when the CGS is used.
- M1.6: Students passing the final test and/or with high marks when the CGS is not used.
- M1.7: Number of students that consider that the CGS is worthy.
- M1.8: Number of instructors that consider that the CGS is worthy.

Following this methodology we will validate the improvement of emotion and motivation (H1.2), worthiness as an educational tool and teaching supporting tool of the IWT (H1.3 and H1.6) as well as the acquisition of collaborative knowledge (H1.5).

2.4.1 *The IWT as a valuable resource*

In this section we analyze the IWT as a valuable educational resource by the evaluation of the worthiness of the IWT as an educational tool (H1.6). To this end, quantitative and qualitative data were collected in sections (iii) and (iv) of the questionnaire by 3 open questions (qualitative) and then 13 test-based questions (quantitative) plus one final open question to provide suggestions for improvement.

In the questionnaire, the rating scales for the three quantitative questions we used a 0-10 point scale, so that students could assess the value of the IWT tool by a scale they felt very familiar with from their experience in the UOC courses. The scale went from the worst mark (0) to the best mark (10) considering a “good” assessment marks from 5.0 to 10 and a “bad” assessment marks from 0 to 4.9.

As for the test-based questions the rating scale ranged from “Not at all” (1); “Somewhat” (2) and “Completely” (3). Despite sometimes these values changed to fit best the expected type of the responses, in all cases 3 options were provided (positive, medium and negative).

Open questions

Three open questions were asked to students about IWT:

1. Evaluate in general the new IWT classroom to support the study of the course “Requirements” (Assess the IWT from this view in the scale 0-10).
2. Indicate how in your opinion the IWT classroom has impacted in your individual learning process as for the topic “Requirements” (assess the IWT from this view in the scale 0-10) (Assess the IWT from this view in the scale 0-10).

3. In comparison to the UOC classroom what advantages and disadvantages do you think IWT provides to study? Indicate in your view what are the main problems, issues and lacks of this tool (Assess the IWT from this view in the scale 0-10).

After calculating the 0-10 scale for each student we got an average of 6.14 (SD=2.27, Md=7). This result is very good considering the IWT tool is still in the first iteration of development.

Students in general liked the IWT system and found it useful for their study (Question 1: M=6.13, SD=2.31, Md=7). The IWT aspect that most liked to students by far was the self-evaluation capabilities by means of on-line tests exercises. Students commented that the on-line tests allowed them to combine study and evaluation while making progress in the learning process. Also students indicated the flexibility and personalization of the study proposed by the search engine from natural language, and they found the effective structure of the course and the connection links provided by IWT as very positive aspects. Finally, students found interesting the possibility to assess the learning material. As negative aspects, they commented the usability problems when reading the PDF material in the study area and also the navigability not being fluent and certain technical problems when studying in IWT. Finally, they indicated the need to read the user manual provided all the way before being able to use the system effectively. Since students have a strong technical background (they belong to the Computer Science or Multimedia degree), this comment is relevant. They provided some hints for improvement following these comments.

Question 2 was slightly better scored than Question 1 (M=6.39, SD=2.33, Md=7). Again, students considered the possibility of self-evaluation very important for their learning process in order to clarify doubts, revise the material where problems arose, guide through the learning path and confirm the concepts learnt about the topic under study. On the other hand, they indicated that even though the self-evaluation exercises were very useful they did not find the study to be easier with IWT nor improve their knowledge significantly. This is in line with the results of the evaluation on assessment reported in Section 2.2.4.3.

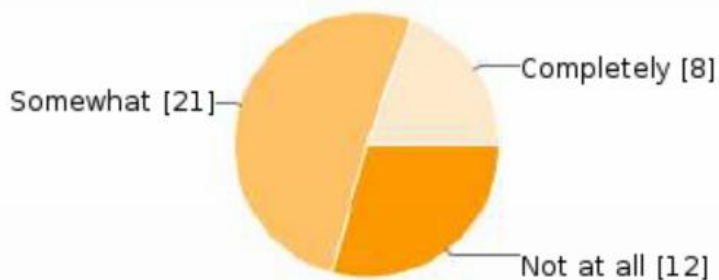
Finally, Question 3 was scored a bit lower than the other 2 questions, though not significantly (M=5.91, SD=2.19, Md=6.5). Students commented that the IWT provided a higher degree of flexibility and personalization than UOC. They also stated that self-evaluation capability provided by IWT was innovative for them and they would like to have it at UOC. Students highlighted the advances in the IWT forum tool and the VCS for in-class communication with respect to the UOC forums (see R4 in Section 5). On the other hand, students commented that the UOC user interface was clearer and easier to use than IWT's. Also they indicated that the UOC materials are more comfortable to read than those in IWT and that at UOC materials were found in different formats, thus making it possible to study from mobile devices.

Test-based questions

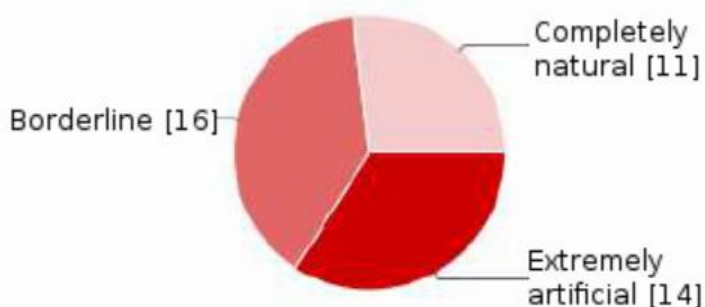
The quantitative results can be checked

13 test-based questions were asked to students:

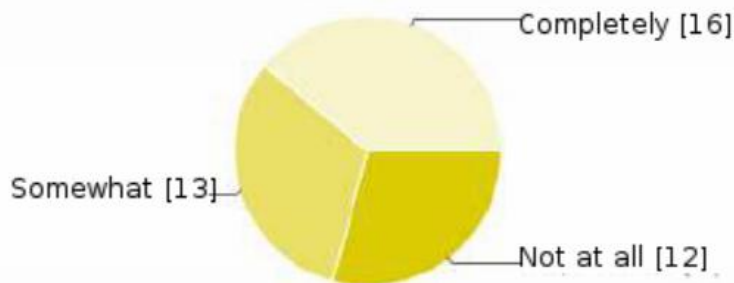
1. The possibility to express your formative needs has allowed you to have more control over your learning?



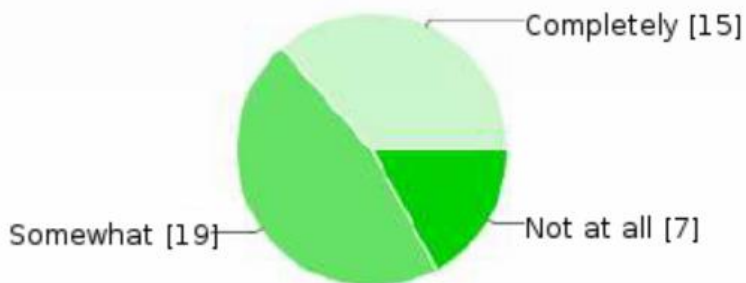
2. Being able to express your needs in a simple language has contributed to motivate your desire to learn?



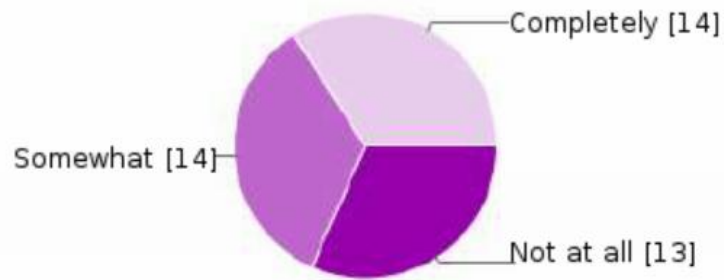
3. Do you think that this solution of asking to take more responsibility over what you need helped you to capture a greater awareness on the right learning path?



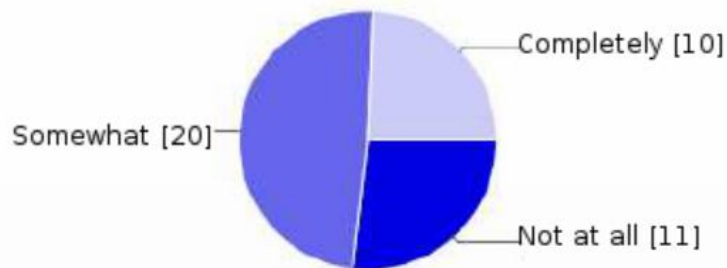
4. Do you think that the answers obtained in terms of learning paths to follow by filling your needs are relevant and effective?



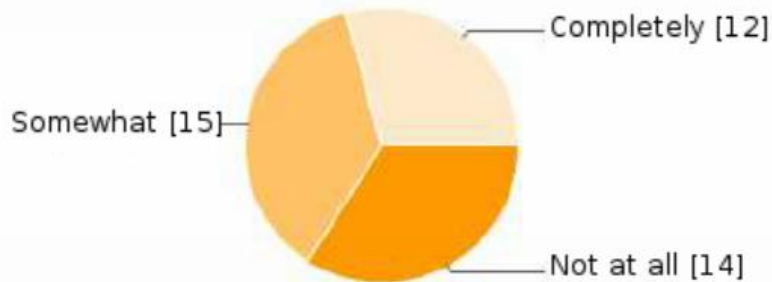
5. Do you think you can speed up learning time by eliminating states to which you are subject when your path is guided by the teacher?



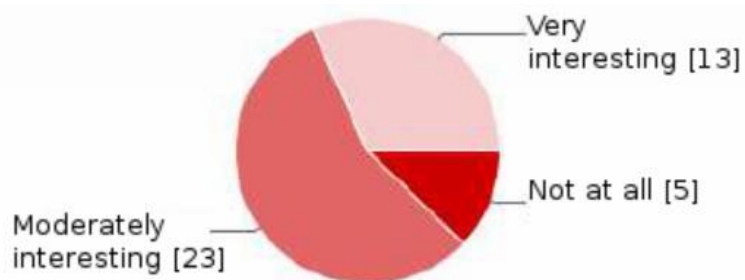
6. The possibility to have specific learning path created ad hoc for filling your need has allowed you to obtain good results in terms of learning?



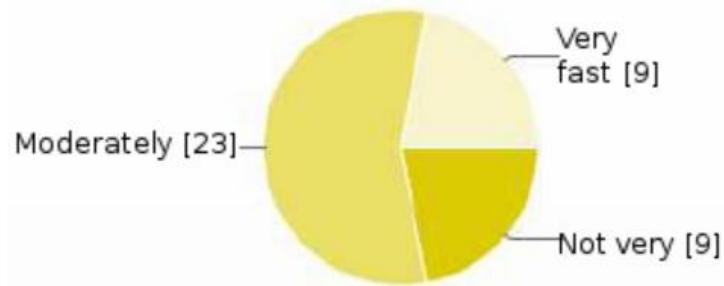
7. Did this learning modality have an impact on your participation in the learning experience?



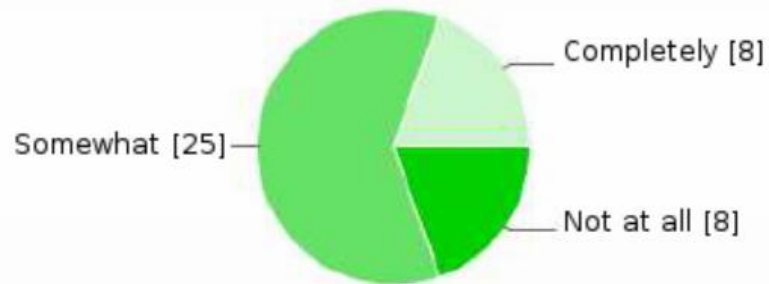
8. How you did you find the interaction with this new method of learning experience?



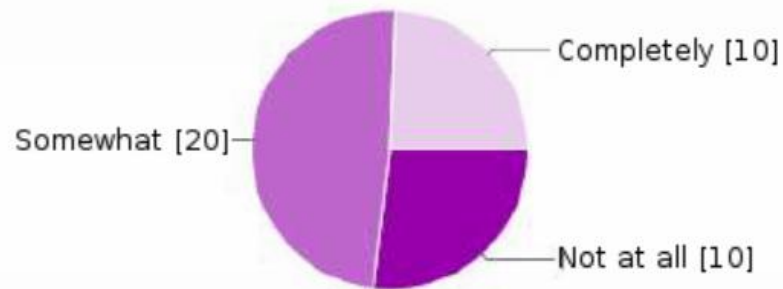
9. How quickly you adapted to this new method of expression through natural language?



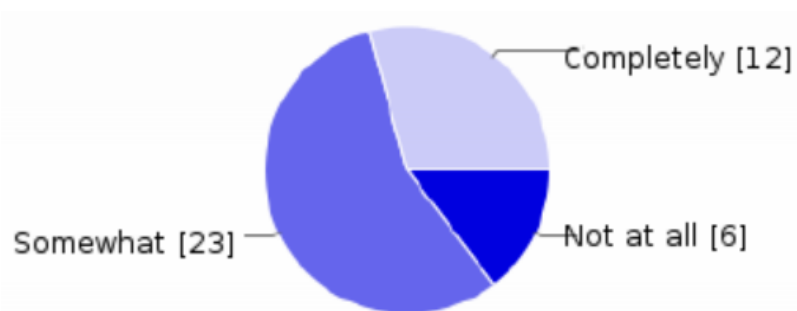
10. Do you think that this new kind of interaction modality student-learning environment can be a step towards a self-regulated learning?



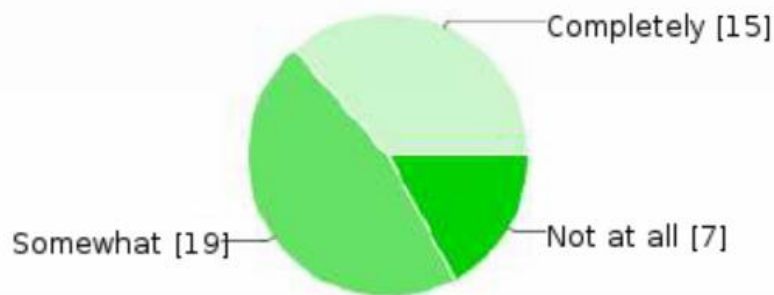
11. Do you think that the recommendations you received in terms of learning path to follow were tailored to your learning style and your profile?



12. How do you consider the learning path predisposed for you for filling your learning needs?



13. The ability to quickly obtain the recommendations brought you to express more than one need?



Final open question for improvements

This open question completed this section of the questionnaire by asking students for giving final hints for potential improvement of the IWT tool. Students proposed to improve usability aspects of the system, such as the small window to read the material and the system navigation (e.g., menus and back-page button). Also, students commented that improvements on system's responsiveness will be interesting.

2.4.2 Motivational aspects

Students' motivation concerning the use of IWT tool (H1.2) was directly investigated naively by including in the Section (iii) of the questionnaire a motivation test, where all students were asked for the amount of motivation they felt when studying by using IWT. The following answer categories were used: "absolutely unmotivated" (1), "unmotivated" (2), "motivated" (3), "very motivated (4)".

Test results provided a score above the mean ($M=2.79$, $SD=0.81$, $Md=3$). This result is in line with the results on the IWT being a valuable resource and also with the usability and emotional results reported in the previous sections. In particular, students indicated to feel very motivated by the dynamic on-line tests found in the course that allowed them to clarify doubts and revise certain parts of the course by following the suggestions of the system.

Finally, clear indications of amounts of motivation came from enthusiastic students who commented that the IWT was a tool "very interesting", "very useful", and "it will change the way to study in the future". However, most of them clarified that the system needed usability improvements and a more fluent navigation before considering IWT to be successful. Eventually, most of students understood it was a pilot trial and IWT was in a beta version and for this reason they overcame some steps of little motivation.

2.4.3 Tutor assessment and knowledge acquisition

All students from both the experimental and the control groups were evaluated on the responses obtained from the questionnaire. To this end section (ii) of all questionnaires included an evaluative assignment with 2 questions about the topic "Requirements" they have studied in either IWT or UOC, as follows:

1. From your experience as a user of social networks (e.g., Facebook, Twitter, etc), indicate 5 functional requirements and 5 non functional requirements implemented in these systems. Classify the non functional requirements according to the Volere template.
2. Indicate what the problems are to identify requirements during their elicitation.

While Question 1 is more general and practical Question 2 is more specific and theoretical. This aim was also to evaluate the impact both on general and on specific acquisition of knowledge.

This part of each questionnaire was assessed by a lecturer who used the standard 10-point scale to score the students’ responses. *Table 5* shows the results.

Evaluative questions	Experimental group (n=41)	Control group (n=27)
Question 1	M=6.38 SD=1.64 Md=6	M=6.11 SD=1.56 Md=6
Question 2	M=7.83 SD=0.78 Md=8	M=6.33 SD=1.28 Md=6
Overall	M=7.11 SD=1.46 Md=7	M=6.22 SD=1.41 Md=6

Table 5: Results of the learning assignment evaluation

From the results of *Table 5*, students from the experimental group (UOC + IWT) scored higher than the control group (UOC) though the overall difference is not significant. However, observing closed the results; while Question 1 got similar marks, Question 2 the marks were significantly different (1.22 out of 10). More interestingly, the SD in Question 2 of the experimental group is considerably lower than in the other group for the same question and also lower than the other questions and groups. This result is in line with the fact that the students could find a specific resource in IWT devoted to answer this question while UOC students had the information related to this question more dispersed in their material.

Both groups got good marks on average and showed a good level of knowledge acquisition. These results are in line with the results from the impact of the IWT in the students’ (see Question 2 in Section 2.4.1) but also in line with the results reported in Section 2.4.2 where students indicated that the IWT did not help acquire new knowledge but consolidate their current knowledge.

In summary, we conclude that IWT provided students with more specific knowledge and according to the needs expressed by them (using the CGS system).

2.5 Conclusion

In this Section, the results are summarized and discussed by considering the goals which were determined at the beginning of the study (see also Section 2.1). Then, based on the results summarized further research and technological directions are proposed.

In general the students liked the IWT tool and found it interesting to have a personalized system to study. From the results of the previous sections it was evident that IWT was able to generate course from the CGS from a need expressed in natural language by the learner (G1.1). In particular, results from Sections 2.4.1 and Section 2.4.3 showed that these courses had been fulfilled the expectations of the learners (G1.2) though not completely.

IWT usability was not a barrier when using the system (G1.3) though it was the most important technical aspect considered by students. Even so, they showed a constructive attitude most of the time that did not have a side-effect in their emotions when using the system. Indeed, comments on usability refer to particular aspects of the system, such as the study area or the navigation button. Eventually, it was noticeable important amounts of resilience to change the e-learning platform from UOC to IWT.

Validation of the impact of IWT in effective learning of scientific concepts was analyzed and evaluated (G1.4) by chiefly Section 2.4.3 on assessment. It was concluded that IWT provided students with more specific knowledge and according to the needs expressed by them (using the CGS system).

Finally, possible ways of improving further the utility of the CGS (G1.5) and a larger extend of IWT were provided in several sections, and mainly at the end of Section 2.4.1 being most of the comments addressed to usability.

The latter conclusion is in line with the current stage of the IWT technological development, which is expected to be further improved during the second stage of the project and especially from the valuable feedback collected from this experiment.

3 R2. Knowledge model contextualization: Experimenting the Knowledge model contextualization

The aim of this scenario is to build an ontological description of a teaching domain that is able to automatically adapt to a context (see [4]). Two pilot sites run trials on this scenario from the instructor's viewpoint. A third trial was run from the students' viewpoint. In summary:

1. Knowledge model contextualization: Experimenting the Knowledge model contextualization from the instructor's viewpoint at TUG (Section 3.1)
2. Knowledge model contextualization: Experimenting the Knowledge model contextualization from the instructor's viewpoint at UOC (Section 3.2)

3.1 R2-1. Knowledge model contextualization: Experimenting the Knowledge model contextualization from the instructor's viewpoint (TUG)

3.1.1 *Research goals and hypotheses*

We conducted a first experiment on this scenario at TUG pilot site in order to test the tool from the instructors' viewpoint. The results of this study give us a first impression of how the tool supports instructors in order to create online courses and whether it needs further enhancements. Therefore, in this study we were primarily interested in the functionality and usability of the tool.

To experiment the knowledge model contextualization from the instructor's viewpoint, we focused on the following goals and hypotheses as described in [4]:

Goals

G2.1: to ensure that the system is able to generate contextualized courses by selecting a domain ontology, a context, a set of target concepts and a learner.

G2.2: to identify possible ways of improving further the utility of the tool and related models and algorithms.

G2.3: To provide a tool that supports the work of the instructors.

Hypotheses

H2.1: a set of feasible courses can be effectively and efficiently created starting from a domain ontology by selecting a context, and a set of target concepts.

H2.2: automatically generated courses are considered as a worthy educational resource by the instructors.

H2.3: automatically generated courses are considered as a worthy educational resource by the instructors.

In order to investigate these goals and hypotheses, we asked two lecturers from two different universities in Graz to create a personalized course about “Scientific Working” using the IWT:

(1) At the Karl-Franzens University (KF), the Institute of Psychology offers courses about scientific working. Lecturer A, who participated in this study, is conducting such a course for about 50 undergraduate students. The course is a face-to face course held in German. The main learning objectives of the course are that students get a deeper insight in scientific working and that they are able to plan and conduct their own (psychological) experiments. Students need basic pre-knowledge both in statistical analysis and research designs in order to enroll the course. Additional material for the course (e.g. slides) is presented on the web site of the lecturer. However, lecturer A is quite favorable to use a learning platform like the IWT to create an online course for the students in order to support them effectively.

(2) Although the topic is very important, at the Technical University (TU) there exists no special course for scientific working. Thus, lecturer B wants to provide a course for PhD students which covers this topic. These students have only minor previous knowledge in scientific working. The course should be an online course because most of the PhD students already work and/or are located in different countries of the world. Furthermore, the learning content has to be provided in English, as some of the students do not speak German.

As both courses have much in common, it stands to reason that the lecturers use the IWT in order to create a course that fits to lecturers’ (and students’) needs.

Regarding the methodological approach of the study, the lecturers were asked to log all their activities concerning the experiment during the study. In their documentation they noted for each step the time they spent on working with the IWT. In addition, the lecturers listed all problems they had to face while working with the system and wrote down advantages and disadvantages.

In addition, both lecturers were asked to fill in the SUS (System Usability Scale; [6]) after the end of the session. For qualitative statistical analysis, we summarized the open answers in the surveys. Note that participants had to sign an informed-consent sheet in order to participate in the study.

3.1.2 Method

3.1.2.1 Participants

Two lecturers, one from the Karl-Franzens University (lecturer A) and one lecturer from the Technical University (lecturer B) participated in our experiment. Both are experienced in higher educational teaching. Lecturer A has been working for three years at the Institute of Psychology at the Karl-Franzens University (KF) and lecturer B has been working for 13 years at the Technical University (TU). According to their experiences with learning platforms, lecturer A has only basic knowledge and lecturer B has advanced knowledge using learning platforms.

3.1.2.2 Apparatus and Stimuli

First of all we asked two instructors to use the IWT (Intelligent Web Teacher) to create a personalized course. The IWT is able to generate contextualized courses by selecting a domain ontology, a context, a set of target concepts and a learner (see [1]).

We used the SUS (System Usability Scale) by [6] in order to investigate the usability of the IWT. The SUS is a simple, ten-item attitude Likert scale giving a global view of subjective assessments of usability. It is generally used after the respondent had an opportunity to use the system being evaluated. SUS scores have a range of 0 to 100 with an average score of 68, obtained from 500 studies. A Score above a 68 would be considered above average and anything below 68 is below average. A score above an 80.3 is considered an A (the top 10% of scores). Scoring at the mean score of 68 gets you a C and anything below a 51 is an F (putting you in the bottom 15%).

3.1.2.3 Procedure

The experiment consisted of two sessions, conducted on two different days within one week: In the first session, each of the lecturers was asked to find concepts and to create an ontology of his/her course on a paper. In order to create the ontology, they used three types of relations which are also provided by the IWT: (a) “*has part*”, (b) “*is required by*” and (c), “*suggested order*”. The relation “*has part*” is needed to indicate that a concept is a sub-concept of another concept. The relation “*is required by*” means that a concept is a requirement of another concept. The relation “*suggested order*” shows that a concept should be explained before another concept.

Then the lecturers were asked to share and discuss their ontologies and also their ideas concerning the “*Scientific Working*” course. As a next step, the lecturers tried to find common concepts for the course together, but also defined individual concepts for each course. Finally, they created a paper-pencil version of an ontology based on these concepts.

In the second session of the study, the lecturers created an online course on the IWT using the concepts they had developed in session one. In order to create such an online course, they had to

- (1) Create a dictionary
- (2) Create an ontology
- (3) Upload the contents
- (4) Create a customized course

(1) Create a Dictionary

The dictionary provides the key concepts for the teaching subjects (see Figure 20). It is possible but not necessary to enter a description of the concepts. After typing the terms in, the dictionary has to be saved.

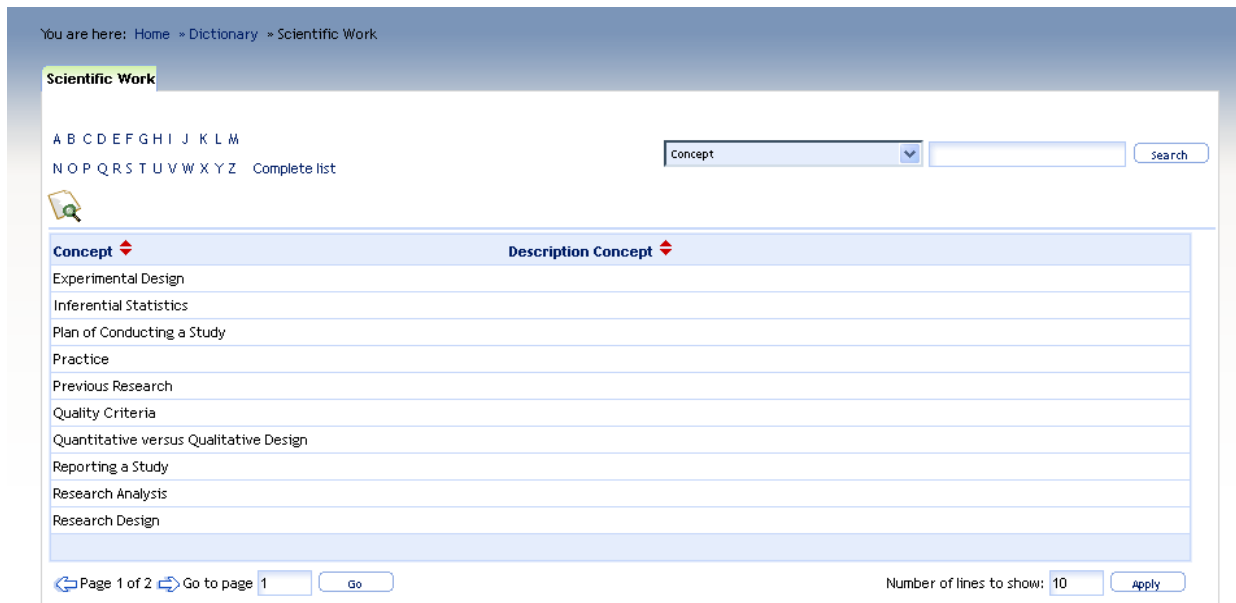


Figure 20: Dictionary “Scientific Working” with concepts

(2) Create an Ontology

The available concepts from the dictionary “Scientific Working” have to be arranged in a specific order to facilitate further steps. Furthermore, the relations “has part”, “is required by” and “suggested order” are added to the concepts. The context data has to be provided for both courses and assigned to the contexts (see Figure 21).

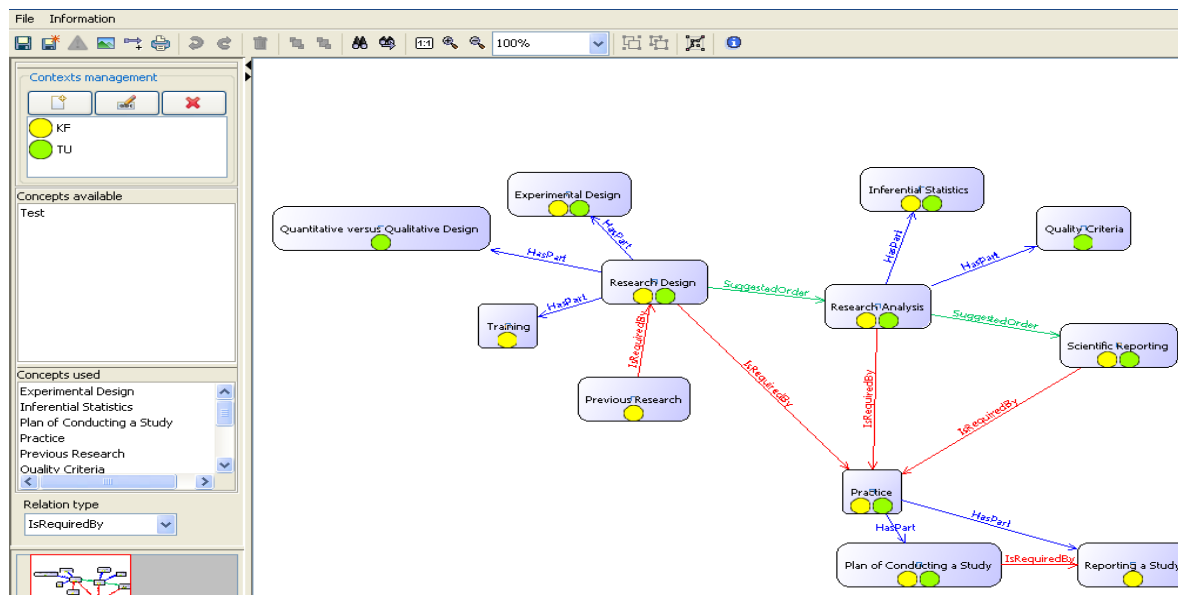


Figure 21: The new created ontology with the concepts, the context and the relations

(3) Upload contents

In this step, the learning contents had to be uploaded to the system. Lecturer A provided parts of the contents of her course on the IWT and lecturer B added some basic contents, which should help the technical students to understand the terms of “Scientific Working” better (see Figure 22). Note that, due to legal reasons, for this study both lecturers did not upload the full contents for the course.

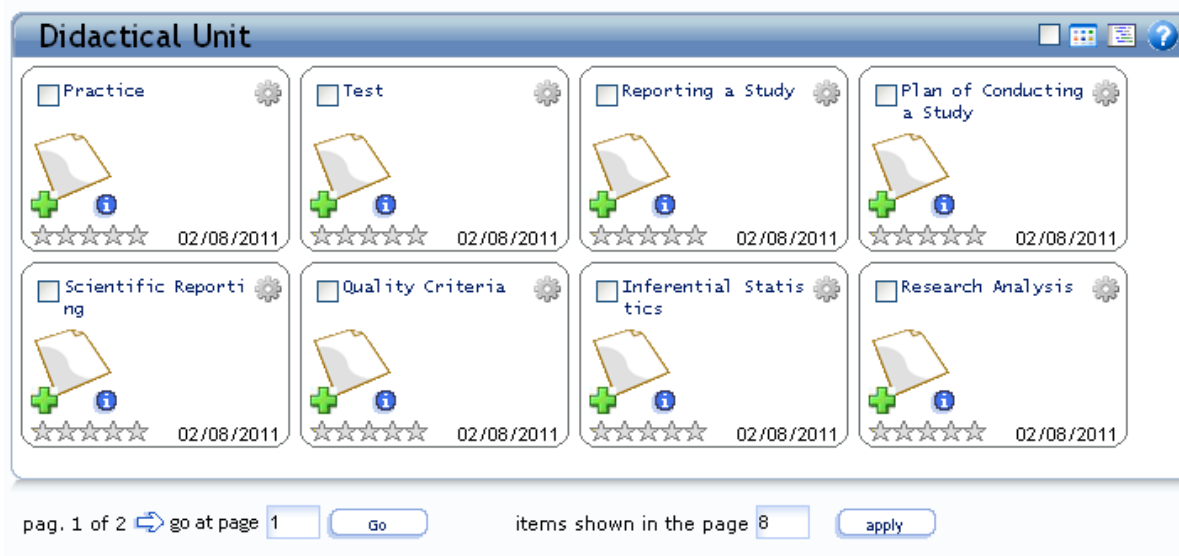


Figure 22: Uploaded contents for the courses

(4) Create a Customized Course

Afterwards, the customized course has to be created and the target concepts for both courses have to be defined (see Figure 23). Also the didactic approach (i.e., the didactic path, the language, the teachers defined profile etc.) have to be settled.

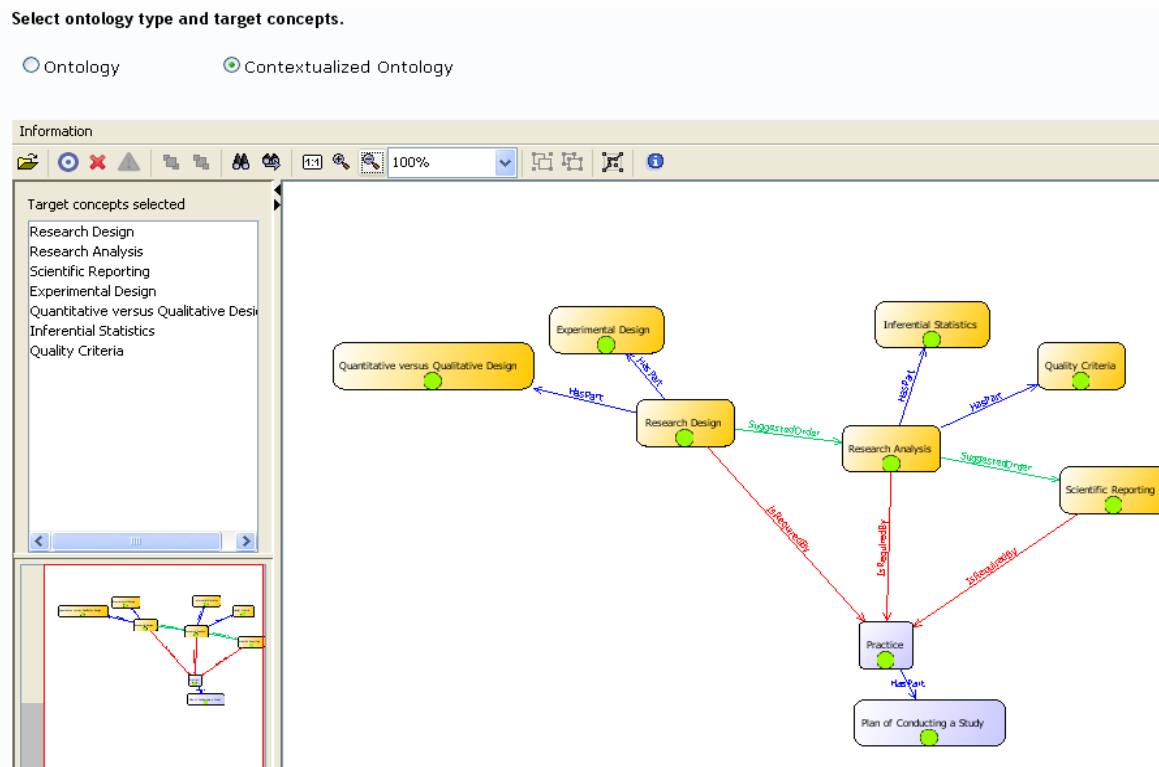


Figure 23: Target concepts for the course “Scientific Working Technical University”

The lecturers were instructed to use the manual of the IWT as provided in Deliverable D7.4.1. They worked every step together on the IWT to support each other in case of doubts. After the task was finished, the teachers were asked to fill out a questionnaire about their experiences with the system, especially concerning the usability of the IWT.

According to the procedure, the lecturers had participated in two sessions. In the first session, they had to create a paper-pencil version of an ontology of the course. In the second session, they used this paper-pencil version in order to create a course with the IWT. In the next sections, the results of the study are presented.

3.1.3 Evaluation Results

In this section we show the validation methodology that includes the following validation criteria and metric extrapolated by [4].

In this section we focus on instructors' perception of the VOE and possible improvements of the tool by evaluating H2.1, H2.2, and H2.3 as they are specified in [4].

Following this methodology we will validate 3 aspects of the scenario: time to run the experience (H2.1), the usability of the IWT (H2.1) and the lecturer's emotions when using the IWT (H2.2).

3.1.3.1 Time to create a course

Session 1: Create a paper-pencil version of the course

In the first meeting, lecturer A and B discussed the most important concepts they have in common and wanted to take part in the courses. It took them 1 hour and 40 minutes to discuss and define the concepts of the course on paper.

All in all, the participants extracted thirteen concepts. Eight out of these concepts could be used for both courses; three were only applicable for the KF course and two for the TU course (see *Table 6*).

Concept	KF	TU
Experimental Design	✓	✓
Inferential Statistics	✓	✓
Plan of Conducting a Study	✓	✓
Practice	✓	✓
Previous Research	✓	
Quality Criteria		✓
Quantitative versus Qualitative Design		✓
Reporting a Study	✓	
Research Analysis	✓	✓
Research Design	✓	✓
Scientific Reporting	✓	✓
Test	✓	✓
Training	✓	

Table 6: Created concepts for the dictionary

Session 2: Creating a course using the IWT

In the second session the lecturers started working on the IWT. First of all they provided the concepts in the dictionary, which took them 12 minutes. Creating the ontologies for both courses took them further 30 minutes.

Lecturer A provided contents of her course on the IWT and the lecturer B added some basic content, which should help the technical students to understand the terms of “Scientific Working” better. Uploading the 13 didactical units and providing additional information took them about 40 minutes.

Afterwards, the customized courses were created and the target concepts for both courses, for the Technical University and for the Karl-Franzens University, were defined. At the same time, the didactic approach was settled. Finally, the courses “Scientific Working Karl-Franzens University” and “Scientific Working Technical University” were finished in about 15 minutes.

Thus, over the sessions it took the instructors 197 min (3h17m) to create the course, i.e. to define concepts, create a dictionary, create ontologies, upload contents, and customize the course. This means that the tool effectively supports teachers in creating courses (see H2.1).

As the focus of this study lies on possible improvements for the tool from the viewpoint of the instructors, we analyzed the usability of the tool (see also H2.2). Both lecturers were asked to fill in the SUS (System Usability Scale; [6]) after the creation of the courses.

3.1.3.2 Usability of the IWT

We calculated the SUS score separately for each lecturer. The score for lecturer A was 65 and the score for lecturer B 57.5 and belong to the bottom 30% to 40%. Regarding the positive aspects of the tool, the teachers would like to use the system more frequently. Additionally, they did not find the system unnecessarily complex or too inconsistent but found the various functions of the system well integrated. Concerning the negative aspects, the teachers needed to learn a lot before they could get going with the system. One teacher would need a technical support to be able to use the tool.

3.1.3.3 Self-generated questions concerning the usability of IWT

In order to find enhancements for the tool, we asked the lecturers to answer five open questions. In this section, we summarize the answers (see H2.2).

1. Please describe what you liked regarding the system.

While lecturer A states that she liked the possibility to create a course together, the idea of the ontology with the relations and the good overview of the course, lecturer B likes the idea of the online tool, the functions and the design. *“It enables an interdisciplinary exchange and a collaborative work. The interface of the tool is pleasing. The graphic design is appealing, especially the selection of the didactical approach. Also, uploading different materials to the contexts is easy. The tool has got many additional functions like evaluation and the entry test. In sum, the tool is well conceived and you can tell that much work was done behind it.”*

2. Please describe what you did not like regarding the system.

Lecturer A did not like the complexity and the difficulty that appears when no support is provided. Lecturer B did not like the search function. *“If you want to search a term or a*

procedure, you have to type in the exactly fitting word to succeed. Other words simply do not suffice. Furthermore, the site takes very long to load Java”.

3. Do you have any suggestions for improvements?

The missing possibility to link the content to a context and the difficulty with finding some features (e.g. contextualized ontology) bothered lecturer A, *“a supportive quick would be nice”*. Lecturer B would optimize the search function of the tool (see also question 2). *“Even keywords should suffice to find specific issues. Furthermore, additional mouse-over-text descriptions would provide more information about some features. Additionally, an online tutorial (e.g. video) on the start page would be helpful.”*

4. Concerning the user manual you have got, how clear was the description of the IWT for you? Did the user manual support you in following the individual steps?

Both lecturers underline the manual’s incorrect order. *“I had to prepone one chapter, because otherwise I had not been able to accomplish the steps for the antecedent chapter”*. Furthermore, lecturer A would prefer a demonstration video on the IWT. Lecturer B adds that the user manual helped him with the tools’ handling. *“The steps are good described and the attached images facilitate the understanding”*.

5. From your point of view, do you think that teachers would like to use IWT to create and plan online courses? What are the pros and cons?

Meanwhile, teacher A repeats the already mentioned pros. *“To create an ontology together is very nice, I really like the idea of the concepts, linking them together, adding contents and the dictionary. A disadvantage is that you cannot link the content to the context- this should be possible, otherwise the teachers have to use the same content. Concerning the relations, the direction is not logical and it should be possible to define relations depending on the context”*. Meanwhile, teacher B would recommend the IWT tool, if the above mentioned issues were corrected. She thinks that teachers would like the tool. It would facilitate the handling of courses and is easy to use. The cons would lie in the implementation of the tool, as described above.

6. Do you think that your students would benefit from the course (please have also in mind that the course would be personalized; i.e., the course would be adapted to the learner’s personal needs)?

Teacher A is convinced that the features provided on the IWT would support the students in their learning process. *“But to provide a platform, where the students can learn in a self-directed way, the system should be improved. Especially to give them a good overview, the system should be structured clearer. In addition the students would need a good briefing before they use the system. A technical support should also be available for them in case of*

problems.” Teacher B thinks that the students would benefit from the course. “It has many advantages to have a personalized course, e.g. through examination of the state of knowledge the didactical units can be adjusted. This saves time, prevents frustration but augments motivation.”

3.1.4 Conclusion

Investigating the usability of the tool showed that the usability refers to the bottom 30% to 40%, meaning that his tool has a higher perceived usability than 30% to 40% of all products tested. It can be interpreted as a grade of a D+ (lecturer B) and a C (lecturer A). Considering that the tool is still in development, the usability is very promising.

The teachers liked the idea of creating a course together, the good overview and the many additional functions. Especially the design was described as appealing. Furthermore the construction of the course allowed an interdisciplinary exchange and collaborative work. The steps of the manual were well described. Both lecturers appreciated the tool and thought that the students would also like it.

Meanwhile, the teachers did not like the complexity of the tool, if no support is available. Therefore they asked for a video tutorial with a good briefing. The search function could be improved and finding features could be easier. Additionally, the manual’s steps are in an incorrect order.

Although several problems occurred, the tool and its user-friendly interface supported the instructors in creating a course. All in all, the lecturers were in favor of the idea and the functions of the tool and they are convinced that the tool can be used in an educational context in order to support students in their learning process. According to their suggestions and comments regarding improvements, it would be helpful for users to facilitate certain things and provide available support.

3.2 R2-2. Knowledge model contextualization: Experimenting the Knowledge model contextualization from the instructor’s viewpoint (UOC)

3.2.1 Research goals and hypotheses

Similarly as in the previous scenario (see Section 3.1.1), the aim of this scenario is also to build an ontological description of a teaching domain that is able to automatically adapt to a context (see [4]). To this end, an experiment was conducted on this scenario at UOC pilot site in order to test the tool from the instructors’ viewpoint. The results of this study give us a first impression of how the tool supports instructors in order to create online courses and whether it needs further enhancements. Therefore, in this study we were primarily interested in the functionality and usability of the tool.

To experiment the knowledge model contextualization from the instructor’s viewpoint, we focused on the following goals and hypotheses as described in [5]:

Goals

G2.1: to ensure that the system is able to generate contextualized courses by selecting a domain ontology, a context, a set of target concepts and a learner.

G2.2: to identify possible ways of improving further the utility of the tool and related models and algorithms.

Hypotheses

H2.1: a set of feasible courses can be effectively and efficiently created starting from a domain ontology by selecting a context, and a set of target concepts.

H2.2: automatically generated courses are considered as a worthy educational resource by the instructors.

In order to investigate these goals and hypotheses, we asked two lecturers from the course “Software Engineering” at UOC to create a personalized course about “Requirements in Software Engineering” using the IWT.

UOC is currently completing the process of adaptation to the European Higher Education Area (EHEA)¹. Over the last two years all the courses and programs at UOC are being designed and deployed in order to be adapted to the new educational system. As a result, profound changes have been made in all courses’ curricula and new programs, degrees have appeared within the EHEA.

Two of the new degrees in the department of Computer Science are the Computer Science Engineering degree and Multimedia degree. Both were designed to provide students with specific skills and competences for each program. As they both belong to the same department, they have much in common, thus sharing some skills and competences. As a result, they share some courses, such as the course “Software Engineering”. This course is attended by students from both degrees since the course provide them with high level skills and competences in developing software system that are required by both type of students.

However, students from de degrees in Computer Science Engineering and Multimedia have slightly different educational and professional profiles in software engineering, such as web development by Multimedia and desktop applications by Computer Science Engineering. Given that the difference is not significant and the cost to separate them is not affordable, the UOC mix them in the same virtual classroom. Hence students from both degrees follow the same curricula with the same activities, material and lecturers.

As both types students have much in common and also certain differences, it becomes a suitable scenario for the lecturers use the IWT in order to create a course that fits to specific students’ needs according their context (i.e., Computer Science Engineering and Multimedia).

¹ http://en.wikipedia.org/wiki/European_Higher_Education_Area

3.2.2 Method

3.2.2.1 Participants

Two experienced and skilled lecturers participated in the experience. Both provide on-line teaching at the UOC in different courses at the Computer Science Degree at UOC in the Software Engineering area. Lecturer A has 6 years of experience in teaching at UOC while lecturer B has 5 years. They also currently teach face-to-face in the same area in the Technical University of Catalonia (UPC) in Barcelona. Finally both are professional developers of software systems, especially e-learning systems and are the owners of a software company settled in Barcelona. Hence they both have a strong background and advanced knowledge developing and using e-learning platforms.

3.2.2.2 Apparatus and Stimuli

First of all we asked two instructors to use the IWT (Intelligent Web Teacher) to create a personalized course. The IWT is able to generate contextualized courses by selecting a domain ontology, a context, a set of target concepts and a learner (see [5]).

Regarding the methodological approach of the study, the lecturers were asked to log all their activities concerning the experiment during the study. In their documentation they noted for each step the time they spent on working with the IWT. In addition, the lecturers listed all problems they had to face while working with the system and wrote down advantages and disadvantages. For this task, the lecturers were provided with technical documentation on this scenario (see [5]).

In addition, both lecturers were asked to fill in the SUS (System Usability Scale [6]) after the end of the session in order to investigate the usability of the IWT. The SUS is a simple, ten-item attitude Likert scale giving a global view of subjective assessments of usability. It is generally used after the respondent had an opportunity to use the system being evaluated. SUS scores have a range of 0 to 100 with an average score of 68, obtained from 500 studies. A score above a 68 would be considered above average and anything below 68 is below average. A score above an 80.3 is considered an A (the top 10% of scores). Scoring at the mean score of 68 gets you a C and anything below a 51 is an F (putting you in the bottom 15%).

In order to investigate in which emotional state the lecturers were when they used the IWT we used the Computer Emotion Scale (CES) [7]. CES scale is used to measure emotions related to learning new computer software. Research showed that the 12 items are describing four emotions:

- Happiness (“When I used the tool, I felt satisfied/excited/curious.”)
- Sadness (“When I used the tool, I felt disheartened/dispirited.”)
- Anxiety (“When I used the tool, I felt anxious/insecure/helpless/nervous.”)
- Anger (“When I used the tool, I felt irritable/frustrated/angry”)

The answer categories and the scores to compute them are “None of the time” (0), “Some of the time” (1), “Most of the time” (2) and “All of the time” (3).

Finally, qualitative statistical analysis, we summarized the open answers in the surveys.

3.2.2.3 Procedure

The experiment consisted of four sessions in a row conducted on the same day:

1. Work session 1: Each lecturer separately proposed a list of concepts that represent/model the selected topic in all contexts. Time spent in this work session was counted.
2. Work session 2: Each teacher separately created an ontology with the concepts proposed and 3 types of possible relations (standard LOM): (a) “*has part*”, (b) “*is required by*” and (c), “*suggested order*”. The relation “*has part*” is needed to indicate that a concept is a sub-concept of another concept. The relation “*is required by*” means that a concept is a requirement of another concept. The relation “*suggested order*” shows that a concept should be explained before another concept. (see [5]). Time spent in this work session was counted.
3. Work session 3: The 2 lecturers met on-line and shared the information (concepts and ontologies) and discussed which were common to both contexts and which were specific of each context. Time spent in this work session was counted.
4. Work session 4: The 2 lecturers created a contextualized course in IWT. Time spent in this work session was counted. Procedure:
 - a. Create a dictionary that incorporates all the key concepts that represent/model the selected topic for all contexts. Time spent was counted.
 - b. Create an ontology with the visual ontology editor (VOE) from the concepts of the dictionary and the 3 types of possible relations. Time spent was counted.
 - c. Set up the contexts and assign to each context the corresponding concepts of the ontology. Time spent was counted.
 - d. Upload contents for each context. Time spent was counted.
 - e. Create a course personalized to each context. Time spent was counted.

The lecturers were instructed to use the manual of the IWT as provided in [5]. Regarding work session 4, they worked every step together on the IWT to support each other in case of doubts. No training sessions on the IWT were programmed given the strong background of the lecturers in developing and using e-learning systems. All the sessions with the IWT were conducted in Catalan language as the targeted students were Catalan speakers.

After the task was finished, the teachers were asked to fill out a questionnaire about their experiences with the system, especially concerning the usability of the IWT.

3.2.3 Evaluation Results

In this section we show the validation methodology that includes the following validation criteria and metric extrapolated by [4].

Following this methodology we will validate 3 aspects of the scenario: time to run the experience (H2.1), the usability of the IWT (H2.1) and the lecturer's emotions when using the IWT (H2.2).

3.2.3.1 *Time to run the experience*

The experiment consisted of four sessions in a row conducted on the same day:

1. Work session 1: Each lecturer separately proposed a list of concepts that represent/model the topic "Requirements" in two contexts: Computer Science Engineering (GEI) and Multimedia (GM).
Time spent in this work session:
Lecturer A: 15 minutes
Lecturer B: 20 minutes
2. Work session 2: Each teacher separately created an ontology with the concepts proposed and 3 types of possible relations.
Time spent in this work session:
Lecturer A: 0 minutes (he preferred to use the tool directly instead of drawing it separately).
Lecturer B: 0 minutes (he preferred to use the tool directly instead of drawing it separately).
3. Work session 3: The 2 lecturers met on-line and shared the information (concepts and ontologies) and discussed which were common to both contexts and which were specific of each context.
Time spent in this work session:
Lecturer A = Lecturer B = 10 minutes
4. Work session 4: The 2 lecturers created a contextualized course in IWT.
Total time spent in this work session:
Lecturer A: 1 hour and 25 minutes (before giving up due to technical problems).
Lecturer B: 3 hour and 20 minutes (before giving up due to technical problems).
 - a. Create a dictionary that incorporates all the key concepts that represent/model the topic "Requirements" for two contexts (GEI and GM) (see Figure 24). Time spent:
Lecturer A: 5 minutes
Lecturer B: 5 minutes

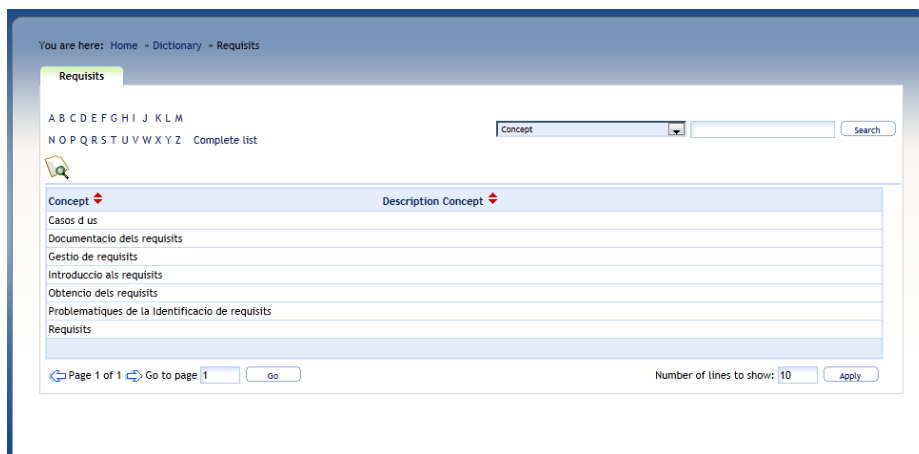


Figure 24: Dictionary “Scientific Working” with concepts

- b. Create an ontology with the visual ontology editor (VOE) from the concepts of the dictionary and the 3 types of possible relations (see Figure 25). Time spent:
 Lecturer A: 15 minutes (10 minutes trying to figure out how to make the application work and 5 minutes of real work)
 Lecturer B: 1h 10min (1 hour trying to make the ontology editor work)

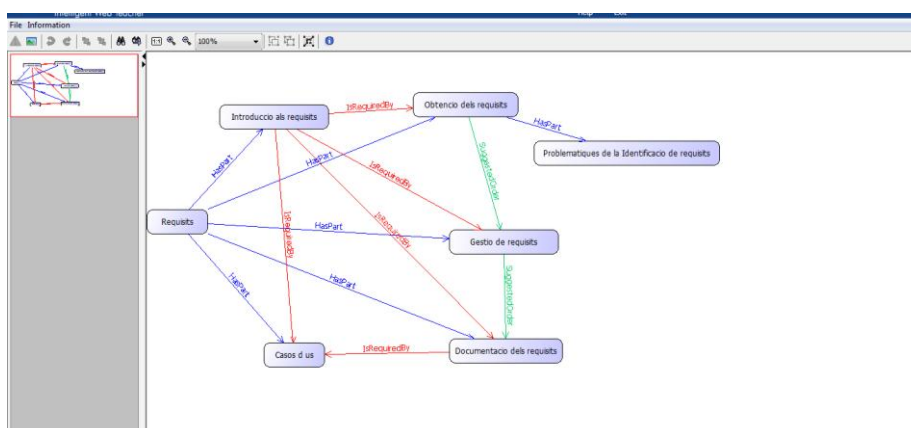


Figure 25: The new created ontology with the concepts and the relations

- c. Set up two contexts “GEI” and “GM”, and assign to each context the corresponding concepts of the ontology. All items were common to both contexts except for one which was concerned with GEI context only (see Table 7 and Figure 26). Time spent:
 Lecturer A: 5 minutes
 Lecturer B: 10 minutes

Concept	GEI	GM
Requirements	✓	✓

Concept	GEI	GM
Introduction to requirements	✓	✓
Requirement elicitation	✓	✓
Issues with identifying requirements		✓
Requirement management	✓	✓
Requirements reporting	✓	✓
Use cases	✓	✓

Table 7: Created concepts for the dictionary

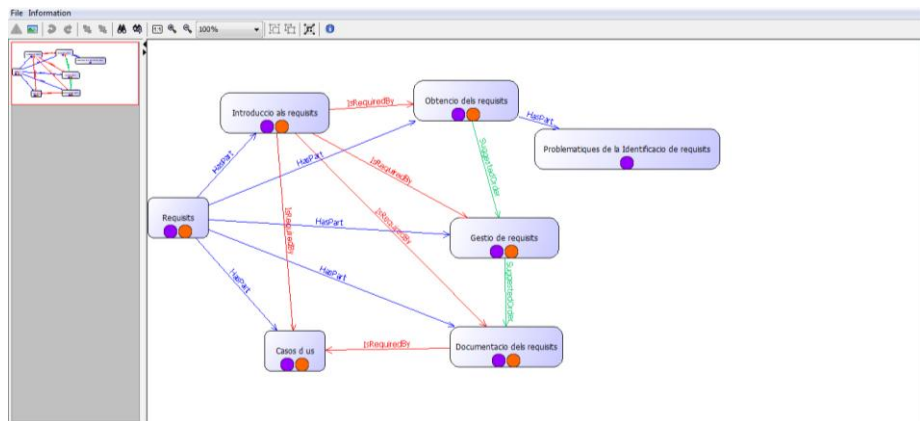


Figure 26: Incorporation of contexts into the ontology

- d. Upload contents for each context (8 materials in all: 5 teaching modules, 1 practical activity and 2 groups of tests) (see Figure 27).
 Time spent:
 Lecturer A: 1 hour (before giving up because a technical problem with creating a multiple choice question, the work was completed by the technical team)
 Lecturer B: 2 hours (before giving up, due to a bug that impeded to create multiple choice questions)

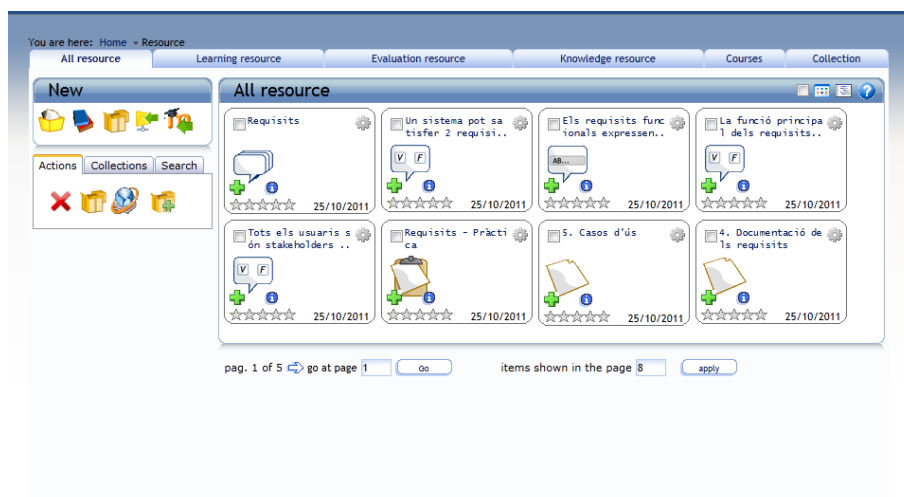


Figure 27: Uploaded contents for the courses

- e. Create a course personalized to each context. Time spent was counted.

Both lecturers could not create the personalized course because of technical problems. Time spent before giving up:

Lecturer A: Give up because could not finalize the previous step.

Lecturer B: Give up because could not finalize the previous step.

Therefore, summing up over the sessions it took the instructors the following time:

Lecturer A: **1 hour and 45 minutes** (before giving up due to technical problems)

Lecturer B: **3 hour and 50 minutes** (before giving up due to technical problems)

Therefore, due to the mentioned technical problems, the tool could not effectively support lecturers in creating courses, thus H2.1 could not be validated. These results are in line with the usability evaluation (see Section 3.2.3.2) where the specific technical problems are reported.

3.2.3.2 Usability of the IWT

In this section, we analyzed the usability of the tool for potential improvements (H2.2). Both lecturers were asked to fill in the SUS report and a questionnaire with open question after the experience.

We calculated the SUS score separately for each lecturer. The score for lecturer A was 11 and the score for lecturer B was 19 belonging both to Grade F (in the bottom 15%). They specially found IWT unnecessarily complex (on average: $M=5$; $SD=0$; $Md=5$), found the system very cumbersome to use (on average: $M=4.5$; $SD=0.7$; $Md=4.5$). As result they would not like to use IWT frequently (on average: $M=1$; $SD=0$; $Md=1$).

As for the positive aspect both lecturers found would not need the support of a technical person to be able to use IWT (on average: $M=2$; $SD=0$; $Md=2$) and needed not to learn a lot of things before to get going with IWT (on average: $M=1.5$; $SD=0.7$; $Md=1.5$).

3.2.3.3 *Emotion of the IWT*

Regarding the lecturers' emotions during the work with the IWT tool, we used the mentioned CES scale. The results from the 4-point rating scale ($n=2$) are as follows:

- Happiness ($M=0.5$, $SD=0.5$, $Md=0.5$)
- Sadness ($M=0.5$, $SD=0.5$, $Md=0.5$)
- Anxiety ($M=2.5$, $SD=0.5$, $Md=2.5$)
- Anger ($M=2.5$, $SD=0.5$, $Md=2.5$)

As shown in the results, lecturers felt high level of anger and anxiety when using the IWT. Happiness emotion was very low. This is in line with previous results on the IWT usability and the procedure to create a contextualized course. In addition, the questionnaires showed in next section reflect specific amounts of frustration and annoyance by both lecturer due to both the technical problems and being the IWT little usable. As a positive aspect, sadness emotion was scored low as they were motivated to use the system for the purpose of the experience and spent significant time before giving up.

3.2.3.4 *Enhancements and improvements of IWT*

Similarly to the previous experience run at TUG, in order to find enhancements for the tool (H2.2), we asked the lecturers to evaluate the experience, especially concerning the usability of the IWT and answer five open questions.

1. Please describe what you liked regarding the system.

Both lecturers liked the idea of using an ontology to structure the course though lecturer A imposed the success of this idea on being implemented in a user friendly fashion

2. Please describe what you did not like regarding the system.

Both lecturers reported many technical problems to describe what they did not like regarding the system. Lecturer A indicated the difficulties to connect to IWT from a different platform than that of MS Windows. He recommended "*it would be better to completely refuse connections from other platforms/browsers than letting me in and then have random errors.*" This was also confirmed by lecturer B who could not use the visual ontology editor from his Mac computer and had to move to a Windows computer.

Lecturer A also did not like that "*I needed to consult the documentation for every little task I tried to accomplish as it slowed me down all the time*". The application does not follow any of the usual UX idioms neither for web applications nor for desktop applications.

Lecturer B thought the system was not very intuitive. For example, he mentioned *“I don’t understand wizards that make me switch to a new tab to continue filling mandatory fields (like when creating a Didactical Unit)”*. Also he did not like the procedure to create test questions by saying *“I didn’t quite understand that I needed to create the questions before creating the test. The usual way would be to create the questions as part of creating the test, even if those questions are reusable for other tests.”*

3. Do you have any suggestions for improvements?

Lecturers’ recommendations were in line with the comments provided in the previous questions. In particular, Lecturer B proposed to make IWT compatible with major browsers and also solve certain bugs that impede the normal functioning of the system. Also both lecturers mentioned that the system is a bit slow and should be more responsiveness.

4. Concerning the user manual you have got, how clear was the description of the IWT for you? Did the user manual support you in following the individual steps?

Lecturer A commented from his vast experience with web applications that *“I never need a manual before using them as long as I knew what I wanted to do. With this assumption in mind, I tried to accomplish the tasks without reading the whole manual first and I failed. Without the manual I would not have been able to complete the tasks so I guess I have to say that yes, the manual supported me following the individual steps.”* Lecturer B almost did not use it. He indicated that *“It was too long for the time I had to spend on the test”*.

5. From your point of view, do you think that teachers would like to use IWT to create and plan online courses? What are the pros and cons?

Lecturers A and B agreed that if technical problems were solved and the application was more intuitive, IWT could be an interesting tool, because it would allow teachers to evaluate students faster. He also found the possibility of creating tests and other automatically evaluable resources very interesting. Eventually, Lecturer A commented that *“I would like to have a tool like that but, right now it is simply too frustrating to use”* whilst Lecturer B indicated that *“Preparing a course is more complicated with this system, but evaluating students could be simpler.”*

6. Do you think that your students would benefit from the course (please have also in mind that the course would be personalized; i.e., the course would be adapted to the learner’s personal needs)?

Both lecturers were very positive with respect to having personalized courses. In particular, Lecturer B indicated that *“I think students would like having such a structured set of resources”*. However, they also indicated that some suggested reading order should be provided if students would prefer a lineal approach to learning.

3.2.4 Conclusion

In contrast to the previous experiment conducted at TUG, this experiment at UOC was conducted by real experts in developing complex computer systems. As professional developers and analysts (and on-line teachers), they are usually very demanding when evaluating a new software, especially if it is from the e-learning domain. Also, having a strong background in web applications as developers and users, they found many technical inconveniences that other people with a different background may miss.

The tool experimented several technical problems that impeded to complete it thus being unable to achieve the main goal (G2.1). It seems the other pilot site did not to have the same technical difficulties and could finalize the experience with success. On view of that, we think that the technical problems faced by our lecturers could be sporadic and exceptional. Next iteration of experiments will confirm or reject these results obtained at UOC.

From the analysis of the usability of the tool it was shown that both lecturers considered usability was not satisfactory and referred this aspect to the bottom 15% meaning that this tool has a lower perceived usability than all products tested. However, since the tool is still under development, the usability can be still far improved by taking the proposals made by the lecturers.

The lecturers' emotions when using the tool were in line with the usability results since lecturers felt angry and anxious most of the time and did not feel satisfaction (happiness), mainly because they could not finish the experiment due to technical problems. Regarding sadness emotion it was proved that the lecturers were enough motivated all of the time in order to make many attempts before giving up.

All in all, the lecturers liked the idea of personalizing a course by an ontology and having structured learning resources to fit the specific students' needs and different contexts. However, they considered the complexity of the tool a barrier for other lecturers and students when using the tool. The user manual was not helpful due to being so long and the fact that web applications are intrinsically and inherently simple and intuitive and usually they do not need to be supported by user manuals.

Finally, the lecturers were very helpful and active and provide many hints and suggestions for improvements at different levels, being the most productive the technical level. This leads to achieve the second goal of this scenario (G2.2).

4 R3. Semantic Connections Between Learning Resources

The aim of this scenario is to provide a set of semantic connections between learning resources and algorithms to automatically activate and deactivate such connections according to teaching and learning preferences as well as to context information.

4.1 Research goals and hypotheses

To experiment with the upper level learning goals, we focused on the following goals and hypotheses as described in [4]:

Goals

G3.1: to build an editor for Compound Learning Resources (CLRs) that allows efficient building of a CLR even in the case of non-expert instructors (i.e. in a friendly way).

G3.2: to playback the generated CLR through a user friendly interface.

G3.3: to ensure that a CLR is able to adapt itself on the basis of the context.

G3.4: to ensure that a CLR is able to adapt itself basing on teaching and learning preferences.

G3.5: to ensure that a CLR allows the effective and efficient learning of scientific concepts in selected domains.

G3.6: to identify possible ways of improving further the utility of the CLR and related tools.

Hypotheses

H3.1: a CLR can be effectively created by instructors as well as stored and played by learners through a user friendly interface.

H3.2: the use of CLRs contribute to support instructors' task.

H3.3: the use of CLRs contribute to improve students' motivation.

H3.4: the use of CLRs contribute to improve students' understanding of key concepts.

H3.5: the use of CLRs contribute to increase students' activity levels.

H3.6: CLRs are considered as a worthy educational resource by both instructors and students.

4.2 Method

4.2.1 Participants

In order to evaluate this scenario and analyze its effects in the learning process, 170 students enrolled in the course Software Engineering from the Computer Science and

Multimedia degrees in the Fall term of 2011 at the UOC participated in the experience. Most of them (154) were from the Computer Science degree and a small group (16) was from the Multimedia degree. Both degrees share the same course “Software Engineering” in its curricula.

The students were equally distributed into 2 classrooms in the UOC virtual campus. Hence, each UOC classroom had 85 students, 77 from the Computer Science degree and 8 from Multimedia degree.

68 out of 170 students (40%) participated actively in the experience. We considered active participation the submission of an evaluation form at the end of the experience. Since the experiment was optional for all students, 60% of them chose not to send the evaluation form and thus they were excluded from the analysis.

41 out of 170 students (25%) also participated in the IWT experience. We considered active participation in IWT the use of the IWT prototypes and the submission of the evaluation form specific to IWT. Hence those 41 students belonged to the group of 68, which left a group of 27 who participated by submitting the form but did not use the IWT prototypes.

From the 68 participants we formed 2 groups for the experiment. One experimental group with 41 students who use IWT (60%) and one control group with 27 students who did not use IWT at all (40%). All of them submitted an evaluation form at the end of the experience.

Therefore, the sample of the experiment was formed by 68 students. For the sake of the experiment, we were only interested in the conglomerate of the experimental group. From this group we formed two sub-groups, 38 from the Computer Science degree (GEI) (95%) and 3 from the Multimedia degree (GM) (5%). 33 students were male (83%) and 7 students were female (17%). The 27 students forming the control group studied at UOC only and did not enter IWT. Hence, whenever referring to IWT we mean the experimental group.

All students of the sample were supervised by one experimented tutor during the experiment.

4.2.2 Apparatus and Stimuli

All students had access to the IWT classroom (where the ALICE prototypes for R3 scenario were installed) from the UOC classroom (see Figure 28 below and Annex A1 for technical details of the integration).

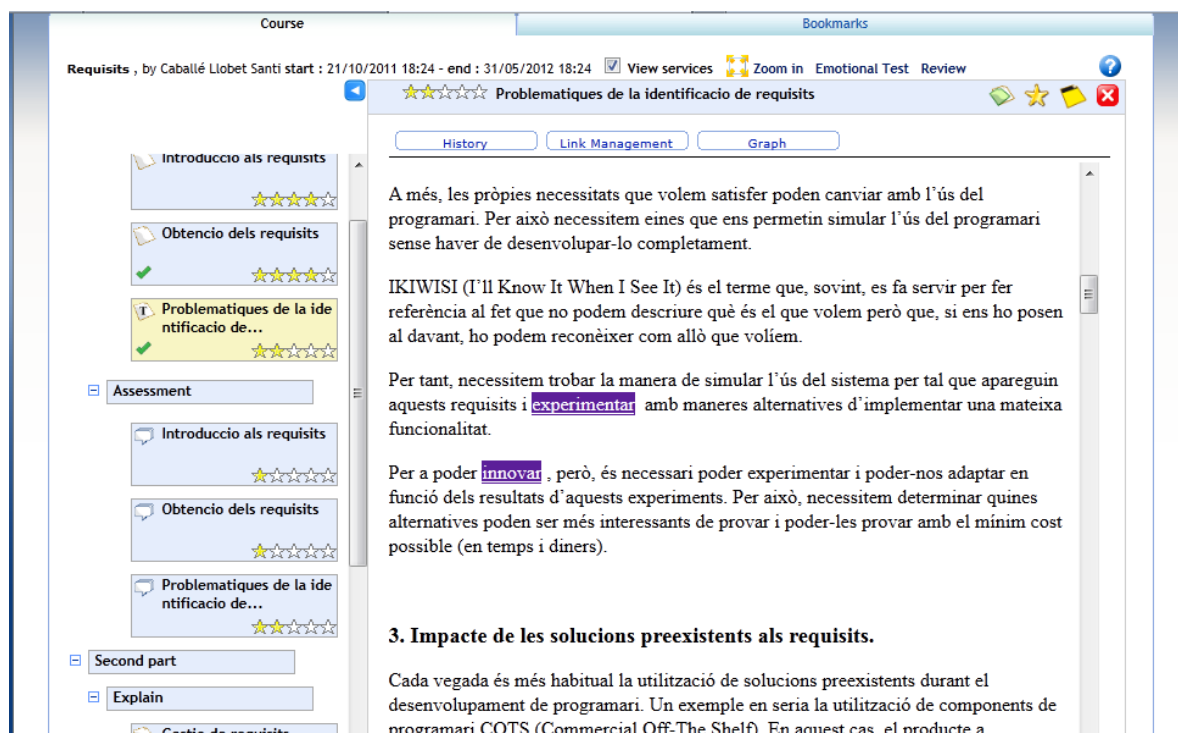
The screenshot shows the UOC classroom interface. On the left, there is a sidebar with 'Salas Estudiantes' and a list of courses. The main area is divided into several sections: 'Comunicació' (Communication) with 'Aula IWT' circled in red; 'Planificació' (Planning) with two calendar views for October and November 2011; and two tables of course activities. The top navigation bar includes icons for 'La meua UOC', 'Comunitat', 'Serveis', 'Aules', 'Tutoria', 'Tutoria IIP', 'Suport docència', 'Secretaria', 'Recerca', 'Biblioteca', 'Notícies', 'Espai de Bolonya', and 'Int'.

Figure 28: UOC classroom with the access to IWT classroom

Once in the IWT classroom, students had access to the R3 scenario (see Figure 29, Figure 30 and [1])

The screenshot shows the IWT classroom interface. At the top, it says 'You are here: Home > Classrooms > Enginyeria del Programari'. Below this, there are three tabs: 'Courses', 'Formative Objectives', and 'Forum'. The 'Courses' tab is active, showing a search bar and a list of courses. The course 'Requisits' is highlighted, with a rating of 4 stars and a session completion of 41%. The interface includes navigation controls like 'Page 1 of 1' and 'Go to page 1', and a 'Number of lines to show: 10' setting.

Figure 29: IWT classroom with a course of Requirements in Software Engineering



Course: Requisits, by Caballé Llobet Santi start : 21/10/2011 18:24 - end : 31/05/2012 18:24 View services Zoom in Emotional Test Review

☆☆☆☆☆ Problematiques de la identificacio de requisits

History Link Management Graph

Introduccio als requisits ☆☆☆☆☆

Obtencio dels requisits ☆☆☆☆☆

Problematiques de la identificacio de... ☆☆☆☆☆

Assessment

Introduccio als requisits ☆☆☆☆☆

Obtencio dels requisits ☆☆☆☆☆

Problematiques de la identificacio de... ☆☆☆☆☆

Second part

Explain

Gestio de requisits

A més, les pròpies necessitats que volem satisfer poden canviar amb l'ús del programari. Per això necessitem eines que ens permetin simular l'ús del programari sense haver de desenvolupar-lo completament.

IKIWISI (I'll Know It When I See It) és el terme que, sovint, es fa servir per fer referència al fet que no podem descriure què és el que volem però que, si ens ho posen al davant, ho podem reconèixer com allò que volíem.

Per tant, necessitem trobar la manera de simular l'ús del sistema per tal que apareguin aquests requisits i **experimental** amb maneres alternatives d'implementar una mateixa funcionalitat.

Per a poder **innovar**, però, és necessari poder experimentar i poder-nos adaptar en funció dels resultats d'aquests experiments. Per això, necessitem determinar quines alternatives poden ser més interessants de provar i poder-les provar amb el mínim cost possible (en temps i diners).

3. Impacte de les solucions preexistents als requisits.

Cada vegada és més habitual la utilització de solucions preexistents durant el desenvolupament de programari. Un exemple en seria la utilització de components de programari COTS (Commercial Off-The Shelf). En aquest cas, el producte a

Figure 30: A CLR with semantic connections to learning resources

We used the SUS (System Usability Scale [6]) in order to investigate the usability of the CLR of IWT. The SUS is a simple, ten-item attitude Likert scale giving a global view of subjective assessments of usability. It is generally used after the respondent had an opportunity to use the system being evaluated. SUS scores have a range of 0 to 100 with an average score of 68, obtained from 500 studies. A Score above a 68 would be considered above average and anything below 68 is below average. A score above an 80.3 is considered an A (the top 10% of scores). Scoring at the mean score of 68 gets you a C and anything below a 51 is an F (putting you in the bottom 15%).

After the assignment, students of the experimental group were required to fill out a questionnaire that included the following 7 sections: (i) identification data (names and program they were enrolled); (ii) evaluation questions about the knowledge acquired with the course "Requisits" (Requirements); (iii) test-based evaluation of the semantic connections of IWT; (v) test-based evaluation on usability of CLR of IWT; (vi) test-based evaluation on the emotional state when using CLR of IWT; and (vii) a test-based evaluation of the questionnaire. Students submitting this questionnaire had the chance to increase their final grade of the course up to 20%. If the questionnaire was not submitted or with wrong responses the final grade would not decrease whatsoever.

For those students of the control group (i.e., they did not enter IWT during the experience), a different questionnaire was sent with only sections (i) and (ii) which had to be filled. Students submitting this questionnaire had the chance to increase their final grade of the course up to 10%. If the questionnaire was not submitted or with wrong responses the final grade will not decrease whatsoever.

For qualitative statistical analysis, we summarized the open answers in the surveys. For the quantitative statistical analysis we employed basic statistics, such as Mean (M), Standard Deviation (SD) and median (Md).

For the section v we used the System Usability Scale (SUS) developed by [6] which contains 10 items and a 5 point Likert scale to state the level of agreement or disagreement. SUS is generally used after the respondent had an opportunity to use the system being evaluated.

Finally, to investigate in which emotional state the students were when they used CLR, section (vi) concerned about the “emotional state” of students when using the CLR, which included 12 items of the Computer Emotion Scale (CES) [7]. CES scale is used to measure emotions related to learning new computer software. Research showed that the 12 items are describing four emotions:

- Happiness (“When I used the tool, I felt satisfied/excited/curious.”)
- Sadness (“When I used the tool, I felt disheartened/despirited.”)
- Anxiety (“When I used the tool, I felt anxious/insecure/helpless/nervous.”)
- Anger (“When I used the tool, I felt irritable/frustrated/angry”)

The answer categories and the scores to compute them are “None of the time” (0), “Some of the time” (1), “Most of the time” (2) and “All of the time” (3).

The data from this experience was collected by means of the web-based forums supporting the discussions in each classroom. Moreover, quantitative and qualitative data were collected from questionnaires containing quantitative and qualitative questions, the answer categories varied between rating scales, multiple choice or open answers. Regarding the rating scales, for the majority of the quantitative questions we used the 5 point Likert scale, so that students could state their level of agreement or disagreement. The rating scale ranged from “I strongly disagree” (1), “I disagree” (2), “neither/nor” (3) to “I agree” (4), “I strongly agree” (5). Finally, quantitative data was also collected from IWT, VCS (see Section 5) and UOC Virtual Campus databases and log files.

4.2.2.1 Procedure

The in-class collaborative formal assignment in both groups lasted three weeks during the second third of the Fall term (October/November 2011) and consisted of studying part of the course “Software Engineering”. The part of the course corresponded with the topic “Requirements” which forms an essential goal of the course.

Students had two options: they either could study the topic “Requirements” only from UOC classroom or, moreover, from the IWT classroom. Hence, all students had to follow the teaching plan at UOC classroom and learn the mandatory material and perform the learning activities planned. In addition, any student who optionally wanted to complement the study of this topic at UOC with the study of the same topic at IWT could do so. The only requirement was to submit the questionnaire at the end of the experience to acknowledge participation in the experiment. Finally, all students could find and study a predefined CLR with semantic connections either by asking a learning resource by expressing their learning needs (see

scenario R1 in Section 2) or by being provided according their context (see R2 scenario in Section 3).

Previous the experience, the topic “Requirements” had been modeled in IWT by using an ontology and concepts. Then it was contextualized into 2 contexts: GEI and GM, and specific contents for each context were then uploaded. Finally a personalized course called “Requirements” was created (see Section 3.1) that may include a CLR. The aim was to provide students with specific learning material in line with the specific needs expressed in the CGS of IWT (see scenario R1 in Section 2) and the context they belonged to (see scenario R2 in Section 3).

After the end of the experience, students received a questionnaire to be filled in order to evaluate the experience with IWT from the viewpoint of the CLR. Whether they belong to the experimental or the control group they received a specific questionnaire. Part of the evaluation consisted in identifying the knowledge acquired on the topic they have studied (in UOC classroom or, also, in IWT classroom).

4.3 Evaluation Results

Following the methodology described in Section 1.3, in this section we focus on the activity, usability and emotional aspects of the IWT tool (H3.1 and H3.4). We also include in this section the evaluation of the questionnaire. On the other hand, the analyses of the tool’s overall impact on student’s learning process are reported in Section 4.4 (Validation Results).

4.3.1 Activity levels in the CLR

In order to give a feedback about how a CLR resource contribute to increase students’ activity levels, we should make a correlation between this kind of resource and some significant parameters (like use and access to the resource, levels of competency acquired) included in IWT database (H3.3-H3.6).

For each parameter we consider the average value in order to make some considerations related to the classroom in the general.

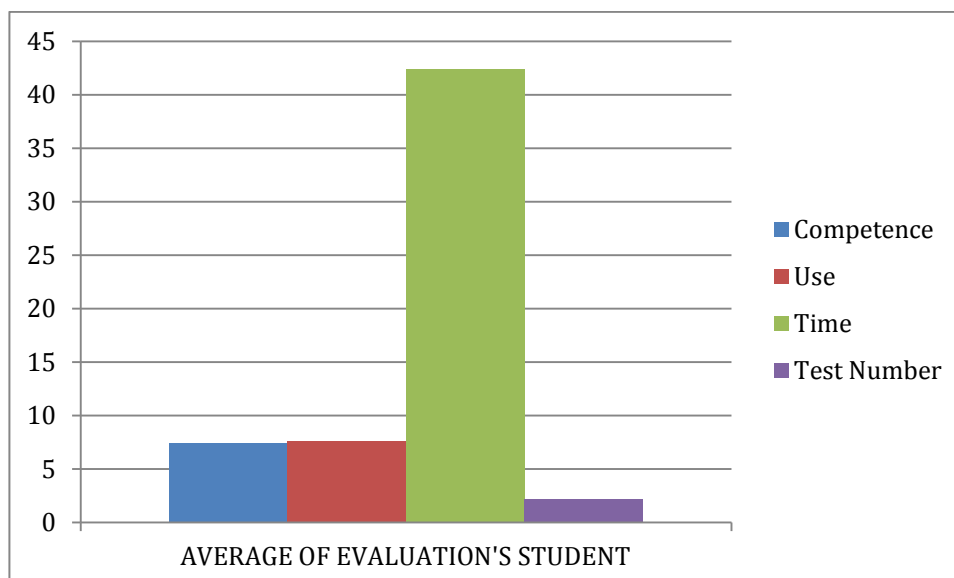


Figure 31: Analysis of IWT Database

As indicated in the Figure 31, we can register a good experimentation results; indeed the levels of competences acquired by exploring a CLR resource is associated to a very short number of test assessment. That denotes that the use of hyperlink within the resource has contributed to improve the students' understanding of key concepts.

Considerable is also the permanence time in IWT with respect to the acquired competences: indeed in the general the competences have been obtained by registering delivery time quite short.

4.3.2 Usability of the IWT

To evaluate student's satisfaction with the tool regarding an efficient and user-friendly management (H3.1), we collected from students' ratings and open comments on the usability/functionality/integration of the CLR with semantic connections.

To investigate the overall usability of the CLR resources, we used the SUS scale (see Section 2.2) and included it in section (v) of the questionnaire. As mentioned, the answers were given on the 5-point Likert scale, so that students could state their level of agreement or disagreement. The rating scale ranged from "I strongly disagree" (1), "I disagree" (2), "neither/nor" (3) to "I agree" (4), "I strongly agree" (5).

After calculating the SUS score for each student, we got an average for 41 SUS scores of 60.78 thus below the SUS mean but nearby, which is a good score considering the first development iteration of CLRs and its integration in IWT. Next, we present the most relevant results of the SUS score by providing several statistics: Mean (M), Standard Desvition (SD) and Median (Md).

4.3.3 Usability of the CLR

To evaluate student’s satisfaction with the tool regarding an efficient and user-friendly management (H3.1), we collected from students’ ratings and open comments on the usability/functionality/integration of the CLR with semantic connections.

To investigate the overall usability of the CLR resources, we used the SUS (see Section 2.2) and included it in section (v) of the questionnaire. As mentioned, the answers were given on the 5-point Likert scale, so that students could state their level of agreement or disagreement. The rating scale ranged from “I strongly disagree” (1), “I disagree” (2), “neither/nor” (3) to “I agree” (4), “I strongly agree” (5).

After calculating the SUS score for each student, we got an average for 41 SUS scores of 40 thus far below the SUS mean (F grade) putting it in the bottom 15%. Next, we present the most relevant results of the SUS score by providing several statistics: Mean (M), Standard Deviation (SD) and Median (Md).

Students found the CLR particularly easy to use (M = 3.44, SD = 1.00, Md = 4) (See Figure 32) and that most people would learn to use CLR very quickly (M = 3.18, SD = 1.20, Md = 3) (See Figure 33).

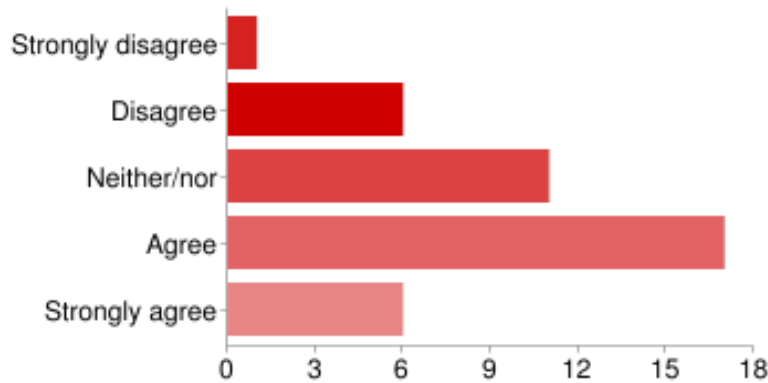


Figure 32: Results on the SUS item “I thought the CLR was easy to use”.

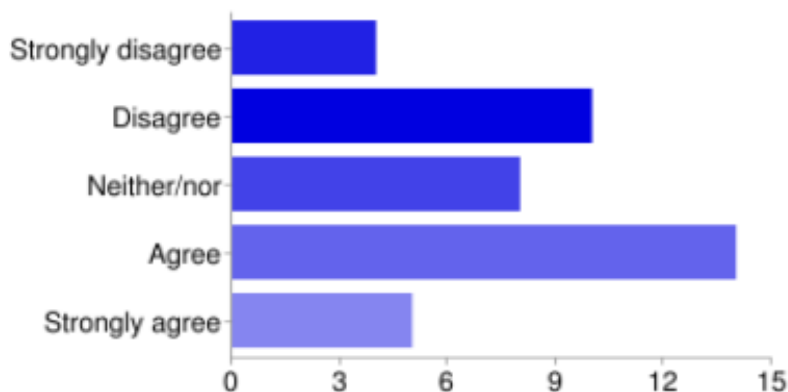


Figure 33: Results on the SUS item “I would imagine that most people would learn to use the CLR system very quickly”.

On the other hand, students thought that most people would not learn to use IWT very quickly ($M = 3.18$, $SD = 1.20$, $Md = 3$) (See Figure 88). In addition, they stated that there was too much inconsistency in the CLR ($M = 3.37$, $SD = 1.10$, $Md = 3$) (See Figure 89) and that the system was cumbersome to use ($M = 3.39$, $SD = 1.22$, $Md = 4$).

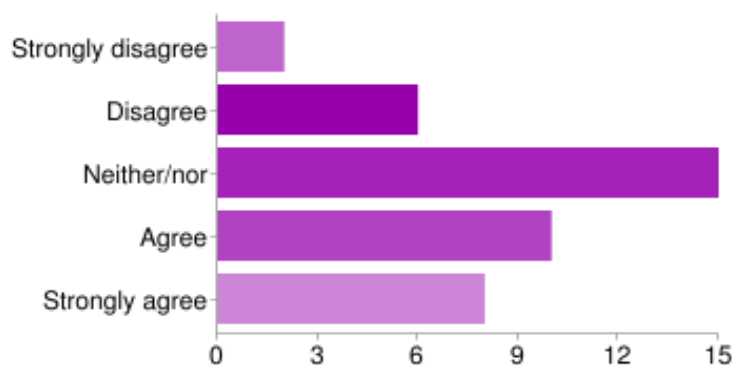


Figure 34: Results on the SUS item “I thought there was too much inconsistency in the CLR”.

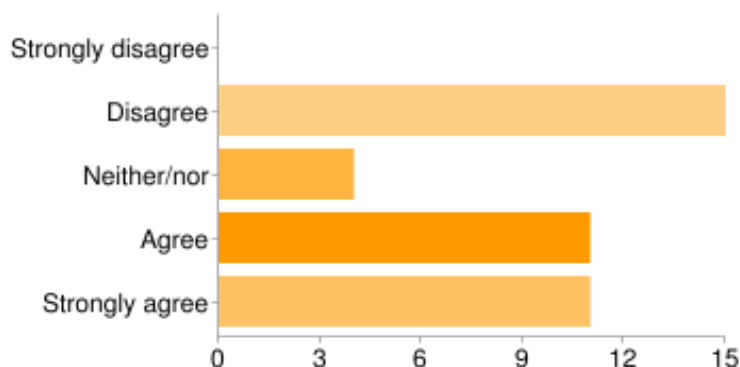


Figure 35: Results on the SUS item “I found the CLR very cumbersome to use”.

Despite students in general liked the CRL resources and the semantic connections a lot (see Section 4.4.1), they all reported a technical issue after visiting the link and coming back to the main thread as the system returned them always at the beginning of the lesson instead of at the point where the learning path was branched. This last point was found very annoying and unpleasant (see emotions in 4.3.3) as it made students momentarily lose the learning path and had to recover it (i.e., find the point in the material where they were and manually go there. This influenced strongly their opinions about the whole idea of the CLR.

In accordance with these results, students indicated in way they would not use the IWT system frequently ($M = 2.50$, $SD = 1.81$, $Md = 1$) in line with the low overall SUS score of 40 and in Figure 36 Students had two fully differentiated opinions. Half of students found the

system and in particular the semantic connections not at all intrusive as they were optional, while the other half considered these internal links very intrusive meaning that visiting the internal links could distract students' attention. No student was indifferent to this question. This binary view is found in the high SD value as a group of students in their questionnaire penalized the CLR because of the usability technical problems while another group of students focused on the purpose of the CLR (See Section 4.4.1).

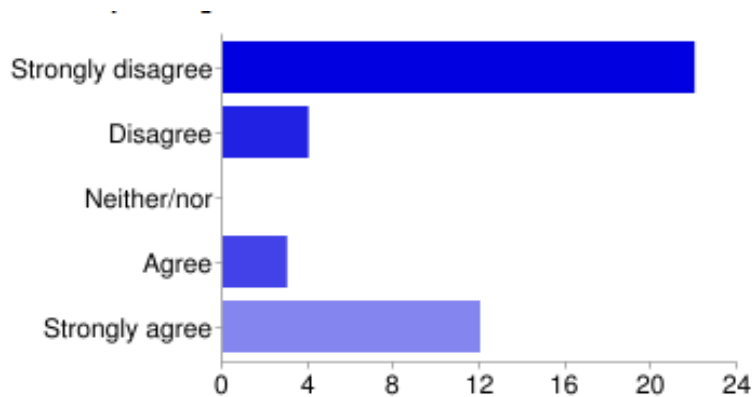


Figure 36: Results on the SUS item “I think that I would like to use this CLR frequently”.

Finally, students stated that the CLR was not very well integrated in the IWT ($M = 2.65$, $SD = 0.96$, $Md = 3$) (see Figure 37). The technical problem reported in the previous paragraph also influenced to this usability dimension since the ill-connections back to the starting point were considered a usability problem. On the other hand, students appreciated the semantic connections a great deal and they considered these links as a result of a good usability.

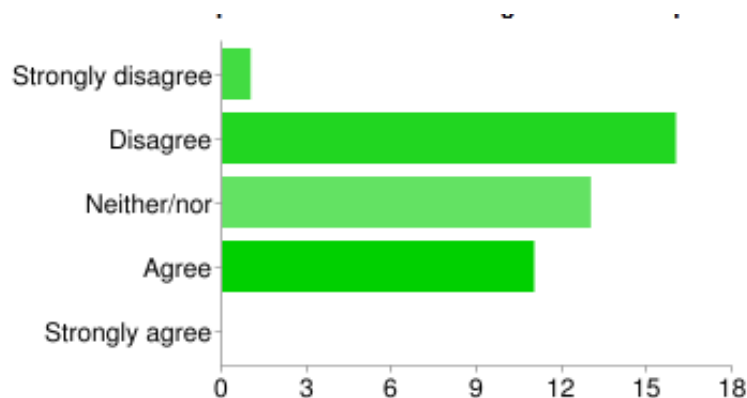


Figure 37: Results on the SUS item “I found the various functions in the IWT were well integrated”.

In overall, this is a good result and promising with challenges to face during the second iteration of the project where this very valuable feedback will be appropriately addressed.

4.3.4 Emotional aspects

Regarding the students' emotions during the work with the IWT tool (H3.1), we used the CES scale to analyze the emotional aspects. The answer categories and the scores to compute them are "None of the time" (0), "Some of the time" (1), "Most of the time" (2) and "All of the time" (3). The results from a 4-point rating scale (n=41) were as follows:

:

- Happiness (M=1.46, SD=0.67, Md=1) (Figure 38)

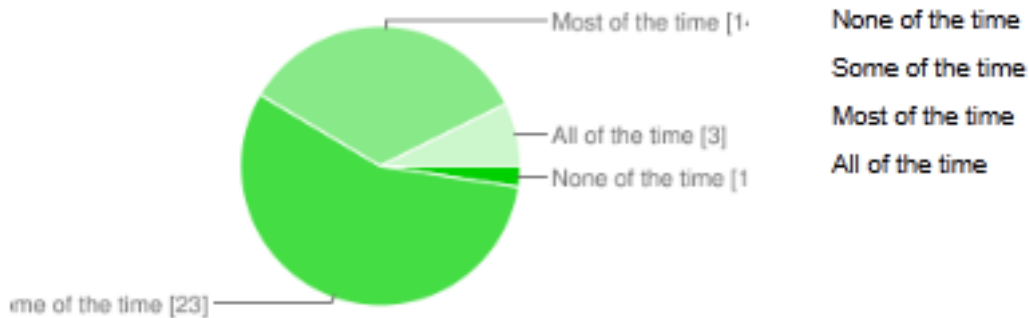


Figure 38: Results on the Happiness emotion

- Sadness (M=0.87, SD=0.67, Md=1) (Figure 39)

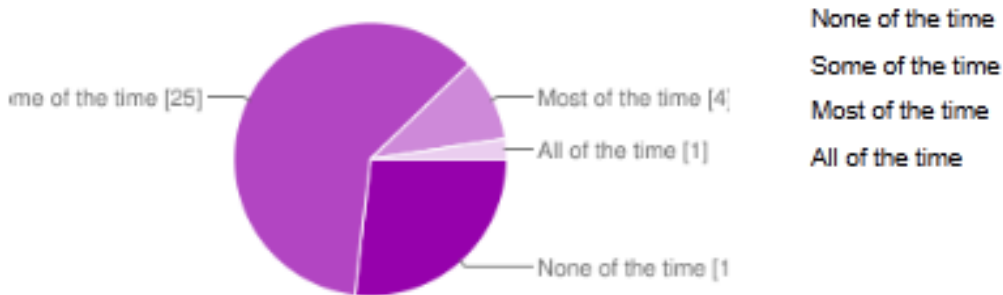


Figure 39: Results on the Sadness emotion

- Anxiety (M=0.90, SD=0.70, Md=1) (Figure 40)

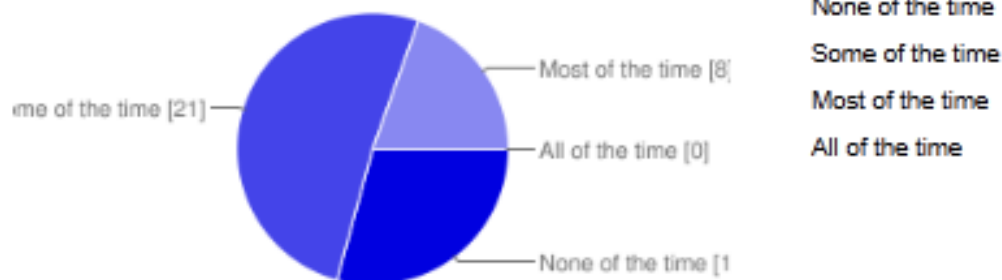


Figure 40: Results on the Anxiety emotion

- Anger (M=1.42, SD=0.63, Md=1) (Figure 41)

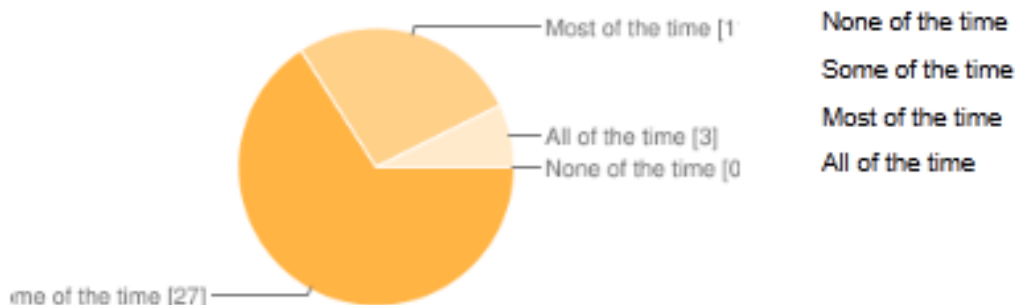


Figure 41: Results on the Anger emotion

Happiness emotion is nearby the mean and students felt more often happiness than sadness, anxiety or anger when learning by means of CLR resources and semantic connections. However, students felt more often anger than sadness or anxiety, being anger especially significant. These results are in line with the results presented above concerning the evaluation of usability of the CLR about the SUS mean (see Section 4.3.2) and with the open comments in the questionnaire (see Section 4.3.4), where high degree of frustration and annoyance emotions were identified due mainly to an ill-navigation issue of the internal links of the CLR reported by students.

In overall, considering the levels founds of anxiety and anger emotions came from the repercussions of a technical issue already identified, which will be soon fixed, this is a good result to face the second iteration of the problem, where this issue will be completely fixed and then it is expected also the happiness emotions to increase.

4.3.5 Questionnaire evaluation

The questionnaire was designed not to be very intrusive in the students' responses by avoiding exceeding the length and/or time employed to fill it. Evaluation results of the

suitability of the questionnaire design confirmed the expectations resulting in most of students (73%) filling and submitting the questionnaire in less than 30 minutes (Figure 42) and 76% of them found it appropriate to evaluate the experience (Figure 43).

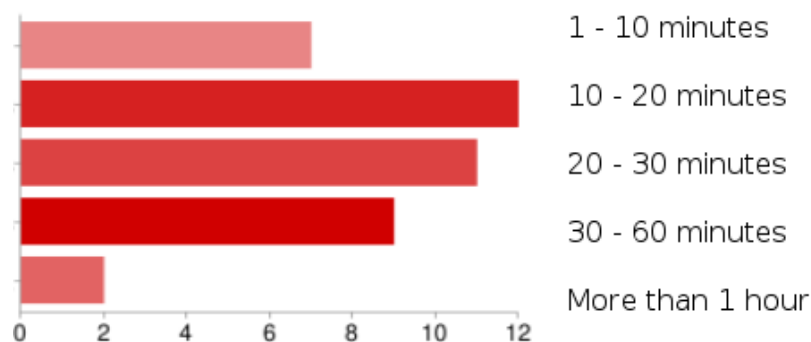


Figure 42: Time employed to fill the questionnaire

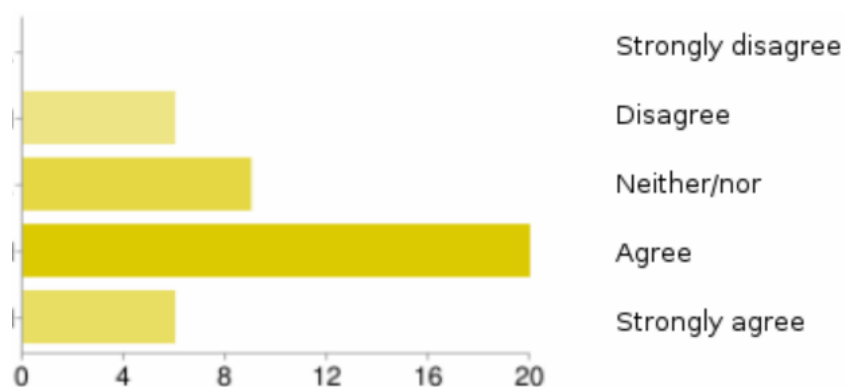


Figure 43: Appropriateness to evaluate the experience with the questionnaire

4.4 Validation Results

Following the methodology described in Section 1.3, in this section we will analyze students' motivation (H3.3), worthiness of the CLR as an educational and teaching supporting resource (H3.6) as well as the acquisition of collaborative knowledge by means of the CLR (H3.4).

4.4.1 The CLR as a valuable resource

In this section we analyze the worthiness of the CLR as an educational resource (H3.6). To this end, quantitative and qualitative data were collected in sections (iii) and (iv) of the questionnaire by 2 open questions (qualitative) and then 4 test-based questions (quantitative) plus one final open question to provide suggestions for improvement. Even though a few of students (3) did not find semantic connections in their CLR (as a result of the

personalization of the system, they responded anyway both qualitative and quantitative questions based on the user manual and their understanding of the whole idea. They missed though the open question for improvements.

In the questionnaire, the rating scales for the two quantitative questions we used a 0-10 point scale, so that students could assess the value of the CLR by a scale they felt very familiar with from their experience in the UOC courses. The scale went from the worst mark (0) to the best mark (10) considering a “good” assessment marks from 5.0 to 10 and a “bad” assessment marks from 0 to 4.9.

As for the test-based questions the rating scale ranged from “Not at all” (1); “Somewhat” (2) and “Completely” (3). Despite sometimes these values changed to fit best the expected type of the responses, in all cases 3 options were provided (positive, medium and negative).

Open questions

Two open questions asked students about the CLR containing semantic connections to learning resources:

1. Evaluate in general the CLR to support the study of the course “Requirements” (Assess the CLR from this view in the scale 0-10).
2. Indicate how in your opinion the CLR has impacted in your individual learning process as for the topic “Requirements”. (assess the IWT from this view in the scale 0-10) (Assess the CLR from this view in the scale 0-10).

After calculating the 0-10 scale for each student we got an average of 6.59 (SD=2.17, Md=7). This result is very good considering the CLR-type resources are still in research evolution and in the first iteration of development.

Students in general liked the CRL resources and the semantic connections a lot. They found these very useful for their study (Question 1: M=7.08, SD=2.87, Md=8.5). Most of them (above 75%) indicated that the internal links between resources allowed them to go deeper and faster into additional information about the topic without having to search for this extra information by themselves. Along with this agreed stance, students had two very differentiated opinions. Half of them commented that the system was not intrusive at all meaning that visiting the internal links was optional and every student could take the decision to check for additional content, while the other half considered these internal links intrusive meaning that the visiting the internal links distracted their attention from the main lesson. This binary view is found in the high SD value as a group of students penalized the CLR because of the usability problem while another group of students focused on the purpose of the CLR. Finally, all of them reported a technical issue after visiting the link and coming back to the main thread as the system addressed them always at the beginning of the lesson instead of the point where the learning path was branched. This last point was found very annoying and unpleasant by many students and influenced strongly their opinions about the whole idea of the CLR. This result is in line with the usability evaluation provided in the 4.3.2 and 4.3.3.

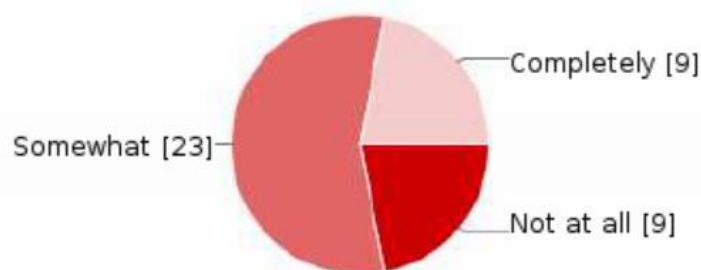
Regarding Question 2, most of students (about 70%) indicated that the CLR and the semantic connections did not have a direct impact in their learning process though the

quantitative results show otherwise ($M=5.86$, $SD=0.86$, $Md=6$). The students insisted on that the internal links just made their study easier and faster. On the other hand, the rest of students mentioned that they had acquired and extended more knowledge and relevant about the study topic by visiting the links. These students stated that the CLR had helped them from the learning perspective.

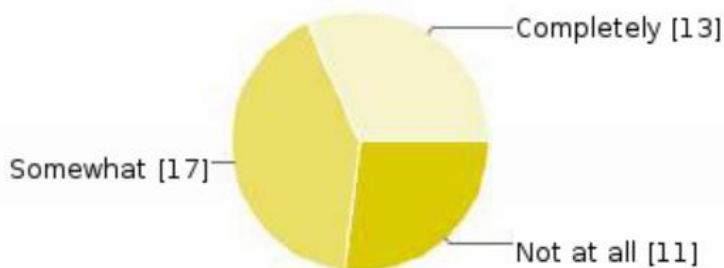
Test-based questions

We evaluated the CLR in the IWT by a test-based questionnaire with 4 questions. The rating scale ranged “Not at all” (1), “Somewhat” (2), and “Completely” (3).

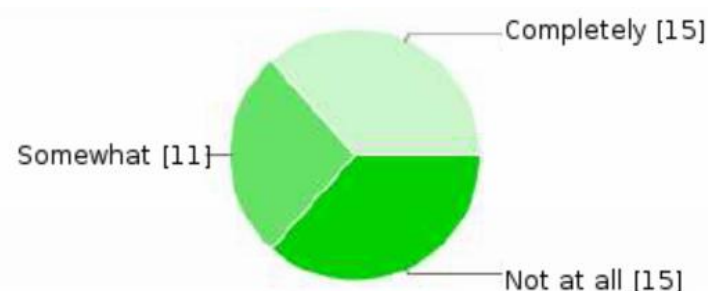
1. The possibility to navigate a learning resource through semantic connections has involved you in a more consistent way to browse the contents?



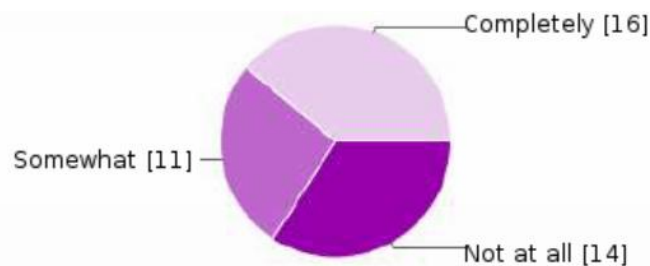
2. Do you think that this solution allows students to read the resource following their own interests or types of reading?



3. The ability to “point” to the external links (as Wikipedia or other important sources) has helped you to maximize your concept of exploration?



4. Do you think that this solution would allow you to explore without always following scattered read paths?



Final open question for improvements

This open question completed this section of the questionnaire by asking students for giving final hints for potential improvement of the CLR resources. In line with their previous comments students claimed to fix the usability problem and enable the internal links return to the main focus after visiting the link and also the new information to appear in anew window.. Some students asked for more semantic connections while others advised not abusing of this resource to avoid messing up the study.

4.4.2 Motivational aspects

Students' motivation concerning the use of IWT tool (H3.3) was directly investigated naively by including in the Section (iii) of the questionnaire a motivation test, where all students were asked for the amount of motivation they felt when studying by using CLR. The following answer categories were used: "absolutely unmotivated" (1), "unmotivated" (2), "motivated" (3), "very motivated (4)".

Test results provided a score far above the mean ($M=3.01$, $SD=0.78$, $Md=3.5$). This result is very good and in line with the previous results on the CLR being a valuable resource and also with the usability and emotional results reported in the previous sections. In particular, students indicated to feel very motivated by the semantic connections that allow them to facilitate their study. They found this a particular valuable innovation of the system.

Finally, clear signs motivation came from enthusiastic students who commented that the semantic links were "really useful", and "all material of UOC should include this type of links" However, most of them proposed improvements on usability.

4.4.3 Tutor assessment and knowledge acquisition

All students from both the experimental and the control groups were evaluated on the responses obtained from the questionnaire. To this end section (ii) of all questionnaires included an evaluative assignment with 1 question about the topic "Requirements" they have studied in either IWT or UOC. This question was purposely designed to provide content on the topic in the form of a CLR resource within IWT. Hence, in combination with the expressing the learning needs (R1 scenario, see Section 2), students eventually obtained this CLR to answer the question. The question was:

- Indicate what the problems are to identify requirements during their elicitation.

This part of each questionnaire was assessed by a lecturer who used the standard 10-point scale to score the students' responses on this question at both IWT and UOC. *Table 8* shows the results.

Experimental group (n=41)	Control group (n=27)
M=7.83	M=6.33
SD=0.78	SD=1.28
Md=8	Md=6

Table 8: Results of the learning assignment evaluation

From the results of *Table 8*, students from the experimental group (material UOC + IWT/CLR) scored higher than the control group (material UOC). More interestingly, the SD in the experimental group is considerably lower than in the control group. This result is in line with the results of R1 scenario (see Section 2.4.3) where students could find a specific resource in IWT devoted to answer this question plus additional information by the semantic connections also related to the question topic, while UOC students had the information related to this question more dispersed in their material and/or had to manually searched for them in case of external information.

Finally, both groups got good marks on average and showed a good level of knowledge acquisition. These very good results are in line with the results from the impact of the CLR (see Section 4.4.1).

In summary, we conclude that IWT/CLR provided students with more specific knowledge and according to the needs and context.

4.5 Conclusion

In this section, the results are summarized and discussed by considering the goals which were determined at the beginning of the study (see also Section 4.1). Then, based on the results summarized further research and technological directions are proposed.

In general the students liked the CLR tool with semantic connections and found it interesting to extend and go deep in certain concepts of Requirements in Software Engineering by means of the semantic connections, and students got better marks when assessed of these concepts (G3.5). The CLR were reported to be reproduced efficiently by students who could use them to find further information about these concepts (G3.1). From the usability point of view, the goals were also achieved by providing CLR with a friendly user interface (G3.2) though a particular technical issue strongly influenced the whole experimentation and prevented users from considering the overall usability of the system satisfactory. Next iteration of the project will fix this particular problem and the usability results are expected to be more objective.

One of the most relevant results was that more than 75% of students indicated that the internal links between resources allowed them to go deeper and faster into additional information about the topic without having to search for this extra information by themselves. This result implicitly achieves G3.3 by providing students of either GEI or GM contexts with the appropriate links and target information suitable to each context. In addition, the levels of competences acquired by exploring a CLR resource denoted that the use of hyperlink within the resource contributed to improve the students' understanding of key concepts. This result also implicitly achieves G3.4.

Finally, possible ways of improving further the utility of the CLR and semantic connections (G3.6) were provided in several sections, and mainly at the end of Section 4.4.1 being most of the comments addressed to a specific problem with usability.

The latter conclusion is in line with the current stage of the IWT technological development, which is expected to be further improved during the second stage of the project and especially from the valuable feedback collected from this experiment.

5 R4. Live and Virtualized Collaboration

The goal of this scenario is to virtualize live sessions of collaborative learning to produce storyboard learning objects embedded in an attractive learning resource (VCS) to be experienced and played by learners. During the resource execution, learners observe how avatars discuss and collaborate, how discussion threads grow, and how knowledge is constructed, refined and consolidated.

Despite the VCS at this stage of the project is fully functional and the development goals have been achieved, it is still far from offering the actual potential to be provided at the of the project. The expected and most distinctive features as for providing a reusable Collaborative Complex Learning Object (CC-LO) as a result of virtualizing recorded live collaborative sessions and augment them with author information is still not available.

Current version of the VCS allows for virtualizing live collaborative sessions at the same time they occur and no augmentation no management of the virtualization process is possible. Hence the result of the virtualization process keeps providing a live collaborative session in a different format. This version was naively tested previously (see [8]) to validate the notion and nature of the approach. Now we proceed with experimenting it in a real context of learning and will validate more complex dimensions of the learning process.

Therefore, the goals and hypotheses formulated for this scenario are related to the current stage of the VCS prototype. In particular, the usability and functionality of the VCS tool to play and observe the current text-based discussion in a multimedia attractive format. To this end, an experiment was run to pilot this scenario in support for a formal in-class assignment of collaborative learning based on a discussion. In this experiment, the VCS acted as the distinctive complement to the underlying discussion tool (IWT forum)..

5.1 Research goals and hypotheses

Goals

G4.1: To build a system that is able to build a Virtualized Collaborative Session (VCS) from a threaded discussion (coming from a forum).

G4.2: To employ the VCS in online courses in order to enhance some aspects of the teaching/learning process.

G4.3: To identify possible ways of improving further the utility of the VCS in online courses.

G4.4: To create, store and playback the generated storyboard through a user friendly interface.

G4.5: To build (automatically) a draft storyboard from a collaborative activity effectively

G4.6: To build (automatically) a draft storyboard from a collaborative activity efficiently

Hypotheses

H4.1: The VCS prototype allows non-expert users to build and use a Story Learning Object (i.e., in a friendly way and efficiently).

H4.2: Use of VCS contributes to significantly improve students' motivation.

H4.3: Use of VCS contributes to support lecturers' task.

H4.4: Use of VCS contributes to significantly increase students' activity levels, both in individual and collaborative activities.

H4.5: Use of VCS contributes to significantly improve students' understanding of key concepts and students' results.

H4.6: VCS are considered as a worthy educational resource by both lecturers and students.

5.2 Method

5.2.1 Participants

The real context of this experience is the virtual learning environment of the Open University of Catalonia (UOC). Given the added value of asynchronous discussion groups, the UOC have incorporated on-line discussions as one of the pillars of its pedagogical model. To this end, great efforts are being made to develop adequate on-line tools to support the essential aspects of the discussion process, which include students' monitoring and evaluation as well as engagement in the collaboration.

In order to evaluate the prototype of the VCS and analyze its effects in the discussion process, the sample of the experiment consisted of 81 graduated students enrolled in the course Organization Management and Computer Science Projects from the Computer Science degree at the UOC were involved in this experience. Students were equally distributed into two classrooms and participated in the experience at the same time.

Despite all 81 students started and participated in the experience, only 69 out of them (85.1%) submitted the final questionnaire, the rest of students (12) dropped out the discussion and the course for several personal reasons. It is worth mentioning here that the 14.9% dropout ratio found is considered rather low in the first third of the academic term when the experience was run². This was caused by the expectations created by the innovative tool that increased the students' motivation as described in section 5.1.4. Eventually this higher number of participants allowed for obtaining more empirical data from the experience.

The students were supervised by two tutors. Each of the tutors was assigned to each group as the official lecturer teaching the whole course.

² Because of the particular profile of the UOC students (students are about 30 years old on average and 95% with a job) the dropout ratio at UOC at the end of the course is 50% on average being about 20% in the first third.

5.2.2 Apparatus and Stimuli

Students from each classroom were required to use standard text-based discussion forums to support the same formal collaborative assignment with the same rules during the same time. In addition, in one of the classrooms (experimental group) the standard forum IWT was equipped with the multimedia-based VCS tool (see Figure 44). In the other classroom (control group) also used a standard discussion forum though the VCS was not available.



Figure 44: Screenshot of a moment of the formal discussion virtualized as a storyboard by the VCS tool from the IWT text forum (note that facial images have been faded and surnames have been removed for private reasons)

After the assignment, the students were required to fill out a questionnaire, which included the following 7 sections: (i) identification data (names and username); (ii) open questions about the knowledge acquired during the discussion; (iii) test-based evaluation of the supporting forum tool (either with or without the VCS), which included a motivation test; (iv) test-based evaluation of the VCS (only in the classroom where the VCS was available); (v) test-based evaluation on the usability of the system (either the VCS or the standard forum without the VCS); (vi) test-based evaluation on the emotional state (no the IWT forum with the VCS and the standard forum without the VCS); (vii) a test-based evaluation of the questionnaire.

Therefore, questionnaire for the classroom with the VCS equipped had all mentioned sections while the other classroom (without the VCS) had all but section (iv). All sections had a final space to express suggestions and further comments about aspects not represented in the questions.

For qualitative statistical analysis, we summarized the open answers in the surveys. For the quantitative statistical analysis we employed basic statistics, such as Mean (M), Standard Deviation (SD) and median (Md). Then we compare these statistics between the control and the experimental group.

Section (iii) included a sub-section with a motivation test which dealt with the amount of motivation the students felt when they were working with the VCS. In this sub-section we used the following answer categories: “absolutely unmotivated” (1), “unmotivated” (2), “motivated” (3), “very motivated”.

For the section v (usability of the forum tools with VCS and without it) we used the System Usability Scale (SUS) developed by [6] which contains 10 items and a 5 point Likert scale to state the level of agreement or disagreement. SUS is generally used after the respondent had an opportunity to use the system being evaluated.

Finally, to investigate in which emotional state the students were when they used the forum tool both equipped with the VCS and without, section (vi) concerned about the “emotional state” of students when using the new system, which included 12 items of the Computer Emotion Scale (CES) [7]. CES scale is used to measure emotions related to learning new computer software. Research showed that the 12 items are describing four emotions:

- Happiness (“When I used the tool, I felt satisfied/excited/curious.”)
- Sadness (“When I used the tool, I felt disheartened/dispirited.”)
- Anxiety (“When I used the tool, I felt anxious/insecure/helpless/nervous.”)
- Anger (“When I used the tool, I felt irritable/frustrated/angry”)

The answer categories in this section are “None of the time”, “Some of the time”, “Most of the time” or “All of the time”.

The data from this experience was collected by means of the web-based forums supporting the discussions in each classroom. Moreover, quantitative and qualitative data were collected from questionnaires containing quantitative and qualitative questions, the answer categories varied between rating scales, multiple choice or open answers. Regarding the rating scales, for the majority of the quantitative questions we used the 5 point Likert scale, so that students could state their level of agreement or disagreement. The rating scale ranged from “I strongly disagree” (1), “I disagree” (2), “neither/nor” (3) to “I agree” (4), “I strongly agree” (5). Finally, quantitative data was also collected from IWT, VCS and UOC Virtual Campus databases and log files.

5.2.3 Procedure

The in-class collaborative formal assignment in both groups lasted three weeks during the first third of the Fall term (October 2011) and consisted of discussing the same issue: “Factors that lead a Computer Science project to failure”. In this assignment, each student was required to post one contribution at least on the issue in hand. Hence, participation in the discussion was mandatory to pass the course.

During the discussion, any student could contribute as many times as needed in the discussion forum by posting new contribution, replying to others as well as start extra discussion threads to provide new argumentations with regards to the issue addressed.

In addition, in one classroom, participants could follow the discussion also by the VCS. The aim was to evaluate the effects of the VCS system in the participation by comparing the activity levels of the discussion between the two groups.

After the assignment, two different questionnaires were sent to students, each to each classroom. Students of the classroom equipped with the VCS tool were asked about questions more focused on this tool. Students from the other classroom were asked about the standard discussion tool used. All students were asked about the results of the discussion in order to identify the knowledge acquired on the topic at hand as well as their emotional state and usability issues when using the tools.

5.3 Evaluation Results

Following the methodology described in Section 1.3, in this section we focus on the activity, usability and emotional aspects of the VCS tool (H4.1 and H4.4). We include an evaluation of the questionnaire. On the other hand, the analyses of the tool’s overall impact on student’s learning process are reported in Section 5.1.4 (Validation Results).

5.3.1.1 Activity level fostered by the VCS

In order to evaluate the students’ activity levels with the VCS (H4.4), we collected and analyzed data by comparing the participation behaviour of the experimental group and the control group as shown in *Table 9*:

Metric / Statistic	Experimental group Standard forum (VCS)	Control group Standard forum
Number of students	41	35
Total of posts	156	119
Mean posts/student	M=3.7	M=3.4
SD posts/student	SD=2.0	SD=1.9
Total words	26669	26591
Mean words/student	M=634.9	M=759.7
SD Mean words/student	SD=406.8	SD=563.1

Metric / Statistic	Experimental group Standard forum (VCS)	Control group Standard forum
Total words	26669	26591
Mean words/post	M=170.9	M=223.3
SD Mean words/post	SD=116.1	SD=111.9
Total visits	1927 (363)	2149
Mean visits/student	M=47 (8.8)	M=53.7
SD visits/student	SD=8.3 (2.4)	SD=6.7

Table 9: Results on activity levels of the discussion in both control and experimental groups. The number of students is higher in both groups than the number of questionnaires received due to some students dropped out during the discussion.

For the posts, words and visits metrics, we computed the mean and its standard deviation. Since no extreme outliers were found, the mean in combination with the standard deviation produced a precise measure. Also for the visits to the forum posts we used the same statistics. Finally, for the “visits” to the VCS (i.e. number of SLO scenes played) we collected information from the VCS log files. In order to compare the post visits (i.e., read) to the scene visits (i.e., seen) we computed the number of SLO created and played (33) multiplied by the average of first scenes seen of each SLO played (11).

Analyzing the results of *Table 9*, they indicate that by using the VCS the participation quantitative behaviour was increased since the number of posts and mean posts/students is higher in the experimental group. On the other hand, the number of views (i.e., readings) of text posts was lower in the forum than in the forum equipped with VCS, pointing out that some of the students found in the storyboard an alternative to the reading of text posts, which was also confirmed by the data collected from the VCS activity logs (363 first scenes seen)

Participation qualitative behaviour is measured in terms of the number of words per post and per student. The lower mean statistics of both words per post and per student in the experimental group indicates that the users of the VCS were more effective and dynamic when communicating their ideas and opinions by either sending new posts or reply posts. As a result, the contributions became more structured and specific whereas the control group promoted larger monolithic one-sided points of view.

5.3.2 Usability of the VCS

To evaluate student’s satisfaction with the tool regarding an efficient and user-friendly management (H4.1), we collected from students’ ratings and open comments on the usability/functionality/integration of the tool.

To investigate the overall usability of the VCS tool, we used the SUS (see Section 2.2) included in section (v) of the questionnaire. As mentioned, the answers were given on the 5-point Likert scale, so that students could state their level of agreement or disagreement. The rating scale ranged from “I strongly disagree” (1), “I disagree” (2), “neither/nor” (3) to “I agree” (4), “I strongly agree” (5).

SUS scores have a range of 0 to 100 with an average score of 68, obtained from 500 studies. A Score above a 68 would be considered above average and anything below 68 is below average. A score above an 80.3 is considered an A (the top 10% of scores). Scoring at the mean score of 68 gets you a C and anything below a 51 is an F (putting you in the bottom 15%).

After calculating the SUS score for each student, we got an average for 38 SUS scores of 63.02 thus nearby the SUS mean, which is a very good score considering the VCS tool is new and still far from being fully developed. Next, we present the most relevant results of the SUS score by providing several statistics: Mean (M), Standard Desvition (SD) and Median (Md).

Students found the tool particularly easy to use (M = 3.47, SD = 1.00, Md = 3) (See Figure 45). Students did not find the VCS unnecessarily complex (M = 2.2, SD = 0,97, Md = 2) (See Figure 46). In addition, students stated that they did not need the support of a technical person to be able to use the VCS (M = 1.89, SD = 0.88, Md = 2) and they thought that most people would learn to use this system very quickly (M = 3,58, SD = 1,00, Md = 4) (See Figure 48).

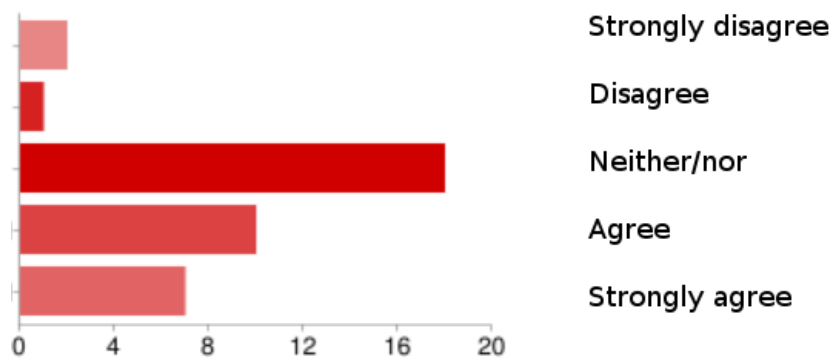


Figure 45: Results on the SUS item "I thought the system was easy to use".

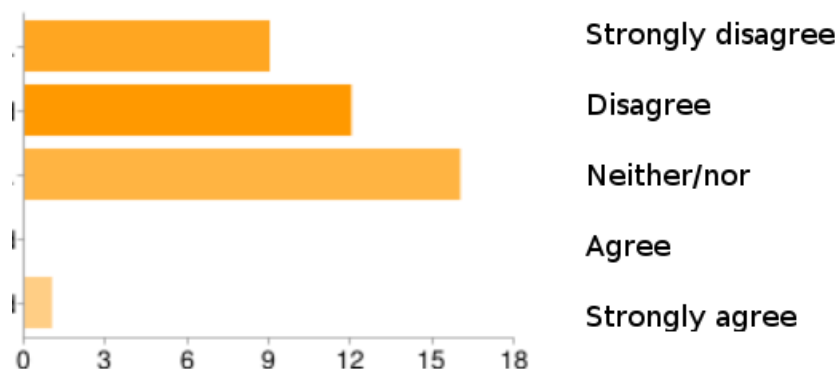


Figure 46: Results on the SUS item "I found the VCS unnecessarily complex".

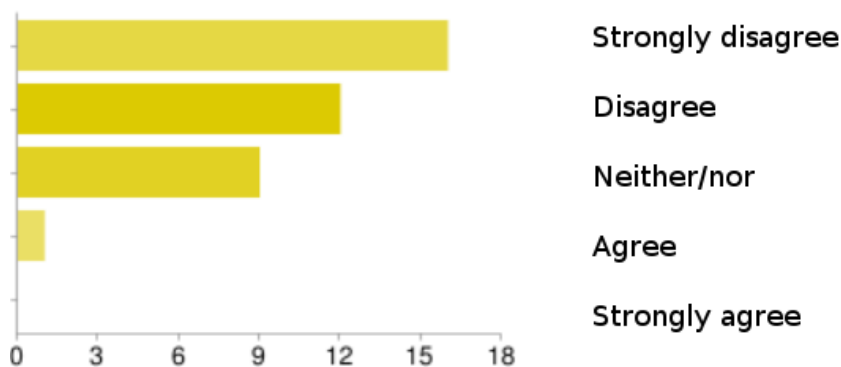


Figure 47: Results on the SUS item “I think that I would need the support of a technical person to be able to use the VCS”.

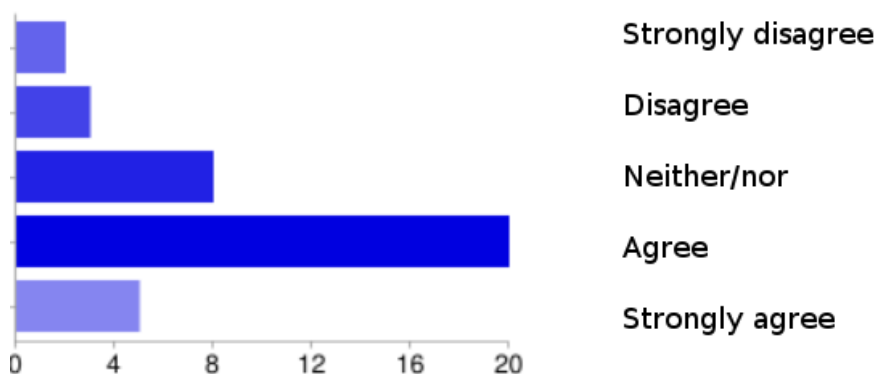


Figure 48: Results on the SUS item “I would imagine that most people would learn to use the VCS system very quickly”.

Some students (about 20%) complained about the VCS being slow to start playing the storyboard as well as the text-to-voice engine sometimes did not reproduce the original contribution perfectly, especially if syntax mistakes were found. As a result, some students preferred to read the forum text messages rather than observe them. On the other hand, students found useful to be able to listen to the discussion while performing other tasks at the same time (e.g., update the agenda, etc.), without being focused only on reading the forum messages. Also they found useful and engaging the possibility to get new ideas and take notes in real time from listening to the discussion in a similar way to a face-to-face discussion.

In accordance with these results, students indicated in a balanced way they would and would not use the VCS system frequently ($M = 2.97$, $SD = 1.16$, $Md = 3$) in line with the overall SUS score of 63.02 and in Figure 49.

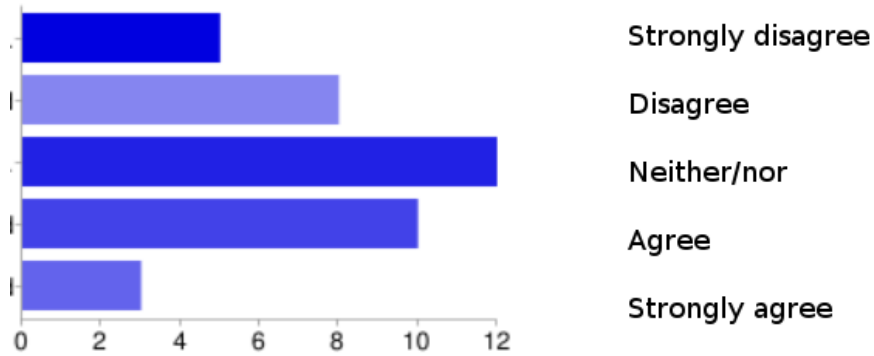


Figure 49: Results on the SUS item “I think that I would like to use this system frequently”.

Finally, students stated that the VCS functionality was well integrated ($M = 3,25$, $SD = 1.01$, $Md = 3$) and the tool itself was adequately integrated in the UOC virtual campus. In particular despite some initial technical problems to gain access, they appreciated to be able to accede to the IWT forum equipped with the VCS directly from the UOC classroom with no reauthentication nor further navigation to the targeted web space..

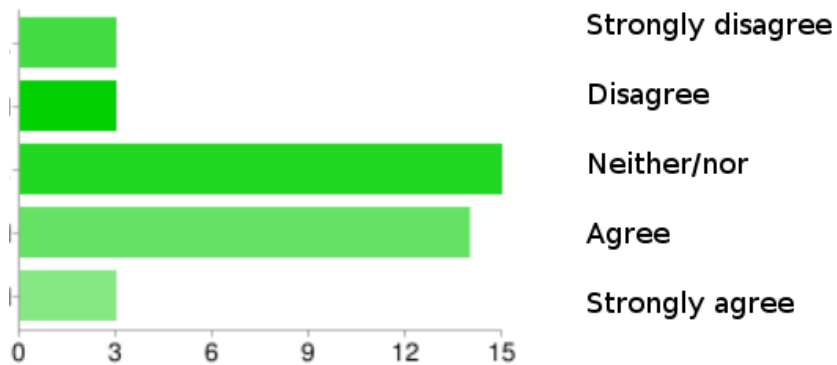


Figure 50: Results on the SUS item “I found the various functions in the VCS were well integrated”.

5.3.2.1 Emotional aspects

Regarding the students’ emotions during the work with the VCS tool (H4.1), the results from a 4-point rating scale (n=38), as follows:

- Happiness ($M=0.95$, $SD=0.89$, $Md=1$) (Figure 51)

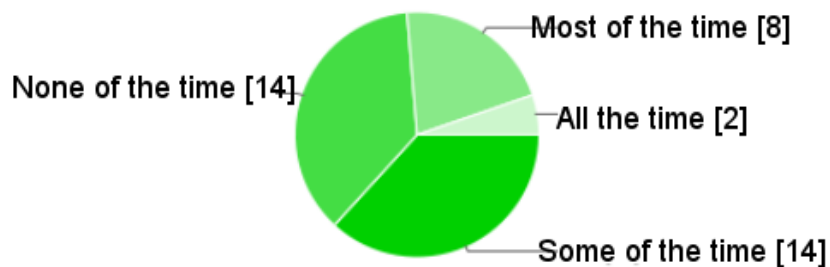


Figure 51: Results on the Happiness emotion

- Sadness (M=0.24, SD=0.49, Md=0) (Figure 52)

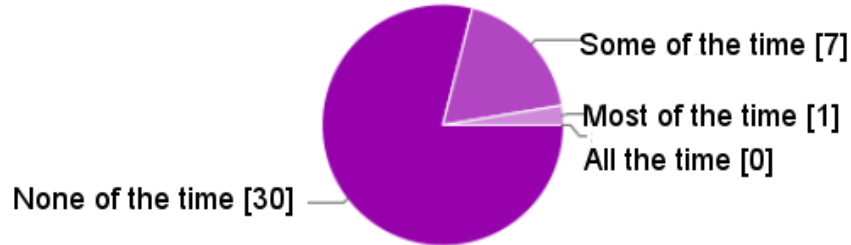


Figure 52: Results on the Sadness emotion

- Anxiety (M=0.21, SD=0.47, Md=0) (Figure 53)

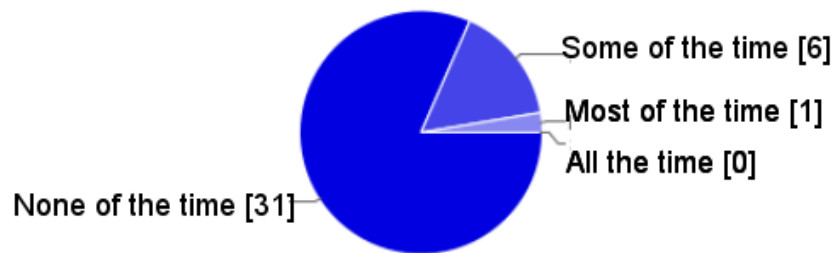


Figure 53: Results on the Anxiety emotion

- Anger (M=0.24, SD=0.49, Md=0) (Figure 54)

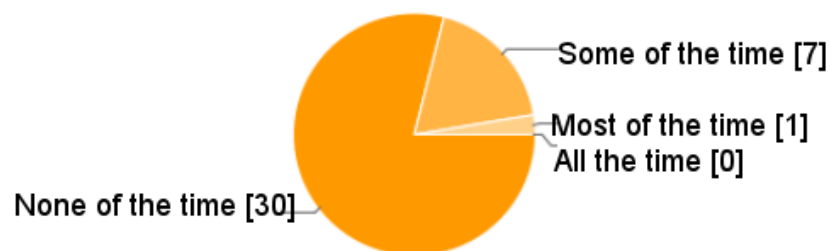


Figure 54: Results on the Anger emotion

Despite the happiness emotion is rather low the students felt more often happiness than sadness, anxiety or anger when learning the new VCS tool. In addition, students felt the same level of sadness, anxiety and anger emotions, which were very low, almost

inappreciable, being anxiety emotion the lowest. These results are in line with the results presented above concerning the evaluation of usability of the VCS tool about the SUS mean (see Section 5.3.2). As already discussed above, no remarkable degree of anger, anxiety and sadness emotions were reported by the students though the level of satisfaction (ie., happiness emotion) was not high due chiefly to some technical problems when uploading the storyboard. Finally, a very few cases of frustration (i.e., anger emotion) were reported by Linux users who could not install the Microsoft Silverlight plug-in to enable the VCS player.

In overall, this is a good result considering the system is far from being fully developed and the user interface needs to take several iterations of improvements before being completed.

5.3.3 Evaluation of the questionnaire

The questionnaire was designed not to be very intrusive in the students' responses by avoiding exceeding the length and/or time employed to fill it. Evaluation results of the suitability of the questionnaire design confirmed the expectations resulting in most of students filling and submitting the questionnaire in less than 30 minutes (Figure 55) and 76% of them found it appropriate to evaluate the experience (Figure 56).

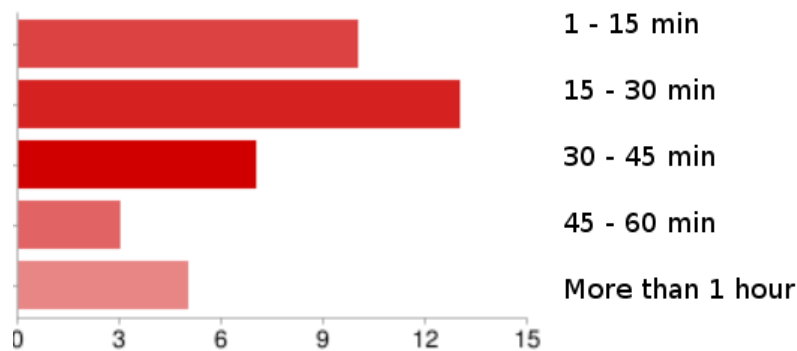


Figure 55: Time employed to fill the questionnaire

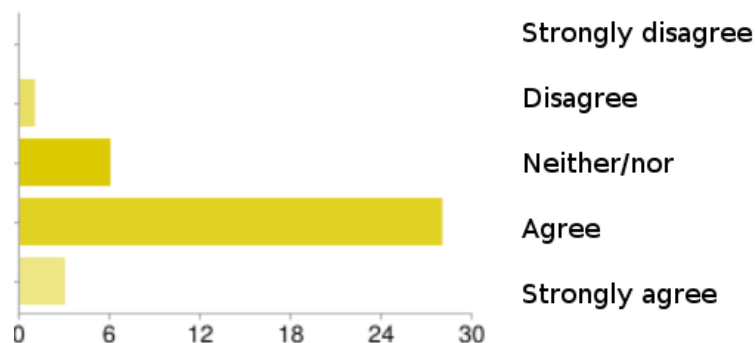


Figure 56: Appropriateness to evaluate the experience with the questionnaire

5.4 Validation Results

In this section we show the validation methodology that includes the following validation criteria and metric extrapolated by [4].

Validation criteria

- C4.1: Level of fulfillment of the VCS features.
- C4.2: Potential increase in students' motivation caused by the use of VCS.
- C4.3: Level of satisfaction of the lecturers with the inclusion of VCS in their courses.
- C4.4: Potential increase in students' activity levels due to the incorporation of the VCS.
- C4.5: Potential increase in students' understanding of concepts and students' results.
- C4.6: Level of satisfaction of students with the inclusion of the VCS in their courses.

Validation metrics

- M4.1: Number of students using the VCS.
- M4.2: Number of visits of the VCS.
- M4.3: Number of visits of the standard forum.
- M4.4: Number of messages submitted by students related to the VCS topics.
- M4.5: Number of messages submitted by students when no VCS is used.
- M4.6: Number of words written by students when the VCS is used.
- M4.7: Number of words written by students when no VCS is used.
- M4.8: Number of students and lecturers that consider that the VCS is worthy.

Following this methodology we will validate the improvement of emotion and motivation (H4.2), worthiness as an educational tool and teaching supporting tool of the VCS (H4.3 and H4.6) as well as the acquisition of collaborative knowledge (H4.5).

5.4.1 *The VCS as a valuable resource*

In this section we evaluate the level of worthiness of the VCS as an educational tool (H4.6). To this end, we collected quantitative and qualitative data in order to know the user's satisfaction with the tool. Both quantitative and qualitative data were collected in section (iv) from 6 open questions of the questionnaire addressed to students. Finally, the lecturer in charge of the classroom also participated by providing his views of the VCS as a supporting tool for teaching (H4.3).

In the questionnaire, the rating scales for the majority of the quantitative questions we used a 0-10 point scale, so that students could assess the value of the VCS tool by a scale they felt very familiar with from their experience in the UOC courses. The scale went from the worst mark (0) to the best mark (10) considering a "good" assessment marks from 5.0 to 10 and a "bad" assessment marks from 0 to 4.9.

The following questions related to evaluate the VCS were asked:

- 1- What did you like and what you did not like from the VCS tool (assess the VCS from this view in the scale 0-10).

- 2- Do you think the VCS tool has fostered your active participation in the discussion in comparison to the text-based IWT forum? (assess the VCS from this view in the scale 0-10)
- 3- Do you think the VCS tool has helped you follow the discussion in comparison to the text-based IWT forum? (assess the VCS from this view in the scale 0-10)
- 4- Do you think the VCS tool has helped you acquire more knowledge about the discussion topics in comparison to the text-based IWT forum? (assess the VCS from this view in the scale 0-10)
- 5- Express your opinion about the storyboard generation by the VCS tool in terms of efficiency and performance (assess the VCS from this view in the scale 0-10)
- 6- Let us know your opinion about the potential of the VCS tool to observe how people discuss and collaborate, and how knowledge is constructed (assess the VCS from this view in the scale 0-10).

About 10% of assessment marks were not provided in the questionnaire due to missing values or because the student could not use the VCS (lack of speakers, technical problems, etc) and followed the discussion by the text messages. We computed a by default value for these questions by the average mark of the rest of responses to the related question where the student's value is missing.

After calculating the 0-10 scale for each student we got an average of 4.98 (SD=1.78, Md=5). This result is good considering the VCS tool is still far from offering the full distinctive features, which influenced in a great deal the responses of those questions related to cognitive benefits that are still not well-supported by the tool.

Students in general liked the VCS tool (Question 1: M=6.07, SD=1.63, Md=5). They indicated to find this resource more attractive and pleasant to follow the discussion than the traditional reading of the text-based messages in a forum. Also students felt the system was more "communicative", meaning they were more engaged in the discussion and they mentioned that the several options to follow the discussion (text and video) motivated them to participate.

On the other hand, while some students appreciated the benefits to navigate among sentences and messages as well as direct access to a certain message (e.g., new message) others found more agile to follow the discussion by the text forum. Students found problematic to understand the VCS voice due to syntax problems of the message source. This will be easily solved in the next development steps by the incorporation of the VCS Editor. Finally, some students indicate the benefits of the VCS tool for disable students.

Analysis from comparing participation with and without the VCS tool scoped Questions 2, 3 and 4. All of them had similar results (M=4.28-4.34, SD=2.63-3.07, Md=5). Students indicated that the VCS did not foster their participation because the VCS allowed them to read the messages but not to write. Also, they mentioned that following the whole discussion only with the VCS could have been more difficult. However, students mentioned that by listening to the messages they could take notes on the contributions in real time, thus enhancing the participation and they reported that they could follow the discussion faster with the VCS, thus leaving time for further participation. In addition students reported to associate

the main discussion concepts faster by combining the text in the balloons with voice rather than just reading the text posts in the standard forum. Finally they could follow the discussion more effectively, especially in large discussions by avoiding the page navigation required by the standard forum and also for review and summary purposes of the most relevant contributions.

Some students reported some performance and efficiency problems during the execution of the VCS tool while other approved the general performance of the system ($M=5.68$, $SD=1.67$, $Md=5$). Also, students reported to have technical problems with the Microsoft Silverlight plug-in while others neglected to install it (ie., Linux users). Students indicated that for short threads it was more efficient to read messages in the text forum than observe them in the VCS.

Finally, students made many advantages of the VCS by exploiting its potential appropriately (Question 6: $M=5.2$, $SD=2$, $Md=5$). In particular, they commented that the VCS could be much more useful if performance and visualization could be improved. Most interestingly, they proposed to “store” or “backup” the storyboard in a repository in order to be able to reuse the most relevant contributions in video or audio format later on by students of next courses. These comments are in line with the actual extension of the VCS for the next development steps in the project that students felt as the next logical step. Also they proposed to link the VCS tool with the IWT forum in order to directly post a message to the forum in response to a contribution read in the VCS. Students indicated the VCS to be particularly useful for large discussions, which can be followed more fluently and comprehensively. Finally, students proposed to foster the use of VCS system at a larger scale, in other courses and programs.

These students’ comments also give many hints for possible improvements of the tool.

Regarding the lecturer in charge of the discussion reported the VCS tool helped him follow and evaluate the discussion more appropriately than the text forum by having direct access to a specific students’ contribution. Even so, he demanded more monitoring tools for the VCS to sort out scenes by student, date and connection between replies, thus following dialogs within a thread. He proposed to turn the VCS session into a learning material so that other students could reuse the knowledge built during the discussion.

5.4.2 Motivational aspects

Students’ motivation concerning the in-class discussion assignment supported by the VCS tool was investigated by comparing the difference in motivation between the experimental and control groups.

Section (iii) of the questionnaire included a motivation test for both the experimental and control groups, where all students were asked for the amount of motivation they felt when collaborating in the discussion by means of the required tools. The following answer categories were used: “absolutely unmotivated” (1), “unmotivated” (2), “motivated” (3), “very motivated (4)”.

Experimental control scored higher ($M=2.85$, $SD=0.69$, $Md=3$) than the control group ($M=2.14$, $SD=0.38$, $Md=2$). The results of the experimental group are in line with the results reported in Section 5.4.1. In particular, students found the VCS more attractive and pleasant to follow the discussion than the traditional reading of the text-based messages in a standard forum. Also students felt the system was more “communicative”, meaning they were more engaged in the discussion and they mentioned that the several options to follow the discussion (text and video) motivated them to participate. Finally, clear indications of amounts of motivation came from enthusiastic students who evaluated the VCS tool as “fascinating”, “impressive”, “very interesting”, “very useful”, “inflection point in e-learning systems”. On the other hand, students who chose not use the VCS tool due to lack of time or technical problems felt unmotivated.

5.4.3 Tutor assessment and knowledge acquisition

All students were evaluated on summarizing the discussion in both the experimental and the control groups. To this end section (ii) of the questionnaire included 3 evaluative questions: 2 first questions to evaluate the discussion topics and the last question to evaluate the knowledge acquisition, as follows:

1. Indicate what are the main factors seen during the discussion, which may lead a software project to fail.
2. Indicate what factors make a project which has been finalized successfully be underused.
3. Comment what you learnt from the discussion than can enrich your personal knowledge.

This part of each questionnaire was assessed by the lecturers of each classroom who used the standard 10-point scale to score the students’ responses. *Table 10* shows the results.

Evaluative questions	Experimental group (n=38)	Control group (n=31)
Question 1	M=6.84 SD=1.48 Md=7	M=6.93 SD=1.15 Md=7
Question 2	M=7.68 SD=1.18 Md=8	M=6.83 SD=1.34 Md=7
Question 3	M=7.21 SD=1.45 Md=7	M=7.12 SD=1.14 Md=7
Overall	M=7.24 SD=1.41 Md=7	M=6.96 SD=1.21 Md=7

Table 10: Results of the discussion evaluation

From the results of *Table 10*, students from the experimental group scored higher than the control group though the difference is not significant. Both groups got good marks on average and showed a good level of knowledge acquisition. These results are in line with the results from the impact of the VCS tool in the students' activity levels, which was higher than in the other classroom (see Section 5.3.1) but also in line with the quantity and quality of the participation reported in Section 5.4.1 where students indicated that the VCS did not foster the quantity and quality of the participation.

In summary, we cannot conclude that the VCS tool had an impact on the knowledge acquisition of the discussion.

5.5 Conclusion

In this Section, the results are summarized and discussed by considering the goals which were determined at the beginning of the study (see also Section 5.1). Then, based on the results summarized further research and technological directions are proposed.

In general the students liked the VCS tool and found it interesting to have another option to follow the in-class discussion-based assignments (G4.3). During this specific assignment, students indicated they could generate the storyboard from the VCS (G4.1) and it was effective to support the discussion for review and summary purposes (G4.5). Despite some initial technical problems the majority of students reported to generate the storyboard efficiently (G4.6) and create, store (transparently) and playback it (usability) as many times as needed (G4.4). Aspects of the learning process, such as motivation and emotional were validated showing an impact of the use of the VCS tool on these aspects (G4.2). In addition, the VCS was proved to become an useful educational resource. Finally, gain in knowledge acquisition by using the VCS could also be validated though not significantly.

Next iteration of the project will provide a full featured version of the VCS prototype. New and essential functionality will be incorporated, such as the VCS Editor that will allow for the building a reusable Learning Object (CC-LO) by eliciting the knowledge acquired in previous live collaborative sessions. From this technology perspective, we plan to provide a learning resource that will have an important impact on the knowledge acquisition and in the learning process.

6 R5. Storytelling

The goal of this scenario is to allow an efficient learning about knowledge and behaviour to be adopted in civil emergency situation (like seismic event in Amusement Park) through the use of complex and innovative learning resource (Storytelling Learning Object). As a result, an Emergency Course has been created for providing suitable learning resources that meet the learners' needs.

6.1 Research goals and hypotheses

To experiment with the Storytelling Learning Object, we focused on the following goals and hypotheses as described in [4].

Evaluation goals

- G5.1: to build digital storytelling methodologies and tools able to let instructors build a Storytelling Learning Object (SLO) on the basis of the defined storytelling design model.
- G5.2: to ensure that the aforementioned methodologies and tools allow efficient building of a SLO even in the case of non-expert instructors (i.e. in a friendly way).
- G5.3: to store and playback the generated SLO through a user friendly interface.
- G5.4: to ensure that a SLO can be played with different roles and can be adapted basing on the role played by the learner and on his/her user model.
- G5.6: to ensure that a SLO allows the efficient transmission of lesson learned inside a learning experience on the theme of the risk managements.
- G5.7: to identify possible ways of improving further the utility of SLOs and related tools in on-line and blended courses.

Evaluation hypotheses

- H5.2: The use of SLOs contributes to improve students' motivation and emotional status.
- H5.3: The use of SLOs contributes to support instructors' task.
- H5.4: The use of SLOs contributes to increase students' activity levels, both in individual and collaborative activities.
- H5.5: The use of SLOs contribute to improve students 'understanding of key concepts as well as related skills.
- H5.6: SLOs are considered as a worthy educational resource by both instructors and students.

6.2 Method

6.2.1 Participants

Two secondary schools have taken part to the experimentation.

In the first school “E. Striano” there were 14 students in the course: gender male and average 16 years old.

In the second school “Pitagora” there were 28 students in the course: 26 were female (98%), 2 students were male (2%) and the participants were on average 14 years old.

The students of the first school have shown a major responsibility and maturity on the topics illustrated in the course with respect to the students of the secondary school.

Each class were supervised by two tutors. Within each class the students have been divided in two groups: experimental and control, in order to make a more comparative analysis of the investigated tools.

6.2.2 Apparatus and Stimuli

We asked to the experimental group to interact with the Storytelling Learning Object related to the risk management in a complex context as the amusement park.

On completion of the session they have filled a Post-Questionnaire, which includes the following sections: demographic data, storytelling learning object activity, usability of the storytelling environment, emotional aspects and further comments or suggestions. Besides, we provided a Questionnaire for the tutors concerning the added value of the complex learning resources (as the storytelling) in a learning course and how it can contribute to ameliorate the knowledge.

For qualitative statistical analysis, we summarized the open answers in the surveys. For quantitative statistical analysis, we performed t-tests.

Regarding the section “Storytelling Learning Object Activity”, the students are asked to assess the work concerning the following questions:

Effectiveness of the methodology

- The combination between the exploration and guide of the storytelling resource, allow you to maintain a good level of motivation?
- Could you measure out the autonomous navigation and exploration of the different paths?
- You have explored different didactic situations characterized by a sequence of 4 educational events. In your opinion, does the sequence allows you to turn attention to the problem, facilitate the discussion, etc.?
- Does the reflection events allowed you to reflect on what you have acquired?
- Are the advancer events useful for resolving the problems?

Validation of the storytelling resource wrt the knowledge objectives

- The explorative logic that characterizes the storytelling allow you to capture different types of knowledge to put in practice in an emergency situation?
- The recovery paths guided you or have been useful in order to recover any gaps?
- The storytelling structure has allowed you to understand the different expected results and their importance?

Originality and innovation in the educational structure

- What do you think about the mix of linear and alternative paths?
- Has the ability to repeat a learning path through different view points got involved you?

Storytelling Interface

- Have the storytelling an user friendly interface?
- Could you browse and operate with the educational content at various levels of detail?
- How have you interacted with the story?
- The visual quality of the experience contents has helped you to have more awareness of the task and tests to overcome?

The answer categories in this section are “In no way”, “Partially”, “Completely” .

For qualitative statistical analysis, we summarized the open answers in the surveys. For the quantitative statistical analysis we employed basic statistics, such as Mean (M), Standard Deviation (SD) and median (Md).

For the section “usability of the storytelling environment” in the Post-Questionnaire and the Questionnaire for the tutors, we used the SUS(System Usability Scale) which contains 10 items and a 5 point Likert scale to state the level of agreement or disagreement, for instance “I think that I would like to use this system frequently”.

To investigate in which emotional state the students were when they used the storytelling tool, we added a section concerning “emotional aspects”, which included 12 items of the Computer Emotion Scale (CES) that measure emotions related to learning new computer software. Research showed that the 12 items are describing four emotions:

- Happiness (“When I used the tool, I felt satisfied/excited/curious.”)
- Sadness (“When I used the tool, I felt disheartened/dispirited.”)
- Anxiety (“When I used the tool, I felt anxious/insecure/helpless/nervous.”)
- Anger (“When I used the tool, I felt irritable/frustrated/angry”)

The answer categories in this section are “None of the time”, “Some of the time”, “Most of the time” or “All of the time”.

All Questionnaires contained quantitative as well as qualitative questions, the answer categories varied between yes/no, rating scales or open answers. Regarding the rating scales, for the majority of the quantitative questions we used the 5 point Likert scale, so that students could state their level of agreement or disagreement. The rating scale ranged from

“I strongly disagree” (1), “I disagree” (2), “neither/nor” (3) to “I agree” (4), “I strongly agree” (5).

Finally, quantitative data was also collected from IWT database and log files and are reported in 6.4 Section.

6.2.3 Procedure

In order to give more emphasis to the experimentation, a learning course, designed to cover a macro concept on emergency management in environments with high levels of risk in case of fires and earthquakes, has been built in the IWT. The course has been delivered by two groups of users having the same learning styles and divided in two groups: experimental and control.

The experimental group has had access to an educational experience created specifically to meet the complex learning topics related to emergency management through the use of Complex Learning Object. The CLOs have been represented a Serious Game, for supporting intuitive learning processes in case of fire in school, and a Storytelling, for promoting the lessons learned through guided explorative processes in the case of a seismic event in a complex structure .

In the following section the individual steps of the experiment are described.

After that each student logged in IWT platform, he see his class and group.

In a first step the two groups were assigned the specific course.

For the experimental group was created a personalized learning path by having as concept objective the acquisition of the behavior to take for managing high risks as the earthquakes in an amusement park through complex learning resources.

The control group has also delivered a personalized learning path with the same concept objective but the kind of learning resources is less interactive and active than the experimental group.

When all groups had finished the delivery of the learning resources, each member of a group had made an assessment test for testing the knowledge acquired by the storytelling for the experimental group and by a passive learning resource for the control group.

6.3 Evaluation Results

In this section we focus on the activity level, usability and emotional aspects of the Storytelling Learning Object delivered by IWT platform (H5.2-H5.6). We also include in this section the evaluation of the questionnaire. On the other hand, the analyses of the tool's overall impact on student's learning process are reported in Section 6.4 (Validation Results).

The evaluation results have been obtained by providing several statistics data, as Mean (M), Standard Deviation (SD) and Median (Md).

The survey is a study of the qualitative questionnaires submitted to all the students of the two schools belonging to the experimental group.

6.3.1 Storytelling Learning Object Activity

Regarding the students activity and interaction with the storytelling learning object, we report some question's category, useful for analyzing this component:

- Storytelling Interface (M= 5.2, SD= 1.1, Md= 5) (Figure 57)

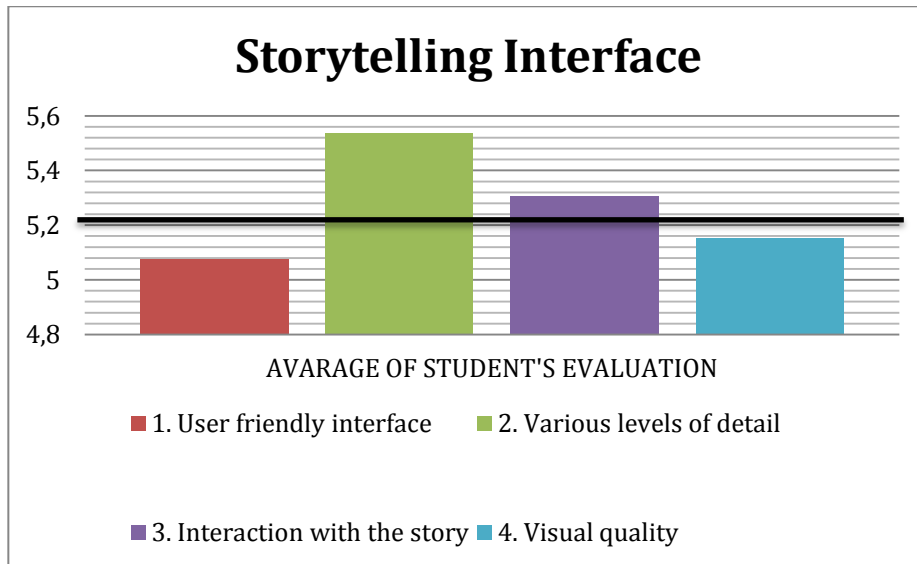


Figure 57: Results on the Storytelling Interface

As indicated by the Figure 57 the students have understood the logic articulation of the storytelling learning object and have analyzed and investigated the different learning path in which the story branches.

- Originality and innovation in education structure (M=5.5, SD=1.2, Md=5.5) (Figure 58)

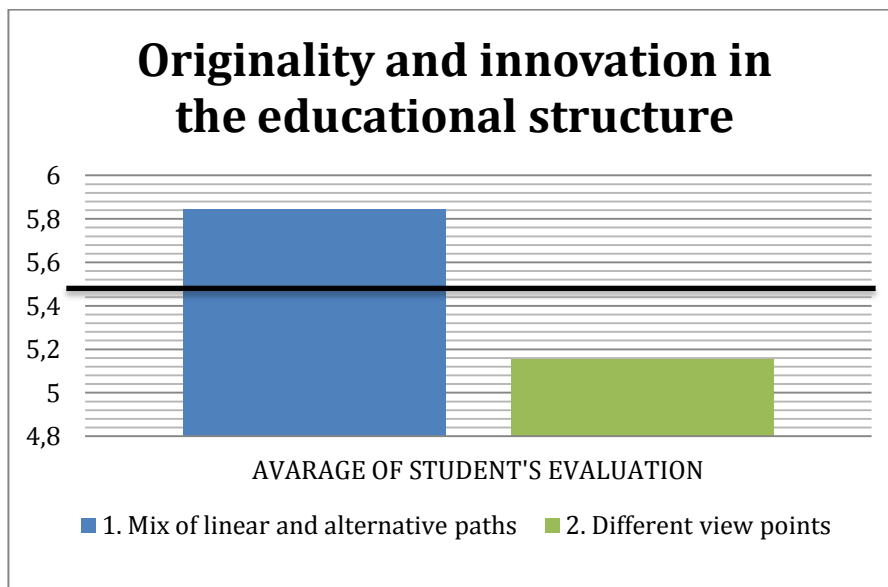


Figure 58: Results on the originality and innovation

As indicated by the *Figure 58* the students have found particularly interesting and innovative the educational structure of the storytelling learning object. In this first experimentation phase they have investigated the micro-adaptivity related to the role change that has given them the possibility to see the story from an other view point. That it has allowed to understand more techniques and evacuation procedure.

6.3.2 Usability of the tool

In order to investigate the overall usability of the Storytelling tool, we collected from students' ratings and open comments on the usability/functionality/ of the tool by using the SUS.

Next, we present the most relevant results of the SUS.

Students found the Storytelling tool particularly easy to use (see *Figure 59*). Students did not find much inconsistency with the Storytelling interface (see *Figure 60*). In addition, students stated that they did not need the support of a technical person to be able to use the tool(see *Figure 61*) and they thought that most people would learn to use the tool very quickly (see *Figure 62*).

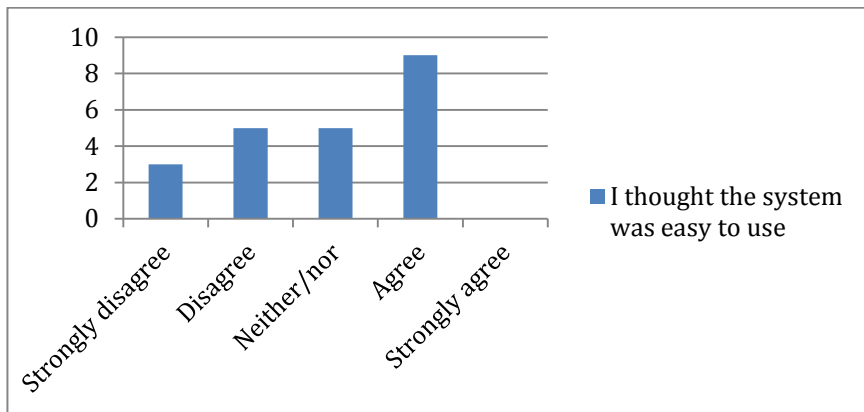


Figure 59. Results on the SUS item “I thought the system was easy to use”

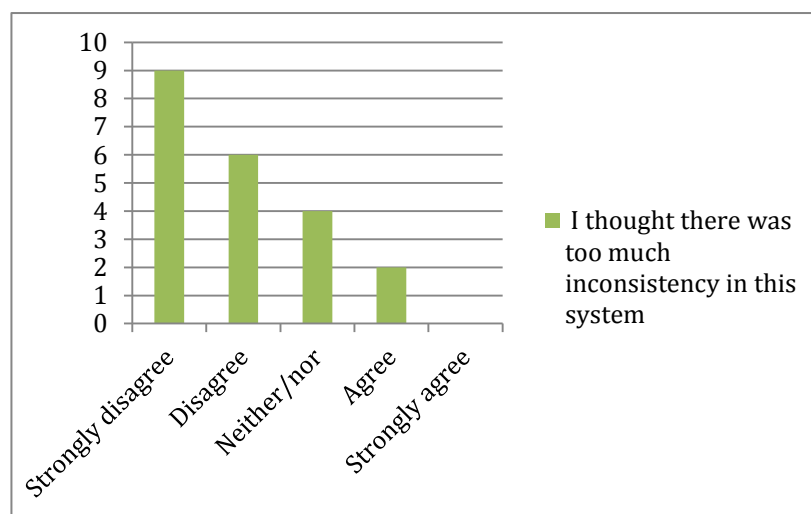


Figure 60. Results on the SUS item “I thought there was too much inconsistency in this system”

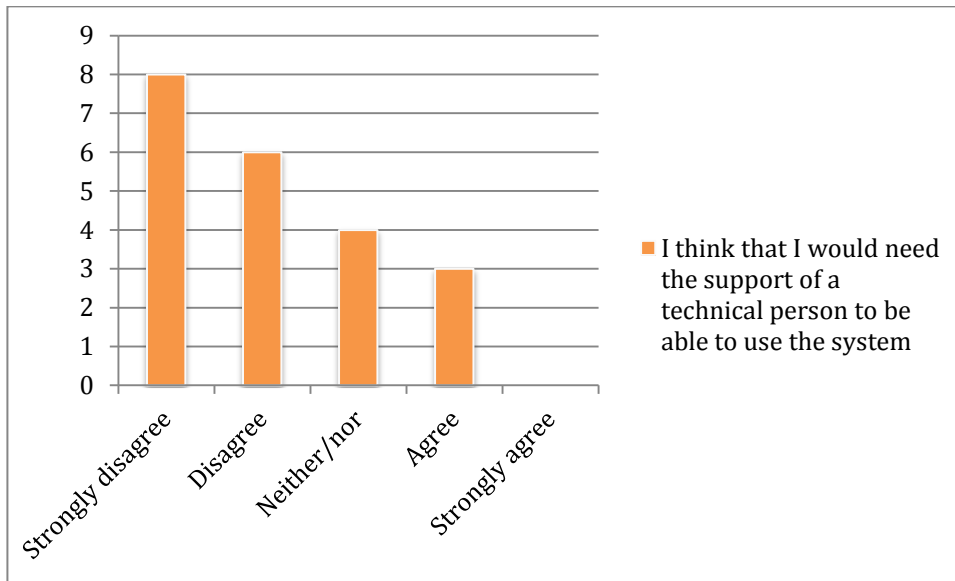


Figure 61. Results on the SUS item “I think that I would need the support of a technical person to be able to use the Storytelling tool”

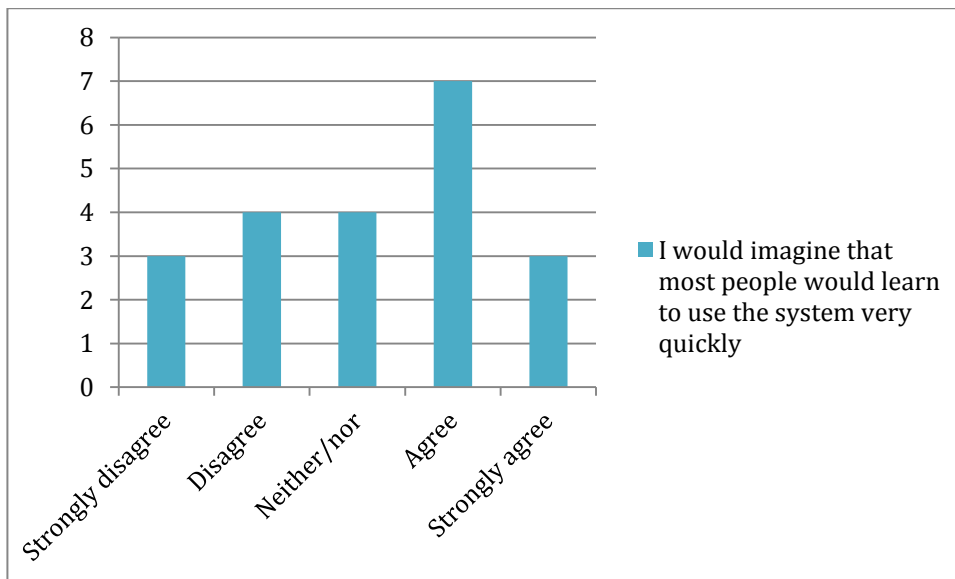


Figure 62. Results on the SUS item “I would imagine that most people would learn to use the Storytelling tool very quickly”

Finally, students stated that the tool was not very well integrated in the course (see Figure 63).

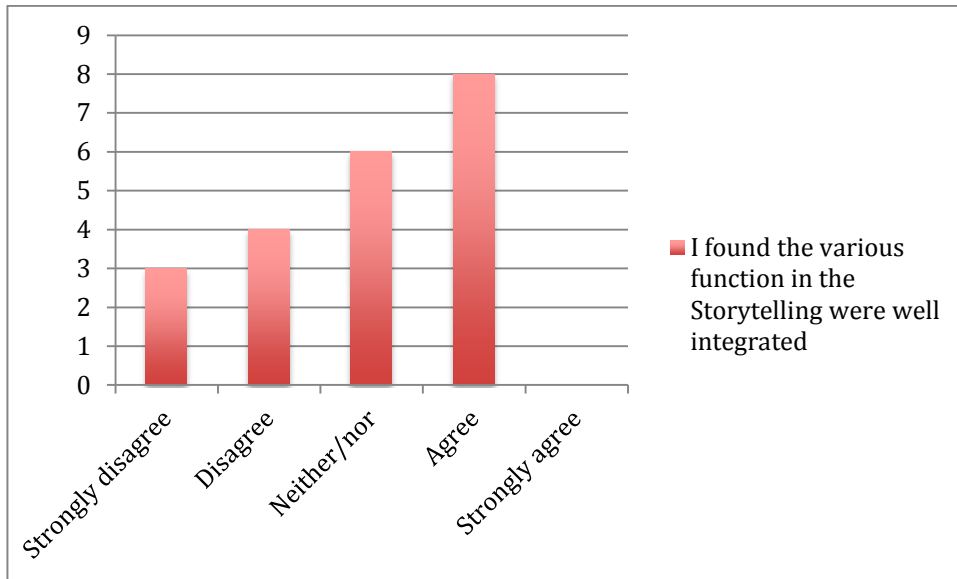


Figure 63. Results on the SUS item “I found the various functions in the Storytelling were well integrated”

In overall, this is a very good result and very promising to face the second iteration of the project having fixed the usability problems found in this first iteration and related to the more flexible management of the story’s structure.

6.3.3 Emotional aspects

Regarding the students emotions during the work with the emotional tool (H5.2), the results from a 4-point rating scale (n=25), as follows:

- Happiness (M=2.3 SD=0.6, Md=2) (Figure 64)

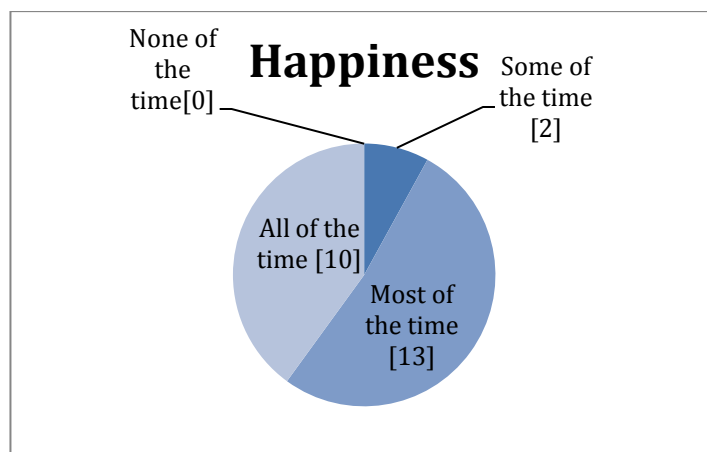


Figure 64. Results on the Happiness emotion

- Sadness (M=0.7, SD=1.1, Md=0) (Figure 65)

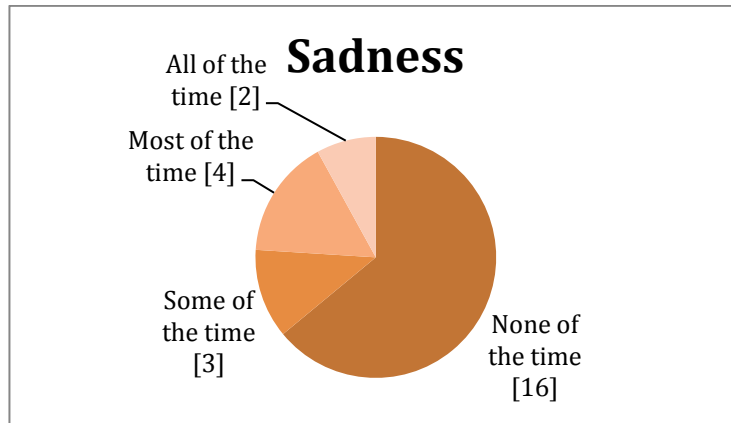


Figure 65. Results on the Sadness emotion

- Anxiety (M=1.04, SD=0.7, Md=1) (Figure 66)

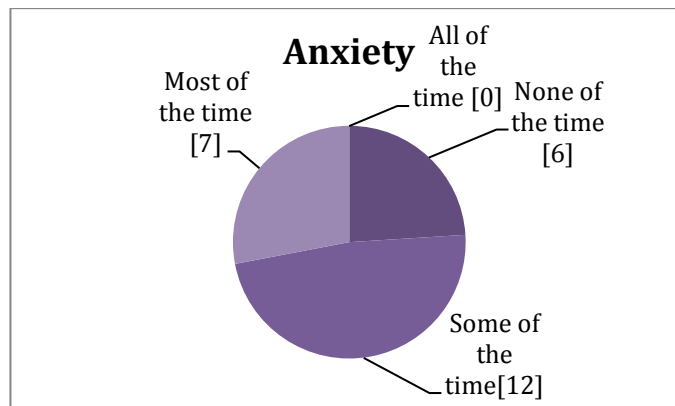


Figure 66. Results on the Anxiety emotion

- Anger (M=0.32, SD=0.47, Md=0) (Figure 67)

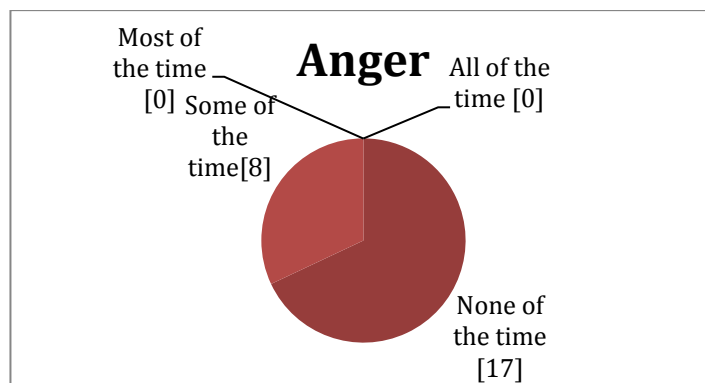


Figure 67. Results on the Anger emotion

6.4 Validation Results

In this paragraph we show the validation methodology that includes the following validation criteria and metric extrapolated by [4].

Validation criteria

- C5.1: To evaluate the level of fulfillment of the tool features.
- C5.3: To evaluate the increase in students' motivation caused by the use of SLOs.
- C5.4: To evaluate the level of satisfaction of the instructors with respect to the inclusion of SLOs in their courses.
- C5.5: To evaluate the increase in students' activity levels due to the use of SLOs.
- C5.6: To evaluate the increase in students' understanding of domain concept.
- C5.7: To evaluate the level of satisfaction of students with the inclusion of the SLO in their courses.

Validation metrics

- M5.5: Number of students using the SLO.
- M5.6: Number of visits of the SLO.
- M5.7: Number of visits of the alternative learning objects.
- M5.8: Students passing the final test and/or with high marks when the SLO is used.
- M5.9: Students passing the final test and/or with high marks when the SLO is not used.
- M5.10: Number of instructors that consider that the SLO is worthy.
- M5.11: Number of students that consider that the SLO is worthy.

Validation techniques both quantitative and qualitative, have included t-test, questionnaire open interview, analysis of the IWT's reporting.

Taking into account the analysis of the questionnaire we can validate the effectiveness of the storytelling's methodology and the storytelling resource. For each category we quoted M, SD and Md.

- Effectiveness of the methodology (M=5.1, SD=1.3, Md= 5.2) (*Figure 68*)

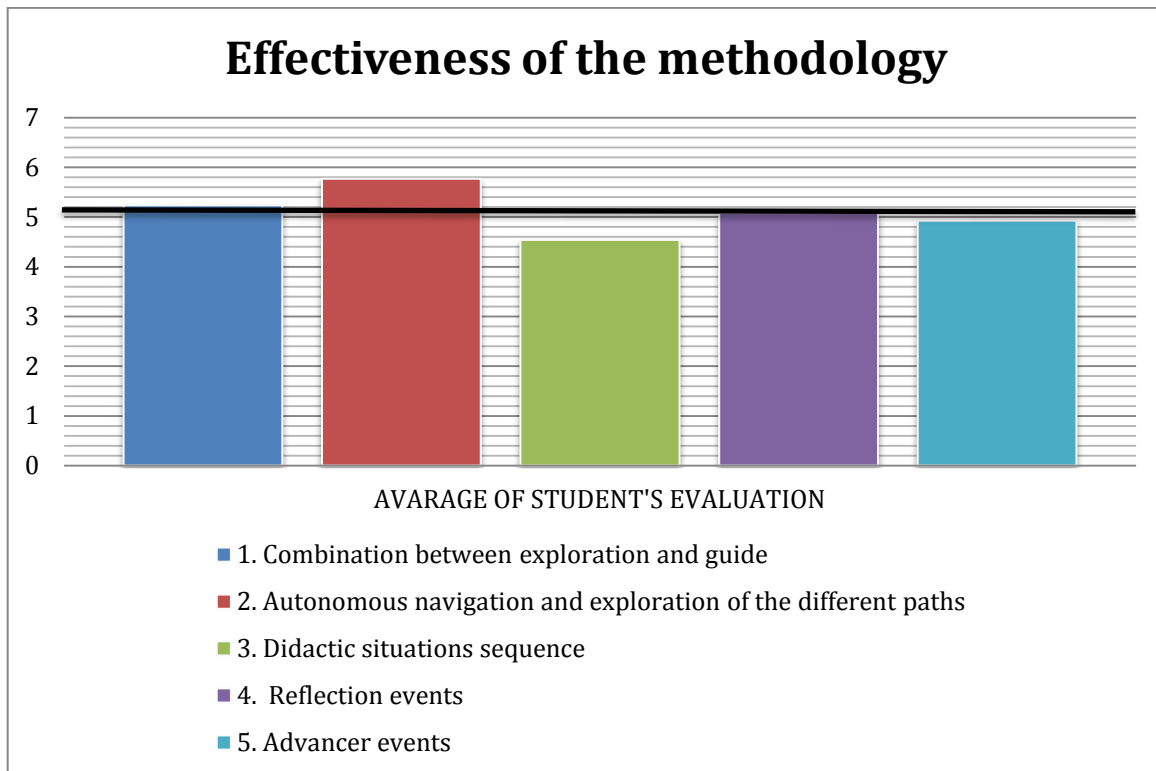


Figure 68: Results on the methodology's effectiveness

In overall, the students have positively valued the new methodology by showing a good percentage to navigate and explore the different learning paths that characterize the Storytelling's structure

- Validation of the storytelling resource wrt the knowledge objectives ($M= 5.1$, $SD= 0.91$, $Md= 5$) (*Figure 69*)

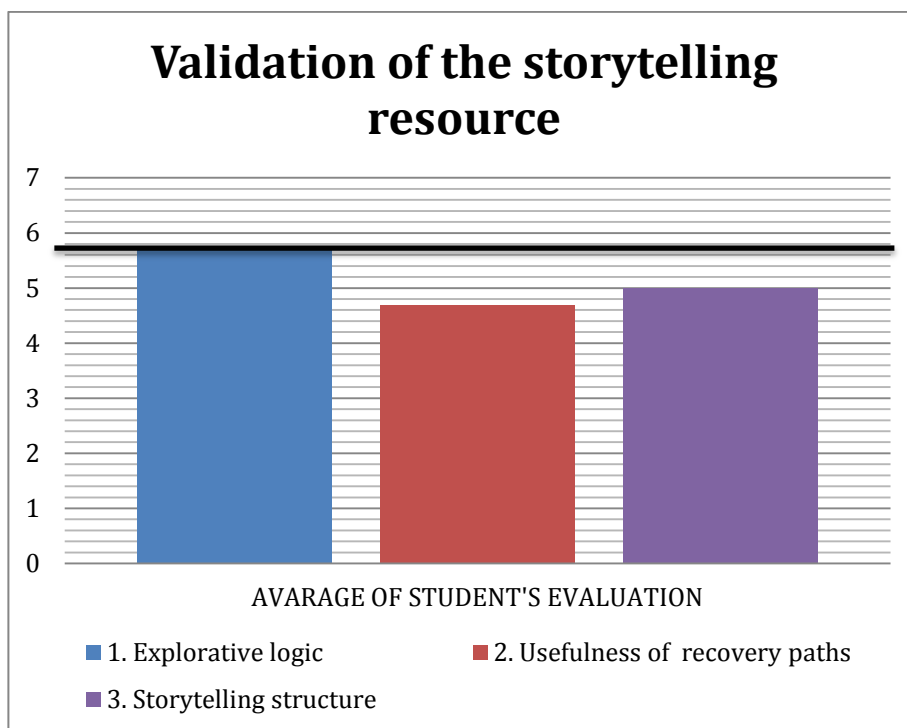


Figure 69: Results on the validation of the storytelling resource

We can affirm a good validation of the storytelling resource, seen as a resource able to meet the different knowledge objectives.

In the following section we report the validation's results for each school involved in the experimentation, done by considering the quantitative data collected from IWT database.

6.4.1 First School "E.Striano"

The class, composed by 14 students, has been divided into two groups: in particular, 7 students form the experimental group and 7 the control group. The students have spent two days for the delivery of the learning course.

6.4.1.1 Experimental group analysis

(1) Startup phase

The experimental group was presented a personalized learning path able to cover, on a one hand, objectives of knowledge dealing with the management of earthquakes, and on the other hand, to match the native profile through the retrieval of resources whose educational metadata was based on intuitive and guided learning approaches, essential for maximising the learned lessons in the prevention of major risks. The group has initially experienced an introductory resource, activating the concept of emergency, then has been allowed to access the storytelling, especially adapted for encouraging the learning through the six situations the storytelling is composed, and linking concepts of motion physics to the dynamic and complex structures of an amusement park.

(2) Correlation between resource efficiency, use and access.

6 out of students from the experimental group had access to the guided-explorative learning resource called “storytelling” that involved them in an adventure whose protagonist (the student impersonates the main character) was asked to know, understand, analyse and act in a situation of seismic event at the amusement park. Such a story, based on an objective-oriented educational model, has kept student’s attention and involvement, so that for 5/7 students the data of use have been higher than 2 hours. As described later, these students have also achieved very good levels of knowledge on the course topics, having the possibility to use the micro-adaptivity of roles. It is interesting to underline that 1/6 students presented low levels of use, insufficient to complete the narrative path and live all the situations opportunely adapted for this.

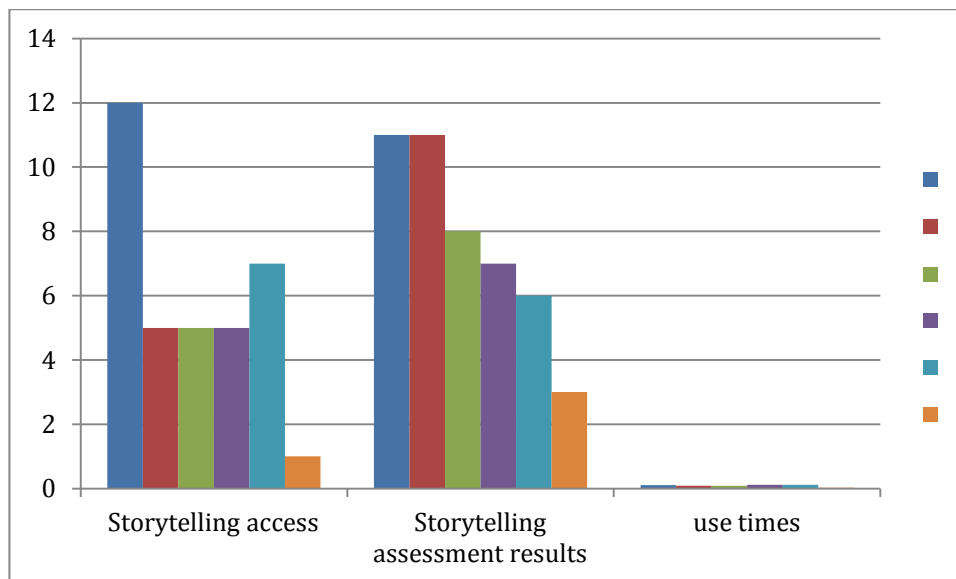


Figure 70: Storytelling feedbacks

(3) Correlation between resource efficacy, test passed and levels of competency acquired.

The didactic narrative resource presented some internal testing steps, suitably inserted to link intermediate trials to different types of target knowledge, whose results would have determined new narrative paths consistent with the branching logic. Passing all the trials would lead the student to new situations until the end of the story. Not passing the trials would lead them to alternative remedial paths, that would bring them to take the narrative sequence again but through different perspectives. Notwithstanding some limits due to the resource complexity, in terms of “challenge” and “effort” the students were asked to cope with, 4 students have completed the adventure path. These students have levels of knowledge higher than the fixed threshold as for “earthquake management”, and have also achieved awareness about the possibility to apply the acquired knowledge to similar contexts (this can be seen from the answers given to final assessment tests).

It should be taken into account that the remaining students composing the experimental group have had access to the test but did not manage to complete and submit it, maybe also owing to the difficulty of some questions which required the student to have spent some time in the story and passed intermediate situations optimally. In this case the learning resource has got an educational value that allows to curb testing mechanisms too easy to be accessed and overcome, due to the weakness of the distractors and to a scarce significance of the questions, which in the storytelling aim at assessing the types of knowledge according to Bloom taxonomy, as for the domain of interest.

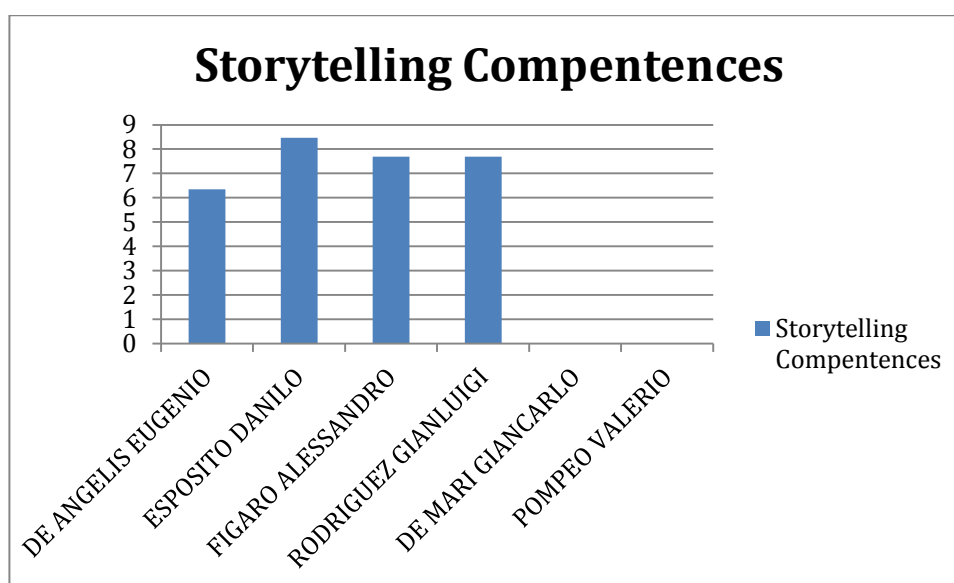


Figure 71: Storytelling Competences

We can observe that 3/4 of those students that had passed the final test has obtained a borderline score with respect to the optimal score for the target concept, while 1/4 had passed with a higher score than the threshold, even though had taken the test twice, and have overcome the knowledge gap. The 3 students that have accomplished and passed the final test at the first attempt not only correspond to those showing high levels of use of the resource (this confirms that the students have spent time attending events and learning the concepts presented through the events), but also to those that have benefited from taking a new role among the ones proposed by one of the 6 situations, thus contributing to learning the concepts through a view that drew them to a greater responsibility.

6.4.1.2 Control group analysis

This group was given the chance to experience a personalized path characterized by the combination of learning resources always effective in covering the target concepts but related to a didactic architecture that offered to the students explanatory resources, far from a kind of interactive and video-based approach.

6/7 students of the group have taken part into the course using the explanatory resources for the target concept “earthquake management”, these resources are also functional, from a communication point of view, and are presented as not interactive LOs. The students have showed a variety of use times, sharing an average limit of 2 hours with minimal accesses. This data also confirms the student’s interest and will to learning, but at the same time the need to end the training sequence as soon as possible and with a minimum number of access. The competences acquired by the students included in the control group are practically of no value, insufficient to consider the competence underlining the concept as fully acquired, and stay below the threshold.

6.4.2 Secondary School “Pitagora”

The class, composed by 28 students, has been divided into two groups: in particular, 14 students form the experimental group and 14 the control group. The students have spent two days for the delivery of the learning course.

6.4.2.1 Experimental group analysis

(1) Startup phase

The group has initially experienced an introductory resource, activating the concept of emergency, then has been allowed to access the storytelling, especially adapted for encouraging the learning through the six situations the storytelling is composed, and linking concepts of motion physics to the dynamic and complex structures of an amusement park. Most students (13/14) have had the opportunity to experiment the new didactic approach and 4/14 were able to finish all the situations of the story, including the micro-adaptivity of the role, in order to achieve a good level of competence in relation to the earthquakes’ management within complex learning environments such as amusement park.

(2) Correlation between resource efficiency, use and emotional state.

The student with a positive emotional state, investigated before the access to the storytelling resource, has reached use times and interaction with the storytelling resource very high. The average time required for understanding key concepts of the interactive story is 1 hours.

3/14 students, in the start up phase, show higher states of indifference, which also associate frustration (1/3) and a combination anxiety and disinterest (2/3). Despite this initial emotional state, these students were involved from the resource storytelling so that 2/3 have maintained a very high level of use with multiple accesses to the resource. The students with a disinterest state (7/14) have also reported an average time to the narrative experience that can lead to think that the structure of the resource has had the ability to maintain attention and arouse curiosity in students not ready.

The students particularly anxious have shown greater concentration especially during the more involving situations so that to arrive at the completion of the intermediate path personalized narrative long before the others. On average the group has accessed at least 4 times in storytelling : that is to be considered a positive sign that the explorative resource has stimulated the desire to relive the experience at 12/14 students.

On the contrary the students with a combination indifference/frustration or indifference/disinterest have obtained use times very low to induce to a passive use of the resource.

	use time	access
Indifference/Anxiety/Disinterest	1.25.45	3
	1.24.18	2
	1.19.11	2
Indifference/Frustration	0.21.23	2
Disinterest	0.46.14	2
	1.25.14	2
	1.25.45	3
Anxiety	1.26.50	2

Figure 72: Correlation between emotional state

(3) Correlation between resource efficacy, test passed and levels of competency acquired.

4/14 students have achieved levels of competency on the concept on the management of the earthquake. The fact that access to the tests for these subjects was also often unique and not repeated also indicates that the content and the presentation modality of the story has allowed students to get to the end of the adventure with a good network of concepts and a good understanding that we can successfully support the time of assessment.

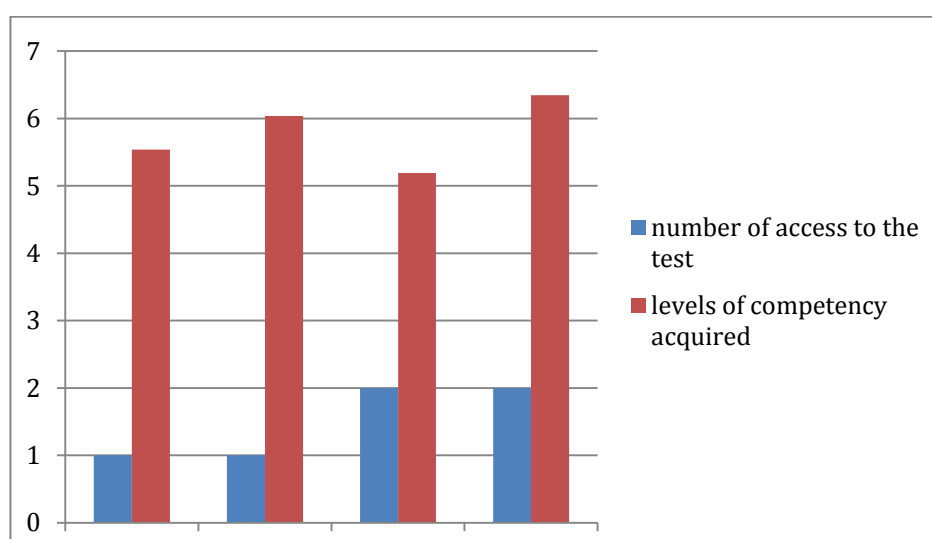


Figure 73: Correlation resource efficacy and levels of competency acquired

1/14 student with a emotional combination of disinterest, anxiety and frustration has reached optimum levels of expertise on the concept on the management of earthquakes that can only mean a good narrative skills educational resource to conduct the student to achieve the fixed objectives.

6.4.2.2 Control group analysis

The control group, composed by 14 students, has delivered, in parallel with the experimental group, a learning path of passive learning resources.

The students did not show involvement, indeed their access to the resources has been very low and the corresponding achieved competences level has been very weak.

So, this way of approaching experience is a confirmation of the fact that there is a gap between how young people prefer to learn and the old ways of teaching. That it is confirmed by the result of the final emotional test that has denoted different emotional indifference or emotional states not altered, a symptom of a liability of the resources that do not support the emergence of emotions that instead the theme for emergencies should stimulate.

6.5 Conclusion

Finally the results are summarized and discussed by considering the goals which were determined at the beginning of the study from evaluation and validation point of view. What goals were achieved and what points should be considered in further work?

In order to investigate these aspects, we reported two interviews submitted to the two tutors of the schools involved within the experimentation:

“Do you think that the storytelling model is didactically well designed and able to meet the needs of learning in a captivating way?”

Tutor A: Well, I think that the experienced educational resources have contributed to students' learning in a motivating and involving way. Participants have experienced potential events recognizing risks and managing ways to face them. Difficult situations have been tackled with right behaviour and I'm sure that all the goals set by the researchers have been reached.

Tutor B: The storytelling learning object is engaging and well developed. The only area for improvement is the language that sometime does not match with the competences level of students belonging to an age range 14-15.

“Do you think that the didactic experience has shown to the students articulated learning path but didactically well guided and able to answer to the different learning style?”

Tutor A:Definetely. The teaching experience has reached the project goals thanks to the digital component according to the students' needs and styles.

Tutor B: The storytelling learning object is a good instrument able to answer to the different learning needs of the students, thanks to the articulation of its structure.

The qualitative data for each school involved in the experimentation lead to the conclusion that the only passive use of didactic resources does not motivate the students to learn and to spend more time studying until the final repetition, and, at the same time, does neither assure the acquisition of knowledge such to allow the students to pass the tests nor a formalization to which the competence make reference.

The storytelling learning resource can offer more variation than the traditional practicing methods. This first experimentation confirms that this innovative and interactive didactic element is more oriented to a student-centered educational approach and it is able to involve emotionally, providing guidance and making more easy the reflection.

Finally, some students have expressed need to can stop the flow of the storytelling for having brainstorming with the tutor and their peers; so, this aspect could be take into consideration in the next experimentation phase in order to give to them the possibility to restart the learning path from the point it was interrupted.

7 R6. A Serious Game for Civil Defence Training in School

The goal of this scenario is to allow an efficient learning about the risk managements through the delivery of a Serious Game (SG) in a personalized learning courses. The use of this kind of resource could contribute to improve the motivation and learning of the students that have a predisposition to the experiential learning.

7.1 Research goals and hypotheses

In this scenario, students of a course held in a secondary school were asked to delivery a complex learning resource (Serious Game) in order to learn the behavior in an emergency situation (like a fire in a building).

The results of this study have been focused on the following goals and hypotheses as described in Deliverable D.1.3:

Evaluation goals

- G6.1: To develop a Serious Game (SG) for Civil Defence that will be deployed alongside IWT within schools
- G6.2: To ensure that the game develops the learners' motivation by placing them in an immersive game environment.
- G6.3: To employ the SG in some online and blended courses in order to enhance some aspects of the teaching/learning process.
- G6.4: To identify possible ways of improving further the utility of the SG in online and blended courses.
- G6.5: To ensure that the SG allows the efficient transmission of lesson learned inside a learning experience on the theme of the risk managements.

Evaluation hypothesis

- H6.1: A SG can be effectively created by instructors as well as stored and played by learners through a user friendly interface.
- H6.2: The use of SGs contributes to improve students' motivation and emotional status.
- H6.3: The use of SGs contributes to support instructors' task.
- H6.4: The use of SGs contributes to increase students' activity levels, both in individual and collaborative activities.
- H6.5: The use of SGs contribute to improve students' understanding of key concepts as well as related skills.

- H6.6: SGs are considered as a worthy educational resource by both instructors and students.

7.2 Method

7.2.1 Participants

Two secondary schools have taken part to the experimentation.

In the first school “E. Striano” there were 14 students in the course: gender male and average 16 years old.

In the second school “Pitagora” there were 28 students in the course: 26 were female (98%) , 2 student were male (2%) ant the participant were on average 14 years old.

The students of the first school have shown a major responsibility and maturity on the topics illustrated in the course with respect to the students of the secondary school.

Each class were supervised by two tutors. Within each class the students have been divided in two groups: experimental and control, in order to make a more comparative analysis of the investigated tools.

7.2.2 Apparatus and Stimuli

We asked to the experimental group to interact with virtual objects in a virtual environment using a physical interface. During this experiment they have been asked to evacuate from a virtual building in an emergency situation.

On completion of the session they have filled a Post-Questionnaire, which included the following sections: demographic data, game experience, usability of the tool interface, further comments or suggestions. Besides, we provided a Questionnaire for the tutors concerning the added value of this kind of resource in a learning course and how it can contribute to ameliorate the knowledge.

For qualitative statistical analysis, we summarized the open answers in the surveys. For quantitative statistical analysis, we performed t-tests.

With regard to the game experience, the students are asked to assess the serious game concerning the following questions:

1. How "in control" did you feel over events in the game?
2. How responsive was the game to actions that you initiated (or performed)?
3. How natural did your interactions with the game seem?
4. How much did the visual aspects of the game involve you?
5. How natural was the mechanism which controlled movement in the game?
6. How much did your experiences with the game seem consistent with your real world experiences?
7. Were you able to anticipate what would happen next in response to the actions that you performed?

8. How completely were you able to actively survey or search the game using vision?
9. How compelling was your sense of moving around the game?
10. How well could manipulate the game?
11. How closely were you able to examine the game?
12. How quickly did you adjust to the experience?
13. How well could you concentrate on the assigned tasks or required activities rather than on the mechanisms used to perform those tasks or activities?
14. How well could you actively survey or search the game using the controls and interface?
15. To what extent did external events distract from your experience of the game?
16. How completely were your senses engaged in this experience?
17. How easy was it to identify the game through physical interaction; such as touching it?
18. Were there moments during the virtual environment experience when you felt completely focused on the task or game?

The answer categories in this section are “Not at all”, “Somewhat”/ “Moderately compelling”, “Completely”/ “very compelling”

All Questionnaires contained quantitative as well as qualitative questions, the answer categories varied between rating scales or open answers.

7.2.3 Procedure

In order to give more emphasis to the experimentation, a learning course, designed to cover a macro concept on emergency management in environments with high levels of risk in case of fires and earthquakes, has been built in the IWT. The course has been delivered by two groups of users having the same learning styles and divided in two groups: experimental and control.

The experimental group has had access to an educational experience created specifically to meet the complex learning topics related to emergency management through the use of Complex Learning Object. The CLOs have been represented a Serious Game, for supporting intuitive learning processes in case of fire in school, and a Storytelling, for promoting the lessons learned through guided explorative processes in the case of a seismic event in a complex structure .

In the following section the individual steps of the experiment are described.

After that each student logged in IWT platform, he see his class and group.

In a first step the two groups were assigned the specific course.

For the experimental group was created a personalized learning path by having as concept objective the acquisition of the behavior to take for managing high risks as the fire in a building through complex learning resources.

The control group has also delivered a personalized learning path with the same concept objective but the kind of learning resources is less interactive and active than the experimental group.

When all groups had finished the delivery of the learning resources, each member of a group had made an assessment test for testing the knowledge acquired by the serious game for the experimental group and by a passive learning resource for the control group.

7.3 Evaluation Results

In this section we focus on the activity level and usability of the Serious Game delivered by IWT platform (H6.1-H6.6). The evaluation results have been obtained by providing several statistics data, as Mean (M), Standard Deviation (SD) and Median (Md).

The survey is a study of the qualitative questionnaires submitted to all the students of the two schools belonging to the experimental group.

Regarding the students activity and interaction with the Serious Game object, we report in the *Figure 74* the results of the questionnaire exposed in the Section 7.2.2.

The related statistics data are: M=4.5, SD= 1.3, Md= 4.5

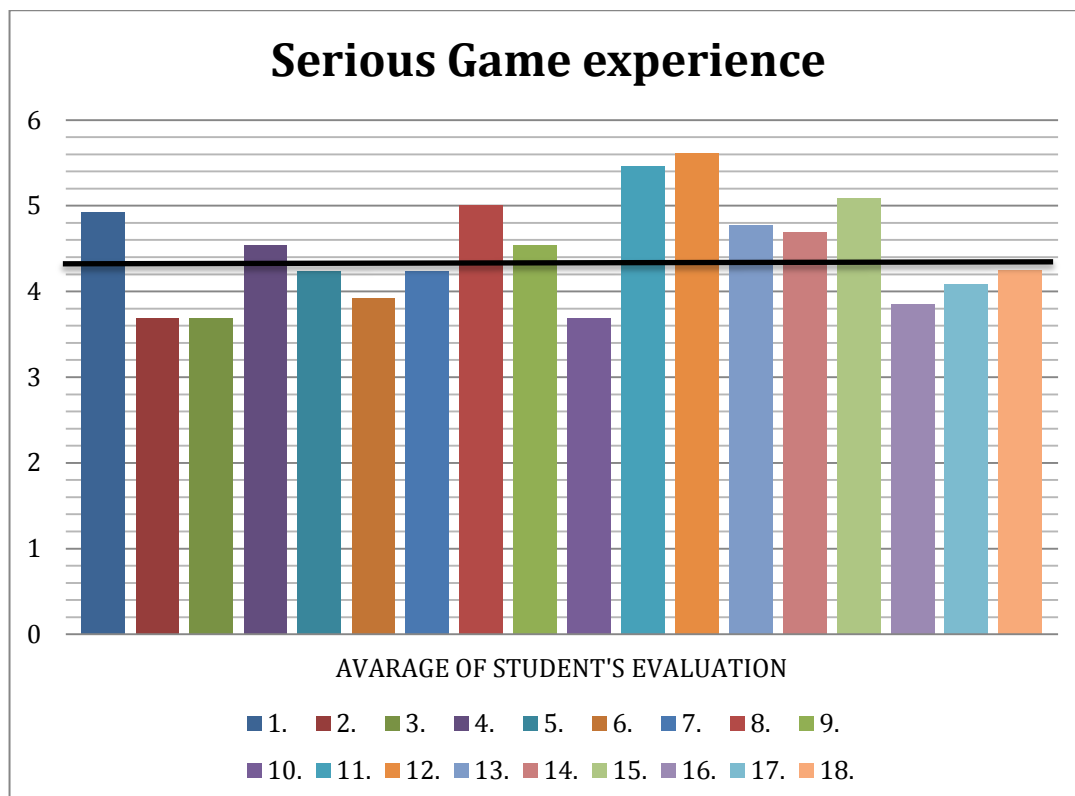


Figure 74: Results on the Serious Game experience

The *Figure 74* allows for doing some consideration in terms of the usability and interaction with the game. Indeed, the answers related to the questions 2-3-10 suggest to improve some game's aspect like the responsiveness to the performed actions, the interaction and the manipulation of the game.

Moderate results have been obtained considering the answers to the questions 16-18 in term of emotion, though this aspect could be improved in the second experimentation by a more involvement of the students in the virtual environment experience.

Overall we can register a good experimentation result since a lot of the students have been able to examine the game as shown by the answers to the question 11.

Validation Results

In this paragraph we show the validation methodology that includes the following validation criteria and metric extrapolated by [4]:

Validation criteria

- C6.1: To evaluate the increase in students' motivation caused by the use of a SG.
- C6.2: To evaluate the level of satisfaction of the instructors with the inclusion of SG in their courses.
- C6.3: To evaluate the increase in students' activity levels due to the use of the SG.
- C6.4: To evaluate the increase in students' understanding of key domain concepts and students' results.
- C6.5: To evaluate the level of satisfaction of students with the inclusion of the SG in their courses.

Validation metrics

- M6.1: Time employed in creating each SG.
- M6.2: Number of students using the SG.
- M6.3: Number of visits of the SG.
- M6.4: Number of visits of the alternative learning objects.
- M6.5: Number of students passing the final test and/or with high marks when the SG is used.
- M6.6: Number of students passing the final test and/or with high marks when the SG is not used.
- M6.7: Number of students passing the final test and/or with high marks when both the SG and the alternative learning objects are used.
- M6.8: Number of instructors that consider that the SG is worthy.
- M6.9: Number of students that consider that the SG is worthy.

Validation techniques both quantitative and qualitative, have included t-test, questionnaire open interview, analysis of the IWT's reporting.

Said that, in the following section we report the validation results for each school involved in the experimentation.

7.3.1 First School “E.Striano”

The class, composed by 14 students, has been divided into two groups: in particular, 7 students form the experimental group and 7 the control group. The students have spent two days for the delivery of the learning course

7.3.1.1 Experimental group analysis

About the participants to the group, 7 students on 7 have had access to the game taking confidence with its structure, communication modalities, interface in order to achieve the objective of the game and can pass the assessment test in a successful way.

(1) Startup phase

5 out of the total number of students have managed to end the game in an optimal way to testify that the new strategy, adopted in a learning course, has correctly supported the learning.

We can now see how different variables are linked to the game in terms of access, use and achieved skills.

(2) Correlation between resource efficiency, use and access.

The students were involved in the game for an average time of 30 minutes; the system has notified different access to the serious game resources showing how the students have had the motivation to enjoy the game several times in order to overcome the various level and relative assessment.

The permanence in IWT platform is also indicative of a more critical lecture of the stimuli shown during the evacuation scene in order that to end the game in a successful way.

The students have interacted with the tutor for analyzing the critical actions and reflecting on possible mistakes that led 5/7 students to not abandon the game. Only 2/7 students have instead reported use times unsatisfactory; it may mean a failure to understand the modalities and the objectives of the game that it has been translated in an abandonment of the game after the first access to the resource.

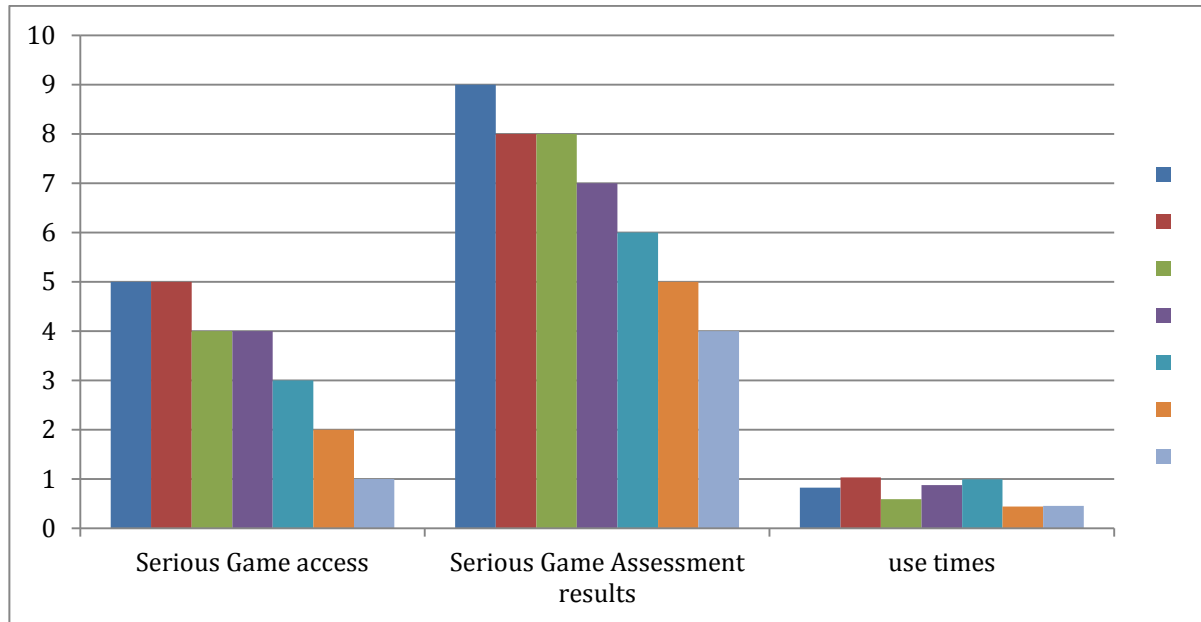


Figure 75: Serious Game feedbacks

(3) Correlation between resource efficacy, test passed and levels of competency acquired.

5 out of the 7 students have managed to end the game in an optimal way by reporting a good competence level. 2 students have not reached satisfactory levels: one of them despite several accesses has not attained sufficient competence; another student may not just passed the assessment test.

We denote that the students with a good competence level were the same that have reported different access to the game. On the contrary, the students who have achieved unsatisfactory levels of the game are those who use times has not sufficient to overcome the assessment test. Consequently the had not acquired the necessary knowledge and the experience for managing emergencies.

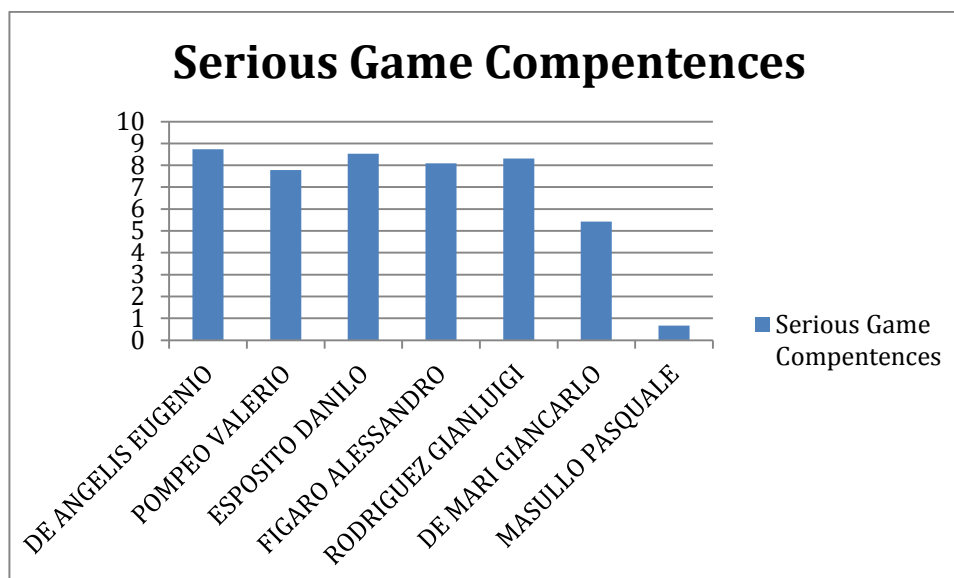


Figure 76: Serious Game Competences

7.3.1.2 Control group analysis

The control group composed by 7 male students has delivered a learning path composed by expositive learning resources that cover the same concepts exposed in the experiential group.

Quantitative analysis of the data has shown that the use times has been rather low for having a significant learning. The passive resources have not convinced the control group so as not to justify a new visualization and delivery of the resources. The achieved competences level by all the member of the group is below to the overcoming threshold; that is due to the fact that the passive delivery (without personal involvement) has not allowed students to get to the test with knowledge relevant to the subject of study.

7.3.2 Second School "Pitagora"

The class, composed by 28 students, has been divided into two groups: in particular, 14 students form the experimental group and 14 the control group. The students have spent two days for the delivery of the learning course

7.3.2.1 Experimental group analysis

Before to start with the delivery of the game, the students have notified different difficulties related to the network connection. For this purpose, the tutor has shown to the group a game's trailer in order to give an idea to the steps of the game. So, a brainstorming phase has anticipated the game, within that the students participation has been so significant to suggest that a blended modality should be a best approach for delivering complex and adaptive learning experience.

After successful installation of the game, students were able to get in a virtual environment and to act individually in order to be able to save themselves in case of fire.

We can now see how different variables are linked to the game in terms of access, use and achieved skills.

(1) Correlation with permanence in IWT and use times

8 out of the total number of students show times higher enjoyment of the game. 4 of these students were those who had recorded high levels of indifference combined with disinterest. The students that within start-up phase have shown a good component of anxiety were also those who have spent more time in the game and its challenges. So, probably, the anxiety within the experiential learning contexts could be considered an interesting variable that feeds the desire to pass the different trial.

(2) Correlation with levels of competency acquired

5/14 have reached a good level of competence on how to behave in case of fire in the school. This indicates a good component of the game that allows the sedimentation of knowledge. In particular, we note that multiple access to the resource by these students is accompanied by a progressive improvement of performance and achieves a higher level of competence.

7.3.2.2 Control group analysis

The control group composed by 14 female students has delivered a learning path composed by expositive learning resources that cover the same concepts exposed in the experiential group.

Quantitative analysis of the data has shown that the use times have been rather low for having a significant learning. The passive resources have not convinced the control group so as not to justify a new visualization and delivery of the resources. Only 2 students have achieved competences related to the game through a positive assessment test. The remaining members of the group have not shown a lot of attention or involvement in the game's issues.

7.4 Conclusion

Finally the results are summarized and discussed by considering the goals which were determined at the beginning of the study from evaluation and validation point of view. What goals were achieved and what points should be considered in further work?

In order to investigate these aspects, we reported two interviews submitted to the two tutors of the schools involved within the experimentation:

“Do you think that the Serious Game model is didactically well designed and able to meet the needs of learning in a captivating way?”

Tutor A: The Serious Game is well developed though the interaction with the learners could be improved.

Tutor B: The Serious Game is engaging and well developed. The only area for improvement is the language that sometime does not match with the competences level of students belonging to an age range 14-15.

The qualitative data for each school involved in the experimentation lead to the conclusion that the only passive use of didactic resources does not motivate the students to learn and to spend more time studying until the final repetition, and, at the same time, does neither assure the acquisition of knowledge such to allow the students to pass the tests nor a formalization to which the competence make reference.

Taking into consideration the quantitative data, some improvement should be taken into account for the next experimentation in term of the usability and interaction with the game.

Anyhow we have to note that these factors are strictly related to the performances of the used PCs. The game in fact exploits many resources of the PCs, and require power PC, so in future we will try to refine also this aspect.

8 R7. Affective and Emotional Approaches

In this scenario, students of a course held in a secondary school were asked to use the affective/emotional tool for testing their condition before and at the end of the course related to the management risk.

8.1 Research goals and hypotheses

The results of this study have been focused on the following goals and hypotheses as described in [4]:

Evaluation goals

- G7.1: to build a system that is able to recognize, evaluate and stimulate the emotions and the affective state of a learner in order to support and improve learning.
- G7.2: to ensure that the system is able to detect alterations of user's emotional/affective state during a learning experience.
- G7.3: to ensure that the system is able to perform an affective/emotional assessment and
to provide a correct estimation of the current learner state.
- G7.4: to assist the learner during affective/emotional assessment through a friendly interface that is easy to use and to understand.

Evaluation hypotheses

- H7.1: it is possible to create a learning system able to stimulate the affectivity and the emotionality of a learner.
- H7.2: by recognizing and assisting emotions and affectivity it is possible to improve students' motivation and to create a predisposition to learning.
- H7.3: by recognizing and assisting emotions and affectivity it is possible to improve students' understanding of domain concepts.
- H7.4: The visualization and interaction of appropriate learning resources improves the emotional state altered.
- H7.5: the system for emotional/affective management is considered as a worthy resource by both instructors and students.

8.2 Method

8.2.1 Participants

Two secondary schools have taken part to the experimentation.

In the first school “E. Striano” there were 14 students in the course: gender male and average 16 years old.

In the second school “Pitagora” there were 28 students in the course: 26 were female (98%), 2 student were male (2%) and the participants were on average 14 years old.

The students of the first school have shown a major responsibility and maturity on the topics illustrated in the course with respect to the students of the secondary school.

Each class was supervised by two tutors. Within each class the students have been divided in two groups: experimental and control, in order to make a more comparative analysis of the investigated tools.

8.2.2 Apparatus and Stimuli

We asked all groups (experimental and control) to edit the affective/emotional test before they started the course and after they ended it.

The students had answered the Pre-Questionnaire concerning their demographic data, and their general attitudes concerning the use of the computerized technology. After the course, they filled out a Post-Questionnaire, which included the following sections: demographic data, affective/emotional interface, usability of the emotional tool, emotional aspects and further comments or suggestions. Besides, we provided a Questionnaire for the tutors concerning the added value of the emotional component in a learning course and how it can contribute to ameliorate the knowledge.

For qualitative statistical analysis, we summarized the open answers in the surveys. For quantitative statistical analysis, we performed t-tests.

Regarding the section “Affective/emotional Interface”, the students are asked to assess the work concerning the following questions:

1. The recognition of your emotional state feels you at the centre of the attention during the learning path?
2. The display of your emotional and affective state leads you to improve your performance levels?
3. Do you think that a collecting of the emotional state during the learning experience could provide useful information for the improvement of learning?
4. Do you think that the emotional test is representative of your emotional state?
5. Do you think that the emotional/affective state impact greatly on the results of your educational experience?
6. Do you think that the data collected can be used to provide additional activities useful for recovering the emotional balance?
7. Do you think that the emotional test should be made visible to the peers in order to trigger a social support?

The answer categories in this section are “In no way”, “Partially”, “Completely” .

For qualitative statistical analysis, we summarized the open answers in the surveys. For the quantitative statistical analysis we employed basic statistics, such as Mean (M), Standard Deviation (SD) and median (Md).

For the section “usability of the storytelling environment” in the Post-Questionnaire and the Questionnaire for the tutors, we used the SUS(System Usability Scale [6]) which contains 10 items and a 5 point Likert scale to state the level of agreement or disagreement, for instance “I think that I would like to use this system frequently”.

To investigate in which emotional state the students were when they used the storytelling tool, we added a section concerning “emotional aspects”, which included 12 items of the Computer Emotion Scale (CES) [7] that measure emotions related to learning new computer software. Research showed that the 12 items are describing four emotions:

- Happiness (“When I used the tool, I felt satisfied/excited/curious.”)
- Sadness (“When I used the tool, I felt disheartened/dispirited.”)
- Anxiety (“When I used the tool, I felt anxious/insecure/helpless/nervous.”)
- Anger (“When I used the tool, I felt irritable/frustrated/angry”)

The answer categories in this section are “None of the time”, “Some of the time”, “Most of the time” or “All of the time”.

All Questionnaires contained quantitative as well as qualitative questions, the answer categories varied between yes/no, rating scales or open answers. Regarding the rating scales, for the majority of the quantitative questions we used the 5 point Likert scale, so that students could state their level of agreement or disagreement. The rating scale ranged from “I strongly disagree” (1), “I disagree” (2), “neither/nor” (3) to “I agree” (4), “I strongly agree” (5).

Finally, quantitative data was also collected from IWT database and log files and are reported in 8.4 Section.

8.2.3 Procedure

In order to give more emphasis to the experimentation, a learning course, designed to cover a macro concept on emergency management in environments with high levels of risk in case of fires and earthquakes, has been built in the IWT. The course has been delivered by two groups of users having the same learning styles and divided in two groups: experimental and control.

The experimental group has had access to an educational experience created specifically to meet the complex learning topics related to emergency management through the use of Complex Learning Object. The CLOs have been represented a Serious Game, for supporting intuitive learning processes in case of fire in school, and a Storytelling, for promoting the lessons learned through guided explorative processes in the case of a seismic event in a complex structure .

All the groups have had access to the Emotional/Affective tools.

In the following section the individual steps of the experiment are described.

After that each student logged in IWT platform, he see his class and group.

In a first step the two groups were assigned the specific course.

Before starting with the delivery of the course each group made the emotional/affective test in order to monitor the students' state.

When all groups had finished the delivery of the course, each member of a group had made a second emotional/affective test in order to check if the learning path has contributed to change or not their emotional condition.

In the post-questionnaire the students were asked about the usability/functionality of the tool. In addition the tutors were also asked to answer a questionnaire about the usability of the tool and their think on the added value of this instrument in the didactic experience.

8.3 Evaluation Results

In this section we focus on the activity level, usability and emotional aspects of the Emotional tool delivered by IWT platform (H7.1-H7.5). The evaluation results have been obtained by providing several statistics data, as Mean (M), Standard Deviation (SD) and Median (Md).

The survey is a study of the qualitative questionnaires submitted to all the students of the two schools belonging to the experimental group.

8.3.1 *Affective/emotional Interface*

In this section we focus on the activity level, usability and emotional aspects of the affective/emotional tool delivered by IWT platform (H7.1-H7.5). The evaluation results have been obtained by providing several statistics data, as Mean (M), Standard Deviation (SD) and Median (Md).

The survey is a study of the qualitative questionnaires submitted to all the students of the two schools belonging to the experimental group.

Regarding the students' activity and interaction with the affective/emotional tool, we report the results of the questionnaire exposed in the Section 8.2.2.

The related statistics data are: M=4.1, SD= 1.3, Md= 4.

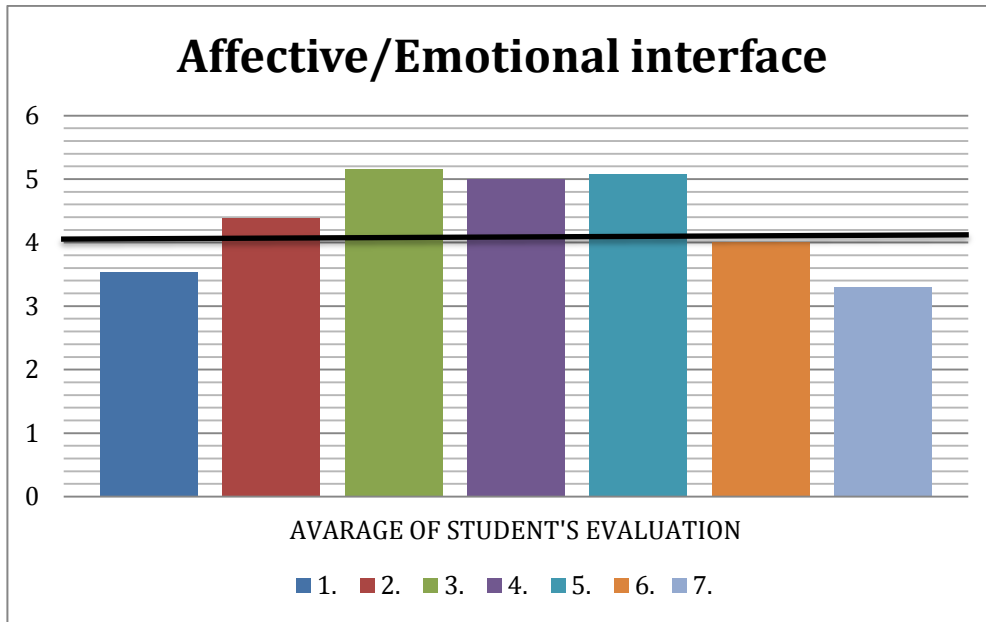


Figure 77: Results on affective/emotional interface

The quantitative analysis of the answers of the questionnaire submitted to the students is quite interesting. Indeed positive results have been obtained from the interaction with the emotional tool both in term of emotional state compliant with the real state of the student and in term of management of emotional interface.

The average answers to the question 1 leads to an improvement for the next experimentation and related to the possibility to define the content for the state equilibration of Emotional aspects.

8.3.2 Usability of the tool

In order to investigate the overall usability of the emotional tool, we collected from students' ratings and open comments on the usability/functionality/ of the tool by using the SUS.

Next, we present the most relevant results of the SUS.

Students found the Emotional tool particularly easy to use (see *Figure 78*). Students did not find much inconsistency with the tool interface (see *Figure 79*). In addition, students stated that they did not need the support of a technical person to be able to use the tool (see *Figure 80*) and they thought that most people would learn to use the tool very quickly (see *Figure 81*).

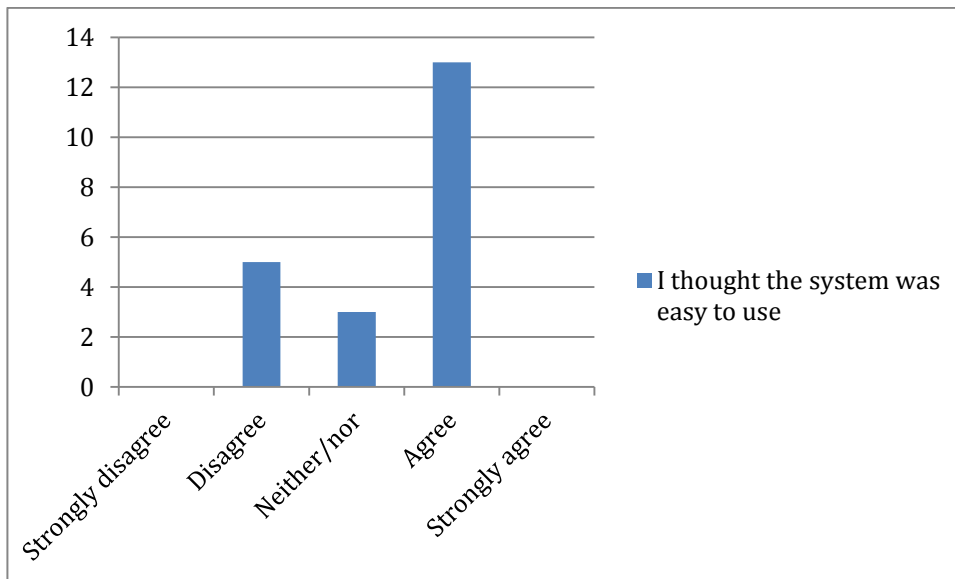


Figure 78. Results on the SUS item “I thought the system was easy to use”

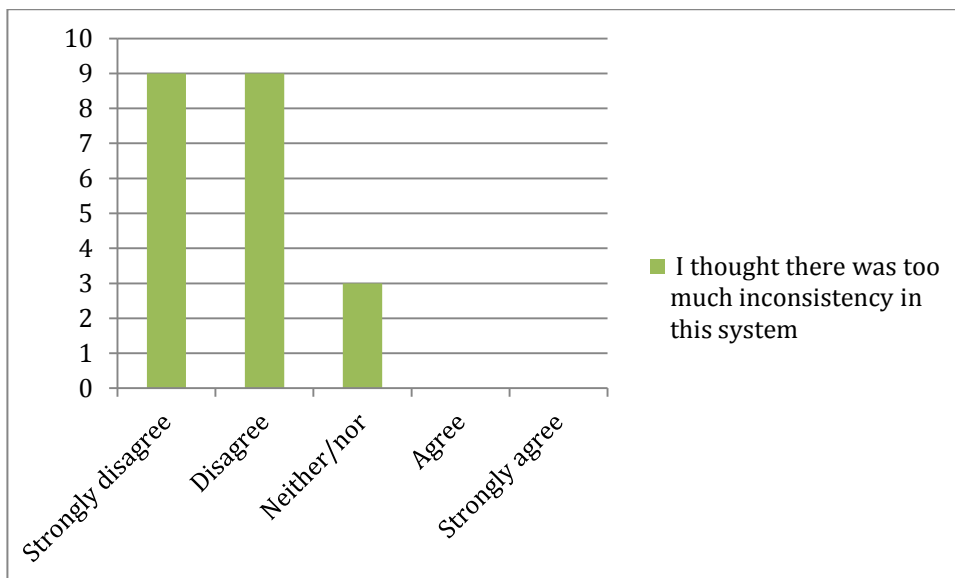


Figure 79. Results on the SUS item “I thought there was too much inconsistency in this system”

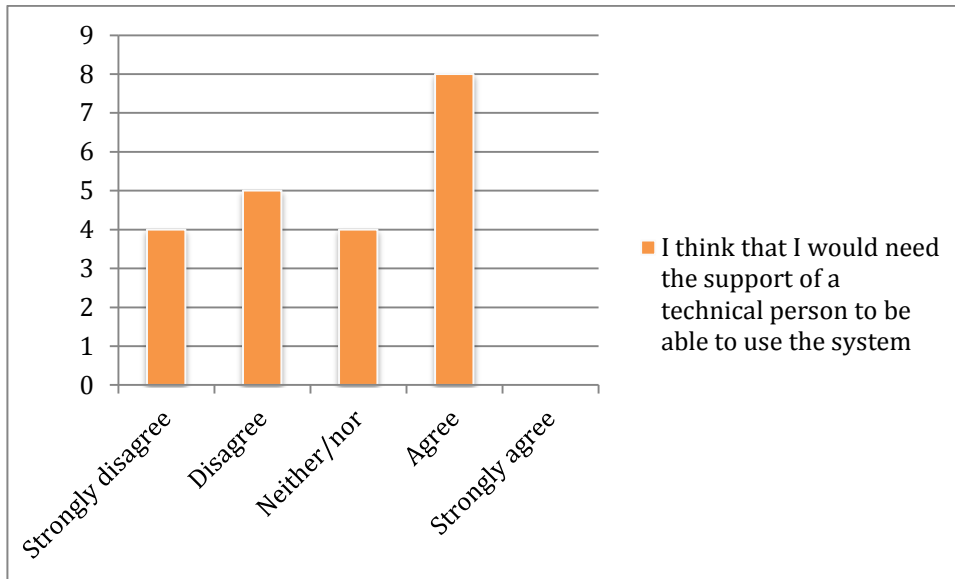


Figure 80. Results on the SUS item “I think that I would need the support of a technical person to be able to use the Emotional tool”

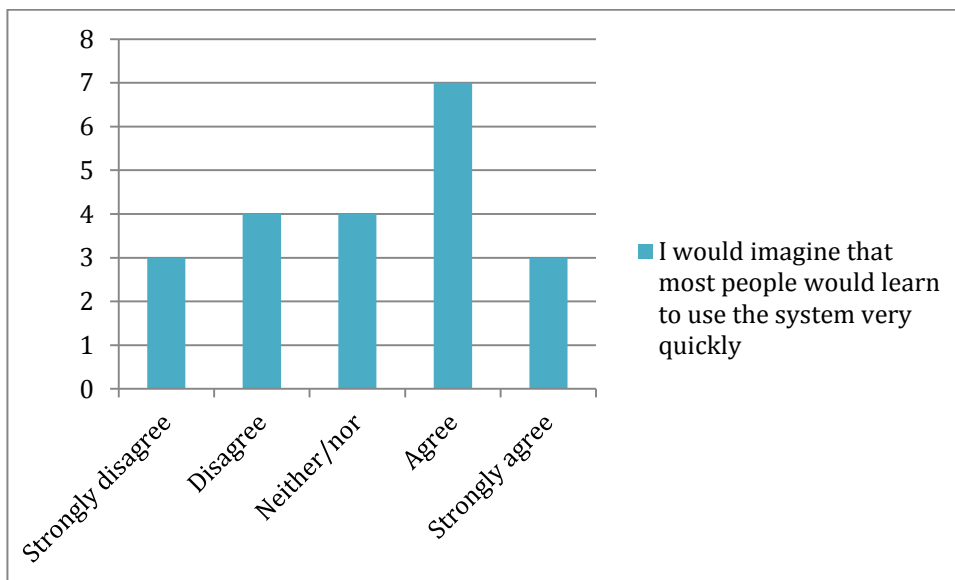


Figure 81. Results on the SUS item “I would imagine that most people would learn to use the Emotional tool very quickly”

Finally, students stated that the tool was not very well integrated in the course (see Figure 82).

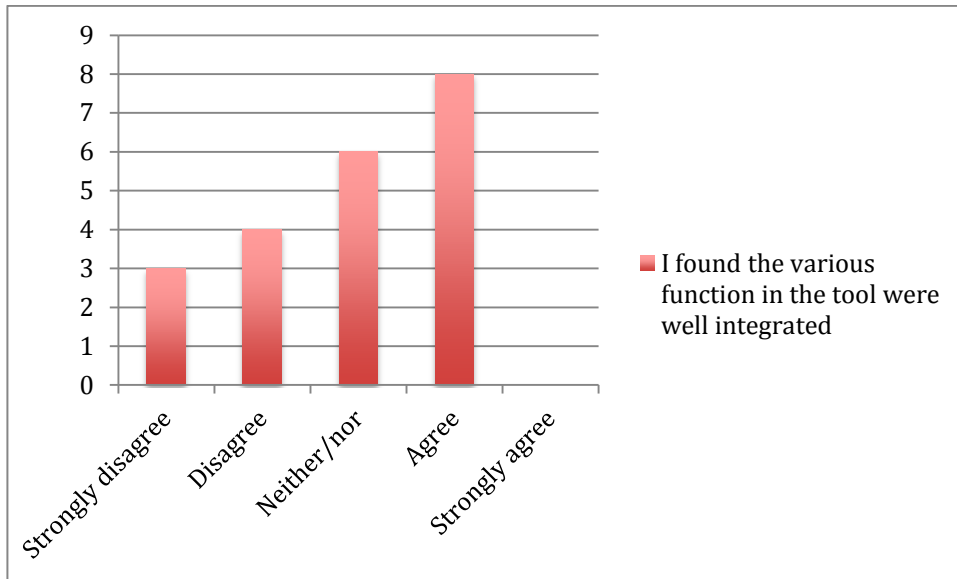


Figure 82. Results on the SUS item “I found the various functions in the tool were well integrated”

In overall, this is a good result though it has been notified a little safeness in the autonomous use of the emotional tool such as to require a physical support as shown by the Figure 82.

For the second iteration of the project, it will be useful to fix this usability problem.

8.3.3 Emotional aspects

Regarding the students emotions during the work with the emotional tool (H7.1), the results from a 4-point rating scale (n=25), as follows:

- Happiness (M=1.3 SD=0.8, Md=1) (Figure 83)

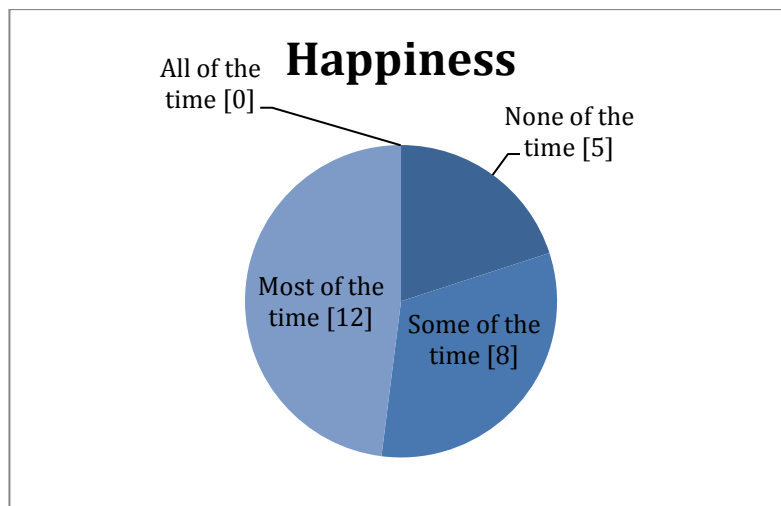


Figure 83. Results on the Happiness emotion

- Sadness (M=1.1, SD=0.7, Md=1) (*Figure 84*)

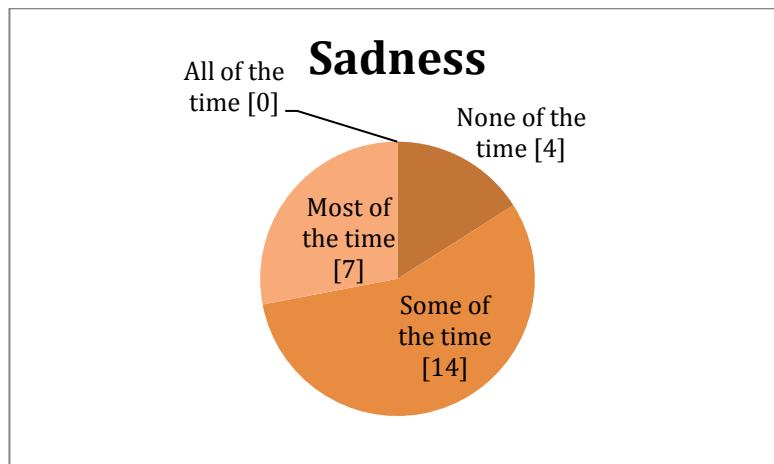


Figure 84. Results on the Sadness emotion

- Anxiety (M=1.3, SD=0.9, Md=1) (*Figure 85*)

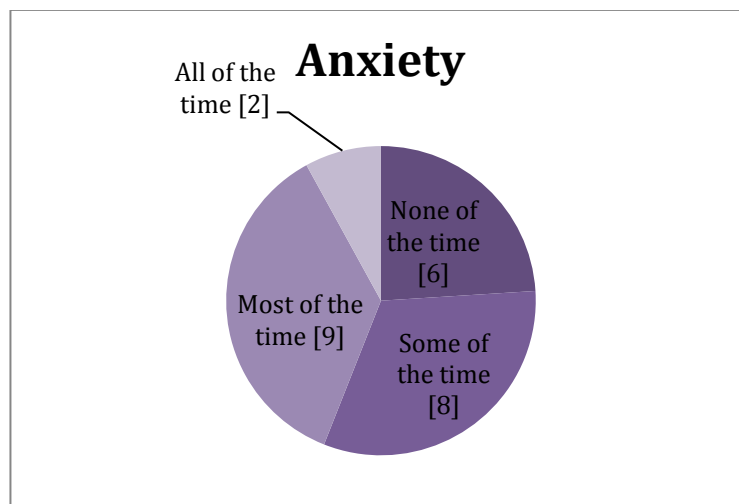


Figure 85. Results on the Anxiety emotion

- Anger (M=0.4, SD=0.5, Md=0) (*Figure 86*)

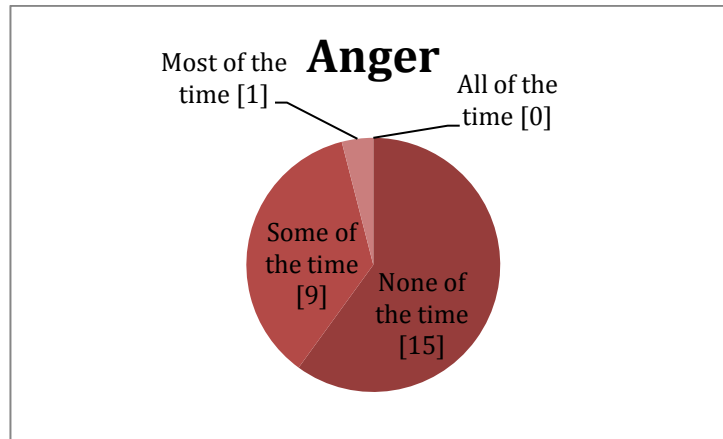


Figure 86. Results on the Anger emotion

In overall, this is a very good result and very promising to face the second iteration of the project.

8.4 Validation Results

In this paragraph we show the validation methodology that includes the following validation criteria and metric extrapolated by [4].

Validation criteria

- C7.1: To evaluate the level of fulfillment of the system features.
- C7.2: To evaluate the level of satisfaction of the learners using the system.
- C7.3: To evaluate the increase in students' motivation due to the affective and emotional support.

Validation metrics

- M7.1: Number of students requiring affective/emotional support.
- M7.2: Number of courses in which it is required the affective/emotional support.
- M7.4: Time spent by the system for evaluation of the emotional/affective state.
- M7.5: Number of students that consider the emotional/affective support worthy.
- M7.6: Number of instructors that consider the emotional/affective support worthy.
- M7.7: Number of students passing the final test and/or with high marks when the emotional/affective system is used.
- M7.8: Number of students passing the final test and/or with high marks when the emotional/affective system is not used.

Validation techniques both quantitative and qualitative, have included t-test, questionnaire open interview, analysis of the IWT's reporting.

Said that, in the following section we report the validation results for each school involved in the experimentation.

8.4.1 First School “E.Striano”

The class, composed by 14 students, has been divided into two groups: in particular, 7 students form the experimental group and 7 the control group. The students have spent two days for the delivery of the learning course

8.4.1.1 Experimental group analysis

About the participants to the group, 6 students on 7 were subjected to the emotional test in input in order to estimate the emotional component before to the delivery of the personalized learning experience related to the management of high emergencies through simulation resources.

(1) Startup phase

The output, obtained from this start-up phase, displays a tendency towards the extreme “1” of the emotion’s class for the Emotivity: indeed, 3/6 students denote a combination “Indifference”-“Disinterest”; 1/6 “Disinterest”; 1/6 “Indifference”;1/6 denotes a combination “Indifference”-“Frustration”:

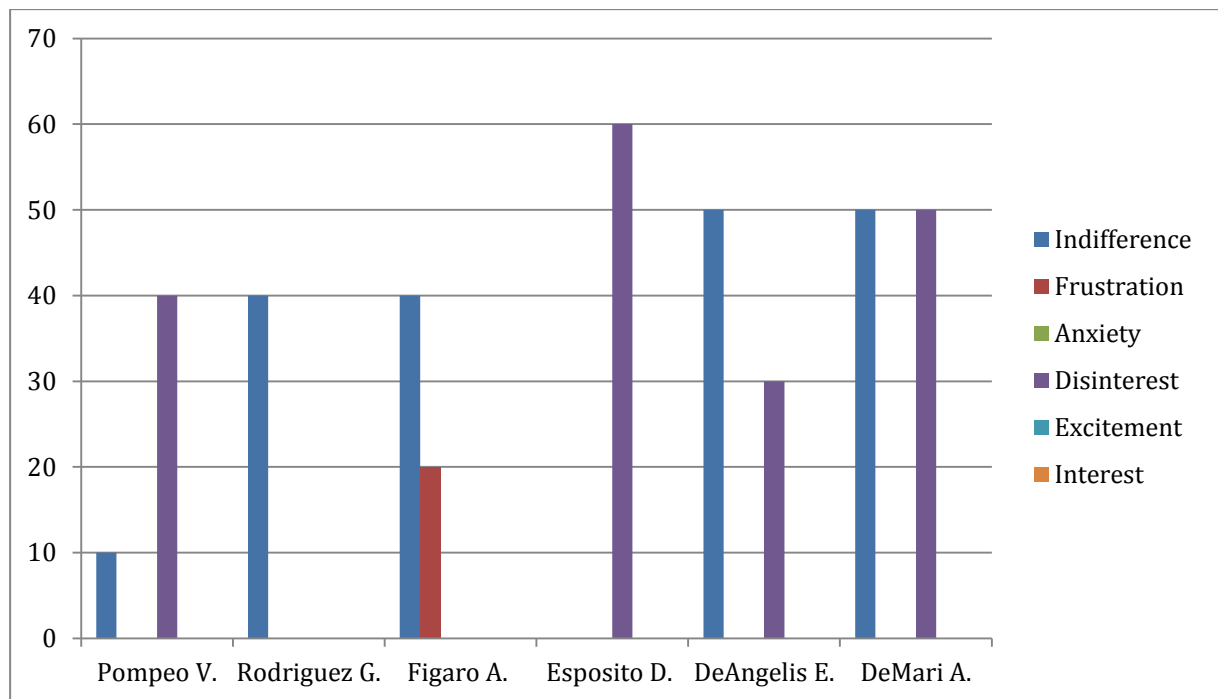


Figure 87: Emotional test results before the delivery of the course

(1) Correlation between resource efficacy, test passed and levels of competency acquired.

In order to analyze the possible impact of the emotional state on the learning path, a correlation between the survey data on the emotional level and the data related to the competence level as well as the time spent by using the learning resources has been made.

In a first day the students were involved in the delivery of the first simulation resource (Serious Game) on the fire management. In a second day they have delivered the storytelling resource.

There were evident correlations between emotional state, level of achieved competence and assessment tests relative to the emergencies themes. From the point of view of skills acquisition on fire by resource Serious Games students who had a combination of emotional input characterized by indifference and disinterest have not reported values below the thresholds are exceeded permitted. The combination of indifference-frustration present in 1 / 6 does not facilitate the predisposition of the student to the acquisitions of the competences. These students have in fact slightly lower results than those with an emotional component unique (only indifference for example).

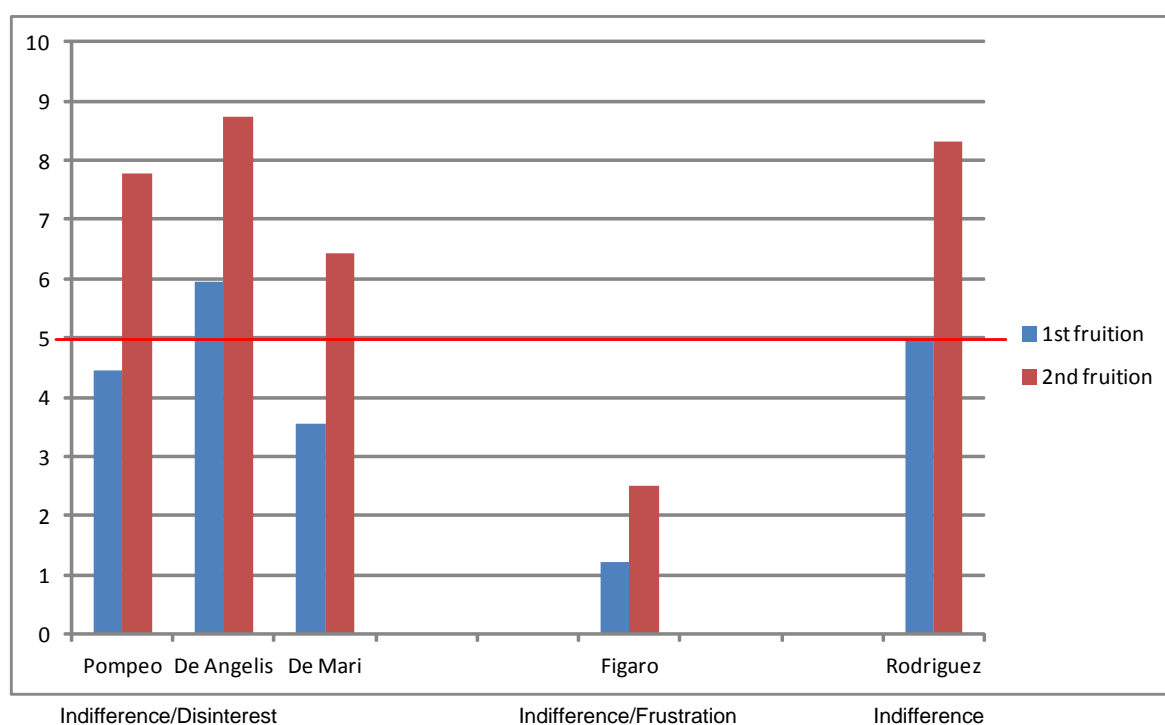


Figure 88: Correlation between emotional data and competence level

(2) Correlation between resource efficiency, use and access.

The frustration component, resulted from the test, has not influenced the number of accesses to the system. The student showing such an emotional state, has maintained levels of use of simulation resources that were above average. In particular, the use of storytelling as a simulation resource has received the most attention and multiple accesses by the students,

who, doing this, have demonstrated their constant interest. Indifference as an emotional component has been detected in 1/6 students and has implied a lower level of anxiety/concern and a total student's disinhibition in accessing the resources, which contribute to achieve some peaks of access for the use of serious game and storytelling. 1 out of 6 students have shown a high disinterest value as well as lower access, permanence and use times. The disinterest value for 3/6 students, even though added to indifference values which were detected together with it, has not negatively impacted. Indeed, for these students not only the permanence time resulted to be on average, but there were numerous accesses to learning resources. This type of resource in fact was used in order to overcome the game, as in the case of serious game, and to follow the story, as in the case of storytelling. It can be observed how the possibility to experience simulation resources has helped trigger a mechanism of a constant improvement, which has led to achieve good levels of knowledge about target concepts.

At the end of this experience, the students have shown some changes in their emotional background attitude. Specifically, 2/6 students showed changes that allow to think that the resources of the experimental course have contributed to increase student's confidence in their own skills and stimulate self-esteem. An anxiety component has also emerged for those students that had shown indifference and frustration at the beginning of the path. This new component should be connected to a condition of waiting for path evaluation and teacher's assessment.

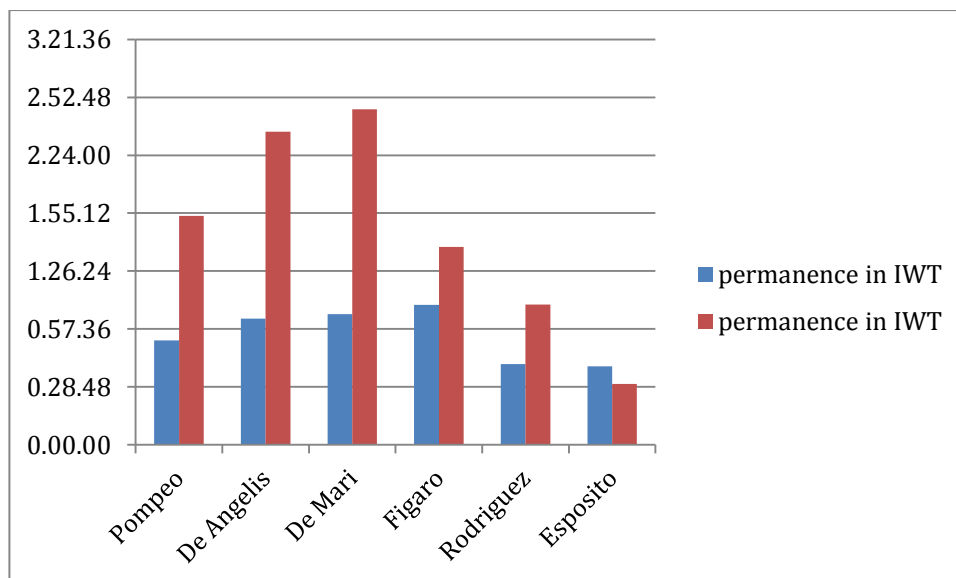


Figure 89: Correlation between emotional data and permanence in IWT

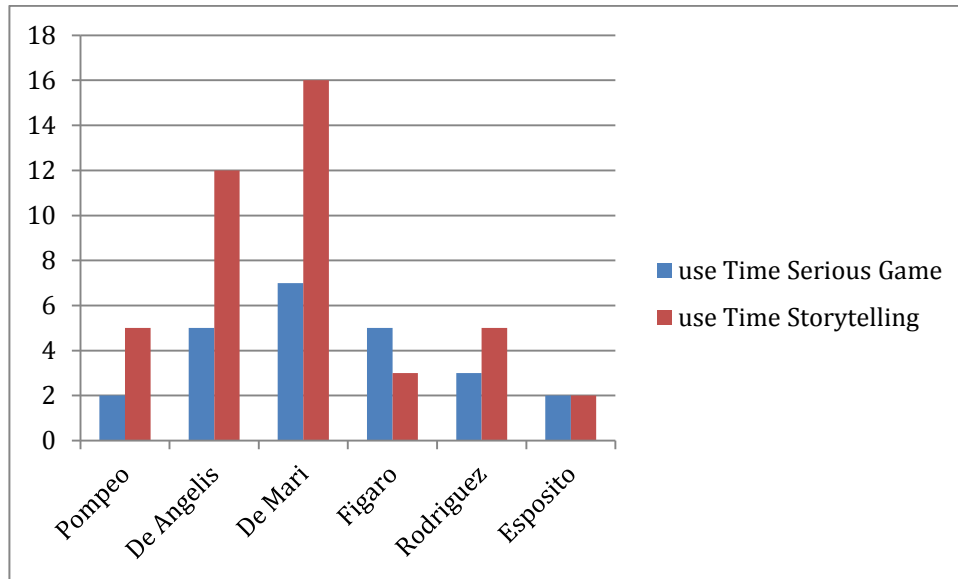


Figure 90: Correlation between emotional data and use times of complex learning resources

8.4.1.2 Control group analysis

The control group was composed of 7 students who had a learning style profile identical to the students included in the experimental group and matched the same generational target (16-17 years old). The emotional states dominant at the early stage fell into the values of frustration and indifference and, unlike what the experimental group presented, here the disinterest component was lower. Only 1 student had an elevated disinterest value combined with an elevated indifference value at the beginning of the experience. These values determined very low times of use and permanence on the resources composing the personalized path, which was characterized by passive LOs.

3 out of the total number of students belonging to the group showed very high levels of frustration while only 1 of them combined this state with indifference. The percentages of use and access have maintained a minimum level compared with the students that had started the experience with high levels of indifference (3/7). These students in fact presented higher permanence and use times, even though their accesses were minimum (indicating passive visualisation of the resources). The tests have been subjected to a quick incursion with the only aim of completing the experience also mechanically. In general, the levels of knowledge achieved in relation to target concepts related to fires and earthquakes are lower than those achieved by the experimental group as well as the attempts to improve the assessment tests performance. This is a symptom of lack of involvement and self-motivation in students to do better.

8.4.2 Second School "Pitagora"

The class, composed by 28 students, has been divided into two groups: in particular, 14 students form the experimental group and 14 the control group. The students have spent two days for the delivery of the learning course.

8.4.2.1 Experimental group analysis

About the participants to the group, all the students were subjected to the emotional test in input in order to estimate the emotional component before to the delivery of the personalized learning experience related to the management of high emergencies through simulation resources.

(1) Startup phase

The output, obtained from this start-up phase, is resumed in the Figure 91:

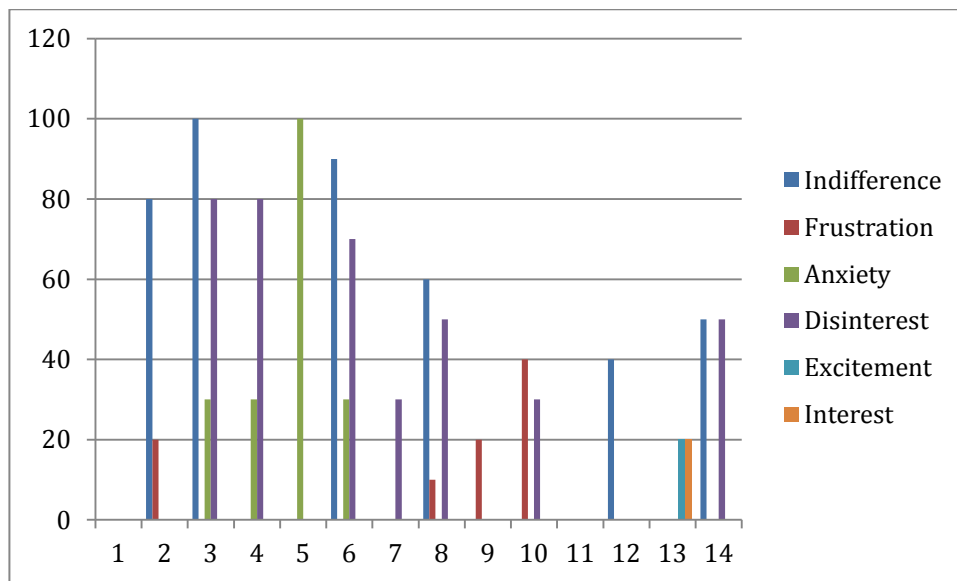


Figure 91: Emotional test results before the delivery of the course

(2) Correlation with emotional data and competence level

In order to analyze the possible impact of the emotional state on the learning path, a correlation between the survey data on the emotional level and the data related to the acquired competence level obtained by delivering the complex learning resources.

In a first day the students were involved in the delivery of the first simulation resource (Serious Game) on the fire management. In a second day they have delivered the storytelling resource.

We can observe how the student that, in the start-up phase is characterized by a disinterest component, has registered the best levels of acquired competences That is also confirmed by the result of the final emotional test that it's changed by disinterest to excited.

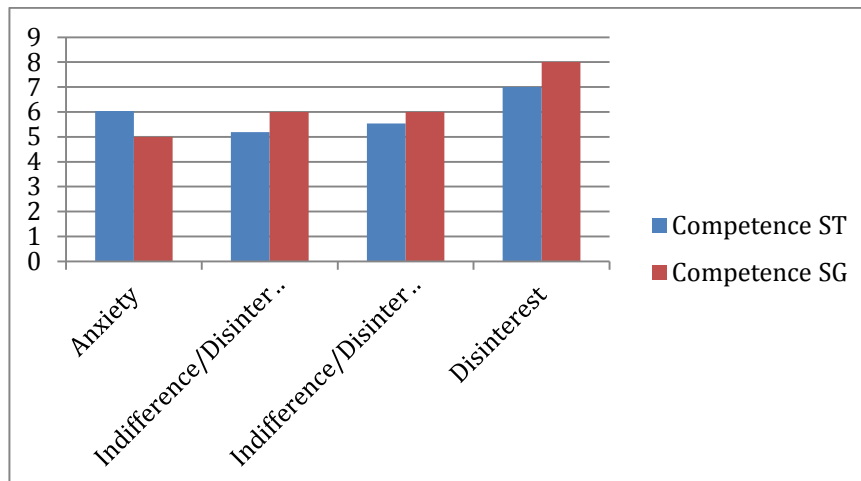


Figure 92: Correlation with emotional data and competence level

8.4.2.2 Control group analysis

The control group, composed by 14 students, have compiled an emotional test ad the end of the didactic experience in order to know their feeling with respect to the sequence of the expositive didactic resources.

The quantitative data have notified a lot of disinterest and indifference. This denotes a lack of satisfaction with the experience and a disinterest with respect to the issues proposed by the passive learning resources. The absence of not altered states denotes a condition of peace to fruition not particularly engaging and not able to stimulate any of the emotions from the axes. In such a way the student remains detached from an emotional point of view and emotional benefited from the learning path.

8.5 Conclusion

Finally the results are summarized and discussed by considering the goals which were determined at the beginning of the study from evaluation and validation point of view. What goals were achieved and what points should be considered in further work?

In order to investigate these aspects, we reported two interviews submitted to the two tutors of the schools involved within the experimentation:

“Do you think that the emotional component provide an added value to a learning course that allow to ameliorate the learning effectiveness?”

Tutor A: According to me, monitoring students’ emotional reactions and state of being is very important to analyze their learning processes.

Tutor B: Very good tool! It is useful both for the teacher and for the same students who develop cognitive skills meta analysis of their emotional state during a learning process.

The qualitative data have out in results a correlation between the emotional state and the acquired competence levels showing that the emotional tool could help the instructional

designer or the teacher to differentiate the learning path taking into account the different learning styles of the students.

The quantitative data have suggested ameliorating the emotional feedback by defining specific contents for the state equilibration of Emotional / Affective aspects. This improvement will be taken into account for the next experimentation phase.

9 R8. Enhanced Wiki-Test and Peer-review for writing assignments

9.1 Research goals and hypotheses

In this scenario, students of a course held at TU Graz were asked to use a co-writing WIKI for collaboratively writing a paper. The performance of the learners had to be assessed by themselves and by their peers. In addition, the learners also assessed the contributions of other groups.

Goals

G8.1: To provide a tool that allows an efficient and user-friendly management.

G8.2: To provide a WIKI system that can be used collaboratively for writing assignments.

G8.3: To identify possible improvements for the tool.

G8.4: To provide a WIKI system with useful actions and contribution graphs in order to enable the students an overview of their learning progress.

G8.5: To provide a peer-assessment that motivates students concerning their learning activity.

G8.6: To provide a feedback out of the peer- and group-assessment that supports the students in their learning process.

G8.7: To provide a tool that facilitates the work for the instructors.

Hypotheses

H8.1: The tool allows an efficient and user-friendly management.

H8.2: Using the tool supports students in working collaboratively.

H8.3: Possible improvements for the tool can be derived from the students' feedback and suggestions concerning its usability.

H8.4: The actions and contribution graphs which are provided in the WIKI system enable the students an overview of their learning progress.

H8.5: The provided peer-assessment motivates the students concerning their learning activity.

H8.6: The feedback provided by the peer- and group-assessment supports the students in their learning process.

H8.7: The tool facilitates the work for the instructors.

9.2 Method

9.2.1 Participants

There were 21 students in the course, 18 out of them participated in the study (3 students did not sign the informed-consent sheet and were therefore excluded from analysis).

Out of this final sample, 15 students were male (83%), 3 students were female (17%) and the participants were on average 26 years old. Regarding the highest level of education, for 6 of the students the “Matura” (= Austrian university entrance diploma) is the highest completed education, 11 students had already reached the Bachelor and one of them finished his Master education.

The students were supervised by three tutors. Each of the tutors was assigned to two groups.

9.2.2 Apparatus and Stimuli

Students used the wiki during a regular course in order to write an essay collaboratively. During the study, the students were asked to fill in two questionnaires. Before they started working with the wiki, they received a Pre-Questionnaire concerning their demographic data, previous experience in group working and working with wiki-tools, and their general attitudes concerning self- and peer-assessment. After the course, they filled out a Post-Questionnaire, which included the following sections: demographic data, experience in the group work, attitudes concerning self- and peer-assessment (based on their experiences), group-assessment, task awareness, usability of the wiki tool, emotional aspects and further comments or suggestions. Besides, we provided a Questionnaire for the tutors concerning the assessment within Co-wiki writing, Co-wiki writing itself, usability of the tool and further comments.

For qualitative statistical analysis, we summarized the open answers in the surveys. For quantitative statistical analysis, we performed t-tests. For statistical analysis, we set the level of significance to $\alpha = .05$. Participants had to sign an informed-consent sheet in order to participate in the study.

Regarding the section “attitudes concerning self- and peer-assessment”, there are four subscales according to [19]:

- The intrinsic motivation scale measures the students’ motivation doing the peer-assessment activity for its own sake, just out of pleasure, e.g. “In a peer-assessment activity I liked opinions from peers because I got more ideas.”
- The extrinsic motivation scale measures the students’ motivation doing the peer-assessment activity in order to get approval from the teacher and a good grade, e.g. “In a peer-assessment activity I think the opinions of my work from teachers were more important than those from peers.”
- The evaluating scale measures the confidence of the students in evaluating the peer’s work, e.g. “In a peer-assessment activity I found the strengths of my peer’s work when I reviewed it.”

- The receiving scale measures how students can handle the peer's-assessment in order to recognize their own weaknesses, e.g. "In a peer-assessment activity I recognized my weakness when I got comments from peers."

For the section "usability of the wiki-tool" in the Post-Questionnaire and the Questionnaire for the tutors, we used the System Usability Scale (SUS) developed by [6] which contains 10 items and a 5 point Likert scale to state the level of agreement or disagreement, for instance "I think that I would like to use this system frequently".

To investigate in which emotional state the students were when they used the wiki tool, we added a section concerning "emotional aspects", which included 12 items. Kay and Loverock [7] developed this scale to measure emotions related to learning new computer software. Research showed that the 12 items are describing four emotions:

- Happiness ("When I used the tool, I felt satisfied/excited/curious.")
- Sadness ("When I used the tool, I felt disheartened/dispirited.")
- Anxiety ("When I used the tool, I felt anxious/insecure/helpless/nervous.")
- Anger ("When I used the tool, I felt irritable/frustrated/angry")

The answer categories in this section are "None of the time", "Some of the time", "Most of the time" or "All of the time".

All Questionnaires contained quantitative as well as qualitative questions, the answer categories varied between yes/no, rating scales or open answers. Regarding the rating scales, for the majority of the quantitative questions we used the 5 point Likert scale, so that students could state their level of agreement or disagreement. The rating scale ranged from "I strongly disagree" (1), "I disagree" (2), "neither/nor" (3) to "I agree" (4), "I strongly agree" (5).

For the group-assessment, we provided assessment rubrics (see Figure 93) with the three categories: literature, content and style. As a rating scale we used 5 stars, in which 1 star is the minimum and means the worst evaluation and 5 stars are the maximum and the best possible evaluation. In the literature section the students are asked to assess the work concerning the following questions:

- Is the literature used for the text relevant? (relevance)
- How is the quality of the literature used in the text? (quality)
- Is the amount of literature used appropriate? (appropriate amount)
- Are the facts/sources presented correctly? (representation of literature/sources)

The content section dealt with the following subcategories:

- Is the content of the text relevant? (relevance)
- Is the topic treated completely? (completeness)
- Is there a common thread and clear line of argumentation in the text? (intelligibility, traceability)
- Is the text good and logical structured? (text structure)

Concerning the style section, we provided the following questions for the students:

- Is the style of writing appropriate and good? (expression)
- Is the outline/format clearly arranged and legibly? (outline/format)
- Is the text free of grammar or spelling mistakes? (grammar/spelling)
- Is the citation of the sources correct? (correct citation)

Literature	Relevance	Quality	Appropriate amount	Representation of literature/sources
	Comment: <input type="text" value="TextComment"/>	Comment: <input type="text"/>	Comment: <input type="text"/>	Comment: <input type="text"/>
	Importance: * <input type="checkbox"/> <input checked="" type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> 60%	Importance: * <input type="checkbox"/> <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> 80%	Importance: * <input type="checkbox"/> <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> 80%	Importance: * <input type="checkbox"/> <input checked="" type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> 60%
Content	Relevance	Completeness	Intelligibility/Traceability	Text structure
	Comment: <input type="text"/>	Comment: <input type="text"/>	Comment: <input type="text"/>	Comment: <input type="text"/>
	Importance: * <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> 100%	Importance: * <input type="checkbox"/> <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> 80%	Importance: * <input type="checkbox"/> <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> 60%	Importance: * <input type="checkbox"/> <input checked="" type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> 40%
Style	Style of writing(expression)	Outline/format	Grammar/spelling	Correct citation
	Comment: <input type="text"/>	Comment: <input type="text"/>	Comment: <input type="text"/>	Comment: <input type="text"/>
	Importance: * <input type="checkbox"/> <input checked="" type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> 60%	Importance: * <input type="checkbox"/> <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> 60%	Importance: * <input type="checkbox"/> <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> 60%	Importance: * <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> 100%

Figure 93: Rubrics used for the self-, peer, and group-assessments

9.2.3 Procedure

In cooperation with a lecturer from the Technical University of Graz, it was possible to run through this study and to test the collaborative WIKI system as part of a course. In the following the individual steps of the experiment are described.

In a first step the six groups were assigned to the topics. The topics were provided by the instructor of the course and each group had to select one topic. Then each group had to provide a short report about their planned activities regarding the tasks. After that the students were asked to answer the Pre-Questionnaire concerning their demographic data, previous experience in group working, and their attitudes to self- and peer-assessment.

After they had finished the Pre-Questionnaire, they started working on the topics using the WIKI-tool. In this second step they had to provide a WIKI-document about the topic they were assigned to.

Each time they logged in to the WIKI, they were asked to review the latest contribution of their peer's first, i.e. they did a peer review within the group by rating the importance of their peer's contribution (peer-assessment activity). Then they were allowed to continue working on the paper.

Each time they stopped working on the paper, they were asked to rate the importance of their own contribution ("How important is your latest contribution in order to reach the final product?")(self-assessment activity).

In the next step, after all groups had finished their papers, each member of a group had to assess the final product of all groups using the group-assessment rubrics, which are shown in Figure 92 (group-assessment activity).

In addition to the students' group-assessment, three tutors and the teacher were asked to assess the groups' contribution using the rubrics of Figure 92 (instructor's assessment activity). Thereafter, feedback was given regarding the results of the group and teacher-assessment.

Finally, the post-questionnaire (see Section 2.2) was sent to the students to gather information on the usability/functionality of the tool, their experiences with the self-, peer- and group-assessments, task awareness, and emotional aspects. In addition the tutors were asked to answer a questionnaire about the assessment within Co-WIKI writing and the usability of the tool.

9.3 Evaluation Results

In this section we focus on students' perception of the WIKI-system itself, whereas the analyses of the tool's impact on student's learning process are reported in Section 9.4 (Validation Results). Thus, we report the evaluation of H8.1, H8.3, and H8.4 as specified in [4].

We analyzed data from 18 participants, but for the Pre-Questionnaire we had to exclude the results of one participant, because he/she did not finish the questionnaire. For each item, we computed the mean and its standard deviation as an exact measure of central tendency. However, in some cases the mean did not allow an interpretation of the data concerning the students' level of agreement or disagreement. Due to some outliers many of the mean values referred to the middle category "neither/nor". Thus, we used the median as additional measure of central tendency to get a better impression of the ratings given by the majority of students. For the mentioned data the median gave a clearer picture of students' level of agreement or disagreement. For that reason, we used the median to interpret the data whenever the mean did not allow a clear interpretation of the data. In these cases, the mean, its standard deviation, and the median are presented. If not noted otherwise, the reported results (means and medians) refer to data from 5-point rating scales.

Note: According to [21] the median is defined as the middle value in a range of scores. As an alternative measure of central tendency, the median is not so sensitive to extreme values. So if there are outliers with extreme scores (like in our case), it is recommended to analyze the median, because in this context it is a much more representative value than the arithmetic average.

9.3.1 Usability of the WIKI-tool

To evaluate student's satisfaction with the tool regarding an efficient and user-friendly management (H8.1), we analyzed students' ratings and open comments on the usability/functionality of the tool.

To investigate the overall usability of the wiki-tool, we used the SUS (see Section 2.2). As by error one out of the 10 items was not provided in the questionnaire, we computed a score for this item by averaging the scores of 4 other items with the same polarity. After calculating the SUS score for each student, we got an average SUS score of 48.53. SUS scores have a range of 0 to 100 with an average score of 68, obtained from 500 studies. A score above an 80.3 is considered an A (the top 10% of scores). Scoring at the mean score of 68 gets you a C and anything below a 51 is an F (putting you in the bottom 15%).

Regarding the wiki-tool in general, the students mentioned that they liked the idea of it. Specifically the students were in favor of the overview page and the possibility to see who of their colleagues did what and to follow the progress. In line with these results the students also stated that they did not have to learn a lot of things before they started working with the system ($M = 2.38$, $SD = 0.96$, $Md = 2$). In their opinion most people would learn to use the system very quickly ($M = 3.38$, $SD = 0.89$, $Md = 4$). However, our SUS score of 48.53 belongs to the bottom 15%. For this low value we assume several reasons. The students complained about the following problems they faced with the wiki.

Almost all participants stated that the system was very slow and that this fact prevented them from working effectively. In addition to that the students mentioned that the star-rating was not working most of the time and that in general a lot of server errors occurred. In this context a student also commented that it was difficult for him/her to find the “colored comparison tool” and another one mentioned that he/she missed an index or a table of contents for the main page and a management for footnotes and references.

In accordance with these statements, the students indicated that they would not use the system frequently ($M = 2.06$, $SD = 0.85$, $Md = 2$) and that they found the system unnecessarily complex ($M = 3.5$, $SD = 0.89$, $Md = 4$). In addition, the students mentioned that the system was not easy to use ($M = 2.56$, $SD = 1.03$, $Md = 2$) and that even if there would be a technical person supporting them it would not be easier to use ($M = 2.25$, $SD = 1$, $Md = 2$).

With regard to the required self- and peer-assessments, students suggested to keep the ratings of importance and comments optional or to remove the rating of importance. A participant proposed to include an option “minor-change” where no rating is required. Another student mentioned that for further work the collision handling when two persons work on the same page should be improved and that an auto generated table of contents and references should be provided. Thus students’ comments give many hints for possible improvements of the tool (see H8.3)

9.3.2 Task Awareness

With regard to H8.4, we found that the actions feed in the assignment homepage supported the students in tracking the activities of their peers effectively ($M = 3.56$, $SD = 0.73$, $Md = 4$). Concerning the contribution graphs in the assignment homepage, the students were aware of who of their colleagues had contributed to the task ($M = 3.25$, $SD = 1.13$, $Md = 4$) and to which amount he/she did this ($M = 2.94$, $SD = 1.29$, $Md = 4$). In addition, the students

reported that the contribution graphs in the assignment homepage gave them a good overview about the progress of the other groups ($M = 3.13$, $SD = 1.15$, $Md = 4$).

9.3.3 Emotional Aspects

Regarding the students emotions during working with the wiki, the results from a 4-point rating scale showed that the students felt more often anger than happiness ($t(15) = 3.25$, $p < .05$), sadness ($t(15) = 3.46$, $p < .05$) or anxiety ($t(15) = 6.46$, $p < .05$). The students stated that some of the time, they felt anger during working with the wiki ($M = 2.44$, $SD = 0.77$). So the students were frustrated, angry, and irritable when they worked with the wiki. These results are in line with the results presented below concerning the low valuation of usability of the wiki and the students' decreasing intrinsic motivation during the course. As already discussed above, it can be assumed that the students felt anger when they faced technical problems with the wiki, because they got a grade for their performance on the wiki. So they did not just explore the wiki, they were dependent on the functionality/usability of the system. As a result, when the students faced any problems, they were kind of frustrated and angry due to their objective to get a good grade.

9.4 Validation Results

In this section we show the validation methodology that includes the following validation criteria and metric extrapolated by [4].

Validation criteria

- C8.2: To evaluate students' experiences regarding working collaboratively by using the tool.
- C8.5: To evaluate the potential change in students' motivation when using the tool.
- C8.6: To evaluate the level of satisfaction of the students with the tool regarding self-, peer, and group assessment activities.
- C8.7: To evaluate the level of satisfaction of the tutors with the tool regarding its usability.

Validation metrics

- M8.2: Ratings of students' extrinsic motivation before/after using the tool.
- M8.3: Ratings of students' intrinsic motivation before/after using the tool.
- M8.4: Ratings of students' experiences regarding receiving feedback.
- M8.5: Ratings of students' self-assessment activities.
- M8.6: Ratings of students' peer-assessment activities.
- M8.7: Ratings of students' group-assessment activities.
- M8.8: Ratings of tutors' satisfaction with the tool.

Following this methodology we will validate the attitudes and experiences concerning peer-assessment, especially whether the WIKI-tool supports student's in working collaboratively (H8.2), and whether it supports student's learning progress (H8.6). Furthermore, students' motivation concerning the learning activity (H8.5) is validated.

As discussed at the beginning of Section 9.3, some validation results are interpreted by referring to the median instead of the mean in order to indicate the students' level of agreement or disagreement. In these cases, the mean, its standard deviation, and the median are presented in brackets.

9.4.1 Attitudes and experiences concerning peer-assessment

By doing the Group-Internal Peer Review, the students agreed that they could feed back the weaknesses ($M = 3.5$, $SD = 0.97$, $Md = 4$) and strengths ($M = 3.75$, $SD = 0.58$, $Md = 4$) of their peers' work. In addition, the comments from their peers supported them in recognizing their own weaknesses ($M = 3.5$, $SD = 0.73$, $Md = 4$), so that they could better examine the problems in their work ($M = 3.19$, $SD = 1.11$, $Md = 4$). In this context the students also stated that they liked to know what the others did, so that they could compare their own work with that of others. Hence, the Wiki-tool supported their learning process and group awareness (see H8.6).

As far as working collaboratively is concerned (H8.2), the students were asked if the internal peer-review allowed them to comment on their peer's contribution. They neither agreed nor disagreed on that ($M = 2.88$, $SD = 1.09$, $Md = 3$). Besides, the students disagreed on the statement that the internal peer-review allowed them to rate the importance of their peer's contribution ($M = 2.56$, $SD = 1.09$, $Md = 2.5$). This result could be explained by the fact that the students didn't like reviewing the work after each change. Even if somebody just added a single word or changed the style, they were asked to review the changes. One participant stated that in such a case of reviewing, the students are possibly just ticking the boxes to be done with it. So it can be assumed that the students were annoyed by commenting and rating the importance of their peer's contribution after each change.

So on the one hand the Group-Internal Peer Review supported the students in recognizing weaknesses and strengths of their peers and the weaknesses of their own work. On the other hand the students didn't like the fact that they were asked to rate the importance of their peer's contribution after each change.

9.4.1.1 Motivational Aspects

Regarding students' motivation as far as the peer-assessment activity is concerned (H8.5), we checked if there is a difference between their intrinsic and extrinsic motivation. To investigate if their motivation changed during the course, we also compared the results of the Pre-Questionnaire (general attitudes concerning peer-assessment) with the findings of the Post-Questionnaire (experiences with peer-assessment during the study).

Analyzing the mean ratings in the Pre-Questionnaire, we got an intrinsic motivation of $M = 3.53$ ($SD = 0.39$) and an extrinsic motivation of 3.04 ($SD = 0.36$), see Figure 94. A t -test

revealed a significant difference ($t(15) = 3.73, p < .05$), thus the students were more intrinsically than extrinsically motivated before the course started. This finding can be explained by the fact that the course was not mandatory for 15 out of 18 students. So almost all students were intrinsically motivated at the beginning of the course and participated out of pleasure. Analyzing the mean ratings in the Post-Questionnaire, we got an intrinsic motivation of 3.18 ($SD = 0.49$) and an extrinsic motivation of 2.88 ($SD = 0.46$). The t-test was not significant ($t(14) = 1.96, p = .07$), thus after the course students' intrinsic and extrinsic motivation was equally high.

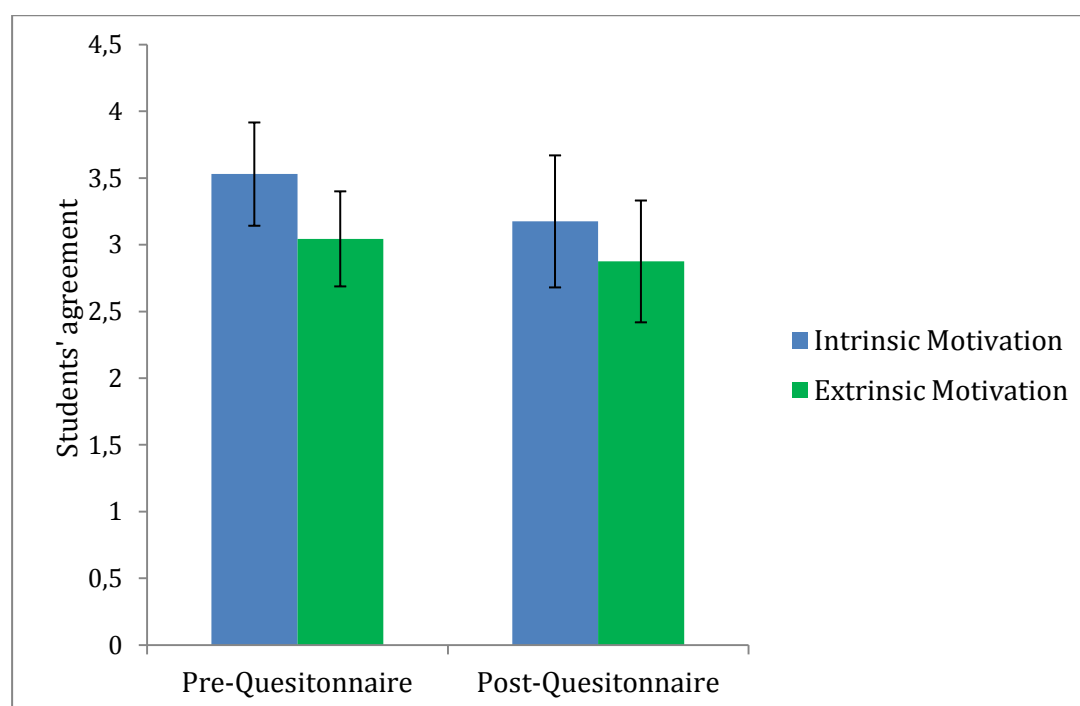


Figure 94: Intrinsic vs. Extrinsic Motivation

Because of the p-value of .07 however, we assume that there was still a tendency that their intrinsic motivation was higher than their extrinsic motivation after assessing their peers. They agreed for instance that the peer-assessment supported them in discussing ideas ($M = 3.69, SD = 0.79, Md = 4$) and sharing opinions with peers ($M = 3.31, SD = 0.95, Md = 4$), which describes their intrinsic motivation. In contrast, the students denied that they only expected to get comments or suggestions back from the teachers when they finished their peer-assessment assignment ($M = 2.81, SD = 0.98, Md = 2.5$), what would concern their extrinsic motivation.

In accordance with the findings above, comparing the Pre-Questionnaire to the results of the Post-Questionnaire, the students' intrinsic motivation decreased ($t(31) = 2.30, p < .05$), whereas their extrinsic motivation did not change during working with the Co-writing wiki ($t(31) = 1.19, p > .05$).

9.4.1.2 Receiving Feedback

It can be assumed that the decrease of their intrinsic motivation is associated with the difference ($t(31) = 2.67, p < .05$) we found between the results of the Pre-Questionnaire and the Post-Questionnaire concerning the term of receiving feedback (H8.6). Before the course started the students stated a higher level of agreement ($M = 3.71, SD = 0.36$) concerning this term than afterwards ($M = 3.22, SD = 0.66$). In the Pre-Questionnaire for instance the students strongly agreed that comments from peers would help them to recognize their weaknesses ($M = 4.06, SD = 0.24$) and examine the problems in their own work ($M = 4.06, SD = 0.43$). In the Post-Questionnaire however, their experiences lead to a decrease of their level of agreement regarding these questions ($M = 3.50, SD = 0.73$ for recognizing weaknesses and $M = 3.19, SD = 1.11$ for examining problems).

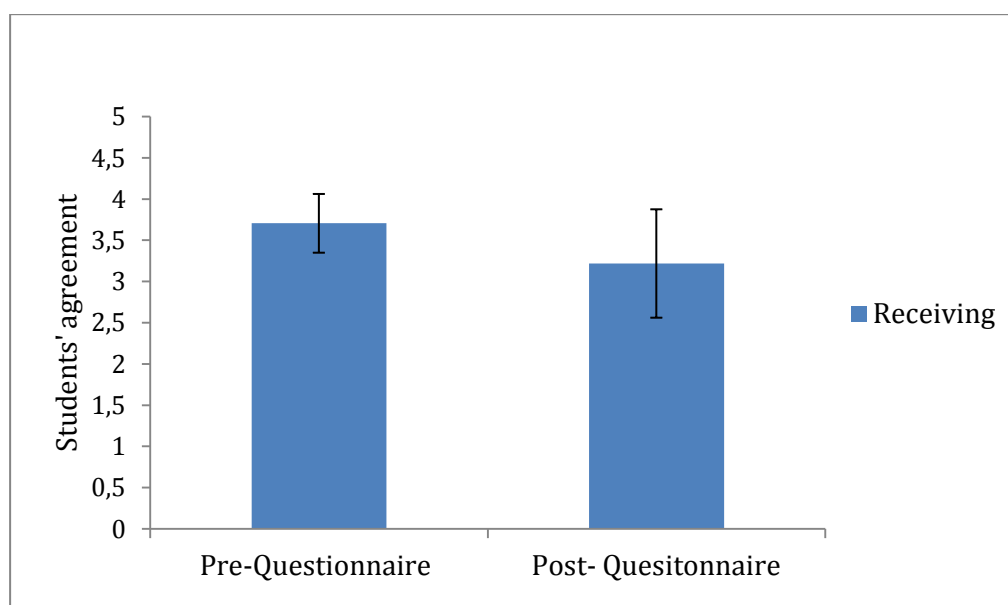


Figure 95: Receiving Feedback

9.4.1.3 Evaluating

Regarding the term of evaluating, we measured the confidence of the students in evaluating their peer's work. Although the mean value of the evaluating scale was higher in the Pre-Questionnaire ($M = 3.71, SD = 0.39$) than in the Post-Questionnaire ($M = 3.42, SD = 0.55$), the t-test showed that there was no significant difference ($t(31) = 1.73, p > .05$) between students' attitudes and their experiences regarding the evaluation of their peers' work. So it seems that students' confidence in evaluating their peer's work did not change during the course.

Summarized, the students were less intrinsically motivated after the course, because their experiences with receiving feedback during the course were worse than the expectations,

they had before the course. It seems that their experience with the WIKI system has reduced their motivation concerning the system. Regarding other results from the questionnaires, an important reason for this is probably that the students faced a lot of problems with the system (see Section 3.1). In addition it has to be considered that the students got grades for their work and hence problems concerning the system will also influence their mood negatively, which is in line with the results of emotional aspects (see Section 4.4 Emotional Aspects).

9.4.2 Group-Assessment

After the students had finished their paper they were asked to evaluate the papers of the other groups. For this group-peer review, the provided assessment rubric supported the students in reviewing the product of other groups ($M = 3.31$, $SD = 0.95$, $Md = 4$) and to learn more about other groups' topics ($M = 3.38$, $SD = 0.89$, $Md = 4$). However the students neither agreed nor disagreed on the statement, that the provided assessment rubric was easy to use ($M = 2.94$, $SD = 0.77$, $Md = 3$).

Regarding the group-assessment the students stated that they benefited from comparing their work with other groups (see H8.6), because they saw other approaches of writing a paper. A participant said that comparing papers was kind of motivational for him/her. In contrast, some of the students mentioned that there were too many rubrics for evaluating and some of them were unclear. In addition, the students were not interested at all in reading other groups' work.

In general the students were in favor of the group-assessment. For most of them the assessment rubric was appropriate in order to review other groups. They really benefited from comparing their papers and became motivated. Some of them, however, showed less motivation and were not interested in the group-assessment. These results can be attributed to the fact that students' intrinsic motivation decreased during the course and that they were not really satisfied with the feedback they received. So it might be that this also affected their attitudes regarding the group-assessment.

9.4.3 Tutor's assessment

We also ask the three tutors (and the instructor) of the course to assess the groups using the rubric and fill in a questionnaire regarding their experience with the tool (see H8.7). Comparing the results from the group-assessment with the tutor's assessment would have allowed us to investigate the quality of the group-assessment. However, only one tutor completed these tasks by assessing the contribution of the two groups he was assigned to. As these results cannot be generalized to all groups, we do not report them here.

9.5 Conclusion

In this Section, the results are summarized and discussed by considering the goals which were determined at the beginning of the study (see also Section 1 of this Chapter). What goals were achieved and what points should be considered for our further work?

Although the students liked the idea of the Co-writing wiki in general and almost all the features, e.g. monitoring the progress, they faced a lot of technical problems (see G8.1).

However, in this context it should be considered that an objective of the study was to find out the weaknesses of the wiki in order to improve the system (G8.3). Hence, the features provided on the wiki are supportive for students to work collaboratively (G8.2). However, the system still needs some technical improvements in order to support the students effectively.

The actions feed in the assignment homepage and the contribution graphs, which were provided on the Co-writing wiki turned out to be very useful for the students (G8.4). Thanks to these tools, the students could track the activities of their peers and knew who and to which amount their colleagues had contributed to the task. As a result, they had always a good overview about the progress of the group.

Regarding their motivation (G8.5), the students' intrinsic motivation decreased during the course. We assume that there are different reasons for that. First, their experiences with receiving feedback during the course were worse than their expectations, they had before the course. Second, the students faced a lot of problems with the Co-writing wiki, so it seems that their experience with the WIKI system has also reduced their motivation. Third, it has to be considered that the students got grades for their work and hence problems concerning the system will also influence their mood negatively, which is in line with the results of emotional aspects.

The feedback out of the peer-assessment supported the students in recognizing weaknesses of their own work (G8.6), so that they could examine eventual problems and improve their work. In addition, the students felt confident to feed back weaknesses and strengths of their peers' work. The students were also in favor of knowing what the others did, so that they could compare their own work with the others. Hence, the Wiki-tool supported them in group awareness. According to the group-assessment, the students really benefited from comparing their papers and became motivated. For most of them the assessment rubric was also appropriate in order to review other groups.

Unfortunately we could not investigate whether the tool facilitates the work for the instructors (G8.7), because the tutors did not fill in the provided questionnaire regarding their experiences with the tool. So this point should be considered in further work.

10 R9. Assessment in Self-Regulated Learning

10.1 Research goals and hypotheses

In general, the aim of this scenario is to investigate the quality of an automatic question creation tool (AQC) that should support students during self-regulated learning.

In order to investigate the research questions, we conducted a **pre-study** in which we evaluated the questions automatically created by the question creation tool (AQC) and a **main study** in which students had the possibility to use the AQC during self-directed learning.

In particular, in the pre-study we compared automatically created questions with manually created questions with respect to several evaluation criteria. Furthermore, we were also interested in the relevance of the concepts extracted by the AQC as the automatically generated questions base on these concepts. The pre-study was divided in two parts: For **pre-study R9-0a**, we asked students in a regular course to evaluate concepts and questions, which were extracted or generated either manually (by human) or automatically (by the AQC). In **pre-study R9-0b**, more experienced students (PhD-Students) were asked to evaluate automatically and manually created concepts and questions. In addition, the questions generated by the AQC for pre-study 0b based on concepts provided by the students in pre-study A. Furthermore, the pedagogical quality of the questions was investigated by categorizing the questions according to Bloom's Taxonomy (Bloom, 1956) which divides the automatically created questions into lower-level (asking for the knowledge and comprehension of a topic) and higher-level questions (asking for deeper understanding of a topic).

In the **main study**, we investigated whether automatically created questions support students in self-regulated learning. The students participated in an online course about "Scientific Working". First, they were asked to study two articles from a provided course material. During reading the articles the students could test themselves with questions provided by the AQC. Then the students were asked to write essays about these articles. After that they received automatically created questions as part of a stage test. Finally they had to collaboratively plan a study. During the course the students used the co-writing wiki for collaboratively working on the essays and the planning of the study.

Through the conduct of these studies we wanted to investigate the goals and hypotheses, which are presented below. In the pre-study we concentrated especially on the goals and hypotheses 9.1 through 9.5, whereas in the main study we focused especially on 9.6 through 9.9. Because the pre-study deals mainly with technical aspects of the tool, the result section comprises only the evaluation of data, whereas the main study includes evaluation as well as validation results.

Goals

G9.1: To provide a tool that generates different types of questions (namely open ended

questions, fill-in-the-blank questions, multiple choice questions and true/false questions) from a text.

G9.2: To ensure that all types of questions provided from the automatic question creator are high in quality.

G9.3: To ensure that the answers provided by the tool are relevant and meaningful.

G9.4: To ensure that the concepts automatically extracted by the tool from a given text are relevant.

G9.5: To provide a tool that creates questions using concepts entered by users.

G9.6: To ensure that the tool is user-friendly.

G9.7: To identify possible improvements for the tool.

G9.8: To provide a tool that motivates students concerning their learning activities.

G9.9: To provide a new form of assessment where automatic question generation is used to create assessments for self-regulated learning style.

Hypotheses

H9.1: The tool generates four types of questions (namely open ended questions, fill-in-the-blank questions, multiple choice questions and true/false questions) from a given text.

H9.2: All types of questions generated from the tool are as high in quality as questions generated by humans.

H9.3: Answers to the questions provided from the tool are relevant.

H9.4: Concepts extracted from the tool are as relevant as concepts extracted by humans.

H9.5: The tool is not only able to generate questions from concepts extracted automatically from a text but also from concepts that are entered by users.

H9.6: The use of the tool is easy even if the user is a non-expert.

H8.7: Possible improvements for the tool can be derived from the students' feedback and suggestions concerning its usability.

H9.8: Using the tool has a positive impact on the users' motivation concerning their learning activities.

H9.9: Using the tool supports students' self-regulated learning; i.e., students benefit from the tool during their learning process.

10.2 Pre-study R9-0a: Evaluation of the automatically created questions

10.2.1 Method

10.2.1.1 Participants

29 participants took part in pre-study R9-0a (4 female, 25 male). They were 25.4 years on average ($SD = 3.3$), ranging from 22 to 39 years. Most of them (93.1%) had a bachelor degree; the rest already had a master degree. The experiment took place within the course “Information Research and Retrieval” at Graz University of Technology. Students were asked to attend a learning activity during the course. Results from the tests delivered during the experiment (see Stimuli) were part of the final grading for the course but note that the participation in the experiment was not a prerequisite for successfully completing the course. Because of the restricted number of computer work places, the participants were divided in two groups that were tested on two consecutive days. All participants gave informed consent before attending the experiment.

10.2.1.2 Apparatus and Stimuli

We used the Automatic Question Creator (AQC) to create questions from a learning content. The learning content was about Natural Language Processing (NLP) and had approximately 2,600 words. It was taken (with slight changes) from Wikipedia (http://en.wikipedia.org/wiki/Natural_language_processing).

The AQC generated questions from the learning content as described in the following (see [15] and [16] or the Deliverable D5.2.1 for a detailed description of the AQC): First of all, the AQC extracted 49 main concepts from the learning content. Example concepts are e.g., “natural language processing”; “modern NLP algorithms”, and “the Georgetown experiment”, respectively. These concepts were automatically ranked regarding their relevance (i.e., the first concept extracted was the most relevant etc.) Afterwards, for each of these concepts, four types of questions were generated (see Table 11).

Question type	Example
Open-ended questions (free text answer)	What do you know about Modern NLP algorithms in the context of Natural language processing? Calculated Region of the Answer: (...)
Fill-in-the-blank questions (one word is missing in a statement)	NLP has significant overlap with the field of computational linguistics, and is often considered a sub - field of artificial intelligence. _____ are grounded in machine learning, especially statistical machine learning.(...) Answer: modern NLP algorithms.
True/false questions (is the statement correct or	Old style NLP algorithms are grounded in machine

incorrect)	learning, especially statistical machine learning. Answer: False
Multiple choice questions (one word is missing in a statement; one correct answer and four distractors are provided).	NLP has significant overlap with the field of computational linguistics, and is often considered a sub - field of artificial intelligence. _____ are grounded in machine learning, especially statistical machine learning. (...) Answer 1: metarule nlp algorithms Answer 2: algorithmic program nlp algorithms Answer 3: modern nlp algorithms Answer 4: heuristic nlp algorithms

Table 11: Examples for different question types

This resulted in 196 questions in total (49 concepts x 4 question types). However, in order to reduce the time effort for the students, only the 20 most relevant questions (as extracted by the AQC) for each question type were evaluated during the study (but all 49 concepts). In order to investigate the quality of the concepts and questions provided by the AQC in more detail, we also added seven concepts and six questions for each question type that were extracted by human (note however that these questions did not base on the concepts provided by the AQC).

In order to collect the data, students had to fill in several surveys using Limesurvey. In the pre-study, the content of the questionnaires were as follows:

Questionnaire1 (Q1): In this questionnaire we asked for demographic data and participants' self-assessment about their English skills. Also pre-knowledge of the topic (NLP) was retrieved. Q1 ended with a short test in which students were asked to summarize the text.

Questionnaire2 (Q2): Q2 included tasks of Learning activity 1(see below). Students were asked to extract the main concepts from the text and to generate two questions per question type.

Questionnaire3 (Q3): Students had to fill in a test. Eight questions were presented. Four of them were based on the AQC, four were generated by human.

Questionnaire 4 (Q4): Q4 included tasks of Learning activity 2(see below). Students were asked to evaluate 56 concepts regarding their relevance. From the 56 concepts, 49 had been extracted by the AQC and 7 by human. Furthermore, they had to evaluate 24 questions, 16 generated by the AQC, 8 by human. The AQC generated four questions for each question type on the basis of the four most relevant concepts. From the eight manually generated questions (two per question type), four were considered to be "good" questions and four were considered to be "bad" questions in at least one of the criteria described in the following. They had to evaluate the questions regarding their pertinence (i.e., relevancy of the question with respect to the topic), level (i.e. is the question trivial or does it express a

significant meaning), and terminology (appropriateness of the words chosen [14]). In addition, when an answer was provided, they had also to evaluate the quality of the answer and the distractors, respectively. These “bad” questions served as control in order to ensure that students work on the task accurately. Note that post-hoc analysis showed that all but open-ended questions were indeed evaluated worse than the manually created “good” questions.

Students were asked to evaluate concepts and questions using a 5-point Likert scale (5 = very relevant/very good; 1 = not relevant at all/very bad).

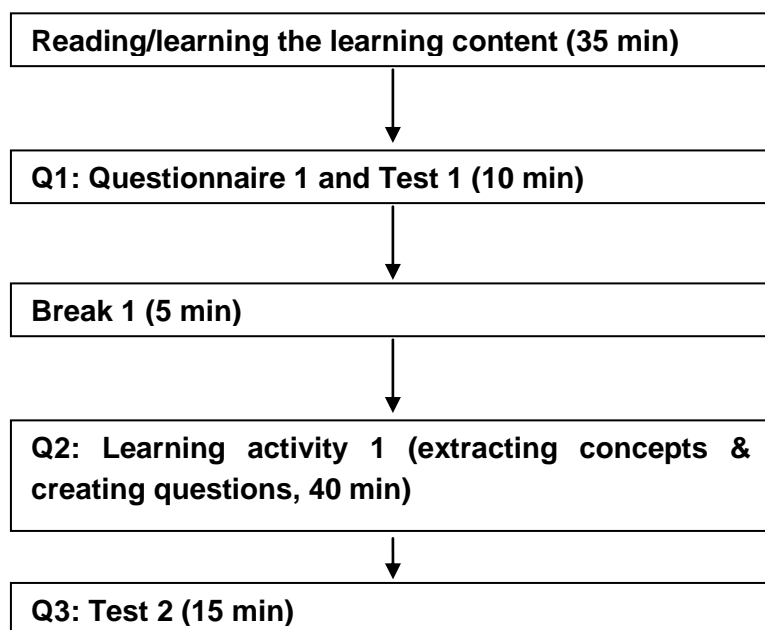
The last part of the questionnaire included general questions about the task itself (e.g., how difficult it was to generate questions; whether there were any problems during the tasks...)

Questionnaire 5 (Q5): After the session, students were asked to fill in this questionnaire as homework. It included 16 questions per question type generated from AQC and 4 questions per question type generated by human.

10.2.1.3 Procedure

Participants were informed that they have to attend several learning activities during the session. To collect the data, we provided the five questionnaires/surveys described above. The learning content, the questionnaires and also the instructions were presented as an online resource.

Figure 96 gives an overview of the course of the pre-study. At the beginning of the study, participants were asked to learn a text about “Natural Language Processing” for 35 minutes. Although most of the students were German-speaking, the learning content and all questionnaires were presented in English. Participants were also asked to provide answers in English. Then participants had to fill in Q1, which also included a short test (10 minutes) in which they had to summarize the previously learnt learning content without consulting it.



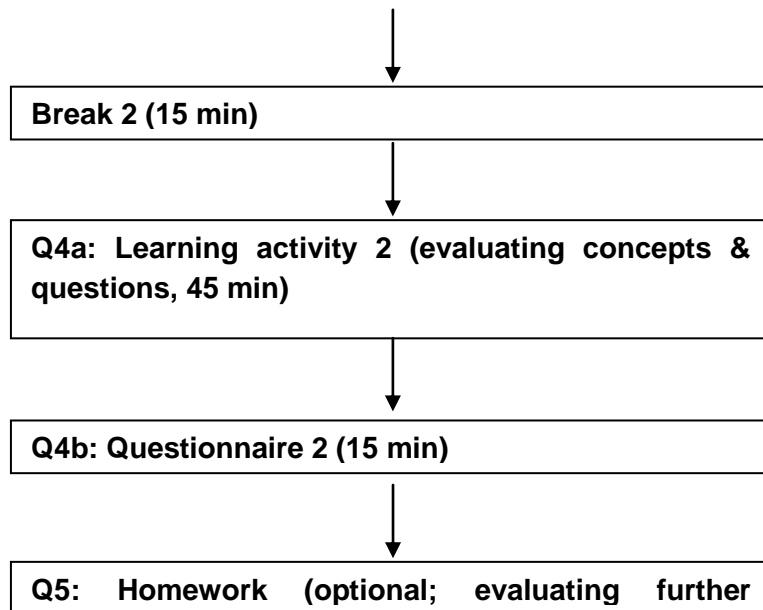


Figure 96: Procedure of pre-studies R9-0a and R9-0b

After a short break (5 minutes), Learning activity 1 started (Q2; 40 minutes). In this learning activity, students were asked to extract relevant concepts from the learning content and then to create eight questions concerning it; two of each question type as described before (open ended questions, fill-in-the-blank questions, true/false questions, and multiple choice questions, respectively). Example concepts and questions concerning a different topic were provided. Participants were allowed to use the learning content while working on the task.

At the end of Learning activity 1, they had to attend a test again (Q3; 15 min). This test included questions created by the AQC and by human.

After a further break (15 minutes), Learning activity 2 (Q4a) started, which lasted approximately 45 minutes. In learning activity 2 participants had to evaluate the 56 concepts and 24 questions that had been extracted beforehand. See Section 1.2 for a detailed description of the concepts and questions. The order of the concepts and questions to be evaluated was randomized. Participants were not informed that the questions were based on the AQC. Finally, they had to fill in a questionnaire (Q4b) in which they were asked to answer more general questions about the task. This questionnaire lasted about 15 min.

The whole experiment lasted approximately three hours. Students were also asked to evaluate further questions as homework (Q5). Note at this point that all concepts and questions were generated by one of the experimenters and evaluated by another experimenter beforehand.

10.2.2 Evaluation Results

In this section we focus on the quality of questions and concepts generated by the AQC. Thus, we report the evaluation of H9.2, H9.3, and H9.4 as they are specified in [4].

10.2.2.1 Extracting concepts

The 29 students who took part in this learning activity extracted 158 different concepts (491 concepts in total) and 17.1 on average ($SD = 10.3$; ranging from 5 to 41 concepts per student). *Table 12* shows the 10 most relevant concepts extracted by the AQC and the 10 most frequent concepts extracted by the students. As can be clearly seen from this table, there is an overlap of several concepts (e.g., *natural language processing*; *machine learning*, but also *word/text segmentation* and *evaluation*). Hence, it can be assumed that the most relevant concepts that were extracted by the AQC reflect the most important concepts extracted from the participants.

Concept by students (frequency and percentage)	Concepts extracted by the AQC
machine learning (28; 96.5%)	natural Language processing
natural language processing (27; 93.1%)	modern NLP algorithms
part-of-speech tagging (21; 72.4%)	those languages text segmentation
NLP evaluation (19; 65.5%)	the first statistical machine translation systems
parsing (14; 48.3%)	linear algebra and optimization theory
statistical NLP (12; 41.4%)	computer science and linguistics
word segmentation (12; 41.4%)	machine learning
topic segmentation and recognition (12; 41.4%)	the Georgetown experiment
history of NLP (10; 34.5%)	evaluation metrics
Word sense disambiguation (10; 34.5%)	an evaluation step

Table 12: Concepts extracted by students and by the AQC.

When the participants were asked to describe their approach on extracting the concepts at the end of the study, most of the students stated that they extracted the relevant concepts by reading or “scanning” the text another time with the specific aim to find them. The majority used the headlines or nouns of the paragraphs as concepts. Some of them tried to remember the most important content at first and added missing concepts by overlooking the text. Others noted in the first step important phrases and extracted the concepts out of them in the second step.

10.2.2.2 Generating questions

Similar to the concepts, we analyzed the self-assessment of the students on this task. The students used different strategies to generate the various types of questions. But in general

they tried to search for the most important information in the text and generated meaningful questions out of that.

Open ended questions. For the majority of the students it was easy to generate open ended questions ($M = 4.03$, $SD = 0.87$). However, two of the participants mentioned that more knowledge and time would be necessary to generate good open ended questions. The students evaluated their self-generated questions on average as quite easy ($M = 2.76$, $SD = 0.79$).

Fill-in-the-blank. Although most of the students found it easy to generate fill in the blank questions ($M = 3.72$, $SD = 0.92$), some of them noted that it is quite hard to find a sentence that contains an exact and unique word that fits for a blank and is at the same time neither too simple, nor too difficult. They rated their own generated questions on average as quite easy ($M = 2.72$, $SD = 0.80$).

True/false questions. In general generating single choice questions was not difficult for the students ($M = 3.90$, $SD = 0.77$). But some of them mentioned that finding something not too obvious or too irrelevant was quite hard. In addition it's important to determine the level of the difficulty of the question. They evaluated their self-generated questions on average as easy ($M = 2.38$, $SD = 0.86$).

Multiple choice questions. On average for the students it was neither difficult nor easy to generate multiple choice questions ($M = 3.00$, $SD = 1.07$). Students who found it difficult to generate multiple choice questions mentioned the problem of finding appropriate distractors. In particular, similar to fill-in-the-blank questions, it was not easy for them to find words which are neither too simple nor too difficult. They rated their own generated questions on average as quite easy ($M = 2.66$, $SD = 0.81$).

10.2.2.3 Evaluation of concepts

49 concepts extracted by the AQC and 7 concepts extracted by human had to be evaluated (see H9.4) using a 5-point rating scale (1= not relevant at all; 5 = very relevant). Mean ratings for all concepts extracted by the AQC was 2.6 ($SD = 0.4$), for concepts extracted by humans 4.0 ($SD = 0.6$; see Figure 97). Concepts extracted by the AQC were evaluated as less relevant compared to concepts extracted by humans. This difference was reliable ($t(28) = 14.87$; $p < .001$). However, when we only use the seven most relevant concepts provided from AQC, mean ratings for the concepts extracted by the AQC increased to 3.9 ($SD = 0.3$) and were equal to concepts by humans ($t(28) = 1.21$, $p = 0.23$) (see Figure 98).

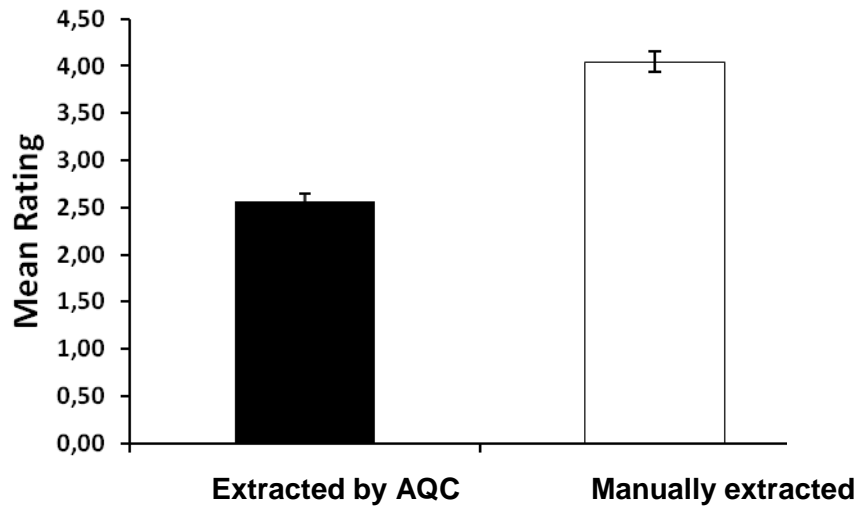


Figure 97: Mean ratings for concepts extracted by the AQC compared to manually extracted concepts. Error bars represent the standard error.

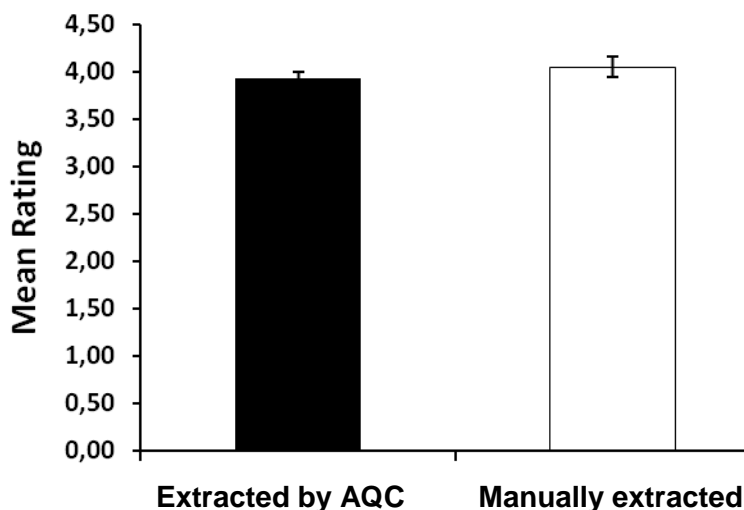


Figure 98: Mean Ratings for the seven most relevant concepts extracted by the AQC and humans. Error bars represent the standard error.

10.2.2.4 Evaluation of questions generated by the AQC

In order to analyze the quality of the questions generated by the AQC (H9.2 and H9.3), we merged data from Learning activity 2 and the homework. Hence, we analyzed 20 questions per question type generated by the AQC. Furthermore, we compared these questions with 5 manually generated questions. Note that data from only 27 participants was included to analysis (because only 27 students finished the homework). In Figure 99, the percentage of answers (averaged across participants) regarding the four evaluation criteria (pertinence,

terminology, level, and answer; see also Section 1.2) for questions generated by the AQC are presented separately for each question type.

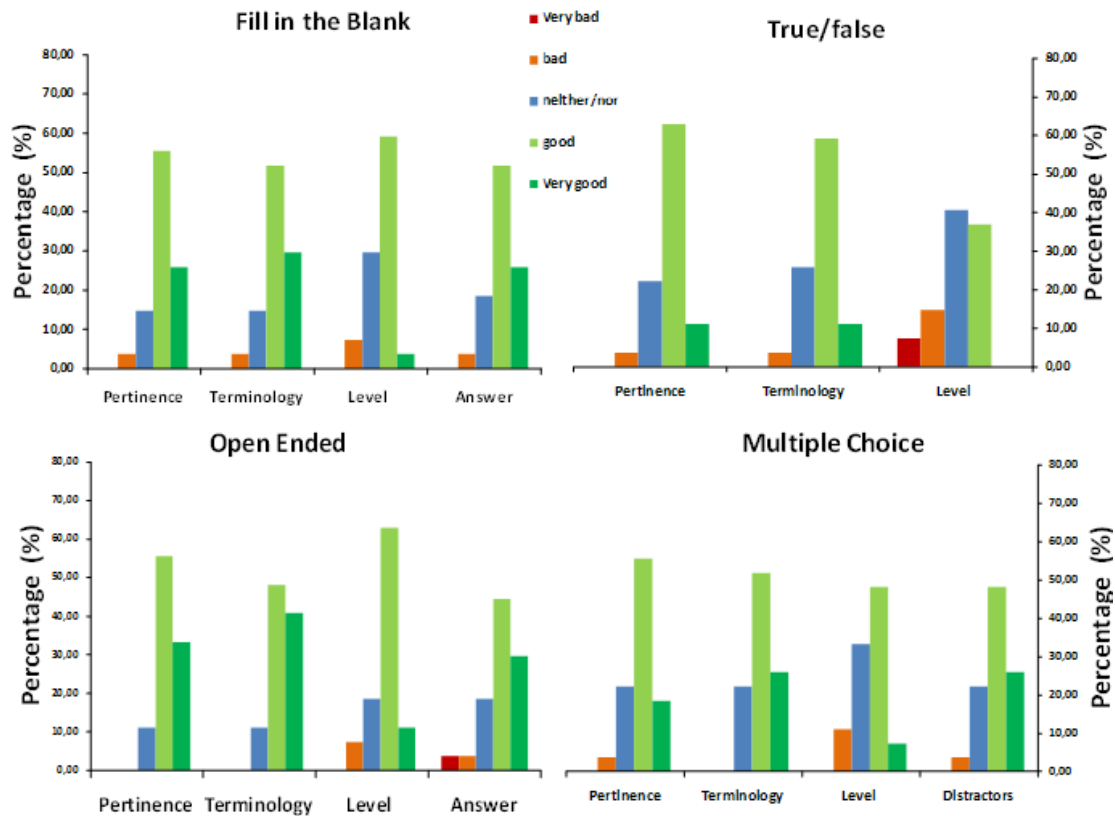


Figure 99: Percentage of answers (averaged across participants) regarding the four evaluation criteria for questions

Table 13 shows the mean ratings (1 = very bad; 5 = very good) for each question type generated by AQC and manually regarding the evaluation criteria (see above).

		Pertinence Mean (SD)	Terminology Mean (SD)	Level Mean (SD)	Answer Mean (SD)	Distractors Mean (SD)
Fill in the blank	AQC	3.4 (0.5)	3.3 (0.6)	3.2 (0.4)	3.5 (0.7)	-
	Manually	4.0 (0.5)	4.1 (0.6)	3.6 (0.5)	4.0 (0.6)	-
True/false	AQC	3.5 (0.4)	3.5 (0.4)	3.1 (0.5)	-	-
	Manually	3.7 (0.5)	3.8 (0.5)	3.2 (0.7)	-	-
Open ended	AQC	3.4 (0.5)	3.3 (0.5)	3.4 (0.5)	3.1 (0.8)	-
	Manually	4.1 (0.4)	4.1 (0.5)	3.7 (0.6)	3.9 (0.7)	-
Multiple choice	AQC	3.4 (0.5)	3.4 (0.6)	3.1 (0.5)	3.4 (0.5)	2.8 (0.5)

	Manually	3.9 (0.5)	4.0 (0.6)	3.5 (0.5)	4.0 (0.6)	3.9 (0.6)
--	-----------------	-----------	-----------	-----------	-----------	-----------

Table 13: Mean ratings for AQC- and manually generated questions for each question type regarding the evaluation criteria. Standard deviations are presented in parenthesis

We compared questions generated by the AQC and manually generated for each question type and each quality criteria using *t*-tests for depended measures. Results showed that fill-in-the blank, open ended and multiple choice questions generated by human were in general better evaluated than questions generated from AQC (all *p*'s < .001, Bonferroni corrected). However, true/false questions did only differ from human generated with respect to terminology (*p* < .01) but not with respect to pertinence or level (all *p*'s > .05). Although comparison between the two conditions should be interpreted with care (because there were less questions generated by human than by the AQC), results nevertheless suggest that the quality of the questions generated by the AQC is less than the quality of questions provided by human.

The level of the questions and the provided answers seem to fulfill the needs of the students. Statistical analysis showed no difference between automatically and manually created questions. However, sometimes the terminology of the questions and the quality of the automatically distractors were worse compared to manually created questions. A closer look at the data suggests that especially for completion exercises and multiple choice questions the terminology was rated worse compared to the terminology of manually created items. It has to be mentioned that, for instance, a completion exercise is created by the AQC using an existing sentence or paragraph of the text, leaving the main concept (= answer) blank. It is possible that students are not familiar with questions constructed in such a way. For example, if students are asked to create completion exercises and multiple-choice questions themselves, they typically construct new sentences. Further experimentation is necessary to investigate this issue in more detail.

Results also showed that the quality of the automatically distractors was worse compared to human created questions. However, automatically creating distractors is still very challenging. For instance, distractors should be as semantically close to the correct answer as possible [17]. Our current approach builds on antonyms and related terms on concept or word level. Improvements could be gained by more carefully choosing distractors. Another alternative for improvements could be the deeply study of the process of distractor creation by subject domains in order to implement a similar process chain in the tool.

10.3 Pre-study R9-0b: Evaluation of the automatically created questions

10.3.1 Method

In pre-study R9-0b, eight participants took part (2 female, 6 male). They were 33.1 years on average (*SD* = 6.6), ranging from 25 to 41 years. Most of them (87.5%) had a master degree;

the rest already had a PhD. All participants gave informed consent before attending the experiment.

Stimuli and Procedure were the same as in the pre-study with the following exceptions: For generating the questions with the AQC we did not only use the concepts provided by the AQC but also concepts that were generated by the participants in the first study (see H9.5). Therefore, we used the 15 most frequently extracted concepts by the students to generate questions with the AQC. However, because it was not possible for the AQC to generate questions for all of these concepts (e.g., it was not possible to generate questions from the concepts “parsing”, “word segmentation”, and “topic segmentation and recognition”), questions were generated for only 10 out of these 15 concepts. In addition, some of the concepts had to be rephrased slightly in order to create questions automatically.

10.3.2 Evaluation Results

In this section we focus on the quality of questions and concepts generated by the AQC. As in pre-study R9-0a, we report the evaluation of H9.2, H9.3, H9.4, and H9.5 with the corresponding criteria and metrics C9.2 through C9.5 as well as M9.1 through M9.11 as they are specified in [4] and outlined in Section 10.2.2. Additionally H9.5 is evaluated. The corresponding criterion and metric is specified as follows:

- C9.5: To evaluate questions generated by the AQC, using concepts created from users.
- M9.11: Ratings for questions when the tool uses human-extracted concepts.

22.2 % of the participants stated that they have previous knowledge about NLP. Most of them (88.8 %) agreed that the text was not difficult to understand and 55.5 % stated that the text gave a good overview about the topic. Thus, it can be assumed that the results are not affected because of the difficulty of the text or insufficient knowledge in English.

10.3.2.1 Extracting concepts

Students extracted 53 different concepts (100 in total) and 12.5 on average (SD = 8.7; ranging from 3 to 24 concepts per student). *Table 14* shows the 10 most frequent concepts extracted by the students in the actual study. We also included the concepts extracted from the students in the first study for comparison. Despite the fact, that there were fewer participants in the actual study, the concepts extracted were quite similar with respect to the first study. In both studies, the most important concepts were “natural language processing” and “machine learning”. Also “NLP evaluation” and “statistical NLP” was mentioned quite frequently in both studies.

Concepts extracted by students in pre-study R9-0b (frequency; percentage)	Concepts extracted by students in pre-study R9-0a (frequency; percentage)
natural language processing (7; 87.5%)	machine learning (28; 96.5%)
machine learning (6; 75.0%)	natural language processing (27; 93.1%)

artificial intelligence (4; 50%)	part-of-speech tagging (21; 72.4%)
linguistics (3; 37.5%)	NLP evaluation (19; 65.5%)
NLP evaluation (3; 37.5%)	parsing (14; 48.3%)
NLP tasks (3; 37.5%)	statistical NLP (12; 41.4%)
Turing test (2; 25%)	word segmentation (12; 41.4%)
hand-written rules (2; 25%)	topic segmentation and recognition (12; 41.4%)
fully automatic translation (2; 25%)	History of NLP (10; 34.5%)
statistical NLP (2; 25%)	Word sense disambiguation (10; 34.5%)

Table 14: Concepts extracted by students in the first and the second study (frequencies/percentages in parenthesis).

10.3.2.2 Analysing the pedagogical quality of the automatically created questions regarding Bloom's taxonomy

As the questions generated by the AQC are expected to support students in self-regulated learning settings, we were interested in the pedagogical quality of the questions. Therefore, we analysed the questions with respect to Bloom's taxonomy. We collapsed the questions from both pre-studies and analysed 290 manually and 120 automatically created test items. Two independent raters categorised the questions with respect to the six levels of Bloom (knowledge, comprehension, application, analysis, synthesis and evaluation). The order of questions was randomised.

Inter-rater agreement was acceptable with $\kappa = 0.71$ [12]. As expected, the vast majority of test items (98.0 %) were categorized as lower-level items (i.e., testing the knowledge, comprehension, or application of the learning content) from both raters. Only a minor amount of the test items (2.0%) were categorized as higher-level items (mostly without accordance of the two raters, however). All of these latter items were created by the students and were categorized as analysis items. Within the lower-level test items, 72.9% were categorized from both raters as knowledge items, and 22.0% as comprehension items. Together, this analysis shows that there was neither a difference in categories with respect to the four different test-item types nor with respect to automatically and manually created items (see H9.2). Interestingly, the main discordance between the two raters was whether single-choice items were rather knowledge or comprehension tasks (see below).

10.3.2.3 Evaluation of concepts

49 concepts extracted by the AQC and 7 concepts extracted by human had to be evaluated using a 5-point rating scale (1= not relevant at all; 5 = very relevant). Mean ratings for all concepts extracted by AQC was 2.5 (SD = 0.5), for concepts extracted by humans 4.1 (SD = 0.3; see Figure 100). The difference between automatically and manually extracted concepts was reliable ($t(7) = 13.08$; $p < .001$). Hence, similar to the first study, concepts extracted by the AQC were less relevant compared to concepts extracted by humans (see H9.4).

When we only investigate the seven most relevant concepts provided from AQC, mean ratings for the concepts extracted by the AQC increased to 3.6 (SD = 0.6; see Figure 101). However, ratings for concepts from AQC were nevertheless worse compared to concepts by humans ($t(7) = 2.86, p < .05$).

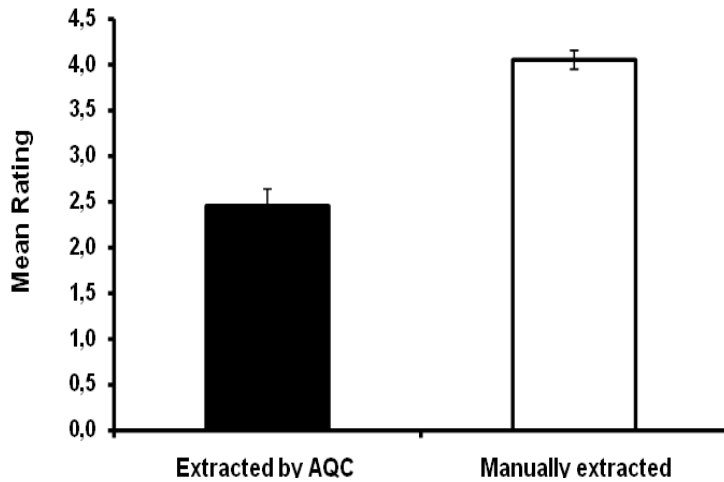


Figure 100: Mean ratings for concepts extracted by the AQC compared and humans. Error bars represent the standard error.

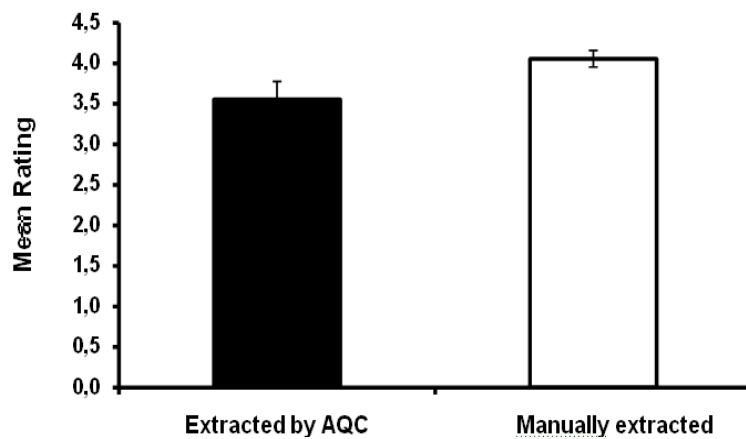


Figure 101: Mean ratings for the seven most relevant concepts extracted by the AQC and humans. Error bars represent the standard error.

10.3.2.4 Evaluation of questions generated by the AQC

In order to analyze the quality of the questions generated by the AQC (H9.2, H9.3, and H9.5), we merged data from Learning activity 2 and the homework. Hence, we analyzed 10 questions for each question type provided by the AQC based on automatically extracted concepts (AQC-a), 10 questions per each question type generated by the AQC using

manually extracted concepts (AQC-m), and 10 manually generated questions per each question type. The evaluation criteria are described in the first study. Table 15 shows mean ratings (1 = very bad; 5 = very good) for each question type regarding the evaluation criteria.

		Pertinence Mean (SD)	Terminology Mean (SD)	Level Mean (SD)	Answer Mean (SD)	Distractors Mean (SD)
Fill in the blank	AQC-a	3.7 (0.7)	3.6 (0.7)	3.4 (0.7)	3.7 (0.7)	-
	AQC-m	3.6 (0.7)	3.6 (0.7)	3.4 (0.6)	3.7 (0.8)	-
	Manually	3.7 (0.9)	3.7 (0.7)	3.6 (0.8)	3.7 (0.7)	-
True/false	AQC-a	3.7 (0.7)	3.8 (0.8)	3.7 (0.7)	-	-
	AQC-m	3.5 (1.0)	3.5 (0.8)	3.4 (0.8)	-	-
	Manually	3.3 (0.8)	3.3 (0.8)	3.2 (0.8)	-	-
Open ended	AQC-a	3.9 (0.7)	3.6 (0.7)	3.9 (0.5)	3.6 (0.6)	-
	AQC-m	3.9 (0.7)	3.8 (0.6)	3.9 (0.6)	3.6 (0.6)	-
	Manually	4.2 (0.7)	4.2 (0.7)	4.0 (0.6)	3.8 (0.6)	-
Multiple choice	AQC-a	3.6 (0.6)	3.6 (0.6)	3.3 (0.7)	3.5 (0.5)	3.1 (0.8)
	AQC-m	3.6 (0.8)	3.5 (0.7)	3.1 (0.7)	3.4 (0.7)	2.9 (0.8)
	Manually	3.8 (0.7)	3.9 (0.6)	3.6 (0.8)	3.9 (0.8)	3.9 (0.7)

Table 15: Mean ratings for AQC- and manually generated questions for each question type regarding the evaluation criteria. Standard deviations are presented in parenthesis

In order to investigate quality differences between the question resource (AQC-a; AQC-m, manually) we computed repeated measures ANOVAs for each question type. For fill in the blank questions, there was no main effect of question resource, $F < 1$. The main effect of evaluation criteria was significant, $F(3, 21) = 4.66, p < .05$. There was no interaction, $F < 1$. Hence, there was no difference between manually and the automatically generated questions. Post-hoc analysis showed, however, that the level of all fill-in-the-blank questions was evaluated slightly worse compared to the other evaluation criteria.

For true/false questions we found a main effect of question resource, $F(2, 14) = 7.78, p < .01$, but no effect of evaluation criteria, $F(2, 14) = 2.94, p = .09$, and no interaction, $F < 1$. Post hoc analysis showed that manually created questions were evaluated even worse compared to both types of automatically created questions.

For open-ended questions we found a main effect of question resource, $F(2, 14) = 5.88, p < .05$, and a main effect of evaluation criteria, $F(1.33, 9.33) = 4.74, p < .05$. There was no interaction, $F = 1$. Post hoc analysis showed that manually created questions were evaluated better compared to automatically created questions that based on human concepts. There

was also a tendency ($p = .09$) that human created question were evaluated better than automatically questions that based on automatically extracted concepts.

For multiple choice questions we found a main effect of question resource, $F(2, 14) = 7.58$, $p < .01$, and a main effect of evaluation criteria, $F(4, 28) = 6.13$, $p < .01$. The interaction was also significant, $F(8, 56) = 2.70$, $p < .05$. Post hoc analysis showed that manually created questions were evaluated better compared to automatically created questions that based on human concepts. There was no such difference between both types of automatically created questions.

10.4R9-1: Investigating the AQC and the co-writing wiki in a complex learning environment

10.4.1 Method

10.4.1.1 Participants

Twelve students participated in the PHD course study, for five of them the course was mandatory. Participants were between 22 and 41 years old, on average they were 32 years old ($SD = 6.53$). Eight of the students are male and four of them are female. Concerning the highest level of education, three students finished their Bachelor, eight of them reached a Master degree and one of the students has already a PHD degree.

Six students finished the entire course, i.e. they did all three phases which are described in detail in Section 1.3. One student almost finished all three phases, he/she just failed to do the group-assessment and to fill in the Post-Questionnaire. Two students participated in Phase 1 and Phase 2 and three students only finished Phase 1.

10.4.1.2 Apparatus and Stimuli

Within the study, the students worked with two separate tools, the AQC-tool and the Co-writing wiki. A detailed description of these tools is given in D 5.2.1.

For our investigation, we provided three questionnaires, a Pre-Questionnaire at the beginning of the course, an Intermediate Questionnaire after Phase 2 and a Post-Questionnaire at the end.

10.4.1.2.1 Pre-Questionnaire

The first section of this questionnaire dealt with “demographic data” of the students and some “general questions”. In addition, the students were asked about their “previous experience in group working and scientific working”.

Regarding the section “General attitudes concerning self- and peer-assessment”, there are four subscales according to [19]

- The intrinsic motivation scale measures the students’ motivation doing the peer-assessment activity for its own sake, just out of pleasure, e.g. “In a peer-assessment activity I liked opinions from peers because I got more ideas.”

- The extrinsic motivation scale measures the students' motivation doing the peer-assessment activity in order to get approval from the teacher and a good grade, e.g. "In a peer-assessment activity I think the opinions of my work from teachers were more important than those from peers."
- The evaluating scale measures the confidence of the students in evaluating the peer's work, e.g. "In a peer-assessment activity I found the strengths of my peer's work when I reviewed it."
- The receiving scale measures how students can handle the peer's-assessment in order to recognize their own weaknesses, e.g. "In a peer-assessment activity I recognized my weakness when I got comments from peers."

Answers were given on a 5-point Likert scale, so that students could state their level of agreement or disagreement. The rating scale ranged from "I strongly disagree" (1), "I disagree" (2), "neither/nor" (3) to "I agree" (4), "I strongly agree" (5).

The last section dealt with "motivational aspects" in general, i.e. how motivated students were regarding the whole course. In order to know how interesting and important the task was for the students, we also took into account the task value. There are three scales developed by [18] to investigate these motivational aspects:

- **Intrinsic Goal Orientation Scale:**
This scale measures the students' intrinsic motivation regarding the course, for instance: "I prefer course material that arouses my curiosity, even if it is difficult to learn." A high value on this scale would mean that the students are doing the course for reasons such as challenges and curiosity.
- **Extrinsic Goal Orientation Scale:**
This scale deals with the extrinsic motivation of the students, e.g. "Getting a good grade is the most satisfying thing for me right now." A student is extrinsically motivated when he/she is rather interested in rewards or a good grade than in the task itself.
- **Task Value Scale:**
This scale is about the task itself, i.e. how important, interesting, and useful the task and the task material are for the students. More interest in the task should lead to more involvement in one's learning. To give an example, one item out of this scale is: "I think I will be able to use what I learn in this course in other courses."

Answers were given on a 5-point Likert scale as already described above.

10.4.1.2.2 *Intermediate Questionnaire*

In the second questionnaire the students gave us feedback about the "quality of the content" they had written and the "quality of the questions" they had received during Phase 2. We provided answer categories from "very bad" (1), "bad" (2), "ok" (3), "good" (4) up to "very good" (5). Besides, we provided a section called "testing", where the students were asked how often they had taken a test. In this case the answers ranged from "never" (1), "seldom" (2), "sometimes" (3), "often" (4).

For the section “usability of the AQC-tool” we used the System Usability Scale (SUS) developed by Brook (1996) which contains 10 items and a 5-point Likert scale to state the level of agreement or disagreement (e.g. “I think that I would like to use this system frequently”).

By providing the section “learning style”, we wanted to investigate if the students prefer the elaborating or the repeating learning style. In our case we concentrated especially on these two learning styles to find out if the students’ learning process is rather superficial or aims at a deeper understanding. For this section we used items developed by [20] and translated them into English (e.g. item regarding the elaborating learning style: “In my mind I try to connect what I have learned with already known issues concerning the same topic.”, item regarding the repeating learning style: “I try to learn the content of scripts or other notes by heart.”). The answers were also given on a 5-point Likert scale.

To investigate in which emotional mood the students were when they used the AQC-tool, we added a section concerning “emotional aspects”, which includes 12 items. Kay and Loverock [7] developed this scale to measure emotions related to learning new computer software. Research showed that the 12 items are describing four emotions:

- Happiness (“When I used the tool, I felt satisfied/excited/curious.”)
- Sadness (“When I used the tool, I felt disheartened/dispirited.”)
- Anxiety (“When I used the tool, I felt anxious/insecure/helpless/nervous.”)
- Anger (“When I used the tool, I felt irritable/frustrated/angry”)

The answer categories in this section are “None of the time”, “Some of the time”, “Most of the time” or “All of the time”.

In the last section, called “further comments” the students stated additional comments and suggestions for improvements.

10.4.1.2.3 *Post-Questionnaire*

At the end of the course the students were asked about the “task difficulty and their learning effort”. In addition, we provided a section regarding the “group-assessment” about students’ experiences in reviewing the contribution of other groups.

For the section “usability of the WIKI-tool” we used the same System Usability Scale (SUS) developed by Brook (1996) [6] which is described above.

A section called “motivational aspects” dealt with the amount of motivation the students felt when they were working on the different tasks. Besides, we also asked them to estimate their peer’s motivation regarding the tasks. In this section we used the following answer categories: “absolutely unmotivated” (1), “unmotivated” (2), “motivated” (3), “very motivated”.

To investigate in which emotional mood the students were when they used the WIKI-tool, we provided the section “emotional aspects” with the items developed by [7] which is described above.

In the section “learning effort” the students were asked the same knowledge questions concerning scientific working again as in the Pre-Questionnaire which should show their learning effort in this context.

This questionnaire also had a last section concerning “further comments” where the students had the possibility to state additional comments and suggestions for improvements.

If not noted otherwise, all sections collect data from 5-point rating scales.

10.4.1.3 Procedure

The PHD course study was divided in 3 different phases.

10.4.1.3.1 Phase 1: Introduction to Scientific Working

The students were asked to learn a provided content about scientific working. The procedure of how to plan a study and the most common statistical tests were described briefly, based on the sections overview, planning experiment and data analysis. After they had finished learning the content, they answered the questions provided by the Knowledge Test section.

10.4.1.3.2 Phase 2: Selected Topics on Experimentation Planning

In a first step, the students are assigned to 6 groups, two group members each. Then, the students were informed that they should work on the activities collaboratively and got in contact with their peer.

In a second step the students were asked to deepen their knowledge in two different topics of scientific working, as a group working activity. One topic was about different research designs, the other topic dealt with the field research analysis. Each topic consisted of 6 articles. An overview of the content of each topic was provided for the students on the AQC-tool. Each group member selected one article from the topic “research design” and one article from the topic “research analysis”. The students could select the two articles per each topic according to their interests. For this task the group members had to discuss internally who of them will work on which topic and articles, so that they did not work on the same articles. For internal communication regarding this decision process, the students were asked to use the discussion forum on the WIKI. After each group member had chosen two different articles, they were asked to post their topics on the WIKI.

During reading the articles, the students could take tests (knowledge assessment for self-testing) to support their learning task. The system offered to take a pre-test on the level of the entire article, intermediate tests on the level of sections, and a post-test again on the level of the entire article. These entire tests could be called several times according to their convenience with the objective to support their learning process. The students were asked to try at least once all options in order to figure out what fits best to their needs. The questions they received during this phase were only for their personal learning progress, the answers were not used for the final performance indication. In addition a Glossary which contained explanations of important words was provided in order to support the students in understanding the articles.

After the students had finished reading the articles and taking the tests, they were asked to fill in a short questionnaire about the quality of the questions they received before. This feedback should give us the possibility to improve the tool afterwards.

In a next step, we asked each group member to write an essay regarding his/her articles (about 1000 words per article; so the group essays had 4000 to max. 5000 words). For writing these summaries, they used the WIKI-tool.

Within the group, they had the possibility to provide feedback to their peer's contribution. So the students could read and learn more about the content provided from their peer and give him/her hints for improvement. For this peer review, a peer-review function was provided in the WIKI-tool due to which they could provide comments on the document and give a rating about the importance of the contribution.

In the last step of phase 2 the students received a second test. This test included questions based on the content the group provided in the WIKI-System, so the students also had to be aware of the content of their peers. This test was taken into account for the final performance indication.

10.4.1.3.3 *Phase 3: Experimentation Planning*

In Phase 3 the students were asked to plan a study, provide a group-assessment, fill out a questionnaire, and finally they received a detailed feedback.

In the last activity of the course the groups were asked to plan their own study. For this task we provided the following research question. "Imagine that you work in a project at your university. You are responsible to design a study with the following research question: Is there a difference between facebook users and non-facebook users concerning their sport activities?" The students had to work out a method section where they described how they would investigate this research question. The students were asked to write maximally 4-5 pages in total (max. 2500 words) for this task. Furthermore, they did not have to provide any introduction with related research (although this would be mandatory in a real scientific paper). Instead, they only focused on the design of the study and gave some ideas how the analysis could be performed.

After they had finished this task, the students were asked to peer review the contribution of two other groups. This should give them also the chance to see how other groups have solved the problem.

At the end of the course we provided a detailed individual feedback for each group concerning their contribution to planning a study and worked out an example method section. When all the activities were completed, the students were asked to fill in the Post-Questionnaire.

10.4.2 *Evaluation Results*

In this section we focus on students' perception of the AQC and the co-writing WIKI itself, whereas the analysis of the tools' impact on student's learning process and motivation are reported in Section 10.5 (Validation Results). After evaluating the quality of automatically generated questions and concepts in the two pre-studies, in this main study we want to

emphasize the evaluation of H9.6 and H9.7 with the corresponding criteria and metrics C9.6, C9.7 and M9.12 as they are specified in [4].

Evaluation criteria and metric

- C9.6: To evaluate the level of satisfaction of the users with the tool.
- C9.7: To identify possible improvements for the tool.
- M9.12: Ratings for functionality/usability of the tool itself.

10.4.2.1 Automatic-Question-Creator (AQC)

10.4.2.1.1 Frequency of using the AQC

At first, they were asked if they had taken a test at all. 7 out of 8 agreed on that and one student said that he has never taken a test, because he/she did not have enough time. Counting the tests, which the students took optionally during the course, 30 tests were taken in total. According to the three different types of tests the students stated on a 4-point rating scale that they seldom took a test before they started reading (pre-test) ($M = 2.13$, $SD = 0.64$) or during reading the articles (sub-sections test) ($M = 2.25$, $SD = 0.71$). Regarding the post tests, the students also stated that they seldom took a test after they finished reading the whole topic ($M = 2.25$, $SD = 0.87$).

10.4.2.1.2 Quality of the automatically-generated questions

A comparison of the different question types shows that the students assessed the quality of the single choice ($M = 3.25$, $SD = 1.04$) and the multiple choice questions ($M = 3.25$, $SD = 1.16$) better than the fill in the blank questions ($M = 2.38$, $SD = 1.19$). The quality of the former as well as the distractors of the multiple choice questions as ok according to the ratings of the students.

In accordance with these results, the students explained that it was difficult for them to answer the fill in the blank questions ($M = 4$, $SD = 1.07$), whereas the single choice ($M = 2.25$, $SD = 1.03$) and multiple choice questions ($M = 2.38$, $SD = 0.92$) were not difficult to solve.

The students were asked what they liked about the three types of tests. The students explained that the different types of questions helped them getting an overview about the topics. Furthermore, they were in favor of the division into smaller modules. Some students also stated that the sub-section and post-tests supported them in observing their learning progress. Afterwards they were asked what they did not like. First of all, the contents of the tests were criticized. In particular, they would focus on memorization, source testing and the syntax and not on the parts of the texts the students would focus on. Additionally, the multiple choice questions were pointed out as not valuable due to the possibility to exclude not suiting answers. Missing graphics and structure as well as a slow interface bothered some students.

10.4.2.1.3 *Usability of the AQC-tool*

As described above, we used the System Usability Scale (SUS) developed by [6] to investigate the functionality/usability of the tool (see H9.6). After calculating the SUS score for each student, we got an average SUS score of 66.88. SUS scores have a range of 0 to 100 and the average SUS score from 500 studies is a 68. So our SUS score of 66.88 is a very good result given that this tool is not a professional one yet.

In addition, the students were asked to describe what they liked regarding the tool. The students stated that they were in favor of the simplicity of the tool and the division of the content into meaningful modules. Furthermore the students liked the consistence and the possibility to have an overview of the learning progress and their own test results. They mentioned that the course was well organized and they appreciated that the course was online, so that they could work from anywhere. Also, the content itself was described as very well compiled and as precise and useful.

Regarding the disadvantages of the tool, students did not like that they were logged out after a short period of time. They also complained about the slow interface. Regarding the AQC questions, the students criticized the difficulty to navigate to different questions, the repetition of questions within one single test, and that sometimes the questions seemed to start or end in an open sentence. Also, they had problems with the fill in the blank questions due to rejections caused by simple misspellings.

Finally the students were asked about suggestions for improvements (see H9.7). Most of the students would prefer a faster responding system and a faster navigation. Furthermore, they would like to download content and print it directly as a handout. Besides, some students suggested a layout for a better overview. The students would improve the text structure and recommended clearer instructions for the assessment parts.

10.4.2.2 *Collaborative Working*

10.4.2.2.1 *Attitudes concerning collaborative working*

After the first Phase, the students were asked about the advantages and disadvantages of collaborative working. In the latter case, students mentioned the possible dependence on others as a disadvantage of collaborative work. Thus, the own working progress depends on the progress of the peer. Furthermore, many students criticized the conflicts that may appear. They are afraid of long discussions with their peer and fear that some colleagues do not want to share their work. In addition, it needs time to coordinate the work and that it could be difficult if students have different schedules. Besides, the students pointed out, that strong personalities could take over the group leadership. It could be also difficult to find the same objective in a group.

Regarding the advantages of collaborative work, almost all students stated the opportunity to learn more from others. In particular, you are able to learn twice by teaching others and getting feedback. In addition knowledge can be shared, so that everyone can benefit from the knowledge of the others. An important role for them plays the discussion which appears on every group work. In this case, different opinions can be contributed and effective groups

gain more productivity. Finally some students mentioned the advantage of more motivation and fun and the opportunity to develop or improve social skills.

10.4.2.2.2 *Collaborative Working using the Co-writing wiki*

By observing the progress in the Co-writing wiki, we noticed that the students did not use the features which are provided to write their study collaboratively. Instead, it was very interesting to see that the students discussed their work in the discussion forum. This forum was in principal provided for the second phase to discuss their topics they would like to choose in order to avoid an overlap. But after they had finished the second phase, they continued discussing their work in the third phase. This finding shows that the students worked collaboratively, although not the way we expected them to do.

10.4.2.3 *Co-writing wiki*

10.4.2.3.1 *Previous Experience with Co-writing wiki*

Before the students started working with the Co-writing wiki, they were asked about their previous experience. The students, who already worked with the wiki tool, used it for a course to submit assignments or for educational purposes at University. Concerning the advantages of the system, they mentioned the simplicity. It was really easy to edit and share data, to produce contents together and it enables collaboration with transparency. Furthermore, one student liked the WYSIWYG editing interface. Regarding the disadvantages, they did not like the verbose syntax, the editors and the difficulty to work in parallel.

10.4.2.3.2 *Usability of the Co-writing wiki*

To investigate the usability of the WIKI-tool, we used the System Usability Scale (SUS) by [6]. After calculating the SUS score for each student, we got an average SUS score of 52.08. SUS scores have a range of 0 to 100 and the average SUS score from 500 studies is a 68. So the result is quite good concerning the fact that an objective of the study was to find out the weaknesses of the Co-writing wiki in order to improve the system.

Almost all of the students stated that the Co-writing wiki is easy to use and focuses on few important functions. They also were in favor of the ability to discuss per topic/per page and creating and modifying pages. In addition, they mentioned that the tool was always available and consistent.

Then we asked them to state what they did not like regarding the tool. Some students complained about the usability of the Co-writing wiki and its slowness. The students also mentioned that they were not aware of all available functions. It was also annoying for some of them providing the type of edits and marking its importance. They also mentioned some editing problems, for instance one participant stated that he had to put all data into a text file first and was not able to edit it in the system. Besides, for some of them it was a little bit confusing to find the pages. The students also complained about the auto-logout without any warnings, the text layout and the navigation of the Co-writing wiki.

Finally the students were asked about suggestions for improvements. Although support was given, it would have been better to provide available support in the tool itself. They also

suggested a user manual of the tool. The students would also like notifications on content available or new discussions available so as to keep the user up to date. Another suggestion was to include all created pages in the menu on the left.

10.4.2.4 Emotional aspects

Regarding the emotional aspects during working with the Co-writing wiki, there was no difference according to the t-tests between the emotions happiness ($t_{sadness} (5) = 0.94, p > .05$; $t_{anxiety} (5) = 0.90, p > .05$; $t_{anger} (5) = 0.23, p > .05$) sadness ($t_{anxiety} (5) = 0.40, p > .05$; $t_{anger} (5) = 2.5, p > .05$), anxiety ($t_{anger} (5) = 0.96, p > .05$) and anger. The results from a 4-point rating scale showed that the students felt equally happy ($M = 1.72, SD = 0.65$), sad ($M = 1.33, SD = 0.41$), anxious ($M = 1.42, SD = 0.34$), and angry ($M = 1.61, SD = 0.53$). By interpreting the mean values, it can be assumed that the students seldom felt consciously happy, sad, anxious or angry. This means that working with the Co-writing wiki did not lead to students' frustration or anger as was the case in R8.

Concerning students' emotions during working with the AQC-tool, we got similar results. The t-tests showed that there was also no difference between the emotions happiness ($t_{sadness} (7) = 0.98, p > .05$; $t_{anxiety} (7) = 1.44, p > .05$; $t_{anger} (7) = 1, p > .05$) sadness ($t_{anxiety} (7) = 0.31, p > .05$; $t_{anger} (7) = 0.17, p > .05$), anxiety ($t_{anger} (5) = 0.57, p > .05$) and anger. According to the comparison of the mean values, it can be assumed, that the students felt equally happy ($M = 1.88, SD = 0.80$), sad ($M = 1.5, SD = 0.60$), anxious ($M = 1.41, SD = 0.65$), and angry ($M = 1.54, SD = 0.31$).

10.4.3 Validation Results

In this section we show the validation methodology that includes the following validation criteria and metric extrapolated by [4].

Validation criteria

- C9.8: To evaluate students' motivation concerning their learning activities.
- C9.9: To evaluate the learning types of the students.

Validation metrics

- M9.13: Ratings of students' extrinsic and intrinsic motivation regarding peer-assessment activity before using the tool.
- M9.14: Ratings of students' extrinsic and intrinsic motivation regarding the course and its tasks before using the tool.
- M9.15: Ratings for the learning types of the students.
- M9.16: Ratings of students' group-assessment activities.

Following this methodology we will validate the motivational aspects regarding students' learning activities (H9.8), and the students' learning type preference (H9.9).

10.4.3.1 Motivational Aspects

10.4.3.1.1 Motivational Aspects regarding the Peer-Assessment

With regard to the students' motivation concerning the peer-assessment, we wanted to investigate if they are rather intrinsic or extrinsic motivated. A comparison of the mean values ($t(11) = 5.99, p < .01$) shows that the students' intrinsic motivation ($M = 3.75, SD = 0.51$) is significant higher than their extrinsic motivation ($M = 2.65, SD = 0.48$). Thus, the students would participate in an assessment for its own sake and out of pleasure and not just for getting a good grade or approval from the teacher. It can be assumed that the students' first aim was to learn something out of the course and that getting a grade does not play such an important role for them. This result stands in accordance with the fact that half of the students participated in the course voluntarily.

The students stated for instance that they like opinions from peers in order to get more ideas ($M = 4.08, SD = 0.67$). In contrast, the students would not feel that they have learned nothing if they get a low peer score on their work ($M = 1.75, SD = 0.75$).

10.4.3.1.2 Motivational Aspects concerning the course and its tasks

The results of the students' motivation regarding the course in general are in accordance with the previously reported results. Comparing the extrinsic - with the intrinsic goal orientation scale, the intrinsic motivation ($M = 3.94, SD = 0.53$) is significant higher than the extrinsic motivation ($M = 2.83, SD = 0.79; t(11) = 3.43, p < .01$). This means that they are interested in the course for reasons such as curiosity and challenges, whereas a good grade or rewards are not so important for them. These findings are supported by the results of the task value scale. A mean value of 3.83 ($SD = 0.74$) shows that the students were really interested in the task itself. The task material was also very useful and important for them. Due to their high interest, it can be assumed that this also leads to more involvement in their learning efforts.

10.4.3.1.3 Motivational aspects concerning their learning activities

By using the mean as an exact measure, we faced problems with interpreting the data in this case. So we decided to have a closer look on the median as an alternative measure of the central tendency (see Section R8 for a more detailed discussion of this issue).

Regarding students' motivation concerning their learning activities (H9.8), they stated that they were motivated up to very motivated while reading the contents ($M = 3.50, SD = 0.55, Md = 3.5$). Also working with the AQC tool and testing themselves with questions motivated the students ($M = 2.67, SD = 0.52, Md = 3.00; M = 2.50, SD = 0.84, Md = 3.00$). Furthermore, students were motivated up to very motivated concerning writing essays ($M = 3.50, SD = 0.55, Md = 3.5$) and planning a study ($M = 3.67, SD = 0.52, Md = 4$). Working with the Co-writing wiki motivated the students ($M = 2.67, SD = 1.03, Md = 3.00$) as well as the assessment activity ($M = 3.00, SD = 0.00, Md = 3$) and filling in the questionnaires ($M = 3.00, SD = 0.00, Md = 3.5$).

10.4.3.2 Learning Type Preferences

As discussed above, we wanted to investigate the learning styles of the students in order to know if the students prefer the elaborating or the repeating learning style. A comparison of the mean values shows that there is a significant difference between the elaborating learning style ($M = 4.05$, $SD = 0.56$) and the repeating learning style ($M = 3.04$, $SD = 0.82$; $t(7) = 2.71$, $p < .05$). The students prefer the elaborating learning style, which means that their learning process aims at deeper understanding and is less superficial.

Concerning elaborating, the students stated for instance that they try to link new terms or new theories to familiar terms and theories ($M = 4.38$, $SD = 0.52$). In contrast to that, the students said that they don't learn the content of scripts or other notes by heart ($M = 2$, $SD = 1.07$) which would indicate a repeating learning style.

An important condition for the elaborating learning style is that people are intrinsically motivated. The results from the first questionnaire showed that they were intrinsically motivated after the first Phase of the course. So due to their learning style preference, it can be assumed that the students were still intrinsically motivated in the second Phase, where they received the questions from the AQC. Thus, the students answered the questions out of pleasure with the aim to deepen their knowledge.

In addition, the students stated that if they learn something, testing themselves with questions often helps them ($M = 3.63$, $SD = 1.50$). This result is in line with the results discussed above. So it can be assumed that testing themselves with the AQC questions supported them regarding their learning progress (see H.9.9).

10.4.3.3 Group-Assessment

After the students had finished their papers they were asked to evaluate the work of other groups. Regarding the assessment rubric provided for the group review, the students stated that the assessment rubric was easy to use ($M = 3.67$, $SD = 1.51$). In addition 50 % of the students agreed on the statement that the assessment rubric supported them to effectively review the product of the other groups ($M = 3.17$, $SD = 0.98$).

The students neither agreed nor disagreed on the statements "The assessment rubric provided for the group review supported me to learn more about other group's topic." and "Using the rate control (stars) was very helpful to assess the student's level of mastery based on the rubric criteria."

In addition, the students were asked what they liked regarding this group-assessment. All of the students mentioned that they liked the group-assessment because of the opportunity to see how other groups approached the problem and solved it in order to improve their own products.

Regarding the question what they did not like, some students answered that they would have preferred to give textual feedback to state suggestions and improvements instead of providing negative feedback by using the assessment rubric. Some of the students also criticized the usability of the Co-writing wiki, especially the low speed.

10.5 Conclusion

In this final section we would like to summarize the results of the pre-studies and the main study based on the goals and hypotheses which were presented at the beginning. According to these goals and hypotheses, within the pre-studies we investigated the quality of the concepts and questions created by the AQC and therefore compared them with human-generated concepts and questions. Additionally we analyzed the relevance of the answers given by the AQC questions. In the main study, we focused on the usability of the AQC and possible improvements which could be deduced from the results. Furthermore, we had a look on students' motivational aspects and their preferences regarding learning styles. As already mentioned above, within the main study we combined AQC and Co-writing wiki in order to create a complex learning environment. This allowed us to investigate (additional to AQC) students' experiences with Co-writing wiki, which is also presented below.

In the **pre-studies** we were interested in the quality of the concepts, questions and answers (G9.1 through G9.5). Comparing the concepts that were extracted by the AQC with concepts extracted by humans, results show that the most relevant concepts provided from AQC were equal to the concepts extracted from the participants. According to the automatically created questions and answers, their quality fulfilled the needs of the students. Furthermore, there was no difference between automatically and manually created questions. Besides, we investigated the pedagogical quality of the automatically created questions regarding Bloom's taxonomy [13]. Two raters categorized the majority of the test-items as lower-level items (i.e., testing the knowledge, comprehension, or application of the learning content). All in all, results show that there was neither a difference in categories with respect to the four different test-item types nor with respect to automatically and manually created items.

According to the results in the **main study**, it can be assumed that the AQC-tool is user-friendly (G9.6). First, we can be satisfied with the functionality because of the high SUS score the tool has reached. So the students were in favor of the various functions of the tool and its simplicity. Second, they stated that the tool gave them a good overview of their learning progress. In improving the tool (G9.7), we should have a closer look on the "fill in the blank" question type, which was not easy to solve and work on a faster interface.

Regarding students' motivation (G9.8), the results show that the students were intrinsically motivated at the beginning of the course. So they were really interested in the course and its tasks, which lead also to more involvement in their learning effort. At the end of the course, the students were asked about their motivation concerning different learning activities. According to the results, students' motivation was high during reading content, writing essays, doing the peer and group assessment, working with the Co-writing wiki and filling in the questionnaires. In addition, testing themselves with questions and working with the AQC-tool also motivated them.

By investigating students' learning styles, we found out that the students' learning process aims at deeper understanding and is less superficial. This result is in line with the results discussed above, because intrinsic motivation is an important condition for this learning type. So it can be assumed, that the students answered the questions out of pleasure with the aim

to deepen their knowledge. Besides, students also stated that testing themselves often supported them in their learning process (G9.9).

In addition to the AQC, we also investigated the usability of the Co-writing wiki. Due to the SUS score, the students assessed the usability of the Co-writing wiki quite good. Almost all of the students stated that the Co-writing wiki is easy to use and focuses on few important functions. They also were in favor of the ability to discuss per topic/per page and used this function to work collaboratively. According to students' comments and suggestions, we have to improve the speed of the tool and work on a technical support, which is always available.

References

- [1] Intelligent Web Teacher (IWT) web site; <http://iwtalice.crimpa.unisa.it/IWT>
- [2] ALICE (Adaptive Learning Via Intuitive/Interactive, Collaborative And Emotional Systems) project, Deliverable D3.4.1 “Sample Collaborative Complex Learning Object v1”, project co-funded by the European Commission within the 7th Framework Programme (2007-2013), n. 257639, 2010.
- [3] ALICE (Adaptive Learning Via Intuitive/Interactive, Collaborative And Emotional Systems) project, Deliverable D1.1 “Requirements”, project co-funded by the European Commission within the 7th Framework Programme (2007-2013), n. 257639, 2010.
- [4] ALICE (Adaptive Learning Via Intuitive/Interactive, Collaborative And Emotional Systems) project, Deliverable D1.3 “Experimentation and validation planning”, project co-funded by the European Commission within the 7th Framework Programme (2007-2013), n. 257639, 2010.
- [5] ALICE (Adaptive Learning Via Intuitive/Interactive, Collaborative And Emotional Systems) project, Deliverable D7.4.1 “Prototype Components for Adaptive e-Learning v1”, project co-funded by the European Commission within the 7th Framework Programme (2007-2013), n. 257639, 2010.
- [6] Brooke, J. (1996). SUS: A “quick and dirty” usability scale. In *Usability evaluation in industry*. London: Taylor & Francis.
- [7] Kay, R.H., & Loverock, S. (2008). Assessing emotions related to learning new software: The computer emotion scale. *Computers in Human Behavior*. 24, 1605-1623.
- [8] Caballé, S., Gañan, D. Dunwell, I., Pierri, A., Daradoumis, T. (2011). CC-LO: Embedding Interactivity, Challenge and Empowerment into Collaborative Learning Sessions. *Journal of Universal Computer Science*. In press. Retrieved from http://cpl.uoc.edu/JUCS_CaballeEtAl
- [9] Bortz, J. & Döring, N. (2006). *Forschungsmethoden und Evaluation*. Heidelberg: Springer.
- [10] ALICE (Adaptive Learning Via Intuitive/Interactive, Collaborative And Emotional Systems) project, Deliverable D1.3 “Annex I – “Description of Work”. V1.0.”, project co-funded by the European Commission within the 7th Framework Programme (2007-2013), n. 257639, 2010.
- [11] ALICE (Adaptive Learning Via Intuitive/Interactive, Collaborative And Emotional Systems) project, Deliverable D3.4.1 “Sample Collaborative Complex Learning Object v1”, project co-funded by the European Commission within the 7th Framework Programme (2007-2013), n. 257639, 2010.
- [12] Brennan, R.L., & Prediger, D.J. (1981). Coefficient κ : Some uses, misuses, and alternatives. *Educational and Psychological Measurement*, 41, 687-699.
- [13] Bloom, B. S. (ed.) (1956). *Taxonomy of Educational Objectives, the classification of educational goals – Handbook I: Cognitive Domain* New York: McKay

- [14] Cannella, S., Ciancimino, E., & Lopez-Campos, M. (2010). Mixed e-Assessment: an application of the student-generated question technique. Paper read at IEEE International Conference EDUCON 2010, Madrid, Spain, April.
- [15] Gütl, C., Lankmayr, K., & Weinhofer, J. (2010). Enhanced Approach of Automatic Creation of Test Items to Foster Modern Learning Setting. Proc. of the 9th European Conference on e-Learning, Porto, Portugal, 4-5th November 2010, 225-234.
- [16] Gütl, C., Lankmayr, K., Weinhofer, J., & Höfler, M. (2011). Enhanced automatic question creator – EAQC: Concept, development and evaluation of an automatic test item creation tool to foster modern e-education. *Electronic Journal of e-Learning*, 9(1), 23-38.
- [17] Mitkov, R., Ha, A. L., & Karamanis, N. (2005). A computer-aided environment for generating multiple-choice test items. *Natural Language Engineering*, 12, 177-194.
- [18] Pintrich, P.R., Smith, D.A.F., Garcia, T., & McKeachie, W.J. (1991). A Manual for the Use of the Motivated Strategies for Learning Questionnaire (MSLQ). Technical Report, 91, 7-17.
- [19] Tseng, S.-C., & Tsai, C.-C. (2010). Taiwan college students' self-efficacy and motivation of learning in online peer-assessment environments. *Internet and Higher Education*, 13, 164-169.
- [20] Wild, K.-P. (2000). *Lernstrategien im Studium. Strukturen und Bedingungen*. Münster: Waxmann.
- [21] Knussen and McQueen (2006): *Introduction to Research Methods and Statistics in Psychology*. Harlow: Pearson Education Limited.

Annex A – Integration of IWT tools with real context of learning

A1 Integration at UOC site

A1.1 Introduction

Alice is an extension of the Intelligent Web Teacher Platform (IWT), which is a commercial LMS built over the Microsoft .NET platform. Hence, the different tools developed in the different working packages are written in .NET and the session mechanisms and parameters used to exchange information with the LMS are specific to the IWT implementation.

One of the experimentation sites is the UOC, which is based on a completely different and open source architecture closely linked to java.

These differences in the base architecture make it difficult to carry out the experimentation, in a direct way in the UOC environment.

A1.1.1 Purpose

The purpose of this report is to show the necessary steps that have been taken to find a software solution to permit the integration between the UOC learning campus and a tool that is running in a different platform that, in this case, it is built using the .NET Microsoft framework. Integration will include a Single Sign-On (SSO) mechanism to control the logging process in both platforms.

A1.1.2 Scope

Integration will be carried out within a specific course. It is important to point out that the tools to integrate won't live within the UOC environment but in the running instance of IWT. So it is, actually, a remote launch. Taking into account this, the integration will cover two possible scenarios:

IWT as a tool

There are several different tools available in a classroom, however, IWT will be considered as a tool in itself in this scenario. That means that, when you click on it, a session will be created in IWT with the same user logged in.

Live and Virtualized Collaboration tool

The other scenario is the typical one. In the list of tools, one can find a link to a IWT-ALICE classroom within a UOC classroom. When you click on the link, a session will be created in IWT platform and a IWT-ALICE classroom will be displayed in a new window as if you had logged in directly on IWT.

A1.2 Survey of tools for interoperability

Although one of the platforms we want to integrate with is a proprietary LMS (IWT), our study has focused on open source solutions for learning tools interoperability.

A1.2.1 Background

The UOC has been working for a long time on innovating and integrating different models, tools and APIs in the campus and its experience has demonstrated that, if you are not updated and do not keep at the same level marked by technology, you become obsolete.

Historically, the software infrastructure of schools has been heterogeneous. This fact has adversely affected them and ultimately the interoperability between different platforms.

After investigating the possible architectures for interoperability, two architectures have been selected to ensure interoperability between the applications and the platforms.

- Open knowledge Initiative (OKI)
- IMS Basic Learning Tools Interoperability (BLTI)

The UOC has adapted both models to its campus and has experience with both architectures.

Indeed, it uses a combination of both to take advantage of both models.

In the following chapters, an overview of both architectures will be provided and we will see how they have been adapted to define a solution architecture.

A1.3 Open knowledge Initiative (OKI)

A1.3.1 Introduction

MIT, through the OKI project, has defined a set of interfaces with the typical services that have been used in different e-learning platforms.

The Open Knowledge Initiative. (O.K.I.) is an open and extensible architecture for e-learning technology specifically targeted to the needs of the higher education community. O.K.I. provides detailed specifications for interfaces (OSIDs) among components of an e-learning management environment and open source examples of how these interfaces work.

OSIDs permit the possibility to have an abstraction layer between the organization infrastructure and the artefacts that implement these interfaces. This can be shown in Fig. A-1:



Figure A-1: OSID interaction

The O.K.I. architecture is intended to be used by commercial product vendors and by higher education product developers. It provides a stable and scalable base that supports the flexibility needed by higher education as learning technology is increasingly integrated in the education process.

A1.3.2 Architecture

OKI architecture, basically, separates e-learning services into two groups: common services and educational services. Fig. A-2 shows the whole architecture:

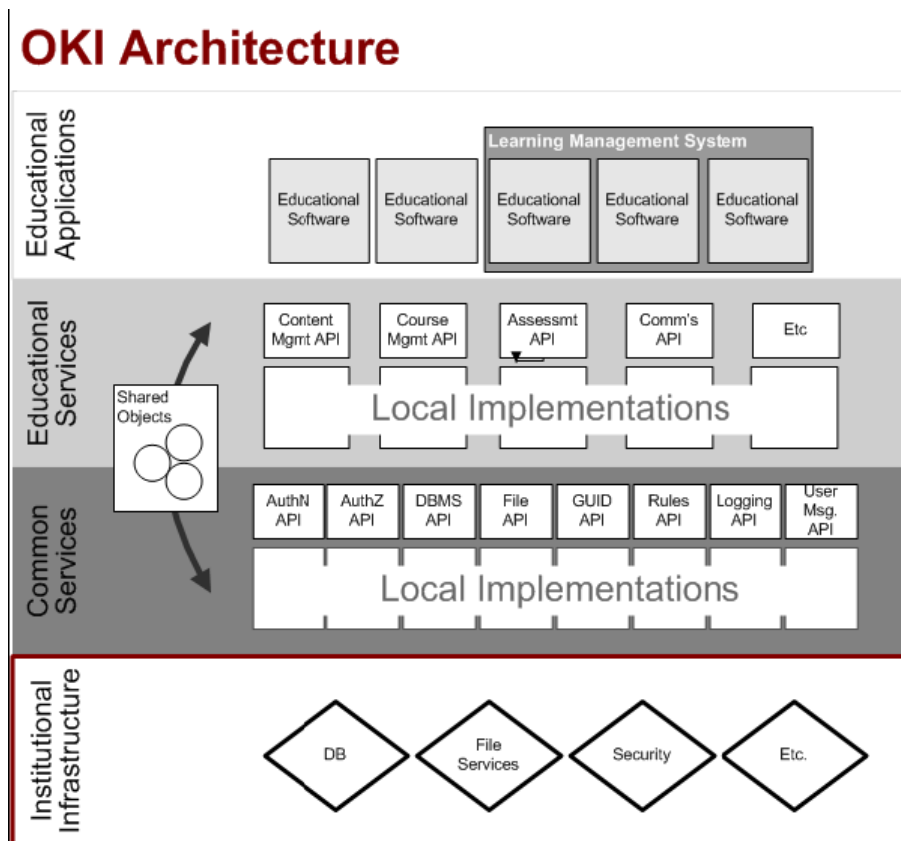


Figure A-2: Tiers of the OKI architecture

A1.4 IMS Basic Learning Tools Interoperability (BLTI)

A1.4.1 Introduction

The other architectural solution for interoperability between the tools of the different platforms is part of the IMS group architectures.

As for interoperability, IMS provides the Learning Tools Interoperability (LTI) architecture, which offers a single framework or standard way of integrating rich learning applications—in LTI called Tools — with platforms like those of learning management systems, portals, or other systems from which applications can be launched — called Tool Consumers. Basic LTI is a subset of the full LTI specification.

Basic LTI allows the integration of a remote application into the current Learning Management System (LMS). The meaning of ‘current’ here is ‘local’. From the point of view of the user, it means that, within a classroom of the course, you could see, in addition to the links of the tools that are available, links to tools that are not, actually, in the local Learning Management System but in a remote one.

A1.4.2 Overview of the architecture

With respect to IMS nomenclature, the local LMS is called Tool Consumer (TC) since it is the part that consumes the external tool or content. The remote application is called Tool Provider (TP) since it is the component that, in fact, provides the application information to the tool Consumer.

Between TC and TP, there is a communication flow through what is called Basic LTI data. This information is passed on in the form of an http POST and it is secured by the OAuth protocol.

All the important pieces are shown in Fig A-3:

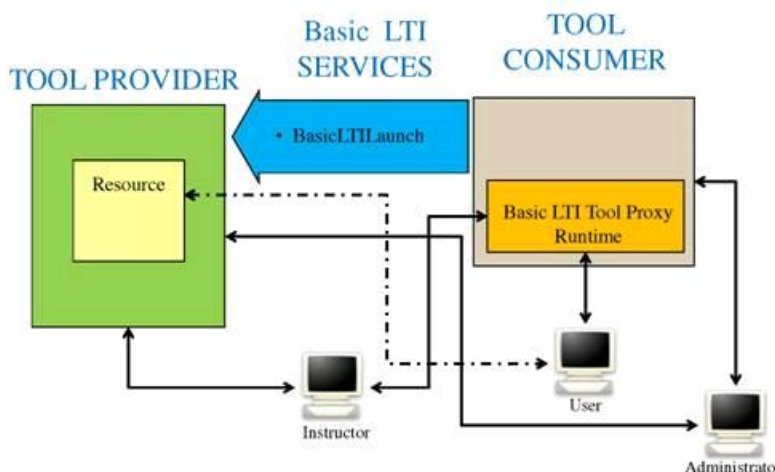


Figure A-3: Overview of Basic LTI

A1.5 Adoption of BLTI for the integration with ALICE

A1.5.1 Requirements

There are two scenarios that must be satisfied when the selected architecture is applied to the integration.

On the one hand, it is necessary to allow the launch of an external tool from a classroom within the UOC campus. Initially, an ALICE instance will be considered as a tool.

The TC and TP are standard connectors and can be applied to any tool.

The differences that could be between the different tools are just the amount of information that is needed to be launched. From the point of view of a teacher or student, there will be a link to ALICE in the classroom like as if it was a local tool.

On the other hand, other tools that are available within the ALICE environment should be capable of being launched directly from the list of tools that are within the classroom, especially, the Live and Virtualized Collaboration tool that has been developed in the WP3.

When the tools are launched, the ALICE context graphical elements should be hidden so that the user keeps the idea that it is still within the UOC classroom.

A1.5.2 Current Architecture

The two environments that are necessary to be integrated have different base architectures:

- UOC campus: It is based on the C language and J2EE containers like TOMCAT and JBOSS. So, basically, the base libraries are written in java and the applications are web applications.
- ALICE LMS: It is built over the .NET framework and uses IIS as a container for the web applications that are written in C#.

The UOC has its own applications and they are integrated with the UOC low level services. Some of them are implemented by using OKI implementations like, for instance Authentication. The UOC has its own session management mechanism.

The IWT has its own course structure and session management.

A1.5.3 Proposed architecture

The solution architecture must permit the opening of a tool that is living in the IWT platform (the detailed scenarios are described in the introduction of chapter 2).

So, basically, some basic information (like language and user information) will be passed on from the UOC to the IWT. This information will be passed on in the way that is specified by BLTI and signed with Oauth. Since the IWT can trust the signed information, it can perform the login in its platform. This mechanism will cover, therefore, the Single Sign-On process.

Thus, the solution architecture includes two software pieces to be included in the two platforms:

- A BLTI consumer in the UOC campus
- A BLTI provider in the IWT environment

The BLTI consumer in the UOC is a web application that can be used as a tool consumer (TC) with other BLTI providers thanks to having different configurations. It uses the OKI authentication service and the Agent OSID to retrieve the necessary user information to be passed on to the IWT.

A1.5.4 Information exchange between UOC and IWT

The list of fields that are used to pass on information to IWT with BLTI are the following ones:

oauth parameters (nonce, signature, version, signature_method, consumer_key, callback)

tool_consumer_instance_guid

launch_presentation_locale

lis_person_parameters (name_given, name_family, name_full, contact_email_primary)

user_id

user_image

lti_version

lti_message_type

tool_consumer_instance_description

basiclti_submit

context_id

roles

key

custom parameters (lti_message_encoded_base64, user_gender, user_birthdate, context_id, username, user_city, service)

NOTES:

- The *lti_message_encoded_base64* field indicates whether the values are encoded in BASE64 to avoid problems with special characters or not
- The *custom_context_id* field contains the id of the group associated to a specific forum

A2 Integration at MOMA site

The experimentation led by MOMA has regarded the testing of the scenarios R5, R6, R7 on the reference e-Learning platform, *Intelligent Web Teacher* (IWT).

The IWT architecture is modular enough to allow the deployment of solutions able to cover application scenarios of different complexity and for different domains. Hence, starting from IWT, different extensions have been made in order to pursue the following key points:

- extension of the IWT adaptivity through the capability of managing the new emotional and affective feedbacks from students;
- generation of new and complex learning resources, like storytelling and serious game, able to assess the progress done in the learning process about scientific themes and the cognitive impact after learning experiences enabling to integrate and manage aspects like adaptivity.

The first point has been obtained integrating in the platform an affective and emotional module, conceived at the aim of permitting a prompt identification, in the background, of the altered emotional states of a student during his learning activities.

The second point has been obtained creating two IWT Drivers for the new Complex Learning Objects.

Finally MOMA has realised a new version of the existing IWT Course driver and prepared a course that highlights the new features.

All these aspects have been experimented within two secondary schools belonging to the networks of secondary schools created by MOMA and that already adopt the IWT platform. In addition, to facilitate the students during the execution of their activities and let them concentrating on the experimentation tasks one of standard features of IWT platform, the customizability of the graphic and layout of pages, has been exploited. Taking into consideration this aspect MOMA has totally customized the web portal used for the experimentation through the following interventions:

- A new private and customized access page has been created for the experimentations;

- The students' home page layout has been designed, setting the menu and modules collocation within the web page and removing the modules not useful in this context;
- Different roles and permission have been defined in order to allow to the different users (students, tutor, teacher) to take part to the learning experience;
- Users can further customize the layout of their home pages directly from the web.

The screen shots of Figure A-4 and Figure A-5 show the customized web portal before and after having made the login:



L'E-LEARNING CHE EMOZIONA
Una Sperimentazione nelle Scuole Italiane
www.aliceproject.eu

Accedi

Nome utente

Password

Entra

[Recupera Password](#)

[Registrati](#)

MOMA s.p.a. Via Aldo Moro 1/P - 84081 Baronissi (SA) - Italy
Se hai problemi ad accedere ad IWT verifica i REQUISITI MINIMI oppure contatta il SUPPORTO TECNICO



Figure A-4: The customized web portal before the login

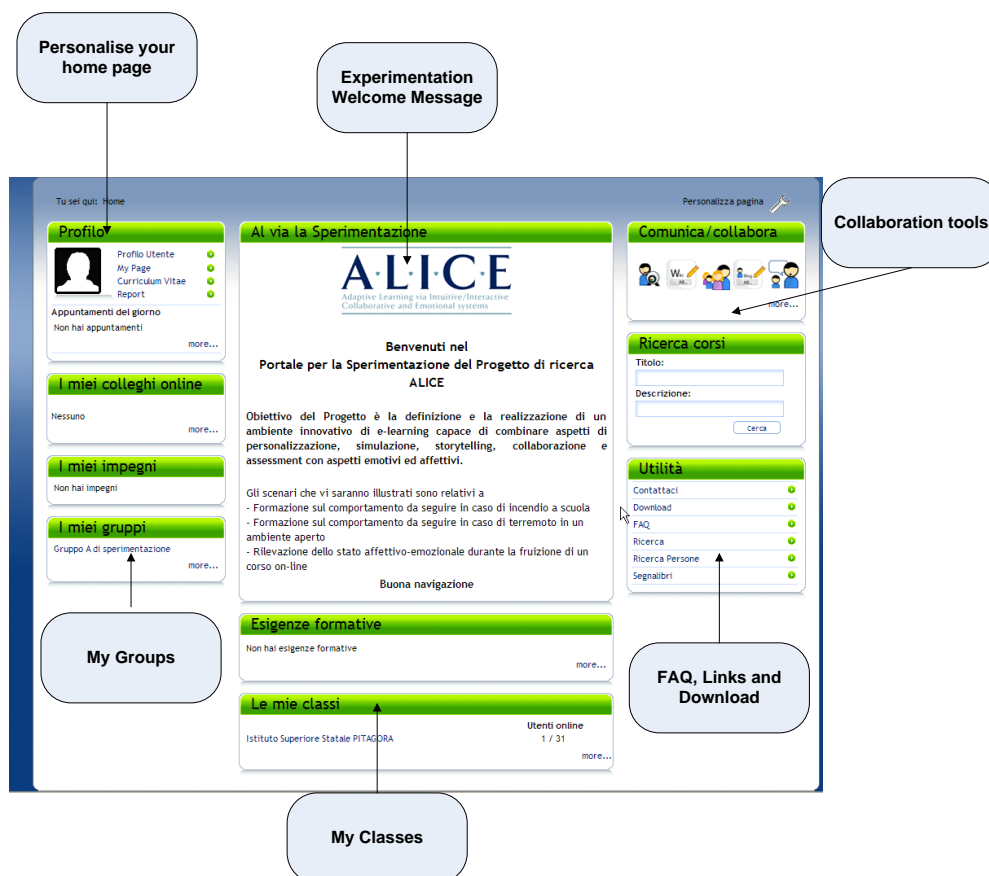


Figure A-5: The customized web portal after the login

A3 Integration at TUG site

In this first experimentation iteration the tools used in R8 and R9, namely the co-writing WIKI and the AQC (Automatic Question Creator), were used stand alone. For R2, the contextualized ontology tool coming from WP7 was experimented directly in IWT, which required neither customization nor integration.

Thus neither customization nor integration activities were performed by TUG for experimentation purposes in this first experimentation step.