



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: **D8.1.2: Final Experimentation and Evaluation Results**
Revision: 2.0
Workpackage: WP8: Experimentation and Validation
Dissemination Level: Public

Due date: 06/30/2012
Submission Date: 06/30/2012
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	29/02/2012	Draft with structure	UOC
0.2	07/03/2012	Draft revised	MOMA
0.3	15/05/2012	Contribution to R2, R8 and R9 scenarios	TUG
0.4	15/06/2012	Contribution to R5, R6 and R7 scenarios	MOMA
0.5	15/06/2012	Contribution to R1, R2, R3 and R4 scenarios	UOC
0.6	20/06/2012	Integration of contributions	UOC
0.7	25/06/2012	Internal review	UOC
1.0	30/06/2012	Final changes	UOC

Table of Contents

1	Introduction	8
1.1	Purpose.....	9
1.2	Methodology.....	10
1.2.1	Experimentation at UOC site.....	12
1.2.2	Experimentation at TUG site	13
1.2.3	Experimentation at MOMA site	14
2	R1. Upper Level Learning Goals.....	15
2.1	Evaluation and validation procedure	15
2.2	Method.....	16
2.2.1	Participants.....	16
2.2.2	Apparatus and Stimuli	17
2.3	Evaluation Results	21
2.3.1	Activity levels.....	22
2.3.2	Usability of the IWT.....	22
2.3.3	Emotional Aspects.....	25
2.3.4	Evaluation of Questionnaire.....	26
2.4	Validation results	27
2.4.1	IWT as a valuable resource	27
2.4.2	Motivational aspects	34
2.4.3	Tutor assessment and knowledge acquisition	34
2.5	Conclusion.....	36
3	R2. Knowledge model contextualization: Experimenting the Knowledge model contextualization.....	37
3.1	R2-1. Knowledge model contextualization: Experimenting the Knowledge model contextualization from the instructor’s and student’s viewpoint (TUG)	37
3.1.1	Evaluation and Validation Procedure.....	37
3.1.2	Method.....	39
3.1.3	Evaluation Results	41
3.1.4	Validation Results.....	50
3.1.5	Conclusion	50
3.2	R2-2. Knowledge model contextualization: Experimenting the Knowledge model contextualization from the instructor’s viewpoint (UOC).....	51
3.2.1	Evaluation and validation procedure	51
3.2.2	Method.....	52
3.2.3	Evaluation and Validation Results.....	54

3.2.4	Conclusion	57
4	R3. Semantic Connections between Learning Resources.....	59
4.1	R3-1. Semantic Connections Between Learning Resources from student’s view ...	59
4.1.1	Evaluation and validation procedure	59
4.1.2	Method.....	60
4.1.3	Evaluation Results	64
4.1.4	Validation Results.....	71
4.1.5	Conclusion	74
4.2	R3-2. Semantic Connections Between Learning Resources from instructor’s view	75
4.2.1	Evaluation and validation procedure	75
4.2.2	Method.....	76
4.2.3	Evaluation and Validation Results.....	78
4.2.4	Conclusion	80
5	R4. Live and Virtualized Collaboration.....	82
5.1	R4-1. Live and Virtualized Collaboration: Experimenting with Collaborative Complex Learning Objects (CC-LO) from the student’s viewpoint.....	82
5.1.1	Evaluation and Validation Procedure.....	82
5.1.2	Method.....	84
5.1.3	Evaluation Results	86
5.1.4	Validation Results.....	94
5.1.5	Conclusion	97
5.2	R4-2. Live and Virtualized Collaboration: Experimenting with Collaborative Complex Learning Resources (CC-LR) from the student’s viewpoint	98
5.2.1	Evaluation and Validation Procedure.....	98
5.2.2	Method.....	99
5.2.3	Evaluation Results	102
5.2.4	Validation Results.....	109
5.2.5	Conclusion	114
5.3	4-3. Live and Virtualized Collaboration: Experimenting with Collaborative Complex Learning Resources (CC-LR) enriched with authoring information from the student’s viewpoint	114
5.3.1	Evaluation and Validation Procedure.....	114
5.3.2	Method.....	116
5.3.3	Evaluation Results	119
5.3.4	Validation Results.....	126
5.3.5	Conclusion	131

5.4	R4-4. Live and Virtualized Collaboration: Experimenting with Complex Learning Resources (CC-LR) from the instructor’s viewpoint	132
5.4.1	Evaluation and validation procedure	132
5.4.2	Method	133
5.4.3	Evaluation and Validation Results	135
5.4.4	Conclusion	141
6	R5. Storytelling	143
6.1	Evaluation and Validation Procedure	143
6.2	Method	144
6.2.1	Participants.....	144
6.2.2	Apparatus and Stimuli	145
6.2.3	Procedure	146
6.3	Evaluation Results	147
6.3.1	The storytelling activity	147
6.3.2	Emotional Aspects	148
6.3.3	Usability of the Storytelling	151
6.4	Validation Results.....	152
6.4.1	The Storytelling as a valuable resource	153
6.4.2	Acquired competences and didactic efficiency of the Storytelling	155
6.5	Conclusion.....	157
7	R6. A Serious Game for Civil Defense	158
7.1	Evaluation and Validation Procedure	158
7.2	Method	159
7.2.1	Participants.....	159
7.2.2	Apparatus and Stimuli	160
7.2.3	Procedure	161
7.3	Evaluation Results	162
7.3.1	The Serious Game Activity	162
7.3.2	Emotional Aspects	163
7.3.3	Usability of the Serious Game	164
7.4	Validation Results.....	166
7.4.1	The Serious Game as a valuable resource.....	166
7.4.2	Acquired competences and didactic efficiency of the Serious Game.....	166
7.5	Conclusion.....	168
8	R7. Affective and Emotional Approaches.....	169
8.1	Evaluation and Validation Procedure	169
8.2	Method	170

8.2.1	Participants.....	170
8.2.2	Apparatus and Stimuli	171
8.2.3	Procedure	172
8.3	Evaluation Results	173
8.3.1	The Emotional Tool Activity	173
8.3.2	Usability of the Emotional Tool.....	174
8.3.3	Emotional Aspects.....	175
8.4	Validation Results.....	177
8.5	Conclusion.....	178
9	R8. Enhanced Wiki-Test and Peer-review for writing assignments	179
9.1	Evaluation and Validation Procedure	179
9.2	General Methodology.....	180
9.2.1	Co-writing Wiki system.....	181
9.2.2	Questionnaires used for evaluation and validation	183
9.3	Study R8.2: Business Course at Curtin University.....	187
9.3.1	Method.....	187
9.3.2	Evaluation Results	187
9.3.3	Validation Results.....	190
9.3.4	Conclusion	191
9.4	Study R8.3: Computer Science Course at TUG (ISR)	192
9.4.1	Method.....	192
9.4.2	Evaluation Results	194
9.4.3	Validation Results.....	199
9.4.4	Conclusion	202
9.5	Study R8.4: Psychology Course at Graz University (KFU)	203
9.5.1	Method.....	203
9.5.2	Evaluation Results	206
9.5.3	Validation Results.....	214
9.5.4	Conclusion	217
10	R9. Assessment in Self-Regulated Learning	220
10.1	Evaluation and Validation Procedure	220
10.2	Method.....	222
10.2.1	Participants.....	222
10.2.2	Design	223
10.2.3	Apparatus and Stimuli	223
10.2.4	Procedure	225
10.3	Evaluation Results	227

10.3.1	Relevancy of extracted concepts	227
10.3.2	Quality of questions	229
10.3.3	Difficulty of questions	233
10.3.4	Usability of the AQC integrated into IWT.....	234
10.4	Validation Results	235
10.4.1	Motivational aspects and task value.....	235
10.4.2	Support of self-regulated learning.....	236
10.5	Conclusion	237
11	Final conclusions	239
11.1	Overview	239
11.2	Scenarios.....	240
11.2.1	R1. Upper Level Learning Goals.....	241
11.2.2	R2. Knowledge Model contextualization	242
11.2.3	R3. Semantic Connections between Learning Resources	244
11.2.4	R4. Live and Virtualized Collaboration	245
11.2.5	R5. Storytelling	249
11.2.6	R6. A Serious Game for Civil Defense.....	249
11.2.7	R7. Affective and Emotional Approaches.....	250
11.2.8	R8. Enhanced Wiki-Test and Peer-review for writing assignments	251
11.2.9	R9. Assessment in Self-Regulated Learning	253
	References.....	256
	Annex A – Integration of IWT tools with real context of learning	258
	A1 Integration at UOC site	258
	A2 Integration at MOMA site	264
	A3 Integration at TUG site.....	266

1 Introduction

This report describes the results of the second round and final experimentation, evaluation and validation activities of the project ALICE within Work Package 8 integrated in IWT [1].

The aim of ALICE [2] 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.

Following this, the aim of this report is to present the results of the execution of the second round of experimentation and validation plan of the research and technology developed in ALICE reported in the Experimentation and Validation Plan in Work Package 1 [3]. 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. In order to evaluate all the scenarios, analyze its effects in the learning process and compare the results with those reported in the first round of the experiments (see [6]), we will follow the same methodology of the first experiments.

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.

Therefore, this document reports the final results of the experimentation, evaluation and validation of ALICE prototypes and 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 scientific work packages base their research goals on [1] 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 [5] and Table 1)

Step	Description/Questions to be considered
1. Evaluation and Validation Procedure	Which are the goals and the corresponding hypotheses to be verified for evaluating the functional requirements? Which are the criteria and the corresponding metrics to be considered for validating the methodology?
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, ...)

Step	Description/Questions to be considered
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),...)
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) [5]

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 sub-section summarizes all 9 scenarios experimented and located in the 3 pilot sites.

1.2.1 Experimentation at UOC site

The following four scenarios (see [3]) 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 [4]).

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 sections.

Study	Description	Schedule
Study R1	Experimenting with the IWT-ALICE classroom on Upper Level Learning Goals	April-May 2012
Study R2	Experimenting the Knowledge model contextualization from the instructor's viewpoint	June 2012

Study	Description	Schedule
Study R3	Experimenting with the IWT-ALICE classroom on Semantic connection between learning resources	April-May 2012
Study R4	Experimenting with the Live and Virtualized Collaboration at UOC	March-June 2012

Table 2: Overview about the studies at UOC

1.2.2 Experimentation at TUG site

The following three scenarios (see [3]) 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 sections.

Study	Description	Schedule
Study R2	Experimenting the Knowledge model contextualization from the instructor's and student's viewpoint	April-May 2012
Study R8	Experimenting the co-writing WIKI at TUG Graz	April-May 2012
Study R9	Experimenting the automatic question creator and the co-writing WIKI in Self-regulated learning	April-May 2012

Table 3: Overview about the studies at TUG

1.2.3 Experimentation at MOMA site

The following three scenarios (see [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 with 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 sections.

Study	Description	Schedule
Study R5	Experimenting the Storytelling Learning Object within an IWT-ALICE classroom on procedure to be performed in case of emergency	May-June 2012
Study R6	Experimenting the Serious Game within an IWT-ALICE classroom on procedure to be performed in case of emergency	May-June 2012
Study R7	Experimenting the Emotional tool within an IWT-ALICE classroom on procedure to be performed in case of emergency	May-June 2012

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 goals 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 [7]). As a result, the ULLG recommendation algorithm provides suitable learning resources that meet the learners' needs.

A similar experimentation process to the previous version of this scenario is performed to validate the improvements made in ULLG recommendation algorithm in this new version.

2.1 Evaluation and validation procedure

The purpose of the second experimentation phase is to satisfy all the scenario goals and criteria that are not completely covered in the first phase.

To experiment with the upper level learning goals, we focused on the evaluation hypotheses in correspondence of the scenario goals and the metrics for fulfilling specific criteria as described in [3]:

Scenario goals

- G1.1: to develop a ULLG recommendation algorithm able to generate a set of feasible learning goals starting from a need expressed in natural language by the learner.
- G1.2: to ensure that generated learning goals cover the expressed needs and the (optionally) selected skills and contexts (taking into account the available learning material).
- G1.3: to ensure that the generated courses are personalized on the basis of learner cognitive state and learning preferences
- G1.4: to provide a user friendly interface for needs expression, learning goals generation, courses preview and course selection.
- G1.5: to ensure that generated courses allow the effective learning of scientific concepts in selected domains.
- G1.6: to identify possible ways of improving further the utility of the ULLG.

Scenario hypotheses

- H1.1: a set of feasible learning goals 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 ULLG contributes to improve students' motivation.
- H1.3: the use of the ULLG contributes to improve students' understanding of domain concepts.
- H1.4: the use of ULLG contributes to increase students' activity levels.

- H1.5: the use of the ULLG 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.

Scenario criteria

- C1.1: To evaluate the level of fulfillment of the tool features.
- C1.2: To evaluate the level of satisfaction of the students that use the ULLG.
- C1.3: To evaluate the increase in students' motivation caused by the use of the ULLG.
- C1.4: To evaluate the increase in students' understanding of key concepts and students' results caused by the use of ULLG.
- C1.5: To evaluate the increase in students' activity levels due to the use of the ULLG.
- C1.6: To evaluate the level of satisfaction of the instructors with the inclusion of the ULLG 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 ULLG.

Scenario metrics

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

2.2 Method

2.2.1 Participants

In order to evaluate this scenario to analyze its effects in the learning process and compare the results with those reported in the first round of the experiments (see [6]), we will follow the same methodology of the first experiments.

The methodology considered 151 students enrolled in the course Software Engineering from the Bachelor in Computing Engineering in the Spring term of 2012 at the UOC participated in the experience. Most of them (142) were from the Bachelor in Computing Engineering and a small group (9) was from the Master in Computing Engineering. Both Bachelor and Master share the same course "Software Engineering" in its curricula.

The students were roughly distributed equally into 2 classrooms in the UOC virtual campus, 77 and 74 students each.

61 out of 151 students (40.3%) 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, 59.7% of them chose not to send the evaluation form and thus they were excluded from the analysis.

29 out of 151 students (19.2%) 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 29 students belonged to the group of 61, which left a group of 32 who participated by submitting the form but did not use the IWT prototypes.

From the 61 participants we formed 2 groups for the experiment. One experimental group with 29 students who use IWT (47.5%) and one control group with 32 students who did not use IWT at all (52.5%). All of them submitted an evaluation form at the end of the experience.

Therefore, the sample of the experiment was formed by 61 students. For the sake of the experiment, we were only interested in the conglomerate of the experimental group formed by 29 students. 27 students were male (93.1%) and 2 students were female (16.9%). The 32 students forming the control group studied at UOC only and did not use IWT at all. 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).

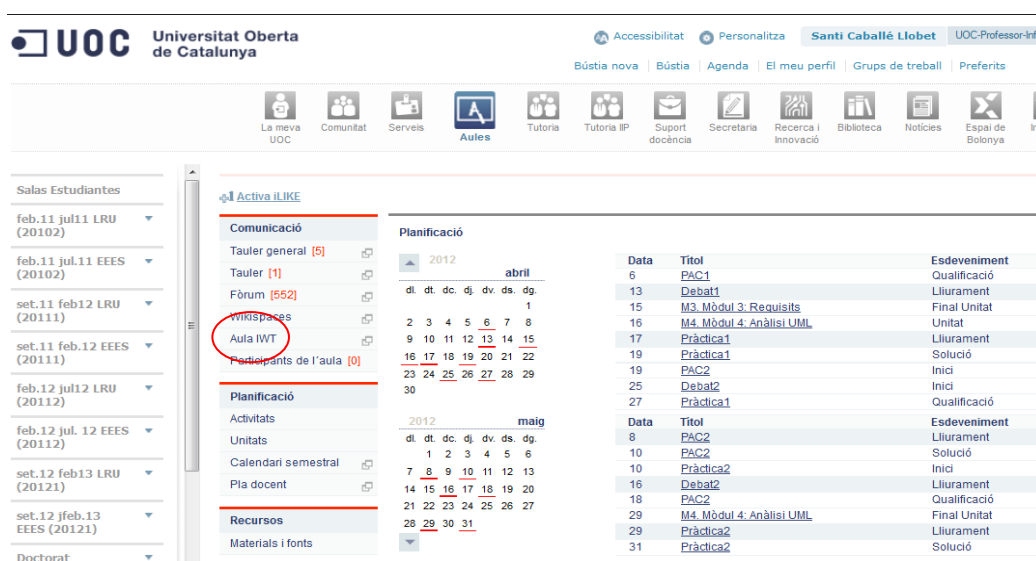


Figure 1: UOC classroom with the access to IWT classroom

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

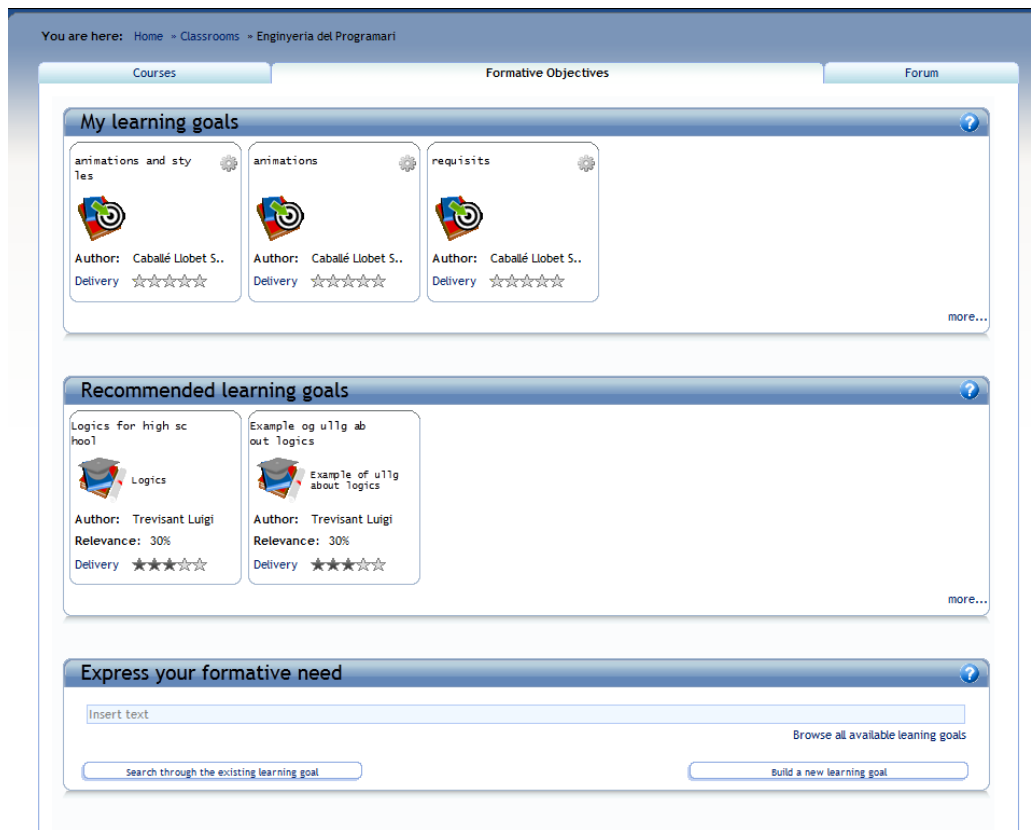


Figure 2: IWT classroom with a list of options to personalize and manage learning goals

In this scenario there are different functionality provided by the R1 prototype (see [7] for a full description):

My learning goals: allows the learner to view their personal formative needs.

Recommended learning goals: allows the learner to view a set of ULLGs the system suggests for him thanks to the recommender system integrated within ALICE.

Express your formative needs: it allows the learner to indicate in natural language the learning goals he/she wants to build (*Build a new learning goal*) and to verify what are the most suitable (see Figure 3 and Figure 4) and also view the complete collection of the available ULLGs (*Search through the existing learning goals*).

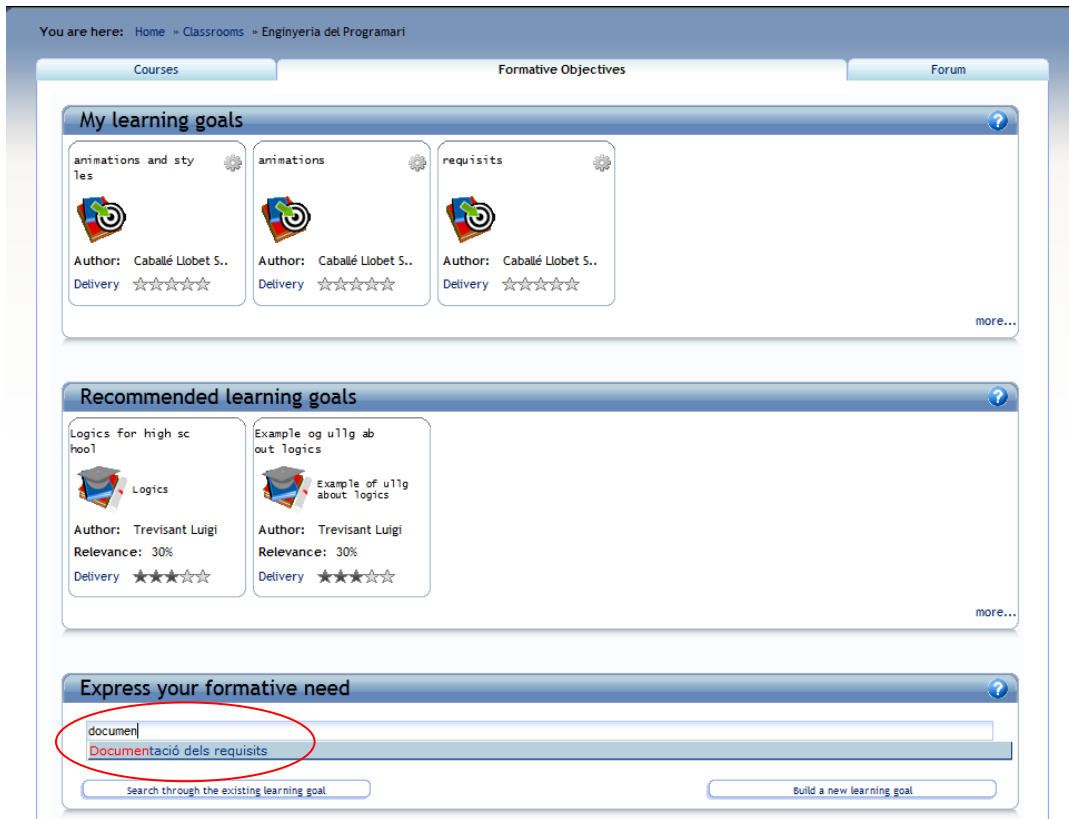


Figure 3: Express your formative need. It allows the learner to indicate in natural language the learning goals he/she wants to achieve and the request suggested by the system

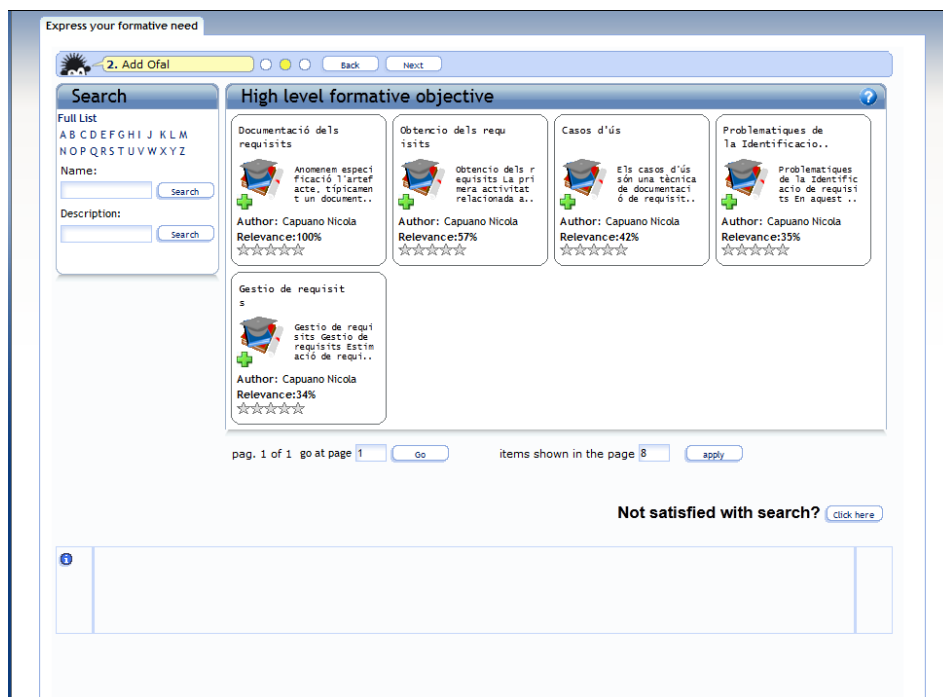


Figure 4: List of the resulting learning resources

We used the SUS (System Usability Scale [8]) 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 [8] 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) [9]. 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 Spring term (April 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 by the ULLG recommendation algorithm 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 ULLG. 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 2.1, in this section we focus on the activity, usability and emotional aspects of the IWT tool (H1.1 and H1.4) by using metrics M1.1 and

M1.4. 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 2.4 (Validation Results).

2.3.1 Activity levels

Expressing a formative need is an additional functionality included in the learning course (see Section 2.2), so, it is activated only if some students intend to more understand a specific topic.

Taking into account this specification, 7 out of 29 students (24%) have used the functionalities included in the R1 prototypes. So, they have expressed in a natural language the learning goals. After that the systems has suggested to the students a set of ULLG to choose for filling the learning gap. That has been obtained taking into account, for each student:

- the specific cognitive state;
- the individual competences also acquired in an other learning context;
- the background of the learning classroom.

The log file of IWT have registered a set of ULLG composed by 2 resources; that is a positive results considering also that the number of the students involved in the experimentation is quite small.

The suggested set of ULLG denotes also a relevance percentage respect to the formative need expressed by the student, that have helped him to choose the learning course more compliant whit the learning objectives.

2.3.2 Usability of the IWT

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

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 note down 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 68 would be considered above average and anything below 68 is below average. A score above 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 51 is an F (placing you in the bottom 15%).

After calculating the SUS score for each student, we got an average for the **29 SUS scores of 53.97**. Next, we present the most relevant results of the SUS scores by providing several statistics: Mean (M), Standard Deviation (SD) and Median (Md).

Analyzing student feedback, we can observe there are more students who think they would like to use the IWT more often than students who wouldn't (46% vs 31%) ($M = 3.13$, $SD = 1.09$, $Md = 3$) (See Figure 5).

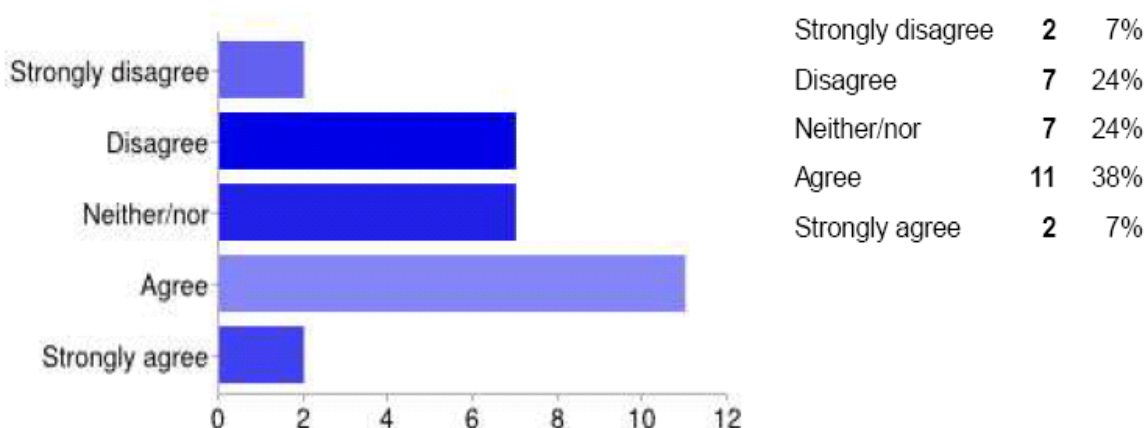


Figure 5: Results on the SUS item "I think I would like to use IWT frequently"

These results are aligned with the amount of people who think that IWT is unnecessarily complex ($M = 3.10$, $SD = 1.11$, $Md = 3$) (See Figure 6). Reasons that could explain this result could be the opinion of many students who think that there is inconsistency in the IWT interface ($M = 3.24$, $SD = 0.98$, $Md = 3$) and that IWT is not well integrated in the UOC campus. Some students reported that the interface is neither user-friendly nor intuitive.

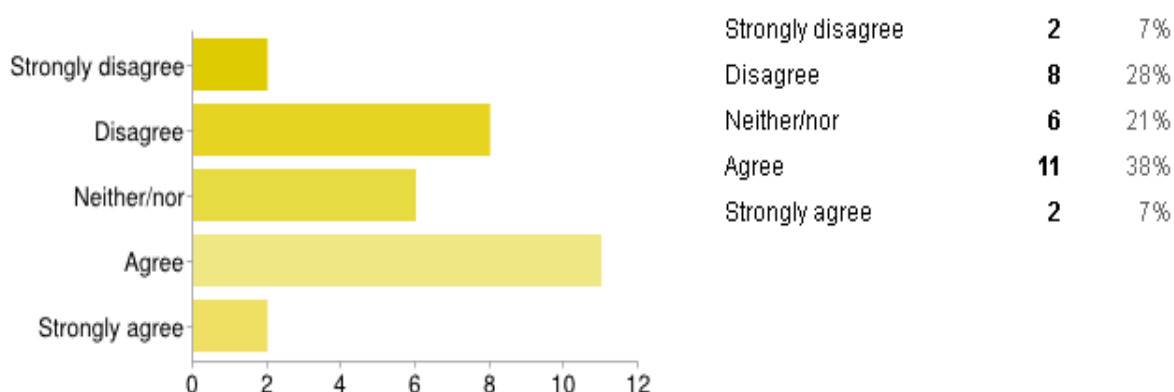


Figure 6: Results on SUS Item "I think IWT is unnecessarily complex"

A lot of students thought that IWT was going to be easy to use ($M = 3.65$, $SD = 0.81$, $Md = 4$) (see Figure 7). In addition, many students stated that they had not needed the support of a technician to be able to use IWT and that people should learn how to use IWT quickly ($M = 2.90$, $SD = 1.01$, $Md = 3$) (See Figure 8) since there is little need to learn too much to be able to use it. ($M = 2.41$, $SD = 0.95$, $Md = 2$) (see Figure 9).

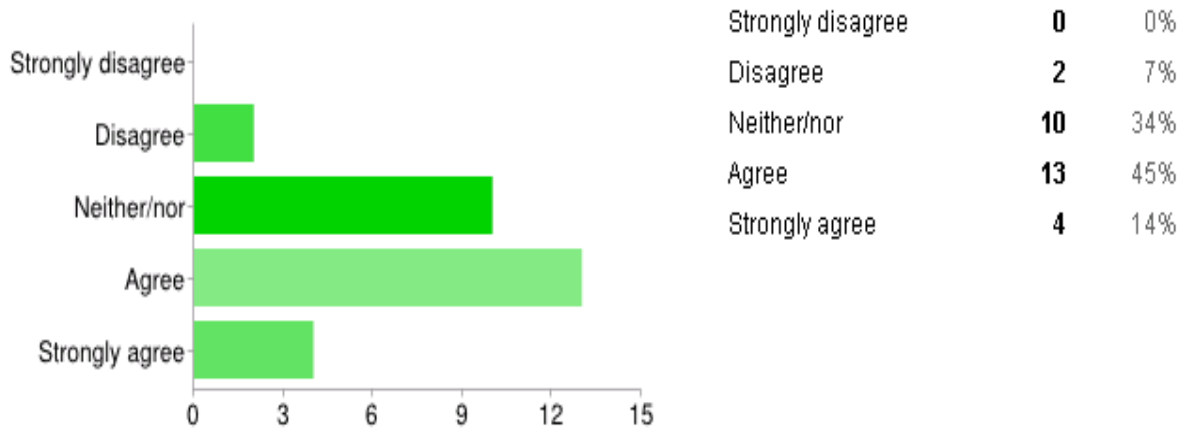


Figure 7: Results on the item "I think IWT was going to be easy to use"

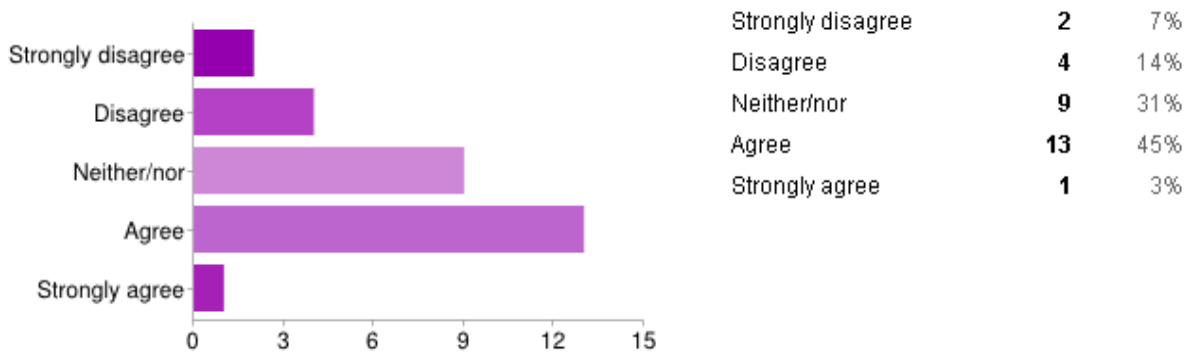


Figure 8: Results on SUS item: "It is thought that people should learn how to use IWT quickly"

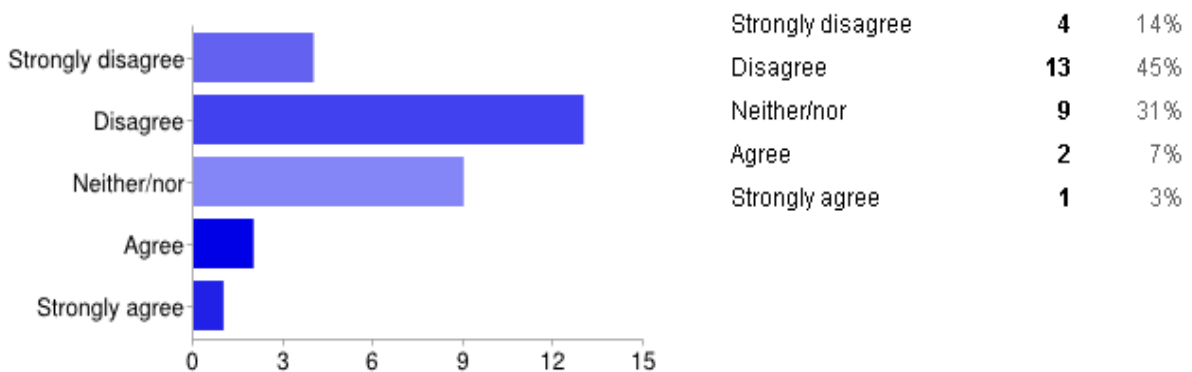


Figure 9: Results on the item "I think I don't need to learn too many things to use IWT"

2.3.3 Emotional Aspects

Regarding student emotion while working with the IWT tool (H1.1), we have 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 in a 4-point rating scale (n=29) have been as follows:

- Happiness (M = 1.51, SD = 0.83, Md = 2)

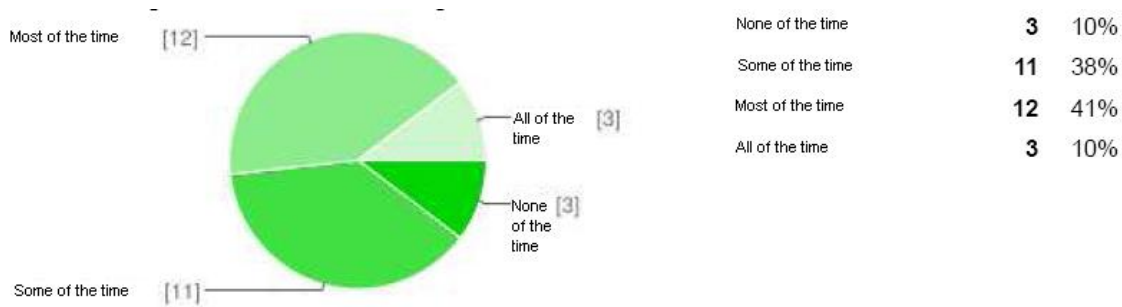


Figure 10: Results on the Happiness emotion

- Sadness (M = 0.62, SD = 0.68, Md = 1)

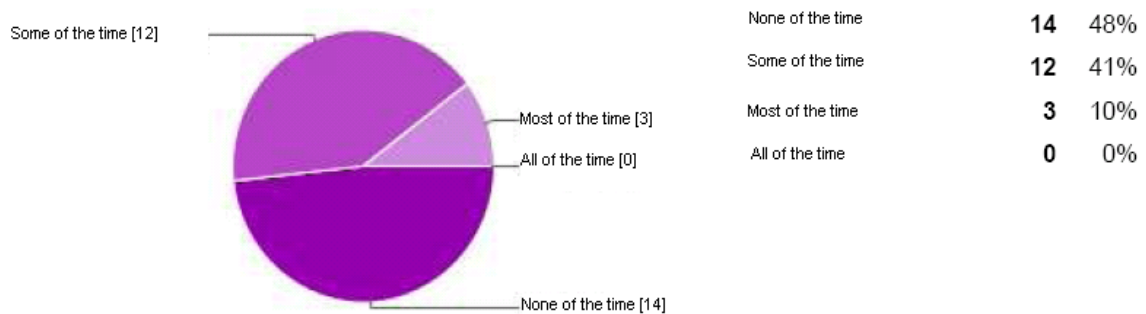


Figure 11: Results on the Sadness emotion

- Anxiety (M = 0.55, SD = 0.63, Md = 0)

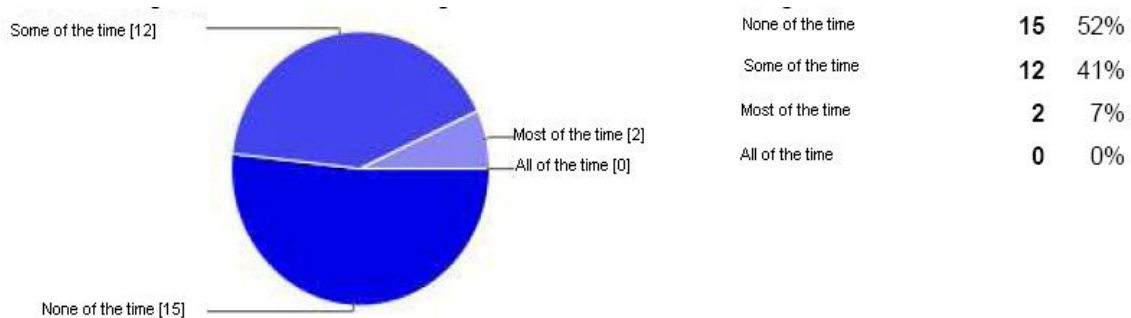


Figure 12: Results on the Anxiety emotion

- Anger (M = 0.34, SD = 0.55, Md = 0)

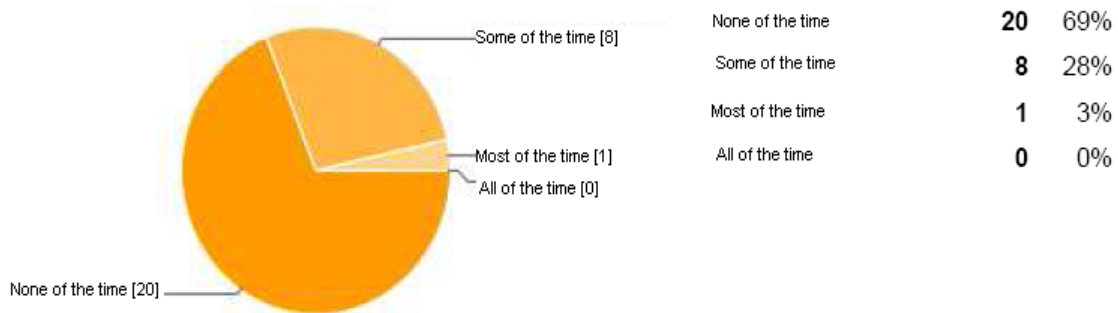


Figure 13: Results on the Anger emotion

The Happiness emotion appears most of the time (Figure 10) and much more than the rest of emotions, which are low in general. Thus, there are more people who are happy most of the time than sad.

Almost 70% of the students have not experienced anger at any time (Figure 13) and anxiety is very low (Figure 12). This result is aligned with the usability results, which indicate that, in general, people have not had problems when dealing with the IWT as a new environment and have managed quite well without any additional help.

If we compare these results with the results of the first iteration, there are now more people who are happy most of the time and so, less people who are sad, anxious or angry. This fact can be explained because IWT has been improved compared to the previous version.

2.3.4 Evaluation of Questionnaire

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

The results of the evaluation of the design of the questionnaire have confirmed, like in the first iteration that the time employed to fill the questionnaire in is less than 30 minutes for most of the students (72%) (Figure 14) and although most of the students think that the questionnaire is appropriate to evaluate the experience (Figure 15), some students stated that the questionnaire was long and heavy.

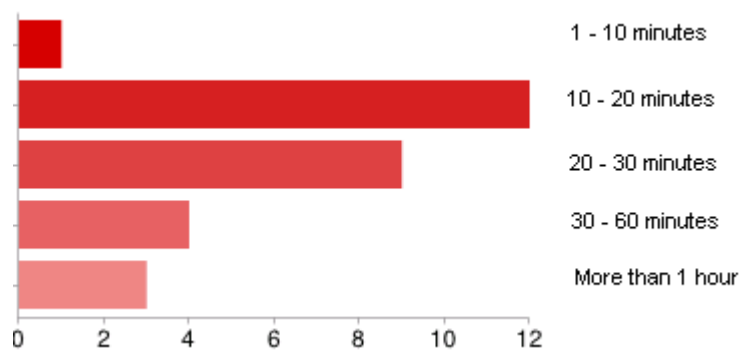


Figure 14: Time employed to fill in the questionnaire

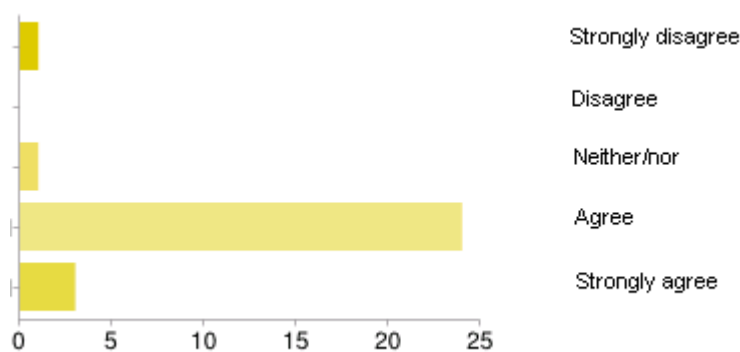


Figure 15: Appropriateness of questionnaire to evaluate the experience

2.4 Validation Results

Following the methodology described in Section 2.1, we validate next 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). For these purposes we used metrics M1.1, M1.3, M1.5, M1.6 and M1.7.

2.4.1 IWT as a valuable resource

This section analyzes IWT as a valuable educational resource through the evaluation of its worthiness as an educational tool (H1.6). To this end, quantitative and qualitative data have been collected in sections (iii) and (iv) of the questionnaire through 3 open questions (qualitative) and then 13 test-based questions (quantitative) in addition to one final open question to provide suggestions for improvement.

With respect to the rating scales of the three quantitative questions in the questionnaire, we have used a 0-10 point scale so that students could assess the value of the IWT tool through a scale that 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 mark from 5.0 to 10 and a “bad” assessment mark 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). Although, these values sometimes changed to best fit the expected type of responses and in all cases, 3 options were provided (positive, medium and negative).

Open-questions

Three open questions were posed 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 made an impact on 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 with the UOC classroom, what advantages and disadvantages do you think IWT provides to the study? Indicate in your view what are the main problems, issues and weaknesses of this tool (Assess IWT from this view in the scale 0-10).

After calculating the 0-10 scale for each student, we got an average of 6.20 (SD=2.01, Md=6). This result is good and is slightly better than in the first iteration.

Regarding Question 1, students have liked, in general, the IWT system and have found it useful for their study (M=6.21, SD=2.02, Md=6). Many students reported that the self-evaluation capabilities such as on-line tests within the same environment have been beneficial to improve the learning process. Students have appreciated these evaluation tests very much to self-evaluate if they have assimilated the content. In addition, some students stated that these tests had been important to foster the learning process.

On the other hand, some students have agreed that the user interface could be improved, especially, in terms of usability but also in complexity. Some of them stated that some explanations in the material of the IWT course were already found in the regular materials of the UOC course.

In Question 2 (M=6.17, SD=2.37, Md=6), students have focused basically on the IWT course. Although some students think that part of the content presented in the environment is redundant with the content they had in the UOC course, most of the students agreed that the study environment complements the UOC material. Students agreed, as well, that the fact of having tests is a good resource to test your knowledge assimilation. They have valued the fact that the information and the tests are integrated in the same environment. However, they indicated that, although the self-evaluation exercises were very useful, they reported that studying with IWT is not easy and they did not improve their knowledge significantly.

In Question 3 (M=6.21, SD=1.97, Md=6), a lot of students once again considered the self-evaluation questionnaires as being very important for their learning process in order to clarify doubts and to assimilate concepts. The separation in sections and subjects and the fact of having self-evaluation exercises per section help them to progress without getting swamped. One of the advantages of IWT is that it is seen as a compact study environment: all the learning resources in the environment are at hand and this is perhaps why many of them

consider it as the most ideal environment to study in a better way rather than separating theoretical content and evaluation.

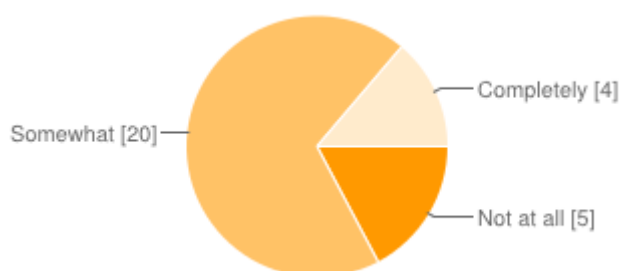
On the other hand, some students agreed that IWT platform is not so intuitive compared to the UOC classroom. This thought is logical because both environments are very different and students need time to get used to it. Some students commented that the UOC campus graphical user interface was clearer and easier to use than IWT. However, some students prefer IWT because they think it is more powerful. Although it is a competent study environment, one of the drawbacks is that you need to be online to study and sometimes this is not possible.

Test-based questions

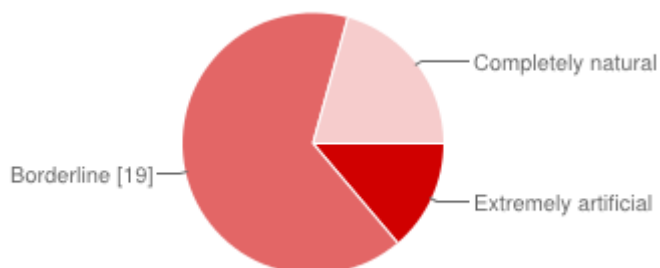
The quantitative results could be checked.

13 test-based questions were posed to students:

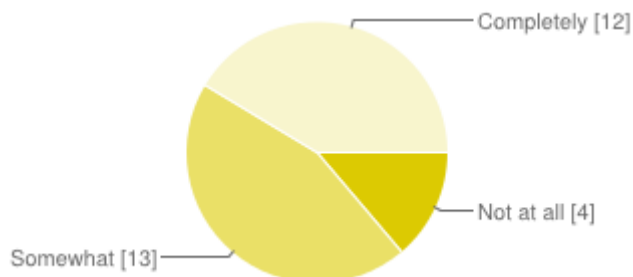
1. The possibility to express your formative needs has allowed you to have more control over your learning?



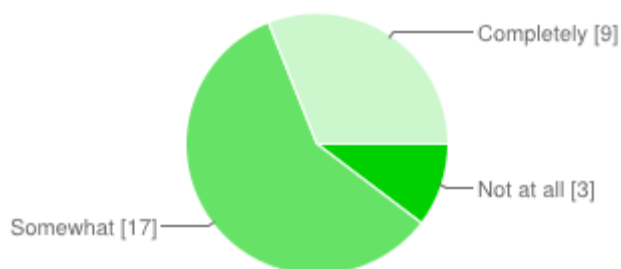
2. Being able to express your needs in a simple language has contributed to motivate your desire to learn?



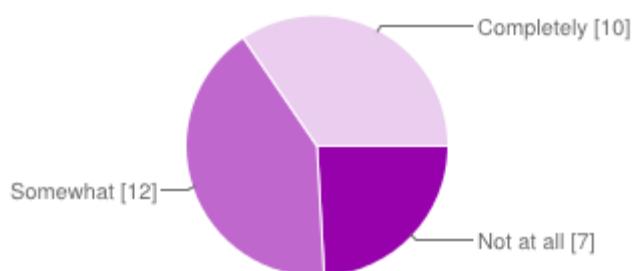
- Do you think that this solution of asking to take more responsibility about what you need has helped you to capture a greater awareness on the right learning path?



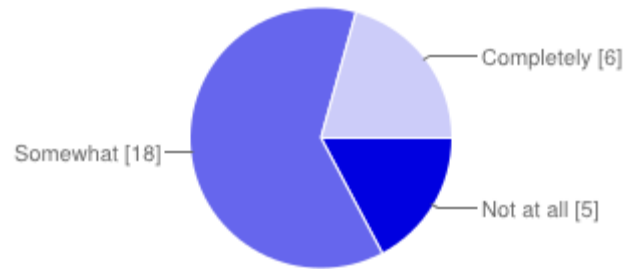
- Do you think that the answers obtained in terms of learning paths to follow by fulfilling your needs are relevant and effective?



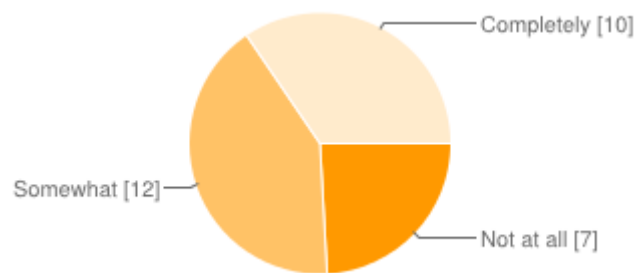
- Do you think you can shorten the learning time by eliminating states to which you are subject to when your path is guided by the teacher?



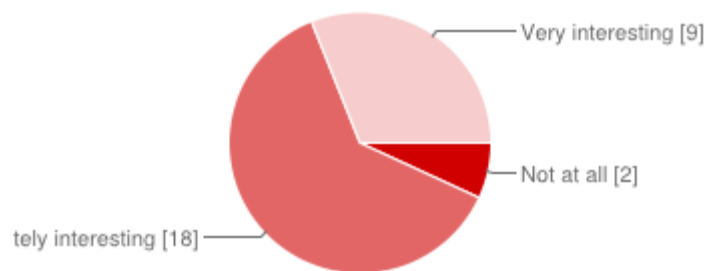
6. The possibility of having a specific learning path created ad hoc to fulfil your needs has allowed you to obtain good results in terms of learning.



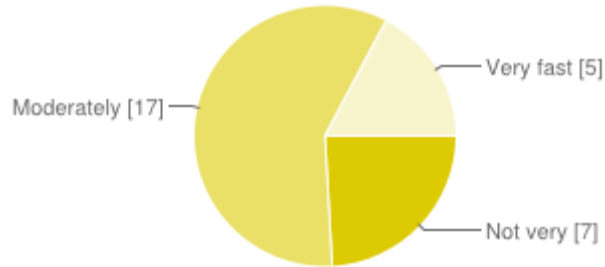
7. Has this learning modality had an impact on your participation in the learning experience?



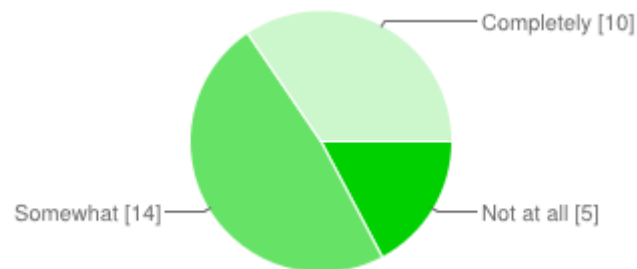
8. How have you found the interaction with this new method of learning experience?



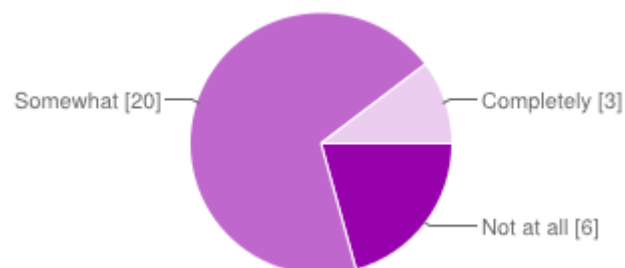
9. How quickly have you adapted to this new method of expression through natural language?



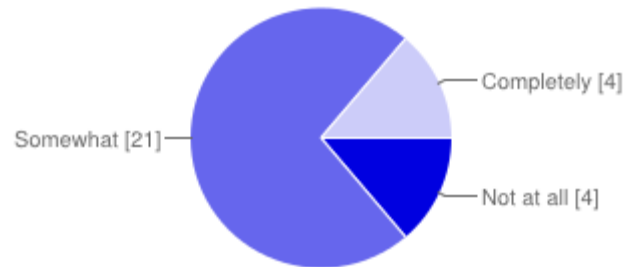
10. Do you think that this new kind of interaction modality of a student-learning environment can be a step towards a self-regulated learning?



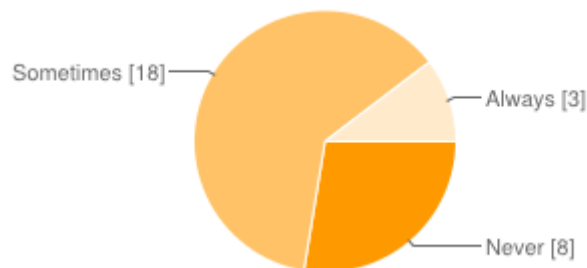
11. Do you think that the recommendations you have received in terms of the learning path to follow were tailored to your learning style and your profile?



12. How do you consider the proposed learning path has fulfilled your learning needs?



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



Final open question for improvement

This open question completed this section of the questionnaire by asking students to give final hints for potential improvement of the IWT tool.

A group of students commented that the environment is not sufficiently intuitive and is not user-friendly. They suggested an online guide to get used to the environment more quickly. They also commented that it is necessary to improve the navigability of the IWT website and solve some problems with the user interface (GUI), however, they think that IWT has a lot of potential.

Students liked to have content and tests within the same study environment. However, some students commented that the drawback is that you need to be online. Structuring the material in separated sections is appreciated to manage the different parts of the course. It was also pointed out that some contents are very similar in the UOC materials. This is considered normal given that both courses were about the same topic.

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 ($M=3.02$, $SD=0.67$, $Md=3$). This result is in line with the results on the IWT being a valuable resource and also in line with slightly improvement from the results of the first phase of experiments ($M=2.79$, $SD=0.81$, $Md=3$). In addition these results are in line with the usability and emotional results reported in the previous sections. In particular, students indicated to feel very motivated by the self-evaluation 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 motivation and engagement came from passionate students who made very positive comments, such as "IWT is a magnificent environment", and "I liked very much the idea to combine study material and self-evaluation, great!". However, most of them clarified that the system needed usability improvements, perform better with a more fluent navigation and compatibility with mobile devices before considering IWT to be successful. Eventually, most of students understood it was a pilot trial and for this reason they showed their motivation from the perspective of a potential tool with needed improvements rather than a consolidated tool.

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=29)	Control group (n=32)
Question 1	M=6.11 SD=1.87 Md=6	M=5.84 SD=1.31 Md=6
Question 2	M=7.81 SD=1.28 Md=8	M=7.32 SD=1.24 Md=7
Overall	M=6.96 SD=1.57 Md=7	M=6.58 SD=1.27 Md=6

Table 5: Results of the learning assignment evaluation

From the results of *Table 5*, students from the experimental group (UOC + IWT) scored slightly higher than the control group (UOC only). The scores are also slightly better in comparison to the first phase of the experiments by passing the same cognitive evaluation process (i.e. same evaluation questions), though the overall difference is not significant (except for Question 2 of control group that scores 1 point more the same question and group of the previous experience).

Observing closer the results, the experimental group got more dispersed marks than the control group (SD=1.57 versus SD=1.27) and also more than the previous experimentation for the experimental group. We suggest that the higher number of participants for the first phase of experiments (n=41 vs. n=27) mitigated the outliers also found in the second phase. Most interestingly, this result uncovers and confirms a higher dispersion of knowledge of the experimental group due to these students having to study with 2 very different environments and different material with the related dispersion of concepts. This result confirms the previous validation “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.”

In line with the results of previous experimentation, both groups got also good marks on average in this second round of experiments 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 that the IWT did not improve their knowledge significantly.

In summary, we conclude that IWT did not provide students with significant amount of new knowledge but it managed to satisfy the needs expressed by them (i.e. students met specific knowledge needs by using the ULLG recommended 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). A comparison with the results of the first round of experiments is also provided.

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 ULLG recommender system 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, especially as for the acquisition of new knowledge. However, the good marks achieved by the experimental group shows that the generated IWT courses from the students' request were personalized on the basis of learner cognitive state and learning preferences (G1.4). These results are in line and confirm the first round of validation.

In addition, in line with the first round of experiments, the system usability was not a barrier when using the system (G1.4) though it was again the most important technical aspect considered by students. Even so, the usability improvements made from the previous phase of the project (e.g., new navigational panels, automatic searching suggestions, etc) were noticeable by students who did not report any more on particular usability aspects that had influenced negatively their emotions during the previous experiences. Finally, it was still noticeable important amounts of resilience to change the e-learning platform from UOC to IWT partially due to the usual learning curve when facing a new system.

Validation of the impact of IWT in effective learning of scientific concepts was analyzed and evaluated (G1.5) by chiefly Section 2.4.3 on assessment. It was concluded that IWT did not provide students with significant amount of new knowledge but it managed to satisfy the needs expressed by them (i.e. students met specific knowledge needs by using the ULLG recommended system).

Finally, possible ways of improving further the utility of the ULLG (G1.6) and al larger extend of IWT were provided in several sections, and mainly at the end of Section 2.4.1 being most of the comments still addressed towards usability, but also towards improving system performance and compatibility with mobile devices.

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 [7]). Two pilot sites run three trials on this scenario: two trials from the instructor's viewpoint and a third trial was run from the students' viewpoint. In summary:

1. Knowledge model contextualization: Experimenting the Knowledge model contextualization from the instructor's and student'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 and student's viewpoint (TUG)

3.1.1 *Evaluation and Validation Procedure*

In the first phase of experimentation regarding knowledge model contextualization, we conducted an experiment at TUG pilot site in order to test the tool from the instructors' viewpoint. So we were primarily interested in the functionality and usability of the tool or rather whether the tool supports instructors in creating online courses.

In the second phase we repeated the experimentation of the first phase in order to indicate improvements of the tool. Apart from the repetition of the first experiment, we also involved students in the second phase of experimentation. To test the knowledge model contextualization, we consulted students with two different contexts, beginner and advanced. In order to assign the students to these contexts, the lecturers provided additionally to the dynamic course also a static course.

Scenario goals

- G2.1.1: to develop a Visual Ontology Editor (VOE) for the definition of domain ontologies and contexts with a user friendly interface.
- G2.1.2: 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.
- G2.1.3: to ensure that generated courses are personalized on the basis of the learner cognitive state and learning preferences.

- G2.1.4: to ensure that generated courses allow the effective learning of scientific concepts in selected domains.
- G2.1.5: to identify possible ways of improving further the utility of the VOE and related models and algorithms.

Scenario hypotheses

- H2.1.1: a set of feasible courses can be effectively and efficiently created starting from a domain ontology by selecting a context, a set of target concepts and a learner.
- H2.1.2: the automatically generated courses are compatible with the selected context and are in line with student needs, previous knowledge and learning preferences.
- H2.1.3: the use of automatically generated courses contributes to improve students' motivation.
- H2.1.4: the use of automatically generated courses contributes to improve students' understanding of domain concepts.
- H2.1.5: the use of automatically generated courses contributes to increase students' activity levels.
- H2.1.6: automatically generated courses are considered as a worthy educational resource by both instructors and students.

Validation Criteria

- C2.1.1: To evaluate the level of fulfilment of the tool features.
- C2.1.2: To evaluate the level of satisfaction of the instructors that use the VOE.
- C2.1.3: To evaluate the level of satisfaction of the instructors with the inclusion of the contextualized courses with their students.
- C2.1.4: To evaluate the increase in students' motivation and understanding of domain concepts caused by the use of contextualized courses.
- C2.1.5: To evaluate the increase in students' activity levels due to contextualized courses.
- C2.1.6: To evaluate the level of satisfaction of the students that use contextualized courses generated by the system.

Scenario metrics

- M2.1.1: Number of instructors using the VOE.
- M2.1.2: Number of courses created with contextualized ontologies.
- M2.1.3: Time employed in creating each course with contextualized ontologies.
- M2.1.4: Number of students passing the final test and/or with high marks when they use contextualized courses.
- M2.1.5: Number of students passing the final test and/or with high marks when they do not use contextualized courses.
- M2.1.6: Instructors that consider that the VOE and contextualized courses are worthy.

- M2.1.8: Students that consider that the VOE and contextualized courses are worthy.

3.1.2 Method

3.1.2.1 Participants

Two lecturers, one from the Karl-Franzens University (KF) (lecturer A) and one lecturer from the Graz University of Technology (TUG) (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 KF University and lecturer B has been working for 13 years at the TUG. According to their experiences with learning platforms, lecturer A has only basic knowledge and lecturer B has advanced knowledge using learning platforms.

Furthermore, 8 students from TUG participated in the experiment. Participants were between 23 and 29 years old, on average they were 25 years old ($SD = 2.03$). Six of the students are male and 2 of them are female. Concerning the highest level of education, six students finished their Bachelor, two of them reached a Master degree. Four students had little previous knowledge about Scientific Working, whereas the other four ones had advanced previous knowledge regarding the topic of the course.

3.1.2.2 Apparatus and Stimuli

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 a Post-Questionnaire concerning the usability of IWT. The Questionnaire included the following sections: SUS (System Usability Scale), open questions regarding the usability of IWT, functions on IWT and emotional aspects.

We used the SUS (System Usability Scale) by Brooke (1996) [8] 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%).

To investigate in which emotional mood the lecturers were when they used IWT, we added the section “emotional aspects”, which includes 12 items. Kay and Loverock (2008) [9] 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”.

The students received two Questionnaires, one after the static course (Questionnaire 1) and the other one after the dynamic course (Questionnaire 2) (see Procedure section). Questionnaire 1 included the following sections: Demographic data, open questions regarding the course, SUS (System Usability Scale) and motivational aspects.

Open questions regarding the course included three questions asking (a) whether the course are line with students’ need, previous knowledge, and learning preferences, (b) whether the course contributed to improve their understanding of domain concepts, and (d) whether the course is considered as a worthy educational resource. Answers were provided on a 5 pt. Likert scale ranging from (1) I strongly disagree to (5) I strongly agree and as open comments (“Why/Why not?”).

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 Pintrich et al. (1991) [10] 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.

3.1.2.3 Procedure

The experiment consisted of three phases.

In the *first phase*, the lecturers were asked to create a contextualized course concerning the topic “Scientific Working”. This course had four contexts. On the one hand the context of the university, KF University or TUG University, on the other hand the context regarding the previous knowledge of the students, beginner or advanced. Besides, the lecturers also provided a static course, called “Introduction to Scientific Working”.

In the *second phase* the students enrolled in the static course, which was presented as a textual learning material on IWT. In order to update students' learner profile, they were asked to take a test on the static course.

In the *third phase* students enrolled the contextualized or rather dynamic course. Based on students' previous knowledge - due to the list of courses, they already participated in - the students were assigned to the context beginner or advanced. Depending on the context (beginner or advanced), the students received different and overlapping learning resources which fitted to their previous knowledge. Finally they were asked to take another test.

3.1.3 Evaluation Results

This section covers the findings from three different phases as discussed earlier in the procedure section. The results cover the findings relevant for Hypotheses H2.1.1, H2.1.2, H2.1.4, H2.1.5, and H2.1.6. The corresponding metrics used for evaluation of the Hypotheses are M2.1.1 through M2.1.8 as they are specified in the previous subsection 3.1.1.

3.1.3.1 Findings from Phase 1 (lecturers)

Two lecturers created a contextualized course with four different contexts on IWT using the concepts they had developed in the first phase of experimentation. In order to create such a dynamic course, they had to

- (1) Create a dictionary and set context
- (2) Upload learning resources
- (3) Create an ontology
- (4) Create a customized course

(1) Create a dictionary and set context

The dictionary provides the key concepts for the teaching subjects (see Figure 16). The lecturers needed 10 minutes to enter their concepts to the dictionary.

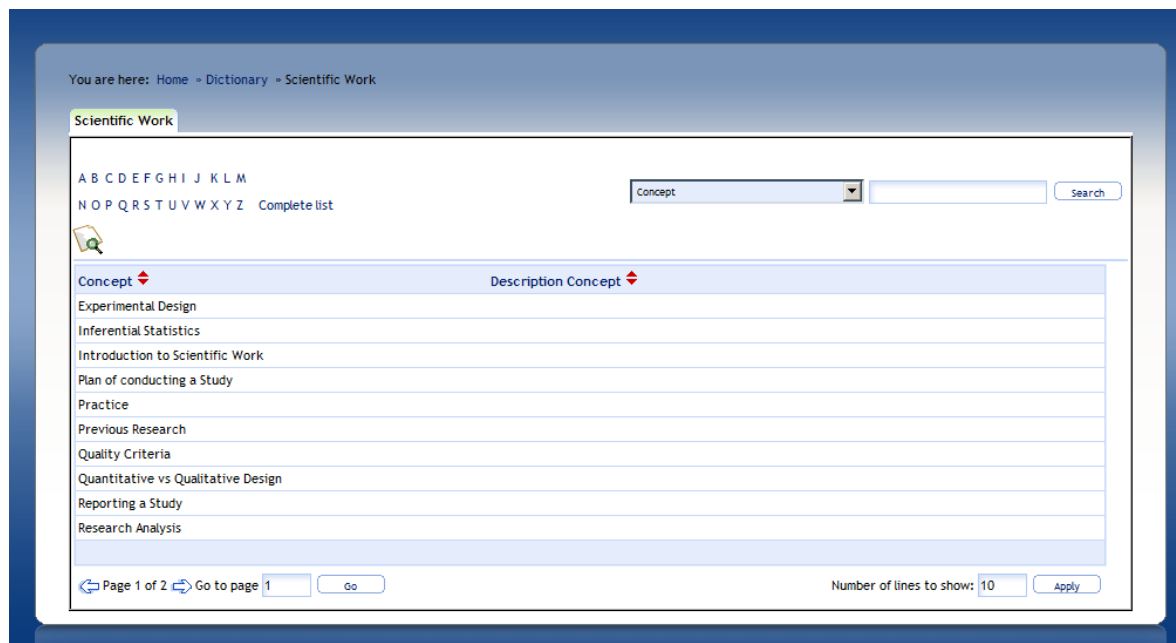


Figure 16. Dictionary “Scientific Working” with concepts

Additionally the lecturers had to set the four different contexts of the course, the context of the university (KF and TUG) and the context regarding the previous knowledge of the students (beginner or advanced) (see Figure 17). The lecturers faced problems with setting the contexts, because there was no button for creating a new context. As this action had to be approved by the technical support of IWT, it causes a waiting period of half an hour.

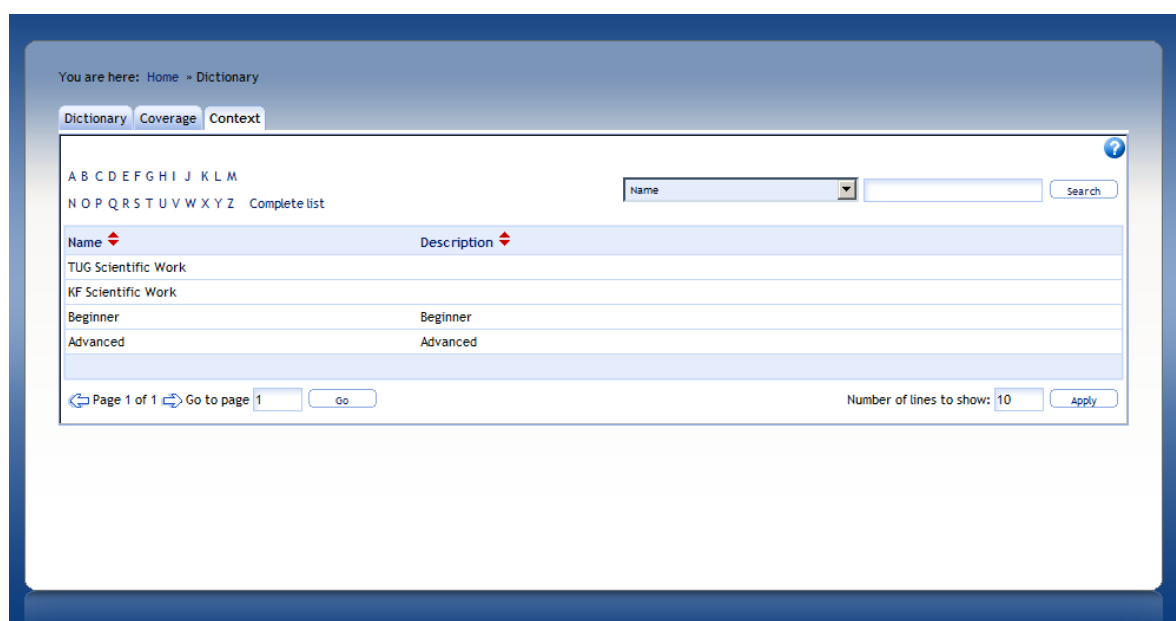


Figure 17. Setting the Contexts of the Course

(2) Upload learning resources

Lecturer A provided the content for the static as well as for the dynamic course and shared these learning resources with Lecturer B (see Figure 18). Lecturer A spent 25 minutes on uploading the learning resources. The lecturers needed technical support in order to find the function of sharing learning resources and transfer or rather copy them to another account.



Figure 18. Uploaded contents for the courses

(3) Create an Ontology

The available concepts from the dictionary “Scientific Working” were arranged in a specific order. Furthermore, the lecturer added the relations “has part”, “is required by” and “suggested order” to the concepts. The context data were provided and assigned to the concepts (see Figure 19).

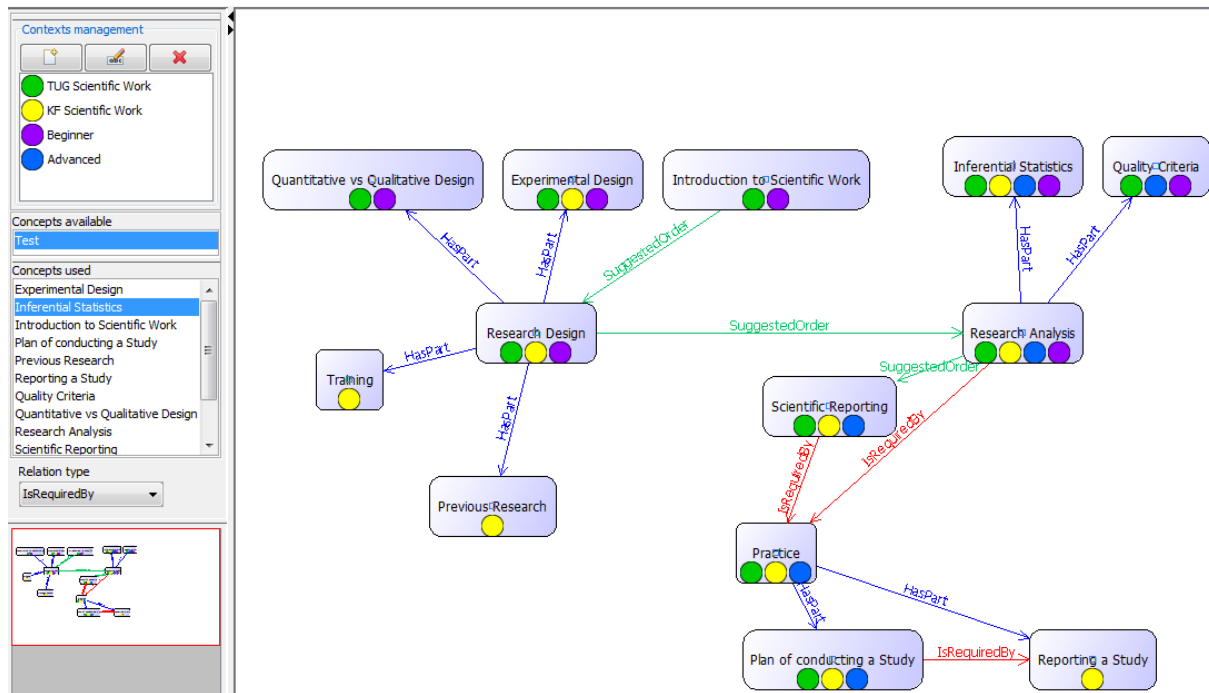


Figure 19. Contextualized ontology with the concepts, the contexts and the relations

Apart from the context regarding the university, the context concerning students' previous knowledge was also assigned to the concepts. In the first phase of experimentation, the lecturers just added the context TUG University or KF University. In this study, the lecturers also considered students' previous knowledge and differed between beginners and advanced. Table 6 shows which concepts have to be learned by beginners and/or advanced students.

For creating the ontology the lecturers needed 1 hour and 40 minutes. First of all, the ontology could not be displayed. After calling the technical support of IWT, some internet options had to be changed and displaying the ontology worked.

Concept	Beginners	Advanced
Research Analysis	✓	✓
Inferential Statistics	✓	✓
Quality Criteria	✓	✓
Research Design	✓	
Quantitative versus Qualitative Design	✓	
Experimental Design	✓	

Plan of Conducting a Study		✓
Practice		✓
Scientific Reporting		✓
Test	✓	✓

Table 6: The contexts “Beginners” and “Advanced” assigned to the concepts

(4) Create a Customized Course

Finally, the lecturers created the customized course and defined the following target concepts: Research Design, Research Analysis and Plan of Conducting a Study (see Figure 20). Also the didactic approach (i.e., the didactic path, the language, the teachers defined profile etc.) were settled.

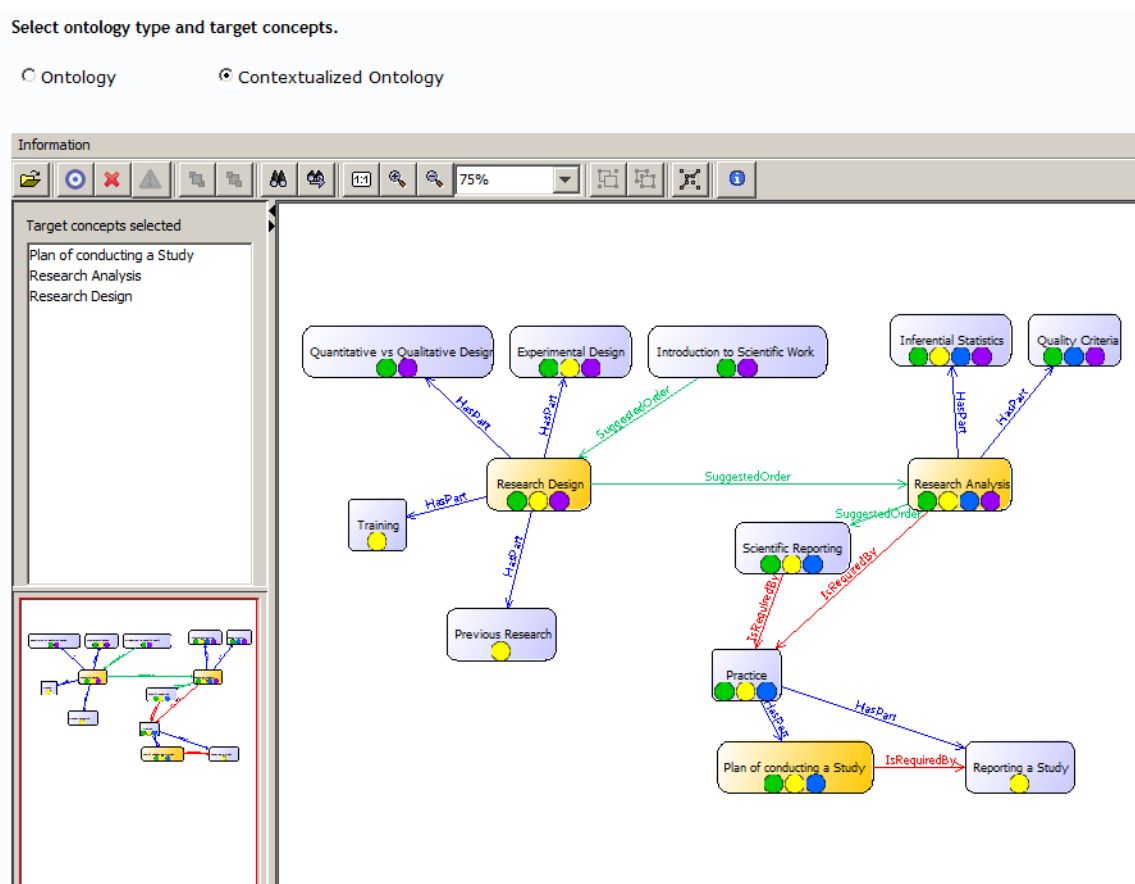


Figure 20. Target concepts for the course “Scientific Working”

The lecturers were instructed to use the manual of the IWT as provided in [7]. 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.

Usability of IWT (from the instructors' point of view)

We calculated the SUS score separately for each lecturer. The score for lecturer A was 45 and the score for lecturer B 37.5 and belong to the bottom 15%. The lecturers also were asked what they liked regarding the system. Lecturer A is in favor of the idea of the online tool and its functions, whereas lecturer B mentions that IWT has a great editor to create ontologies. Regarding the disadvantages of the system, both lecturers criticized that they needed the support of the technical stuff to create the course with all its properties. In addition, lecturer A stated that it was very difficult for her to find all functions and that she spent a lot of time on searching for the functionalities. Concerning improvements, lecturer B states that popup for ontology creation should also work in IE and that in case of errors the ontology editor should be editable. For lecturer A the system doesn't seem user-friendly and in her opinion the system should be designed less complex and with detailed information explaining the functions. Furthermore, the "Help" button should be available and they mention that in general more instructions would be helpful (H2.1.1).

Regarding the user manual, lecturer B states that it was supportive, but although he needed the help of the technical stuff, so he thinks that it should be improved. Lecturer A explains that she needed time to understand everything in the user manual, because the order of the actions are incorrect and some important actions are not described. So in her opinion, the manual should be ordered correctly and the actions should be explained in more detail.

Moreover, the lecturers were asked whether teachers would like to use IWT to create and plan online courses. Lecturer B thinks that with a few improvements, teachers will be able to use the system, though the current version is too complex. Lecturer A assumes that it would be too difficult for teachers to use the system even if they have advanced knowledge using learning platforms.

They were also asked whether they think that their students could benefit from the course. While lecturer B is convinced of that, lecturer A is not sure if their students could work with the system without technical support. In addition, she fears that the students could get frustrated if something doesn't work and for her IWT seems very unreliable.

Difficulty of the activities

Regarding the several steps the teachers had to follow on IWT, the lecturers were asked to state whether it was easy or difficult to do these steps. Creating the dictionary and providing the learning material was quite easy for them. The lecturers mentioned that they faced just few problems observing the students' progress. Furthermore, they stated that it was neither difficult nor easy to assign the students to their context. This answer is in line with the fact

that the technical staff assigned the students to their context, because the lecturers didn't have these rights on IWT. Creating the ontology, the contextualized course and add tests to the course seemed quite difficult for the lecturers. Finally they stated that monitoring the course without technical support is very difficult (H2.1.1).

Emotional Aspects

Concerning lecturers' emotions during working with IWT, the results from a 4-point rating scale showed that the teachers felt almost equally happy ($M = 2.5$, $SD = 1.18$), sad ($M = 1.5$, $SD = 0.71$), anxious ($M = 2$, $SD = 0.35$), and angry ($M = 1.83$, $SD = 0.24$). By interpreting the mean values, it can be assumed that the teachers seldom felt consciously happy, sad, anxious or angry. A closer look shows that most of the time the lecturers were curious ($M = 3$, $SD = 1.41$) while working with the system, but some of the time they also felt insecure ($M = 2.5$, $SD = 0.71$), helpless ($M = 2.5$, $SD = 0.71$) and frustrated ($M = 2.5$, $SD = 0.71$).

3.1.3.2 Findings from Phase 2 (students)

In the static course one learning material regarding the topic "Introduction to Scientific Work" was presented. After the students finished reading and learning this content, IWT generated a "on the fly" Test. The Test has six questions automatically created using the tools developed for R9 scenario. From six possible points (maximum score), the students achieved on average 4.75 points ($SD = 2.12$), the beginners achieved an average 5 points ($SD = 1.15$) and the advanced students 4.5 points ($SD = 3.0$). Regarding their activity level, students worked on average 35 minutes for this course on IWT ($SD = 13.61$).

At the end of this unit the students were asked whether the static course fitted their needs, previous knowledge and learning preferences (H2.1.2). On a 5pt. rating scale, students did not agree that the static course fitted their needs, previous knowledge and learning preferences ($M = 2.75$, $SD = .89$). They stated that the course content was well structured and gave a good introduction. Moreover, they also liked the idea of an additional test. However, the students criticize the quality of the questions, because some of them were the same and didn't make sense. With regards to their previous knowledge, half of the students (advanced) mention that they already knew most of the learning matter. Students also stated that they missed possibilities to interact with the system and add for example notes or content to the text. Students, who like learning by reading a text agreed that the course fit their learning preferences. Those who prefer to mark something and add notes disagreed that the course fit their needs.

Furthermore the students were asked whether the static course improved their understanding of domain concepts (H2.1.4). With an average rating of 3.13 ($SD = .84$), they agreed on that. One student also mentioned that due to the awareness of having an additional test she focused more on understanding the concepts. Some of the students already had previous knowledge about the content and state that the text refreshed their knowledge, although for a real improvement they would have to learn with paper and pencil.

Finally the students were asked if they would consider the course as a worthy educational resource (H2.1.6). With $M = 3.13$ ($SD = 1.25$) they indicated an average agreement on that

question. The students further stated that the content was very interesting and that there were good explanations. Although the course was helpful, the students wouldn't consider the course as a worthy educational resource. One of them mentions that the course is rather for a lower learning level, another one suggests using the course as an additional opportunity to strengthen the learning content. Students explained that the test raised their concentration and focused onto the learning process, but they were not satisfied with the quality of the questions and the evaluation of their answers. They mention for example that a lot of important knowledge was missing and that the questions were too easy (H2.1.6).

3.1.3.3 Findings from Phase 3 (students)

In the dynamic course, the students were assigned to beginners and advanced. Due to their previous knowledge they received different learning material. After the students finished reading and learning this material, they got a provided test with AQC (see scenario R9) and instructor questions. From 6 possible points (maximum score), the overall average of score is 4.75 points ($SD = 1.03$), whereas the beginners achieved on average 5 points ($SD = 0.8$) and the advanced students 4.5 points ($SD = 1.3$). Regarding students' activity level, they spent on average 56 minutes in IWT ($SD = 21.45$). Hence, compared to students' activity level in the static course, the students spent more working hours on the dynamic course (H2.1.5).

At the end of the dynamic course the students were also asked whether this course fitted their needs, previous knowledge and learning preferences. Some of the students stated that the content fit to their needs, others mentioned that they already knew parts of the content as their average agreement of 2.88 ($SD = .84$) indicates. This result does not differ from the one found for the static course ($t_{(6)} = 0.32$, $p = .763$). They also said that they would prefer a different presentation of the content and not only a set of documents. Besides, a few students explain that they have different learning preferences. They prefer, for instance, to provide own questions or add notes to the text. Another student also mentioned that the test was quite difficult for him (H2.1.2). Here and in the following, see Figure 21 for a comparison of the ratings in the static and the dynamic course.

Moreover the students were asked whether the dynamic course improved their understanding of domain concepts. Regarding this question on a five-point scale all participants indicated their level of agreement with 4. In contrast to the static course students significantly rated this question higher in the dynamic course as a repeated measures t -test analysis showed ($t_{(6)} = 3.1$, $p = .021$). They were in favor of the explanations and stated that the texts were easy to read and explained the domain concepts quite good. Nevertheless, a student would have preferred more self assessment possibilities. Another one is convinced that he has to read more secondary literature about the subject in order to improve his knowledge (H2.1.4). Moreover, for the same hypothesis, analyzing the average scores from the students automated tests in both phases - i.e. static and dynamic - there is no significant difference among them. Therefore it cannot be argued that the dynamic provision of contextualized course improves the students' knowledge acquisition, although students had the opinion that they had a better understanding of domain concepts.

After the dynamic course, the students were also asked if they would consider the course as a worthy educational resource. Their average level of agreement was 3.63 ($SD = .52$), which

does not differ significantly from their opinions about the static course ($t_{(6)}=1.56$, $p=.17$). According to the students a combination of learning material and questions always supports students' understanding. Additionally they also stated that the content was interesting and that the dynamic course was better than the first course. However, some of them suggest using the course only as an additional opportunity. Others mentioned that both courses didn't fit his learning type (H2.1.6).

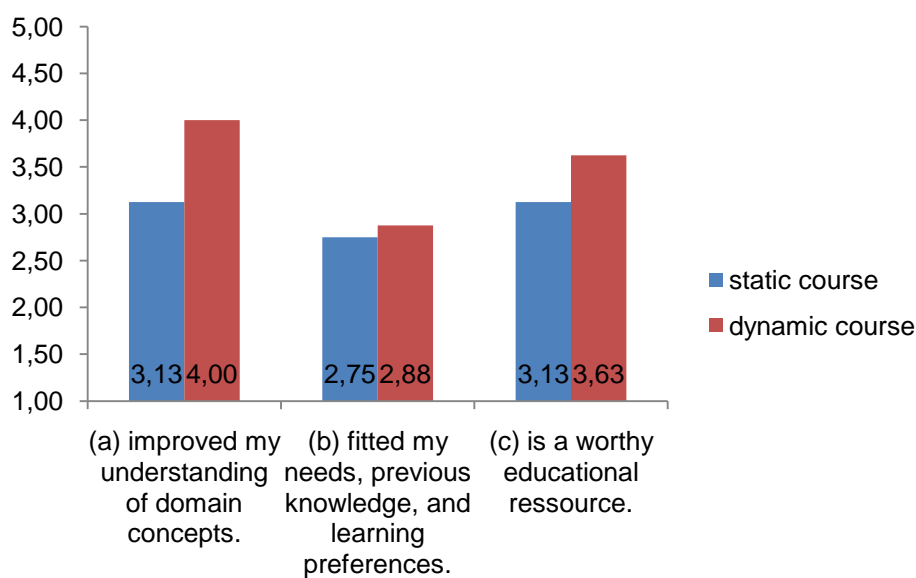


Figure 21. Average ratings for three questions targeting static or dynamic course ($N= 8$).

Usability of IWT (from the students' point of view)

As described above, we also used the System Usability Scale (SUS) to investigate students' experiences with the functionality of the system. After calculating the SUS score for each student, we got an average SUS score of 62.19 ($SD = 17.90$) for the static course and an average SUS score of 62.81 ($SD = 14.09$) for the dynamic course. So the students evaluated the usability of IWT in both courses equally ($t(7) = 0.88$ $p > .05$). As SUS scores have a range of 0 to 100 with an average score of 68, a SUS score of 62.5 is below average and is considered as C.

Almost all of the students like the idea of the system and state that it was easy to use. They are also in favor of the layout, the overview of the single learning parts and the useful links to Wikipedia which enhanced their learning experience. According to the students the system integrates most aspects of a course. Some students mention that they appreciate the stimulating learning atmosphere.

Then we asked them to state what they did not like regarding the tool. According to the students the icons are not representative for the functionalities and they also had to wait for the tooltips. The students suggest using a toolbar/menu with functionality names. The different styling formats, fonts and so on were confusing and the messages from the system

often came too late. Some of them also criticize the overall design, especially the color scheme and the slowness of the system. They also mention that there was no automatic test start. For the students, the system seems complicated and cannot be used without support. Another student didn't like the way the system communicated with him, he would rather prefer pop-ups or something more visible and effective.

Finally the students were asked about suggestions for improvements. According to the students there should be more interactive possibilities, such as editing the text or adding notes. All students state that the viewing port should be bigger. In addition, the back button should work and the links to Wikipedia should open in an extra window, so that the text is visible all the time. For the students it is also important to improve the design and performance of the system. They also suggest giving more possibilities to access additional information about specific parts of the learning content (like other websites, videos, pictures). According to the students it is necessary to improve the usability of the system with a direct integration into the frame menu lists in addition to the icons (H2.1.1).

3.1.4 Validation Results

In this Section the results regarding the pedagogical aspect of the tool is reported by looking at students' motivation while taking the static and dynamic course (H2.1.3). The corresponding metrics used for validation are M2.1.1 through M2.1.8 as they are specified in the previous subsection 3.1.1.

3.1.4.1 Motivational Aspects concerning the course and its tasks

Finally we evaluated students' motivation regarding the course in general. Comparing the extrinsic - with the intrinsic goal orientation scale, the intrinsic motivation ($M_{static} = 4$, $SD_{static} = 0.53$; $M_{dynamic} = 3.94$, $SD_{dynamic} = 0.72$) is significantly higher than the extrinsic motivation ($M_{static} = 2.47$, $SD_{static} = 1.04$; $M_{dynamic} = 2.53$, $SD_{dynamic} = 0.95$). These results were found in the static ($t(7) = 3.68$, $p < .05$) as well as in the dynamic course ($t(7) = 3.31$, $p < .05$). This means that the students were interested in both, the static and the dynamic course for reasons such as curiosity and challenges, whereas a good grade or rewards were not so important for them. These findings are supported by the results of the task value scale. A mean value of 3.23 ($SD = 0.94$) in the static course and 3.38 ($SD = 0.88$) in the dynamic course showed 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 (H2.1.3).

3.1.5 Conclusion

This section reports a study which was conducted to evaluate the scenario R2 concerning Knowledge Contextualization and its impact on students learning. For the sake of this, a study consisting of three phases was conducted. In the three phases 2 lecturers and 6 life-long learners have participated. The study provided a contextualized course in the topic of Scientific Working where two groups of life-long learners namely beginners and advanced learners took part to learn aspects related to scientific research. The students were provided two courses a static one which is provided without considering the students knowledge level - i.e.

context - and another dynamic one created by R2 tools to provide learning material fit with the learner knowledge state. To this end, the first results indicate that, from a view point of lecturers the current version of R2 tool is complex and teachers would require technical support thus to be able to create contextualized ontologies (G2.1.2). Nevertheless, a closer look shows that most of the time the lecturers were curious ($M = 3$, $SD = 1.41$) (G2.1.1) while working with the system, but some of the time they also felt insecure ($M = 2.5$, $SD = 0.71$), helpless ($M = 2.5$, $SD = 0.71$) and frustrated ($M = 2.5$, $SD = 0.71$).

Findings from the student phase - dynamic course - indicate that the students learning activities increased in terms of working time as the time spent on the dynamic contextualized time was more than the time they spent on the static course provided in the second phase (G2.1.3). However, analyzing the knowledge acquisition based on an automated test provided after the two phases does not lead to a significant difference that the contextualized courses help students to learn better. Despite that analyzing the students motivation regarding the course activities, results show that students were intrinsically motivated towards the course phases and this means that the students were interested in both, the static and the dynamic course for reasons such as curiosity and challenges, whereas a good grade or rewards were not so important for them. In addition the student's motivation towards the tasks show that they were motivated and 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 (G2.1.4). Finally, from the user's feedback and suggestions, potential ways to improve the systems were identified (G2.1.5).

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

3.2.1 Evaluation and validation procedure

Similarly as in the previous scenario (see the instruction view experimentation at TUG site reported in 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 [7]). 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 provide relevant feedback of how the Visual Ontology Editor (VOE) tool of IWT supports instructors in order to create online courses with the tool and can confirm the improvements made from the first experimentation phase of the project. Therefore, in this second phase of experiments 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 [3]:

Scenario goals

- G2.2.1: to develop a Visual Ontology Editor (VOE) for the definition of domain ontologies and contexts with a user friendly interface.
- G2.2.2: 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.
- G2.2.3: to identify possible ways of improving further the utility of the VOE and related models and algorithms.

Scenario hypotheses

- H2.2.1: a set of feasible courses can be effectively and efficiently created starting from a domain ontology by selecting a context, a set of target concepts and a learner.
- H2.2.2: automatically generated courses are considered as a worthy educational resource by instructors.

Validation Criteria

- C2.2.1: To evaluate the level of fulfilment of the tool features.
- C2.2.2: To evaluate the level of satisfaction of the instructors that use the VOE.
- C2.2.3: To evaluate the level of satisfaction of the instructors with the inclusion of the contextualized courses with their students.

Scenario metrics

- M2.2.1: Number of instructors using the VOE.
- M2.2.2: Number of courses created with contextualized ontologies.
- M2.2.3: Time employed in creating each course with contextualized ontologies.
- M2.2.4: Instructors that consider that the VOE and contextualized courses are worthy.

3.2.2 Method

3.2.2.1 Participants

In order to investigate the above goals and hypotheses, we asked one lecturer from the course “Other Technologies of Microsoft.NET” of the Computer Science postgraduate program of the UOC to create a personalized course about the theme “Styles and Animations” using the VOE tool of IWT.

3.2.2.2 Apparatus and Stimuli

First of all we asked the instructor to use the IWT (Intelligent Web Teacher) [1] 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 [7]).

Regarding the methodological approach of the study, the lecturer was asked to log all his activities concerning the experiment during the study. In his documentation he annotated for each step the time he spent on working with the IWT. In addition, the lecturer listed all problems he had to face while working with the system and wrote down advantages and

disadvantages. For this task, the lecturer was provided with technical documentation on this scenario (see [7]).

In addition, the lecturer was asked to fill in the SUS (System Usability Scale [8]) 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 lecturer was when he used the IWT he used the Computer Emotion Scale (CES) [9]. The 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, as qualitative statistical analysis, we summarized the open answers in the surveys.

3.2.2.3 Procedure

The experiment consisted of five sessions in a row conducted on the same day during the second week of June 2012:

1. **Work session 1:** The lecturer proposes a list of concepts that represent/model the topic “Styles and Animations” of the course “Other technologies of MS .NET” in both context (basic and advanced). Count the time invested.
2. **Work session 2:** The lecturer proposes an ontology drawing it on paper with the concepts proposed and 3 types of possible relations (standard LOM): *hasPart*, *isRequiredBy*, *suggestedOrder*. Count the time invested. **Work session 3:** The lecturer thinks over the concepts and decides which are common to the two contexts and which are specific of each context. Count the time invested.
3. **Work session 3:** The lecturer thinks over the concepts and decides which are common to the two contexts and which are specific of each context. Count the time invested.

4. **Work session 4:** The lecturer creates a contextualized course in the IWT platform. Procedure (see IWT user manual)::
 - a) Create a dictionary that incorporates all the key concepts that represent/model the topic “Styles and Animations” of the course “Other technologies of MS .NET” in both context (basic and advanced). Count the time invested.
 - b) Create an ontology with the IWT visual editor (VOE) from the concepts of the dictionary and the 3 types of possible relations. Count the time invested.
 - c) Create and configure two contexts:”basic” and “advanced”, and assign each context to the concepts corresponding of the ontology. Count the time invested.
 - d) Upload suitable material on IWT and tests for the topic and for each context Count the time invested.
 - e) Create a course personalized to each context. Time spent was counted.
5. **Work session 5:** Conduct a survey to evaluate the experience. Count the time invested.

The lecturer was instructed to use the manual of the IWT as provided in [7]. No training sessions on the IWT were programmed given the strong background of the lecturer in developing and using e-learning systems. All the sessions with the IWT were conducted in Spanish language as the targeted students were Spanish speakers. The, the comments and all the information to be reported were translated into English.

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

3.2.3 Evaluation and Validation Results

In this section, we show the evaluation and validation methodology that includes the criteria and metric extrapolated by [3]. Following this methodology we will evaluate and validate 3 aspects of the scenario by using metrics M2.2.1 through M2.2.4: time to run the experience, usability and emotions with the IWT (H2.2.1) as well as the IWT as a valuable resource (H2.2.2).

3.2.3.1 Time to run the experience

Next the time invested in each session is shown below:

1. **Work session 1** (Figure 22a): 5 minutes.
2. **Work session 2** (Figure 22b): 7 minutes
Time spent in this work session: 7 minutes.

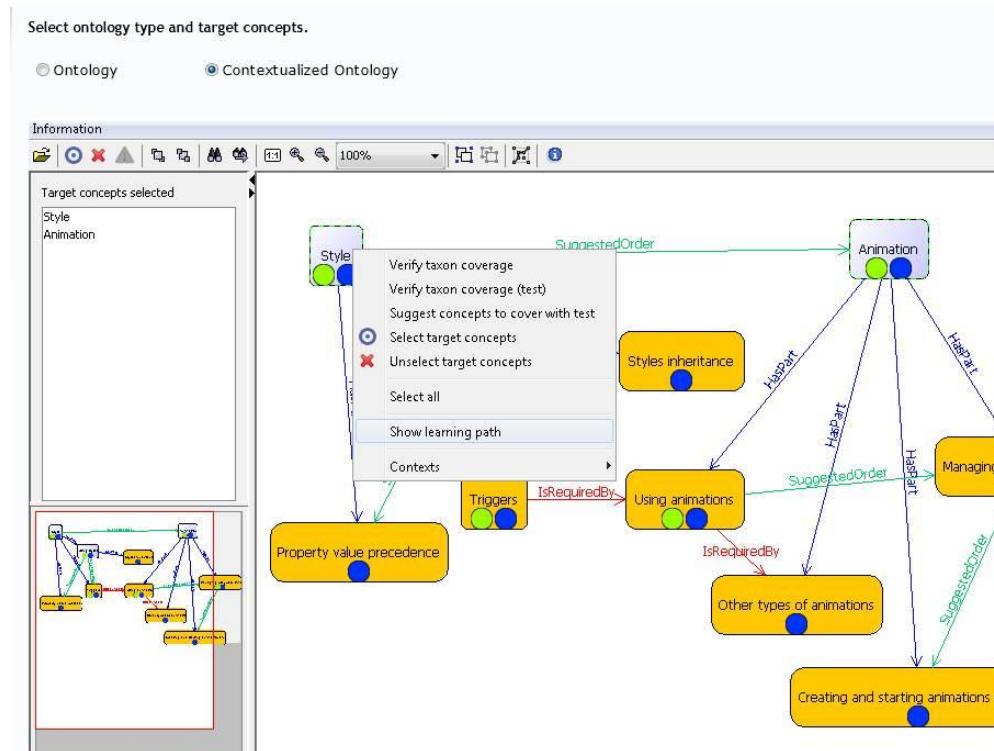


Figure 22a/b: *The newly created ontology with concepts (a) and relations (b) on the theme “Styles and Animations”.*

3. **Work session 3:** 5 minutes.
4. **Work session 4:** 5h and 17 minutes
 - a) Time spent: 5 minutes.
 - b) Time spent: 7 minutes.
 - c) Time spent: 5 minutes.
 - d) Time spent: 5 hours.
5. **Work session 5:** 7 minutes.

The time invested in the whole experiment was quite high (5h and 41 minutes), mainly because of the time of preparing the materials, and creating the tests. As only one topic of the course (“Styles and animations”) was created with R2, the lecturer was also asked to estimate the time required to create the rest of the topics (i.e., the whole course). The answer was that time would increase linearly.

3.2.3.2 Usability of the IWT

In this section, we analyzed the usability of the tool for potential improvements (H2.2.2). The lecturer was asked to fill in the SUS report (see Section 3.2.2.2) and a questionnaire with open questions after the experience. The **SUS score was 55**, despite it is below the SUS mean score (68), it is above of the SUS bottom level (51).

The lecturer had in general a good opinion of the IWT usability and in particular the CLR Editor. Although he admitted needing to learn a lot of things before using the IWT, he did not need the support of a technical person.

As for the positive aspects, the lecturer found the IWT well integrated and with little inconsistency.

These comments are in line with the SUS score achieved.

3.2.3.3 *Emotion of the IWT*

Regarding the lecturer's emotions, during the work with the IWT tool, a four-answer test question has been used for each feeling with the following answers:

- None of the time
- Some of the time
- Most of the time
- All of the time

The answers have shown that the lecturer was happy most of the time. Some of the time, his feelings were sadness, anxiety and anger. This can be explained because of the difficulties that the lecturer found when using the IWT (he reported that IWT is not very intuitive) and the lack of contextualized information that would have been useful in those situations. These difficulties are commented in the next section in the analysis of the open questionnaires.

3.2.3.4 *IWT as a valuable resource*

In order to find improvements for the tool, we asked the lecturer to evaluate the experience, especially concerning the usability of the IWT and answer five open questions.

1. Please describe what you liked regarding IWT.

The lecturer found IWT a very complete and useful suite of tools and functionalities for helping teachers and students in e-learning activities.

2. Please describe what you did not like regarding IWT

The lecturer found that the elements and tools you can find within IWT are not very intuitive and, for this reason, it is hard to use them the first time. He says that, perhaps, the practice can help in this sense.

3. Do you have any suggestions for improvements?

The lecturer thought that there are a lot of improvements that could be done regarding performance and usability. He has suggested the following ones:

- Dictionary description is too short and do not allow characters like double quotes or support for tildes or other language special characters.
- When opening a form, main field should be focused.

- Pop-up forms take a bit long to open.
 - Ok buttons usually go in the left and close or cancel ones on the right.
 - Contextualized help about the options in each form would be great.
 - Some forms and dialogs are not translated into English (Italian text appears).
 - There is no possibility (or not found) to create a resource of type external link (a URL to link).
 - When uploading big files it should be great to include an uploading progress control.
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?

The lecturer stated that the manual helped him in some parts but, for other parts, he required external assistance to configure the desired course structure.

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?

The lecturer recognized that IWT is a great platform, with a lot of useful functionalities. However, he pointed out that the learning curve to exploit its potential efficiently is rather high.

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)?

The lecturer answered affirmatively to this question. However, he highlighted the matter of the learning curve. He added that Usability was critical to take advantage of the course.

As a final remark about the whole experimentation, the lecturer considered that the experiment was in general fine and that the tested tools were proved to be useful (from the point of view of the teacher). However he found technical problems related to the java version installed in the system and some performance problems.

3.2.4 Conclusion

In contrast to the previous experiment conducted at TUG, this experiment at UOC was conducted by a real expert in developing complex computer systems. As a professional developer and analysts (and on-line teacher), he is 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 user, he found many technical inconveniences that other people with a different background may miss.

From the usability analysis, the lecturer considered the tool was satisfactory and confirmed the improvements made in the second stage of experiments from the valuable feedback collected (G2.2.1).

The lecturer' emotions when using the tool is also in line with the above mentioned satisfaction and usability, and confirms from this perspective the improvements made in the tool.

The tool did not experience any technical problem during the experiment and could be completed, thus achieving the main goal (G2.2.2). This is in line with the other pilot site that could finalize the experience with success. This confirms that the technical problems faced by some lecturers in the first round of experiments were sporadic and exceptional as no relevant technical problem was reported at this final stage of the experiments.

Finally, the lecturer was very helpful and active, and provided 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 (G2.2.3).

To sum up, the lecturer liked the idea of personalizing a course by an ontology and having structured learning resources to fit the specific students' needs and different contexts. He still considered the complexity of the tool a barrier for other lecturers and students when using the tool and the learning curve to exploit its potential efficiently is rather high. All in all, the lecturer believes the IWT is a great platform. The user manual was not very helpful and sometimes the lecturer needed external assistance.

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 (see [7]).

Two trials were run on this scenario at UOC pilot site: on trial from the student's perspective and another from the instructor' viewpoint. In summary:

1. Semantic Connections Between Learning Resources from the student's viewpoint (Section 4.1)
2. Semantic Connections Between Learning Resources from the instructor's viewpoint (Section 4.2)

4.1 R3-1. Semantic Connections Between Learning Resources from the student's viewpoint

4.1.1 Evaluation and validation procedure

To experiment with the Semantic Connections between Learning Resources from the student's viewpoint, we focused on the following scenario goals and hypotheses as well as criteria and metrics as described in [3]:

Scenario goals

G3.1.1: to playback the generated CLR through a user friendly interface.

G3.1.2: to ensure that a CLR is able to adapt itself basing on learning preferences.

G3.1.3: to ensure that a CLR allows the effective and efficient learning of scientific concepts in selected domains.

G3.1.4: to identify possible ways of improving further the utility of the CLR and related tools.

Scenario hypotheses

H3.1.1: a CLR can be effectively played by learners through a user friendly interface.

H3.1.2: the use of CLRs contribute to improve students' motivation.

H3.1.3: the use of CLRs contribute to improve students' understanding of key concepts.

H3.1.4: the use of CLRs contribute to increase students' activity levels.

H3.1.5: CLRs are considered as a worthy educational resource by students.

Scenario criteria

C3.1.1: To evaluate the increase in students' motivation caused by the use of CLRs.

C3.1.2: To evaluate the increase in students' understanding of key course concepts and students' results caused by the use of CLR.

C3.1.3: To evaluate the increase in students' activity levels due to the use of CLR.

C3.1.4: To evaluate the level of satisfaction of students with the inclusion of the SLO in their courses.

Scenario metrics

M3.1.1: Students passing the final test with high marks when CLR are used.

M3.1.2: Students passing the final test with high marks when CLR are not used.

M3.1.3: Students that consider that the CLR is worthy.

M3.1.4: Number of students using CLR.

M3.4.5: Number of visits to CLR.

4.1.2 Method

4.1.2.1 Participants

In order to evaluate this scenario to analyze its effects in the learning process and compare the results with those reported in the first round of the experiments (see [6]), we will follow the same methodology of the first experiments.

The proposed methodology for this experiment considered 151 students enrolled in the course Software Engineering from the Bachelor in Computing Engineering in the Spring term of 2012 at the UOC participated in the experience. Most of them (142) were from the Bachelor in Computing Engineering (BCE) and a small group (9) was from the Master in Computing Engineering (MCE). Both Bachelor and Master share the same course "Software Engineering" in its curricula.

The students were roughly distributed equally into 2 classrooms in the UOC virtual campus, 77 and 74 students each.

61 out of 151 students (40.3%) 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, 59.7% of them chose not to send the evaluation form and thus they were excluded from the analysis.

29 out of 151 students (19.2%) 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 29 students belonged to the group of 61, which left a group of 32 who participated by submitting the form but did not use the IWT prototypes.

From the 61 participants we formed 2 groups for the experiment. One experimental group with 29 students who use IWT (47.5%) and one control group with 32 students who did not use IWT at all (52.5%). All of them submitted an evaluation form at the end of the experience.

Therefore, the sample of the experiment was formed by 61 students. All students of the sample were supervised by one experimented tutor during the experiment. For the sake of the experiment, we were only interested in the conglomerate of the experimental group formed by 29 students. 27 students were male (93.1%) and 2 students were female (6.9%). The 32 students forming the control group studied at UOC only and did not enter IWT. Hence, whenever referring to IWT we mean the experimental group.

4.1.2.2 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 23 below and Annex A1 for technical details of the integration).

The screenshot shows the UOC (Universitat Oberta de Catalunya) classroom interface. At the top, there are navigation links for 'Accessibilitat', 'Personalitza', 'Santi Caballé Llobet', and 'UOC-Professor-Info'. Below this is a menu with 'Bústia nova', 'Bústia', 'Agenda', 'El meu perfil', 'Grups de treball', and 'Preferits'. A row of icons represents various services: 'La meua UOC', 'Comunitat', 'Serveis', 'Aules', 'Tutoria', 'Tutoria IP', 'Suport docència', 'Secretaria', 'Recerca i Innovació', 'Biblioteca', 'Notícies', 'Espai de Bolonya', and 'Intr'. On the left, a sidebar lists 'Salas Estudiantes' with various course codes and dates. The main content area is titled 'Activa iLKE' and contains several sections: 'Comunicació' (with 'Aula IWT' circled in red), 'Planificació' (with a calendar for April and May 2012), 'Recursos', 'Avaluació', and 'Activitats d'avaluació continua'. The 'Activitats d'avaluació continua' section includes a table with columns for 'Títol', 'Enunciat', 'Lliurament', 'Solució', and 'Nota'.

Títol	Enunciat	Lliurament	Solució	Nota
Formació de...	29/02/2012	27/03/2012	-	-
PAC1	08/03/2012	27/03/2012	29/03/2012	06/04/2012
Debat1	28/03/2012	13/04/2012	-	-
Pràctica1	29/03/2012	17/04/2012	19/04/2012	27/04/2012

Figure 23: UOC classroom with the access to IWT classroom

Once in the IWT classroom, students had access to the R3 scenario (see Figure 24, Figure 25 and [1])

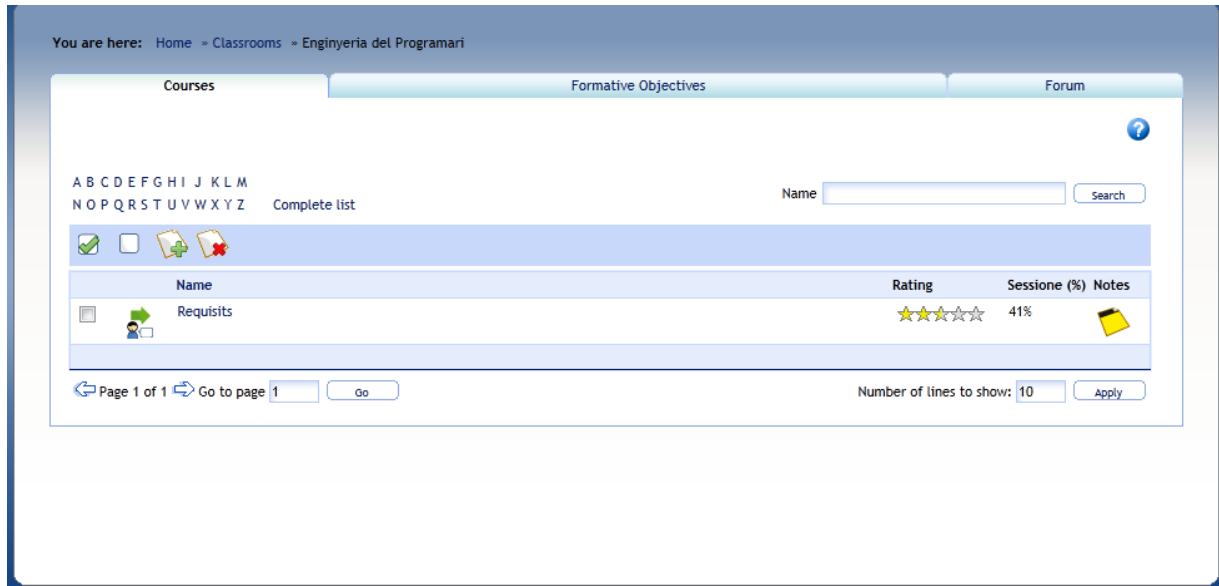


Figure 24: IWT classroom with a course of Requirements in Software Engineering

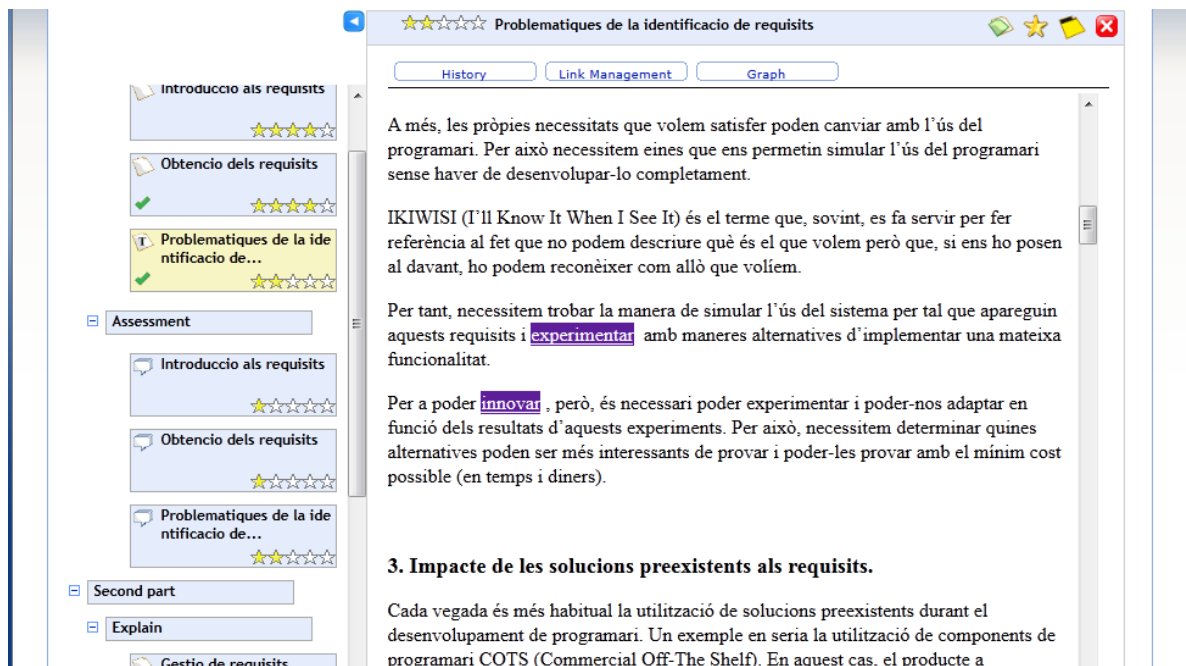


Figure 25: A CLR with semantic connections to learning resources

We used the SUS (System Usability Scale [8]) 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 “Requisites” (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) [8], 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 CLRs, section (vi) concerned about the “emotional state” of students when using the CLR, which included 12 items of the Computer Emotion Scale (CES) [9]. 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).

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.

4.1.2.3 Procedure

The in-class collaborative formal assignment in both groups lasted three weeks during the second third of the Spring term of 2012 (April 2012) 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. 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 ULLG recommendation system of IWT (see scenario R1 in Section 2).

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).

4.1.3 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) by using metrics M.3.1.5. 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 4.1.4 (Validation Results).

4.1.3.1 Activity levels in the CLR

In order to give a feedback about how a CLR resource contributes 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).

81 out of 151 students (54%) have delivered the CLR; among them, 29 students have not delivered the assessment test associated to the CLR that allows for analyzing the acquired level of competence.

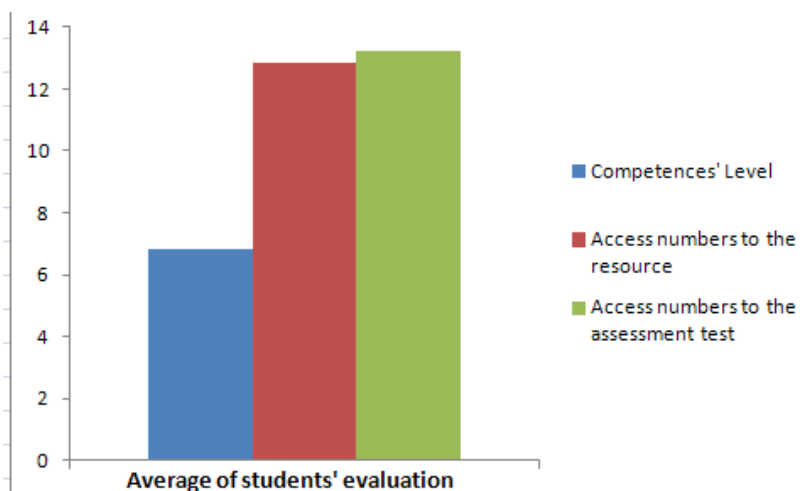


Figure 26: Competence level with the access number

The Figure 26 shows the average competence level acquired by the interaction of the students with the CLR resource. Taking into account that the number of accesses both to the CLR resource and to the corresponding assessment test is very similar, we can register a great interest of the students to the resource and their interest to institutionalize the implicit knowledge by making an assessment test. For consequence it has been obtained an average competence level quite high ($M=6.80$).

The activity levels of the students in the CLR have been also identified by the permanence time (expressed in second) in IWT (Figure 27).

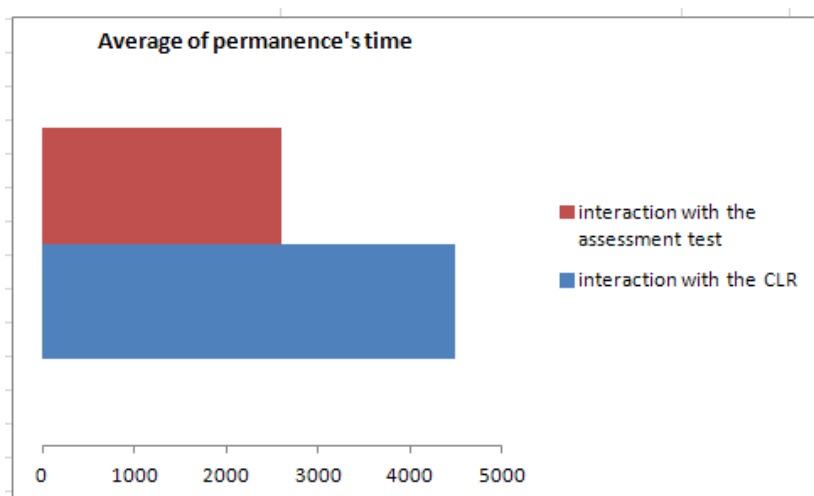


Figure 27: Permanence's time in IWT

Figure 27 shows a great interest of the students respect to the CLR that has also been confirmed by a high interaction with the assessment resources.

4.1.3.2 Usability of the CLR

To evaluate student’s satisfaction with the tool regarding an efficient and user-friendly management (H3.1.1), 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 53.97**, thus below the SUS mean score (68). Next, we present the most relevant results of the SUS score by providing several statistics: Mean (M), Standard Deviation (SD) and Median (Md).

Analyzing student feedback, we can observe there are more students who think they would like to use CLRs more often than students who wouldn’t (46% vs 31%) (M = 3.13, SD = 1.09, Md = 3) (See Figure 28).

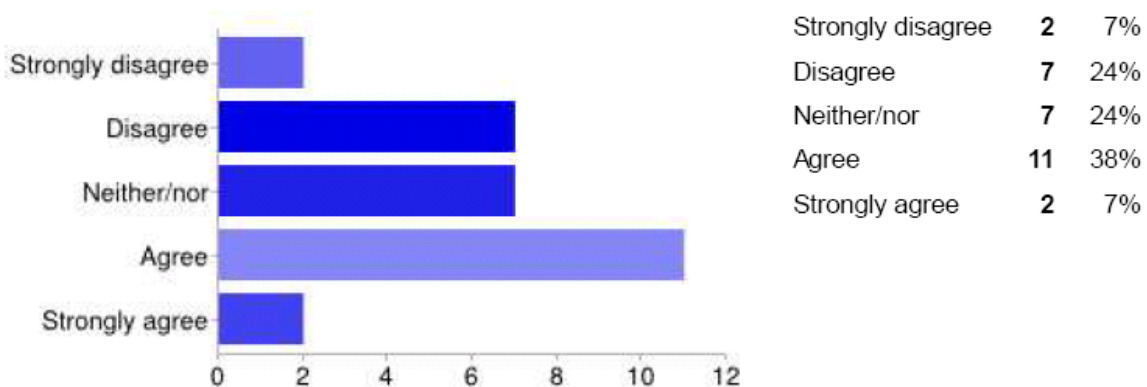


Figure 28: Results on the SUS item "I think I would like to use CLRs frequently"

These results are aligned with the amount of people who think that CLRs are unnecessarily complex (M = 3.10, SD = 1.11, Md = 3) (See Figure 29). Reasons that could explain this result could be the opinion of many students who think that there is inconsistency in the CLR interface (M = 3.24, SD = 0.98, Md = 3) and that CLR are not well integrated in the IWT environment. Some students reported that the interface is neither user-friendly nor intuitive.

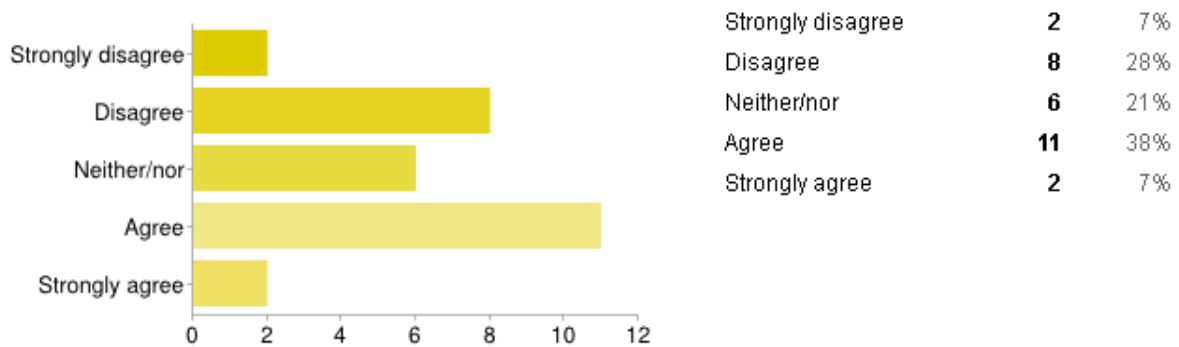


Figure 29: Results on SUS Item "I think CLR's are unnecessarily complex"

A lot of students thought that CLR's were easy to use ($M = 3.65$, $SD = 0.81$, $Md = 4$) (see Figure 30). In addition, many students stated that they had not needed the support of a technician to be able to use CLR's and that people should learn how to use CLR's quickly ($M = 2.90$, $SD = 1.01$, $Md = 3$) (See Figure 31) since there is little need to learn too much to be able to use it. ($M = 2.41$, $SD = 0.95$, $Md = 2$) (see Figure 32).

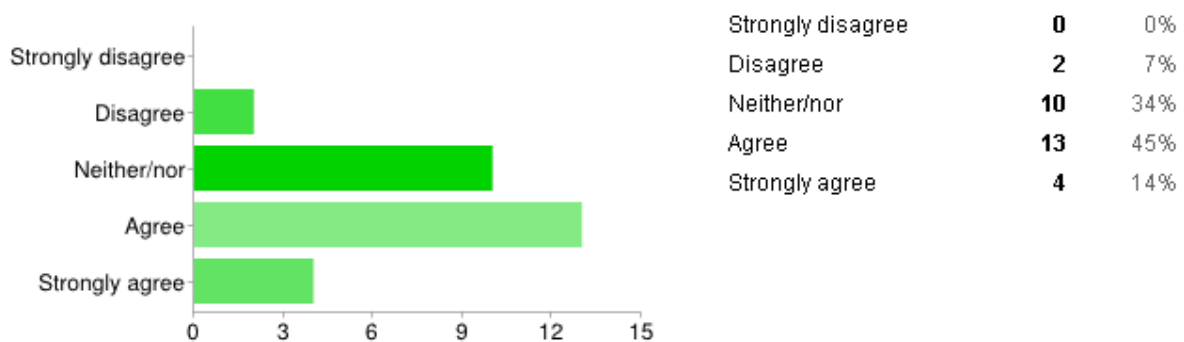


Figure 30: Results on the item "I think CLR's were going to be easy to use"

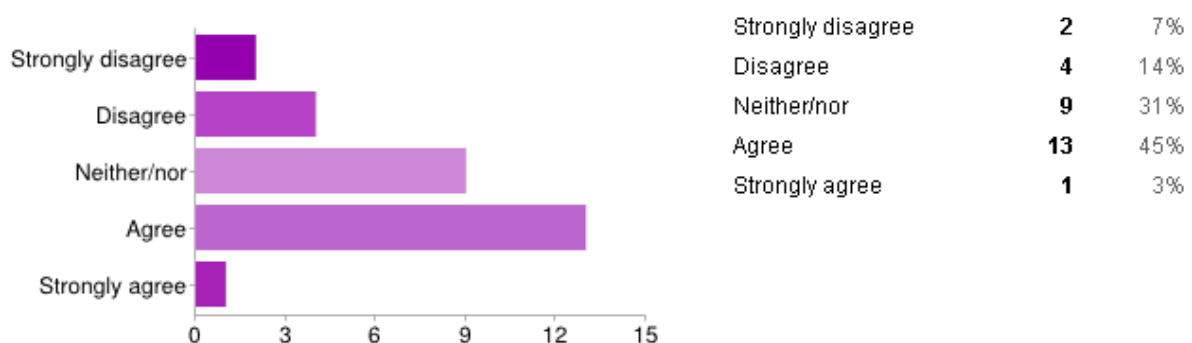


Figure 31: Results on SUS item: "It is thought that people should learn how to use CLR quickly"

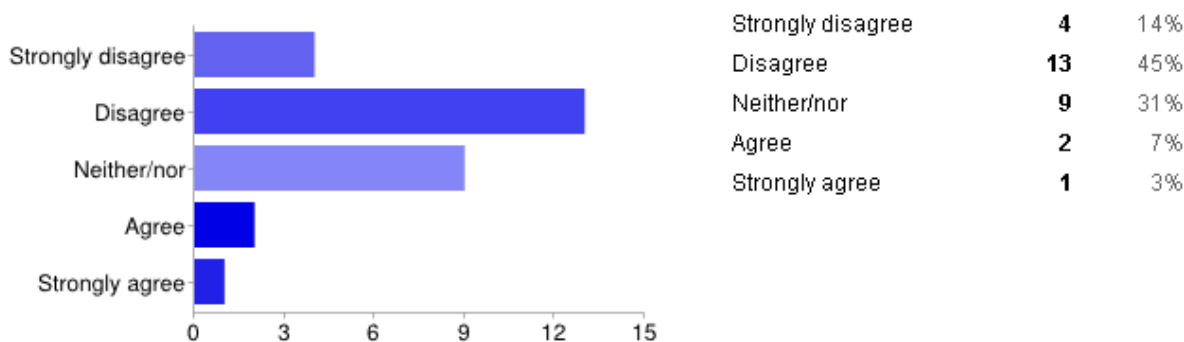


Figure 32: Results on the item "I think I don't need to learn too many things to use CLR"

Unlike the first iteration of experiments with the CLR, students did not report a technical navigational problem found in previous experiments and from the improvements made in the current prototype, students could visit internal and external links in the material and then go back the point where the learning path was branched. In general students liked the CLR resources and the semantic connections a lot (see Section 4.1.4.1). Therefore, no strong opinions against the CLR usability were found this time.

4.1.3.3 Emotional aspects

Regarding student emotion while working with the IWT tool (H1.1), we have 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 in a 4-point rating scale (n=29) have been as follows:

- Happiness (M = 1.51, SD = 0.83, Md = 2)

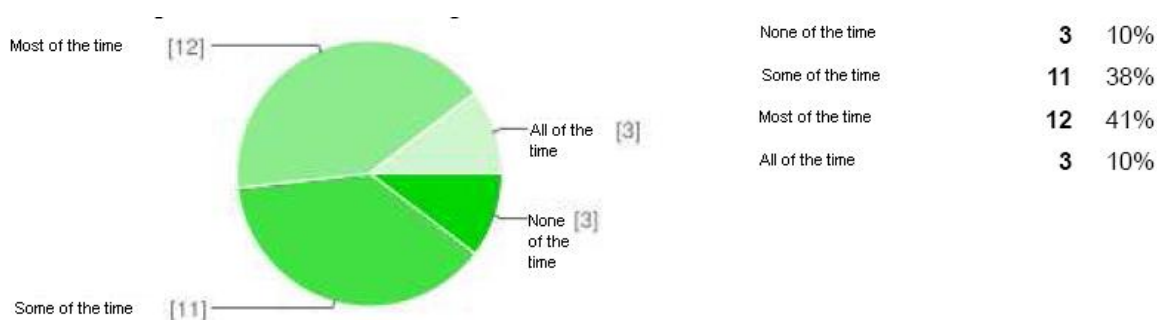


Figure 33: Results on the Happiness emotion

- Sadness (M = 0.62, SD = 0.68, Md = 1)

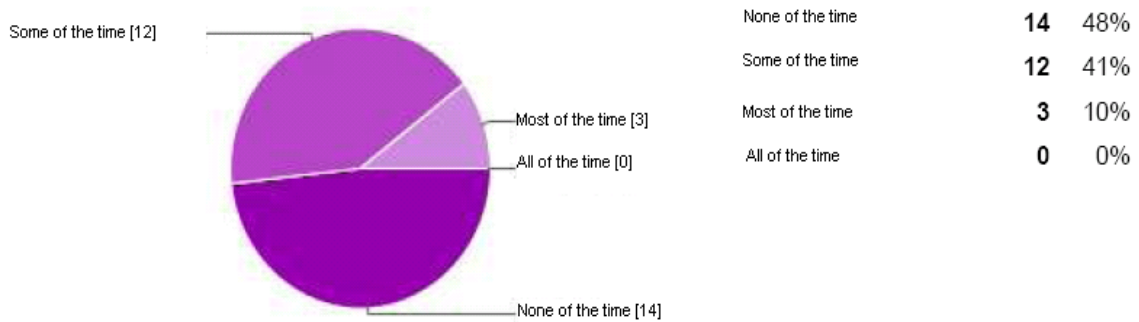


Figure 34: Results on the Sadness emotion

- Anxiety (M = 0.55, SD = 0.63, Md = 0)

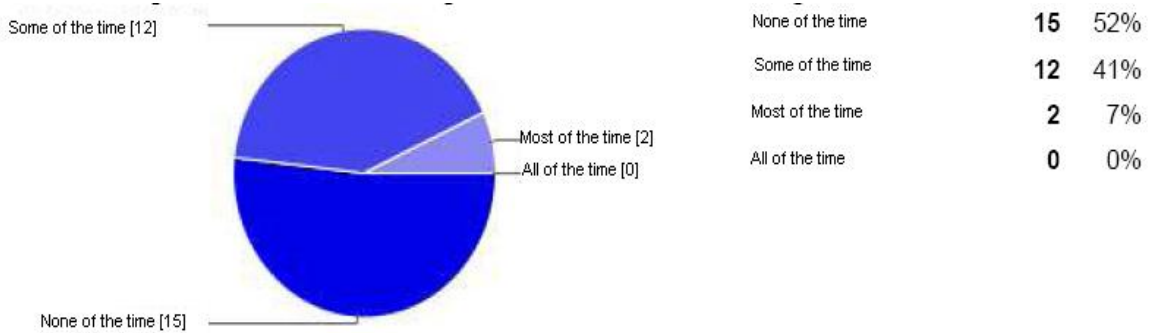


Figure 35: Results on the Anxiety emotion

- Anger (M = 0.34, SD = 0.55, Md = 0)

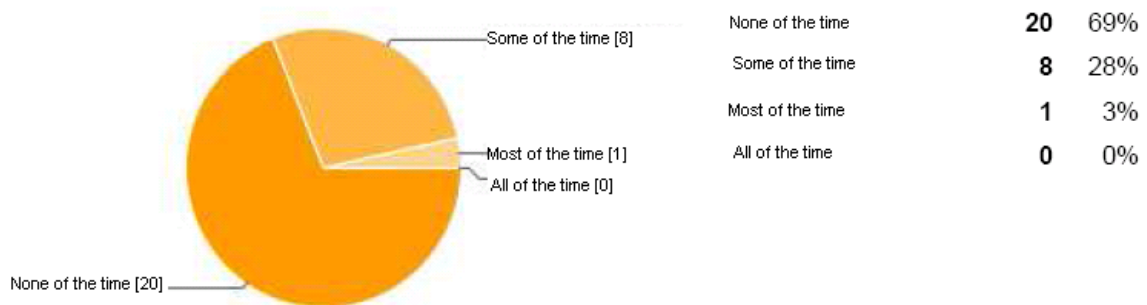


Figure 36: Results on the Anger emotion

The Happiness emotion (Figure 33) appears most of the time and much more than the rest of emotions, which are low in general. Thus, there are more people who are happy most of the time than sad.

Almost 70% of the students have not experimented anger at any time (Figure 36) and anxiety is very low (Figure 35). This result is aligned with the usability results, which indicate that, in general, people have not had problems when dealing with the CLR as a new type of learning material and have managed quite well without any additional help.

If we compare these results with the results of the first iteration, there are now more people who are happy most of the time and so, less people who are sad, anxious or angry. This fact can be explained because CLR prototype has been improved compared to the previous version.

4.1.3.4 Questionnaire evaluation

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

The results of the evaluation of the design of the questionnaire have confirmed, like in the first iteration that the time employed to fill the questionnaire in is less than 30 minutes for most of the students (72%) (Figure 37) and although most of the students think that the questionnaire is appropriate to evaluate the experience (Figure 38), some students stated that the questionnaire was long and heavy.

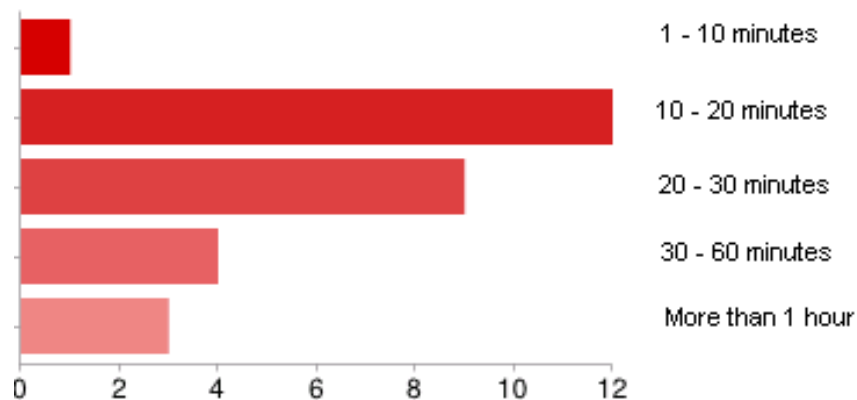


Figure 37: Time employed to fill in the questionnaire

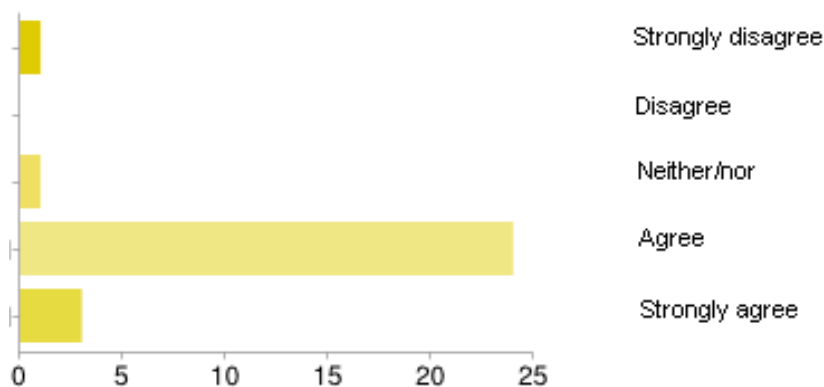


Figure 38: Appropriateness of questionnaire to evaluate the experience

4.1.4 Validation Results

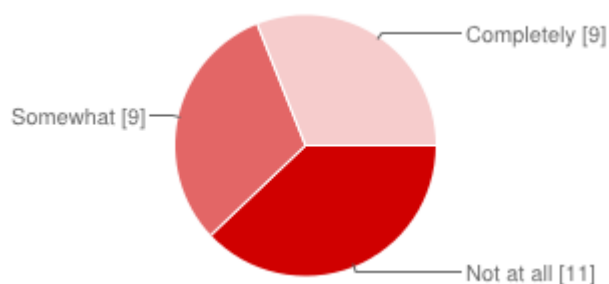
Following the methodology described in Section 4.1.2, in this section we will analyze students' motivation (H3.1.2), worthiness of the CLR as an educational and teaching supporting resource (H3.1.5) as well as the acquisition of collaborative knowledge by means of the CLR (H3.1.3). For these purposes we will use the metrics M3.1.1 through M3.1.4 as specified in Section 4.1.1.

4.1.4.1 The CLR as a valuable resource

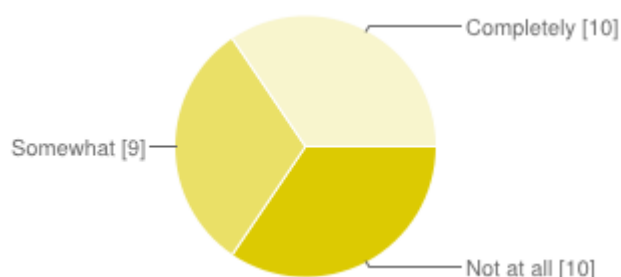
Test-based questions

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).

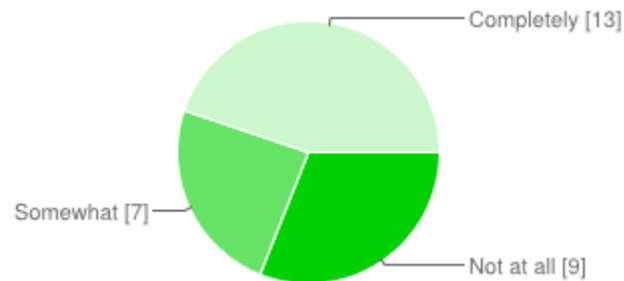
1. The possibility to navigate a learning resource through semantic connections has involved you in a more consistent way to browse the contents?



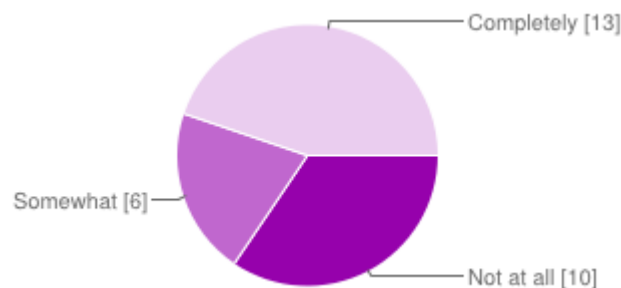
2. Do you think that this solution allows students to read the resource following their own interests or types of reading?



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 improvement

This open question completed this section of the questionnaire by asking students for giving final hints for potential improvement of the CLR resources.

Students have liked, in general, the semantic connections between learning resources. They have found it a good idea to clarify concepts and enrich their knowledge about them. Indeed, some of them pointed out that there were few semantic connections and that they would have liked to have more. Many students stated that the impact of the semantic connection in their study is very low and that they have not used them.

On the other hand, some students agreed that, although semantic connections give value to the learning process, they could make lose track of their study.

4.1.4.2 Motivational aspects

Students' motivation concerning the use of IWT tool (H3.1.2) 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=3.83, SD=0.45, Md=4). This result is very quite good and better than the equivalent of the first phase of the experiment (M=3.01, SD=0.78, Md=3.5) and also in line with the usability and emotional results reported in the previous sections. They are also in line with the previous results on the CLR being a valuable resource, especially considering that many students scored high because of the potential of CLR rather than the direct benefit achieved. In particular, students indicated they liked the semantic connections and would like to have more in the material. They indeed found this a particular valuable innovation of the system.

This result is in line with the results on the IWT being a valuable resource and also in line with slightly improvement from the results of the first phase of experiments (M=3.01, SD=0.78, Md=3.5). In addition these results are in line with the usability and emotional results reported in the previous sections. In particular, students indicated to feel very motivated by the self-evaluation 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 motivation and engagement came from passionate students who made very positive comments about the semantic connections, such as “the semantic connections are very useful, it is the best I saw in the IWT classroom”, “I found the semantic connections great, and actually in my study I use the strategy manually”. However, most of them clarified that they found very few connections while other said that some external connections to very different material make one self feel lost. Eventually, most of students understood it was a pilot trial and for this reason they found normal the few connections found for experimentation purposes and the disturbing external connections could be changed easily into good ones.

4.1.4.3 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. *Table 7* shows the results.

Experimental group (n=29)	Control group (n=32)
M=7.81	M=7.32
SD=1.28	SD=1.24
Md=8	Md=7

Table 7: Results of the learning assignment evaluation

From the results of *Table 7*, students from the experimental group (material UOC + IWT/CLR) scored higher than the control group (material UOC) and in line with the previous experiments. In comparison to the first round of experiments, the SD for the experimental group is significantly lower. This uncovers the reason behind this result and overrides the validation reported in the previous stage. In particular, the higher SD is produced by the lower number of participants in this experimentation (41 vs. 29) that makes it more sensible to the outlier data. However, this result is in line with the results of R1 scenario (see Section 2.4.3) where students from the experimental group 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 (control group) 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 4.1.4.1).

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

4.1.5 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). A comparison with the results of the first round of experiments is also provided.

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 (G3.1.2). The recurrent comment on finding few semantic connections positively confirms that students liked this approach for their study.

The CLR were reported to be reproduced efficiently by students who could use them to find further information about these concepts (G3.1.3). From the usability point of view, the goals were also achieved by providing CLR with a friendly user interface (G3.1.1) and also it was noticeable the improvements after in the second stage of the project. In particular, a technical issue reported in the first round of experiments related to navigation problems, which strongly influenced the whole experimentation, was now not reported thus considering the overall usability of the system satisfactory. However, the whole usability did not improve from the first experiments and stay still in the SUS mean Grade, and even lower in absolute terms.

One of the most relevant results was that many 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. Some of them even reported on that they had used in other studies at UOC the same strategy manually. This result implicitly reinforces the achievement of G3.1.2 by providing students of either BCE or MCE 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 implicitly achieves also G3.1.3.

Finally, possible ways of improving further the utility of the CLR and semantic connections (G3.1.4) were provided everywhere, mainly at the end of Section 4.1.4.1, being most of the comments addressed to a sense of disorientation when using external connections.

The positive results are in line with the first round of experiments and from above all the improvements made in the second phase of the project from the valuable feedback collect in the previous experiments confirm the level of satisfaction in this second round of experiments.

4.2 R3-2. Semantic Connections Between Learning Resources from the instructor's viewpoint

4.2.1 Evaluation and validation procedure

Similarly as in the previous scenario (see the instruction view experimentation reported in Section 4.1.1), the aim of this scenario is also to create semantic connections between learning goals that is able to automatically adapt to a teaching and learning preferences (see [7]). To this end, an experiment was conducted on this scenario at UOC pilot site in order to test the CLR designer tool of IWT from the instructors' viewpoint. The results of this study provide relevant feedback of how the tool supports instructors in order to create semantic connection with the tool. Therefore, in this second phase of experiments we were primarily interested in the functionality and usability of the CLR designer/editor tool.

To experiment with the Semantic Connections between Learning Resources from the instructor's viewpoint, we focused on the following scenario goals and hypotheses as well as criteria and metrics as described in [3]:

Scenario goals

- G3.2.1: the Compound Learning Resources (CLR) designer that allows efficient building of a CLR even in the case of non-expert instructors (i.e. in a friendly way).
- G3.2.2: to identify possible ways of improving further the utility of the CLR and related tools.

Scenario hypotheses

- H3.2.1: a CLR can be effectively created by instructors as well as stored and played by learners through a user friendly interface.
- H3.2.2: the use of CLRs contributes to support instructors' task.
- H3.2.3: CLRs are considered as a worthy educational resource by instructors.

Validation Criteria

- C3.2.1: To evaluate the level of fulfilment of the tool features.

- C3.2.2: To evaluate the level of satisfaction of the instructors that use the VOE.
- C3.2.3: To evaluate the level of satisfaction of the instructors with the inclusion of the contextualized courses with their students.

Scenario metrics

- M3.2.1: Number of instructors using the VOE.
- M3.2.2: Number of courses created with contextualized ontologies.
- M3.2.3: Time employed in creating each course with contextualized ontologies.
- M3.2.4: Instructors that consider that the VOE and contextualized courses are worthy.

4.2.2 Method

4.2.2.1 Participants

In order to investigate the above goals and hypotheses, we asked one lecturer from the course “Other Technologies of Microsoft.NET” of the Computer Science postgraduate program of the UOC to create a semantic connections with the CLR designer about the theme “Styles and Animations” specialized in two different contexts: basic and advanced, using the CLR editor tool of IWT.

4.2.2.2 Apparatus and Stimuli

First of all we asked the instructor to use the CLR Editor of the IWT (Intelligent Web Teacher) [1] to model the theme “Styles and Animations” of course “Other Technologies of Microsoft.NET” by using semantic connections between resources within this theme.

Regarding the methodological approach of the study, the lecturer was asked to log all his activities concerning the experiment during the study in a Notebook. In the documentation he annotated for each step the time he spent on working with the IWT. In addition, the lecturer listed all problems he had to face while working with the system and wrote down advantages and disadvantages. For this task, the lecturer was provided with technical documentation on this scenario (see [7]).

In addition, the lecturer was asked to fill in the SUS (System Usability Scale [8]) 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 lecturer was when he used the IWT we used the Computer Emotion Scale (CES) [9]. The 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.

4.2.2.3 Procedure

The experiment consisted of four sessions. All sessions (see below) were scheduled during the second week of June 2012 and the lecturer could run the sessions at his most convenient time.

The lecturer was instructed to use the manual of the IWT as provided in [7]. No training sessions on the IWT were programmed given the strong background of the lecturer in developing and using e-learning systems. All the sessions with the IWT were conducted in Spanish language as the targeted students were Spanish speakers, then the comments and information to be reported were translated into English.

The work procedure was as follows:

- **Work session 1:** The lecturer proposes a set of resources and URLs associated to one of the aspects of the topic “Styles and Animations” of the course “Other technologies of MS .NET”. Count the time invested.
- **Work session 2:** The lecturer designs relationships between the proposed resources using the main keywords of the topic as linking concepts. Count the time invested.
- **Work session 3:** The lecturer creates a CLR resource in the IWT platform. Procedure (see IWT user manual):
 1. Create a CLR resource that incorporates the proposed materials for the selected aspect to the topic “Styles and Animations” of the course “Other technologies of MS .NET”. Count the time invested. Select one of those resources as the index page for the CLR.
 2. Define navigation relations between related resources using the corresponding keywords.
 3. Associate the new resource to the personalized course. Count the time invested.
- **Work session 4:** Conduct a survey to evaluate the experience. Count the time invested.

4.2.3 Evaluation and validation results

In this section, we show the validation methodology that includes the criteria and metric extrapolated by [3]. Following this methodology we will validate 3 aspects of the scenario by using metrics M3.2.1 – M3.2.4: time to run the experience and the usability of the IWT (H3.2.1) as well as the lecturer's emotions when using the IWT (H3.2.2).

4.2.3.1 Time to run the experience

Next, the time spent in each session is shown:

- **Work Session 1** (see Figure 39): 1h

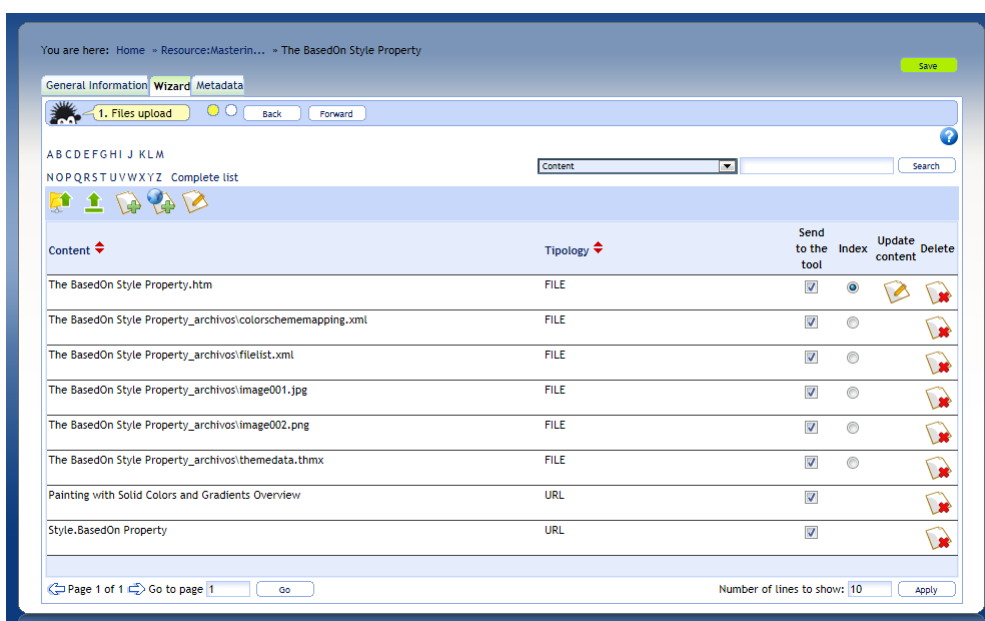


Figure 39: A list of resources to link related to the topic

- **Work Session 2** (Figure 40): 30 min

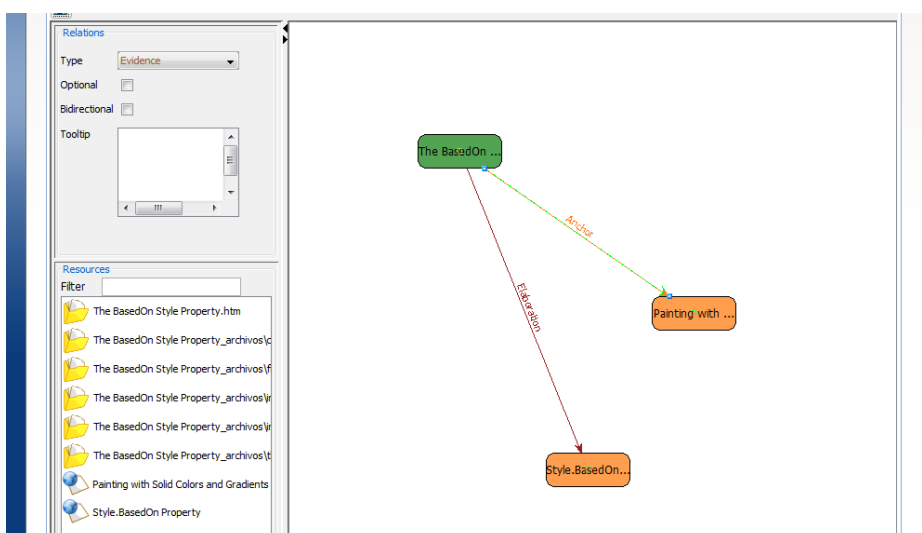


Figure 40: A sample of some semantic links between resources for this topic

- **Work Session 3** (Total): 1h 10min
 1. 30 min
 2. 30 min
 3. 10 min
- **Work Session 4:** 1h

The total time invested in the whole experiment was quite high (**4h and 50 minutes**), mainly because of the time of preparing the materials, and creating the tests. Similarly to R2 scenario (see Section 3.2), as only one topic of the course (“Styles and animations”) was created with R3, the lecturer was also asked to estimate the time required to create the rest of the topics (i.e., the whole course). The answer was that time would increase linearly.

4.2.3.2 Usability of the IWT

In this section, we analyzed the usability of the tool for potential improvements (H2.2.2). The lecturer was asked to fill in the SUS report (see Section 3.2.2.2) and a questionnaire with open questions after the experience. The SUS score was 70, thus above the SUS mean score (68) and considered as a very good result.

Despite the time spent in creating the course, the lecturer considered normal to spend time to personalize a course and this was not to do with the usability. Therefore the lecturer was satisfied from the usability perspective with the CLR editor tool of IWT and in line with the SUS score achieved.

4.2.3.3 Emotion of the IWT

Regarding the lecturer’s emotions, during the work with the IWT tool, a four-answer test question has been used for each feeling with the following answers:

- None of the time
- Some of the time
- Most of the time
- All of the time

The answers have shown that the lecturer was happy most of the time. Moreover, none of the time he felt anxiety nor sadness. Only some of the time he felt anger. The latter can be explained by some minor technical problems found with the tools as commented in the next section in the analysis of the open questionnaires.

4.2.3.4 Enhancements and improvements of IWT

In this final section the lecturer’s answers to an open questionnaire about the usability of the tool, potential for teaching and learning, and comments and suggestions for improvements, are shown:

1. Please describe what you liked regarding the CLR Editor.

I like the easy way to create linked materials from a set of individual resources.

2. Please describe what you did not like regarding the CLR Editor

The tools had some problems with the java version. The WYSWYG editor too.

3. Do you have any suggestions for improvements?

Ease of use can be improved. Also some further explanation of the type of the different available types of binding between resources would be welcome.

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

Yes, the user manual was enough.

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

Maybe, depending on how difficult to find appropriated materials and define the corresponding links between them is (so I think it depends on the topic being modeled).

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

I think it is a good tool for the learner too, but also it depends on the kind of topic being taught. Also depends from the structure defined by the lecturer, if it is too complicated or deep it would be difficult to follow by students. So it depends on the ability of the lecturer of composing the most convenient resources in the best way.

4.2.4 Conclusion

Similarly to the previous experiment with R2 prototype (Section 3.2 - instructor's view) this experiment at UOC was also conducted by a real expert in developing complex computer systems. As professional developer and analysts (and on-line teacher), he is usually very demanding when evaluating a new software, especially if it is from the e-learning domain. Despite having a strong background in web applications as developers and user, he did not find many technical inconveniences with the IWT tools in the R3 scenario.

From the analysis of the usability of the lecturer considered the CLR Editor tool was very satisfactory. The lecturer' emotions when using the tool is also in line with the level of satisfaction and usability, and confirms from this perspective the tool is working very well. Therefore, the tool did not experience any major technical problem during the experiment and could be completed, thus achieving the main goal (G3.2.1).

Finally, the lecturer was very helpful and active, and provided some hints and suggestions for improvements at different levels. This leads to achieve the second goal of this scenario (G3.2.2).

To sum up, the lecturer liked the idea of personalizing a course by semantic connections that link different learning resources to fit the specific students' needs and different contexts. Depending on the topic to be taught, the lecturer thought that the lecturers and students could be benefitted from incorporating semantic connections between learning resources.

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 to be experienced and played by learners (VCS) [17]. During the resource execution, learners observe how avatars discuss and collaborate, how discussion threads grow, and how knowledge is constructed, refined and consolidated.

The goals and hypotheses formulated for this scenario are related to the final stage of the VCS prototype by following an iterative approach driven by empirical phases. First phase sets out a solid basis for the next experiments by evaluating the VCS with a CC-LO embedded already evaluated in the first round of experiments in order to validate the improvements made to this respect from then. Second phase is to evaluate the usability and functionality of the VCS tool to edit and play the current text-based discussion in a multimedia attractive format (CC-LR). To this end, an experiment was run to pilot this scenario from both the student' view in support for a formal in-class assignment of collaborative learning based on a discussion. Finally, the third phase was focused purposely on the cognitive and emotional aspects of the CC-LR as complex aspects. To this end an experiment was run to pilot this scenario from the student's view and also from the lecturer's view.

In overall, 4 trials were run at UOC on the R4 scenario, including the 3 mentioned experimental phases plus the experiment from the lecturer's view to fully experiment all the features of the WP3 prototypes and the impact in the learning and teaching process, as follows:

1. Live and Virtualized Collaboration: Experimenting with Collaborative Complex Learning Objects (CC-LO) from the student's viewpoint (Section 5.1)
2. Live and Virtualized Collaboration: Experimenting with Collaborative Complex Learning Resources (CC-LR) from the student's viewpoint (Section 5.2)
3. Live and Virtualized Collaboration: Live and Virtualized Collaboration: Experimenting with Collaborative Complex Learning Resources (CC-LR) enriched with authoring information from the student's viewpoint (Section 5.3)
4. Live and Virtualized Collaboration: Experimenting with Collaborative Complex Learning Resources (CC-LR) from the instructor's viewpoint (Section 5.4)

5.1 R4-1. Live and Virtualized Collaboration: Experimenting with Collaborative Complex Learning Objects (CC-LO) from the student's viewpoint

5.1.1 Evaluation and Validation Procedure

In the first phase of experimentation regarding live and virtualized collaboration, an experiment was conducted at UOC pilot site in order to test the virtualization of live sessions

of collaborative learning to produce storyboard learning objects (CC-LO) embedded in a virtualized collaborative session system (VCS) to be experienced and played by learners. During the resource execution, learners observed how avatars discuss and collaborate, how discussion threads grow, and how knowledge is constructed, refined and consolidated.

In the second phase we repeated the experimentation of the first phase in order to validate improvements of the VCS tools. In particular, the usability and functionality of the VCS tool to play and observe the 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).

In this experiment, we focused on the following scenario goals and hypotheses as well as criteria and metrics as described in [3]:

Scenario goals

- G4.1.1: To build a VCS system that is able to build a CC-LO from a threaded discussion (coming from a forum).
- G4.1.2: To employ the VCS in online courses in order to enhance some aspects of the teaching/learning process.
- G4.1.3: To identify possible ways of improving further the utility of the VCS in online courses.
- G4.1.4: To create, store and playback the generated storyboard through a user friendly interface.
- G4.1.5: To build (automatically) a draft storyboard from a collaborative activity effectively
- G4.1.6: To build (automatically) a draft storyboard from a collaborative activity efficiently

Scenario hypotheses

- H4.1.1: The VCS prototype allows non-expert users to build and use a Story Learning Object (i.e., in a friendly way and efficiently).
- H4.1.2: Use of VCS contributes to significantly improve students' motivation.
- H4.1.4: Use of VCS contributes to significantly increase students' activity levels, both in individual and collaborative activities.
- H4.1.5: Use of VCS contributes to significantly improve students' understanding of key concepts and students' results.
- H4.1.6: VCS are considered as a worthy educational resource by students.

Scenario criteria

- C4.1.1: Level of fulfillment of the VCS features.
- C4.1.2: Potential increase in students' motivation caused by the use of VCS.
- C4.1.4: Potential increase in students' activity levels due to the incorporation of the VCS.

- C4.1.5: Potential increase in students' understanding of concepts and students' results.
- C4.1.6: Level of satisfaction of students with the inclusion of the VCS in their courses.

Scenario metrics

- M4.1.1: Number of students using the VCS.
- M4.1.2: Number of visits of the VCS.
- M4.1.3: Number of visits of the standard forum.
- M4.1.4: Number of messages submitted by students related to the VCS topics.
- M4.1.5: Number of messages submitted by students when no VCS is used.
- M4.1.6: Number of words written by students when the VCS is used.
- M4.1.7: Number of words written by students when no VCS is used.
- M4.1.8: Number of students that consider that the VCS is worthy.

5.1.2 Method

5.1.2.1 Participants

In the same way as in the first phase of the experiments [6], the real context of this experience is the virtual learning environment of the Open University of Catalonia (UOC).

In order to evaluate the prototype of the VCS and analyze its effects in the discussion process, the sample of the experiment consisted of 44 undergraduate students enrolled in the course Organization Management and Computer Science Projects from the Bachelor in Engineering Computing degree at the UOC were involved in this experience. These 44 students formed the experimental group and participated in the current Spring term (the experiment took place in March-April 2012) while the control group participated in the Fall term of 2011 during the first phase of the experiments using a previous version of the VCS tool (see [6] for full details of the control group). None of the students belonged to both groups. Therefore, the results of the experimental group with the improved VCS tool will be compared to the control group already evaluated and validated in the initial experiments.

Despite all 44 students participated in the experience, only 40 out of them (91%) submitted the final questionnaire, the rest of students (4) dropped out the discussion and the course for personal reasons. It is worth mentioning that the 9% dropout ratio found is considered very low in the first third of the academic term when the experience was run¹. This was caused by the expectations created by the innovative tool that increased the students' motivation as described in section 5.1.4.2. Eventually this higher number of participants allowed for obtaining more empirical data from the experience.

Each group was supervised by one tutor as the official lecturer teaching the whole course. The lecturer's view of the R4 scenario is analyzed in detail in Section 5.4.

¹ 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.

5.1.2.2 Apparatus and Stimuli

Students of the experimental group were required to use standard forum IWT was equipped with the multimedia-based VCS tool (see *Figure 41*).



Figure 41: 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 (with the VCS), which included a motivation test; (iv) test-based evaluation of the VCS; (v) test-based evaluation on the usability of the system; (vi) test-based evaluation on the emotional state; (vii) a test-based evaluation of the questionnaire.

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 actual experimental group en the last semester experimental group.

For the section v (usability of the forum tools with VCS) we used the System Usability Scale (SUS) developed by [8] 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 using the new system, which included 12 items of the Computer Emotion Scale (CES) [9]. The 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/dispirted.”)
- 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 the 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.

5.1.2.3 Procedure

An in-class collaborative formal assignment in the experimental group lasted three weeks during the first third of the Spring term (March-April 2012) and consisted of discussing the 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.

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.

This procedure is the same for the control group that participated in the Fall term of 2011 in the same type of in-class collaborative discussion with the same topic of the discussion, same questionnaires and same evaluation and validations rules. All this data was already reported in D8.1.1 (see [6]).

5.1.3 Evaluation Results

Following the methodology described in Section 5.1.2, in this section we focus on the activity (H4.1.4), usability and emotional (H4.1.1) aspects of the VCS tool of both the control and experimental group. For these purposes we used metrics M4.1.1 – M4.1.7. 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.

5.1.3.1 Activity level fostered by the VCS

In order to evaluate the students' activity levels with the VCS, we collected and analyzed data by comparing the participation behaviour of the actual experimental group and the last semester's control group as shown in Table 8:

Metric / Statistic	Control group [6] 1 st Experimentation (Fall 2011) Standard forum (+VCS)	Experimental group 2 nd Experimentation (Spring 2012) Standard forum (+VCS)
Number of students	41	40
Total of posts	156	121
Mean posts/student	M=3.7	M=3.03
SD posts/student	SD=2.0	SD=1.56
Total words	26669	18941
Mean words/student	M=634.9	M=473
SD Mean words/student	SD=406.8	SD=235.28
Total words	26669	18941
Mean words/post	M=170.9	M=156.54
SD Mean words/post	SD=116.1	SD=66.98

Table 8: Results on activity levels of the discussion in both experimental groups.

For the posts and words 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.

We can observe that for the experimental group, the number of both posts and words are lower than the control group though the posts are less worded, thus becoming much tighter (i.e., concise). Also the SD statistic improves a great deal in the experimental group. Therefore the improvements made in the VCS influence the levels of activity in terms of writing fewer posts with fewer words but more homogenous and concise (i.e. more quality).

5.1.3.2 Usability of the VCS

To evaluate student's satisfaction of the experimental group with the tool as for an efficient and user-friendly management (H4.1.1), we collected data 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 5.1.2.2) 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 **40 SUS scores of 64.87**, thus nearby SUS mean and above the control group (38 SUS scores of 63.02). Next, we present the most relevant results of the SUS score by providing several statistics: Mean (M), Standard Deviation (SD) and Median (Md).

The results of the experimental group and control group are the following:

- Students of the **experimental group** (n=40) did not find the VCS unnecessarily complex (M = 2.15, SD = 0,88, Md = 2) (See *Figure 42*). Students found the tool particularly easy to use (M = 3.34, SD = 1.04, Md = 3.5) (See *Figure 43*). In addition, students stated that they did not need the support of a technical person to be able to use the VCS (M = 1.84, SD = 0.91, Md = 2) (*Figure 44*) and they thought that most people would learn to use this system very quickly (M = 3.73, SD = 0.89, Md = 4) (See *Figure 45*).
- Students of the **control group** (n=41, see [6] for the graphical results) found the tool particularly easy to use (M = 3.47, SD = 1.00, Md = 3). Students did not find the VCS unnecessarily complex (M = 2.27, SD = 0.97, Md = 2). 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).

At this point, the results of both groups are very similar, being the experimental group slightly better than the control group in all the above aspects and in accordance with the SUS score. This confirms the students of the experimental group realized the improvements made on usability in the second phase of the project considering the feedback received in the initial experiments.

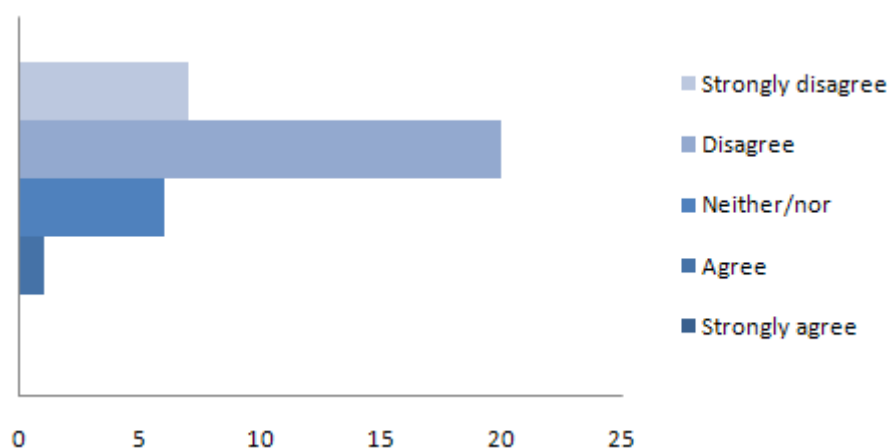


Figure 42: Results on the SUS item "I found the VCS unnecessarily complex".

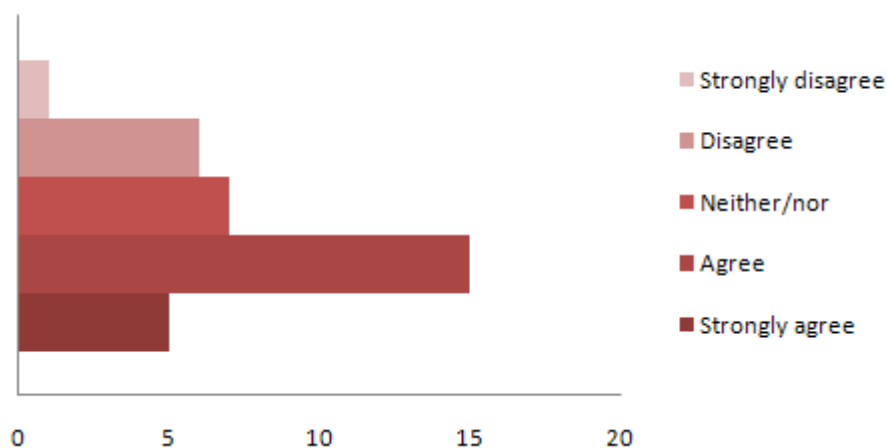


Figure 43: Results on the SUS item “I thought the system was easy to use”.

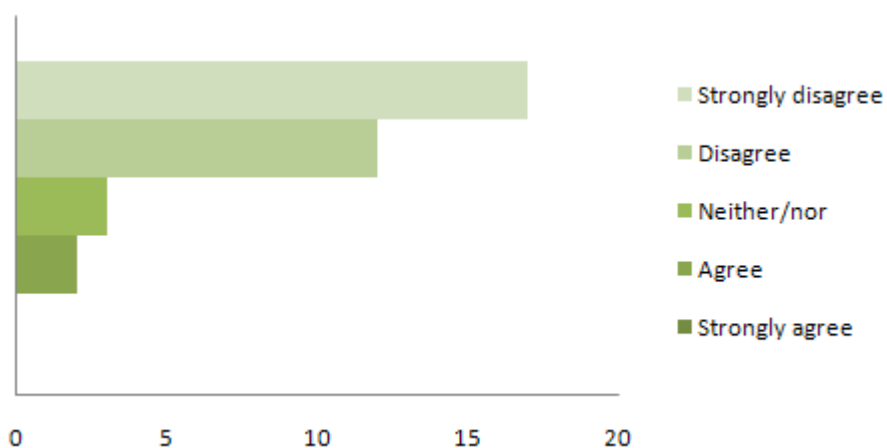


Figure 44: Results on the SUS item “I think that I would need the support of a technical person to be able to use the VCS”.

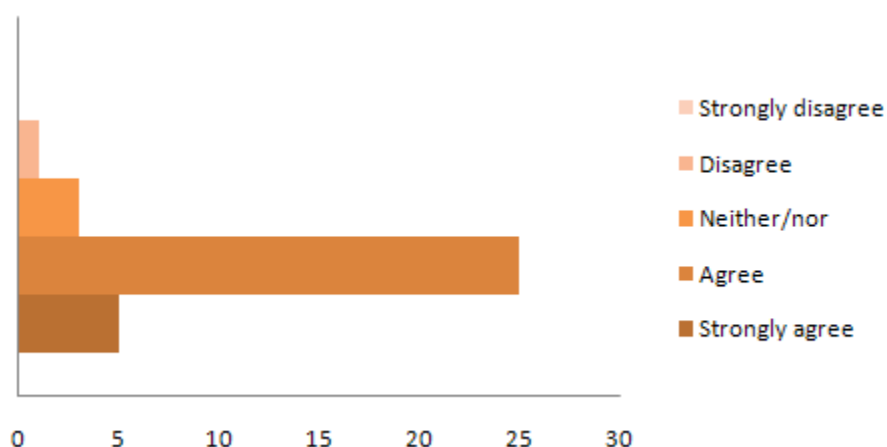


Figure 45: Results on the SUS item “I would imagine that most people would learn to use the VCS system very quickly”.

Moreover, students of the experimental group stated that the VCS functionality was well integrated ($M = 3.47$, $SD = 1.00$, $Md = 4$) (Figure 46) and the tool itself was adequately integrated in the UOC virtual campus. This result is clearly better than the control group ($M = 3.25$, $SD = 1.01$, $Md = 3$) In particular, students of the control group reported technical problems to gain access that the experimental group did not. Both groups 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.

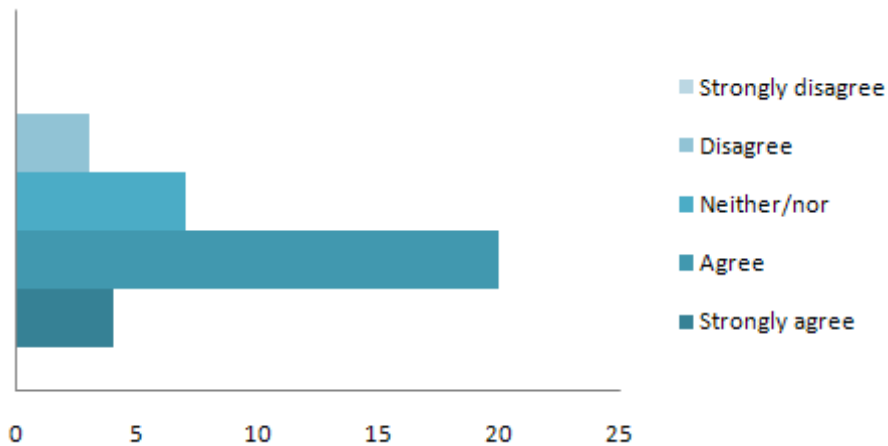


Figure 46: Results on the SUS item “I found the various functions in the VCS were well integrated”.

Finally, students indicated in a balanced way they would and would not use the VCS system frequently ($M = 2.92$, $SD = 0.95$, $Md = 3$) (Figure 47). Despite this result is slightly lower than the control group ($M = 2.97$, $SD = 1.16$, $Md = 3$), is in line with the overall SUS score of 64.87 nearby SUS mean (68).

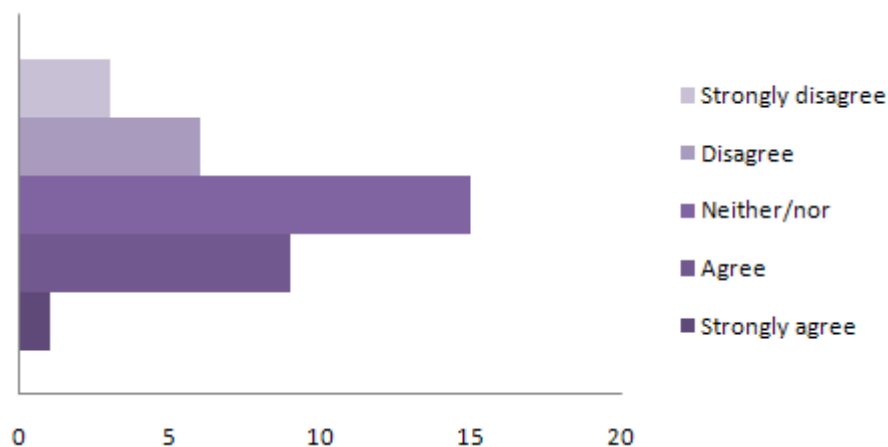


Figure 47: Results on the SUS item “I think that I would like to use this system frequently”.

In summary, the improvements made to the VCS tool on usability were noticeable by the students of the experimental group who in general did not report the problems found by the control group.

5.1.3.3 Emotional aspects

Regarding the students' emotions of the experimental group during the work with the VCS tool (H4.1.1), the results from a 4-point rating scale (n=40) are as follows and they are compared to the results of the control group (n=41, see [6] for a graphical results):

- Happiness (M=1.05, SD=0.81, Md=1) (Figure 48). This result is clearly better than the control group (M=0.95, SD=0.89, Md=1) showing they were especially curious and satisfied with the new system.

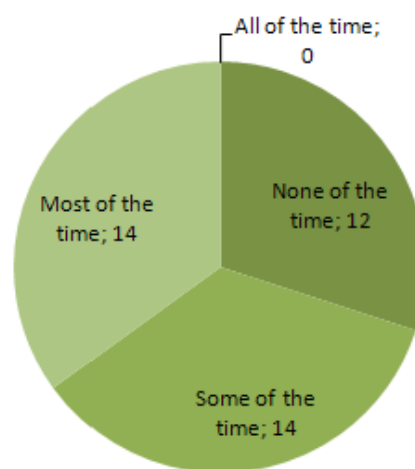


Figure 48: Results on the Happiness emotion

- Sadness (M=0.32, SD=0.65, Md=0) (Figure 49). This result is slightly worse than the control group (M=0.24, SD=0.49, Md=0). However, both results are very good with Md=0, which means that students of both groups did not experienced this bad feeling.

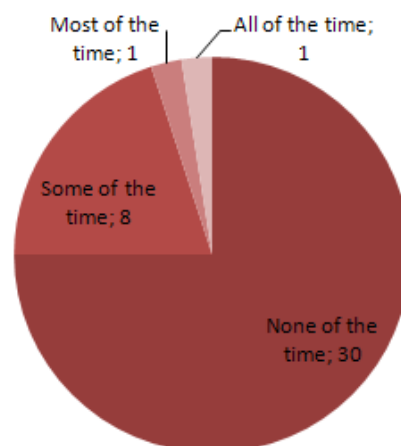


Figure 49: Results on the Sadness emotion

- Anxiety ($M=0.15$, $SD=0.36$, $Md=0$) (Figure 50). This result is slightly better than the control group ($M=0.21$, $SD=0.47$, $Md=0$). However, both results are very good with $Md=0$, which means that students of both groups did not experienced this bad feeling.

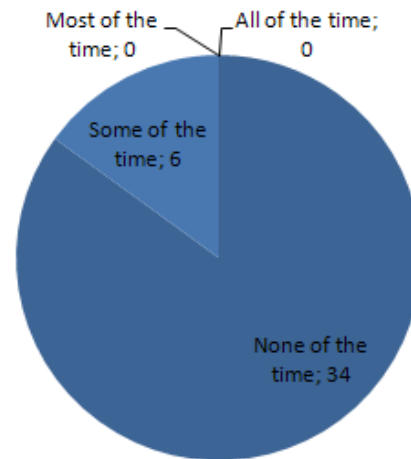


Figure 50: Results on the Anxiety emotion

- Anger ($M=0.22$, $SD=0.42$, $Md=0$) (Figure 51). This result is slightly better than the control group ($M=0.24$, $SD=0.49$, $Md=0$). However, both results are very good with $Md=0$, which means that students of both groups did not experienced this bad feeling.

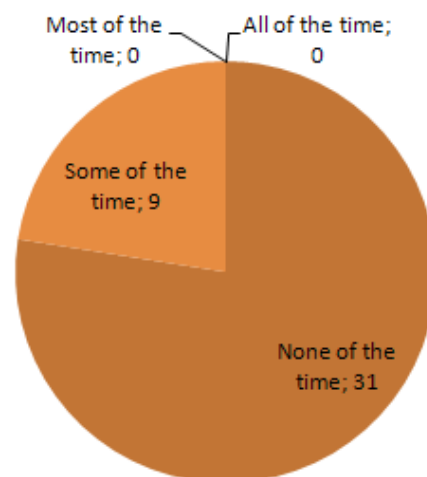


Figure 51: Results on the Anger emotion

In summary, students felt more often happiness than sadness, anxiety or anger when learning with the new VCS tool. The results in general are better in the experimental group than the control group though, being the most noticeable result the increase in happiness.

The students felt the same level of sadness, anxiety and anger emotions, which were very low, almost inappreciable (Md=0), being the anxiety emotion the lowest.

In overall, this is a good result and is in line with the results presented above concerning the improvement of usability of the VCS tool from the SUS mean (see Section 5.1.3.2). We can conclude with a co-relation on improvements between usability and emotions, both slightly better in the experimental group than in the control group.

5.1.3.4 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 out and submitting the questionnaire in less than 30 minutes (Figure 52) and 97% of them found it appropriate to evaluate the experience (Figure 53) (n=40).

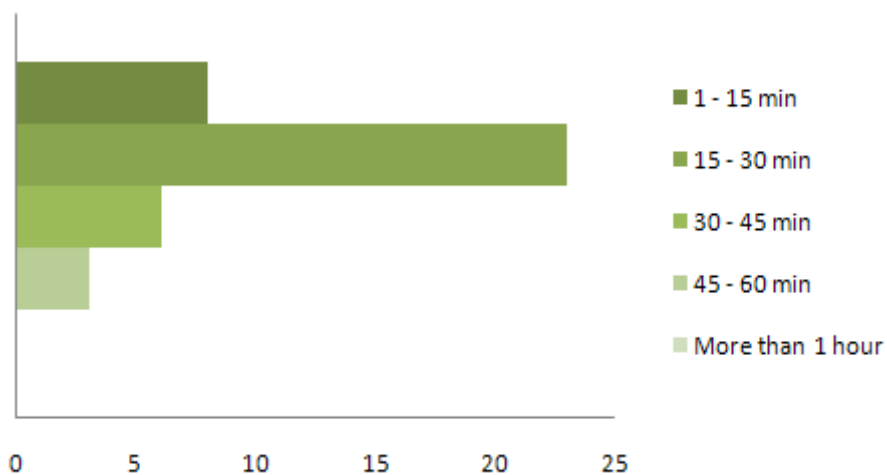


Figure 52: Time employed to fill the questionnaire

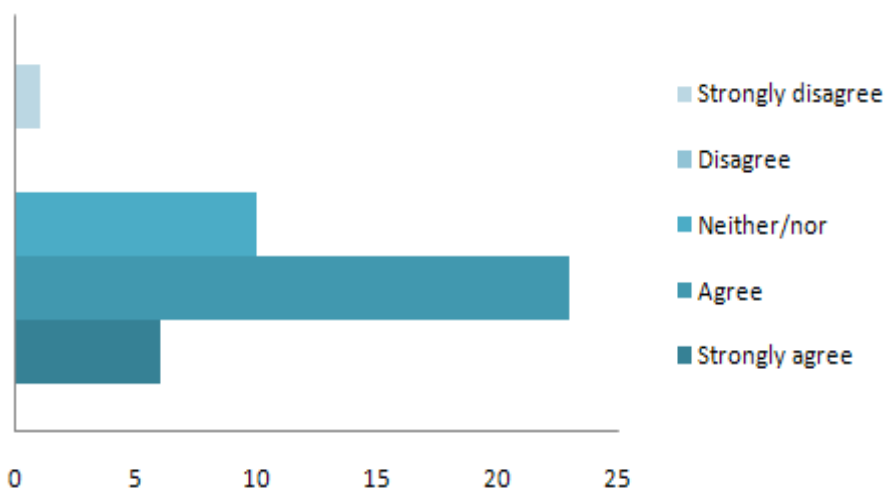


Figure 53: Appropriateness to evaluate the experience with the questionnaire

5.1.4 Validation Results

Following the methodology set out in Section 5.1.2 we will validate the improvement of motivation (H4.1.2), worthiness from the control group of the VCS as an educational tool (H4.1.3 and H4.1.6) as well as the acquisition of knowledge achieved with this tool (H4.1.5). For these purposes we used the metrics M4.1.1 and M4.1.8.

5.1.4.1 The VCS as a valuable resource

In this section we evaluate the level of worthiness of the VCS as an educational tool (H4.1.6). To this end, we collected quantitative and qualitative data in order to know the user's satisfaction in the experimental group 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 (H4.1.3). All this data was also collected with the same questionnaire and questions from students of the control group (see [6] for details). This will make it possible a fair comparison between both groups.

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 under 5.0.

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).

Six students did not provide assessment marks (15%) because either the student could not use the VCS (lack of speakers, technical problems, etc.) and thud followed the discussion by the text messages or the student was not interested in that part of the questionnaire. We computed a by default value for these questions by the average mark of the rest of responses to the related question, thus not affecting the overall results.

After calculating the 0-10 scale for all the questions of the experimental group we got a general mean score of 5.43 (SD=2.20 and Md=5.60). This result is significantly better than the control group (M=4.98, SD=1.78, Md=5) [6] and in line with the previous results on usability and emotions, hence these results confirm the improvements made in the VCS tool.

In particular, students of the experimental group liked the VCS tool (Question 1: M=6.34, SD=1.80, Md=6) and liked it more than the control group (Question 1: M=6.07, SD=1.63, Md=5) [6]. Similarly to the control group [6], 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 (more comfortable). Also students felt the system was more “real”, meaning that it was closer to a real discussion with the presence of the discussants.

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 engine due to syntax problems in the posts and especially the “robotic” voice of the VCS was found quite annoying for some students. The problems with syntax will be easily solved in the next development steps by the incorporation of the VCS Editor (see Section 5.2). Finally, some students indicate the benefits of the VCS tool for disable students.

The analysis from comparing participation with and without the VCS tool scoped Questions 2, 3 and 4. All of them had similar results (M=4.38 – 4.88, SD=2.26 - 2.69, Md=5). These results are better significantly than the control group (M=4.28 - 4.34, SD=2.63 - 3.07, Md=5) [6], and though they are low, they are on the average score (Md=5). Students in general indicated that they preferred the text format of the posts and the VCS did not foster their participation because “listening to” the posts was slower than reading them. However, some students admitted that they used the VCS to follow the discussion in a natural order close to reality, which was not possible with the text format and this fostered their participation. Others mentioned that both text and video formats were compatible and had different purpose: while the text format is good to review fast the whole discussion, the video format facilitates knowledge retention by listening to the entire posts and understanding better. These comments about the influence of the VCS in the participation and the benefits with respect to text-based discussion were either unique or much more emphasized than the control group [6].

The improvements in performance and efficiency were particularly good (Question 5: M=6.53, SD=1.97, Md=7) and significantly better than the control group (M=5.68, SD=1.67, Md=5) [6]. Students reported fewer technical problems related to installation and execution of the VCS tool than the control group. Most of the students mentioned the system was easy to use and fast. The graphical interface was found simple, pleasant and intuitive. Only a very few students noticed the video generation performed slowly at the beginning but for the rest of the video it performed fast. These results and comments are in line with the results achieved about usability and emotions in the Section 5.1.3.

Finally, students found many advantages of the VCS by exploiting its potential appropriately (Question 6: M=5.98, SD=2,17, Md=6). This result is also significantly better than the control

group ($M=5.20$, $SD=2$, $Md=5$) [6]. In particular, unlike the control group, the experimental group commented that the VCS was useful and performance and visualization was quite good. This confirms the improvements made in the VCS tool. Also, these students confirmed the VCS helped them observe the knowledge construction more effectively than the text-based discussion, thus increasing the knowledge retention. A particular comment summarized this benefit by the following: “A picture speaks a thousand words”. Finally, some students proposed to promote the use of VCS system in other courses of the UOC.

These students also gave many comments and hints for possible improvements of the tool.

5.1.4.2 *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 (see [6] for the data on motivation in the initial experiments).

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 ($M=3.07$, $SD=0.75$, $Md=3$) than the control group ($M=2.85$, $SD=0.69$, $Md=3$) [6]. The results of the experimental group are in line with the results reported in Section 5.1.4.1. In line with the control group [6], the students of the experimental group 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 “real” and that “A picture speaks a thousand words” meaning that the video format helped them understand better the posts. As a result they were more engaged and self-motivated in the discussion than the control group, in line with the results achieved on emotions in Section 5.1.3.3. Finally, clear indications of amounts of motivation came from enthusiastic students who evaluated the VCS tool as “I liked it a lot!”, “spectacular!”, “very interesting”, “nice”, “surprising”. On the other hand, students who chose not use the VCS tool due to lack of time or technical problems felt unmotivated. These comments are similar to the control group [6] though the experimental group put more emphasis in the benefits and good aspects and less in the problems, which is in line with the better results achieved.

5.1.4.3 *Tutor assessment and knowledge acquisition*

All students were evaluated on summarizing the discussion in both the experimental and the control groups (see [6] for the data on tutor assessment). 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 9 shows the results.

Evaluative questions	Control group [6]	Experimental group
	1 st Experimentation (Fall 2011) Standard forum (+VCS) (n=41)	2 nd Experimentation (Spring 2012) Standard forum (+VCS) (n=40)
Question 1	M=6.84 SD=1.48 Md=7	M=6.43 SD=1.84 Md=7
Question 2	M=7.68 SD=1.18 Md=8	M=7.71 SD=1.52 Md=8
Question 3	M=7.21 SD=1.45 Md=7	M=7.26 SD=1.17 Md=7
Overall	M=7.24 SD=1.41 Md=7	M=7.13 SD=1.51 Md=7

Table 9: Results of the discussion evaluation

From the results of Table 9, students from the experimental group scored similar to the control group [6]. Both groups got good marks on average and showed a good level of knowledge acquisition. These results confirm the initial experimentations that the VCS used to create animated storyboard does not have an evident impact in the knowledge acquisition.

5.1.5 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 5.1.1). Then, based on the results summarized they are compared with the initial experiments.

In general the students confirmed in this final experiment that they liked the VCS tool and found it interesting to have another option to follow the in-class discussion assignments (G4.1.3). During this specific assignment, students could generate the storyboard from the VCS (G4.1.1) and it was effective to support the discussion for review and summary purposes (G4.1.5). These results were better than in the initial experimentation, which validate the improvements made in the prototypes in the second phase of the project.

Unlike the initial experiments, no relevant technical problems were reported and most of students could generate the storyboard (SLO) efficiently (G4.1.6) and create, store and playback it as many times as needed (G4.1.4). Aspects of the learning process, such as

motivation and emotional were validated by showing an impact of the use of the VCS tool on these aspects (G4.1.2) and by comparing them with the results obtained in the first phase, noticing an increase in usability, emotions when using the tools and also motivation.

Finally, the VCS was proved to become an worth educational resource by assessing several aspects of the learning process, such as knowledge construction and participation, and then compare the results with the initial experiments. The gain in knowledge acquisition by using the VCS could also be validated by comparing the gain of knowledge with the initial experiments, though the results obtained were not significant.

5.2 R4-2. Live and Virtualized Collaboration: Experimenting with Collaborative Complex Learning Resources (CC-LR) from the student's viewpoint

5.2.1 Evaluation and Validation Procedure

In the previous experiment regarding live and virtualized collaboration (see Section 5.1), an experiment was conducted at UOC pilot site in order to test the virtualization of live sessions of collaborative learning to produce storyboard learning objects (CC-LO/SLO) embedded in a virtualized collaborative session system (VCS) to be experienced and played by learners. During the resource execution, learners observed how avatars discuss and collaborate, how discussion threads grow, and how knowledge is constructed, refined and consolidated. The experimentation results were compared to the same experiment performed during the initial experimentations in order to validate the improvements made in the VCS tool in the second phase of the project.

The purpose of this new experiment is to validate the use of CC-LO as complex learning resources (CC-LR) to be provided to students as regular learning material. A CC-LR (also called video-debates) leverages live collaborative sessions as animated storyboards (CC-LOs) such that learners can observe with the VCS Player how people discuss and collaborate, and how knowledge is constructed. The development of a VCS Editor tool provides lecturers and experts with edition capabilities of the CC-LRs, such as cutting scenes, modifying involved characters, selecting emotional states, dialogues and connected concepts. See Figure 54 bellow and [17] for more details.

To experiment with the CC-LR approach and evaluate and validate it as for the usability, emotional, impact in the learning process, etc., from the student's viewpoint, we focused on the following scenario goals and hypotheses as well as criteria and metrics derived from [3]:

Scenario goals

- G4.2.1: The VCS Editor system that is able to build a CC-LR from a threaded discussion (coming from a forum).
- G4.2.2: To employ the CC-LR in online courses in order to enhance some aspects of the teaching/learning process.

- G4.2.3: To identify possible ways of improving further the utility of the CC-LR in online courses.
- G4.2.4: To create, store and playback the generated CC-LR through a user friendly interface.
- G4.2.5: To build (automatically) a draft CC-LR from a collaborative activity effectively
- G4.2.6: To build (automatically) a draft CC-LR from a collaborative activity efficiently

Scenario hypotheses

- H4.2.1: Use CC-LR by non-expert users (i.e., in a friendly way and efficiently).
- H4.2.2: Use of CC-LR contributes to significantly improve students' motivation.
- H4.2.4: Use of CC-LR to significantly increase students' activity levels, both in individual and collaborative activities.
- H4.2.5: Use of CC-LR contributes to significantly improve students' understanding of key concepts and students' results.
- H4.2.6: CC-LR is considered as a worthy educational resource by students.

Scenario criteria

- C4.2.1: Level of fulfillment of the VCS Editor features.
- C4.2.2: Potential increase in students' motivation caused by the use of CC-LR.
- C4.2.4: Potential increase in students' activity levels due to the incorporation of the CC-LR.
- C4.2.5: Potential increase in students' understanding of concepts and students' results.
- C4.2.6: Level of satisfaction of students with the inclusion of the CC-LR in their courses.

Scenario metrics

- M4.2.1: Number of students using the CC-LR.
- M4.2.2: Number of visits of the CC-LR.
- M4.2.3: Number of students passing the course and/or with high marks when the CC-LR is used.
- M4.2.4: Number of students passing the course and/or with high marks when CC-LR is not used.
- M4.2.5: Number of students that consider that the CC-LR is worthy.

5.2.2 Method

5.2.2.1 Participants

In the same way as in the previous experiment (see Section 5.1), the real context of this experience is the virtual learning environment of the Open University of Catalonia (UOC).

In order to evaluate the CC-LR and analyze its effects in the discussion process, the sample of the experiment consisted of 44 undergraduate students enrolled in the course Organization Management and Computer Science Projects from the Bachelor in Engineering Computing degree at the UOC were involved in this experience.

These same 44 students formed two groups and both participated during the Spring term of 2012 in the same course: the control group participated at the beginning of the course (March-April 2012) while the experimental group participated at the middle of the course (May 2012). All details about the experiment and results of the control group are found in Section 5.1.

Despite all 44 students participated in this experience, only 25 out of them (57%) submitted the final questionnaire, the rest of students (19) dropped out the discussion and the course for personal reasons. It is worth mentioning that the 43% dropout ratio found is considered normal in at the end of the academic term when the experience was run².

Each group was supervised by the same tutor as the official lecturer teaching the whole course.

5.2.2.2 Apparatus and Stimuli

Students of the experimental group were required to use standard forum IWT was equipped with the multimedia-based VCS tools, SLO Editor and Player (see Figure 54 and [7]).

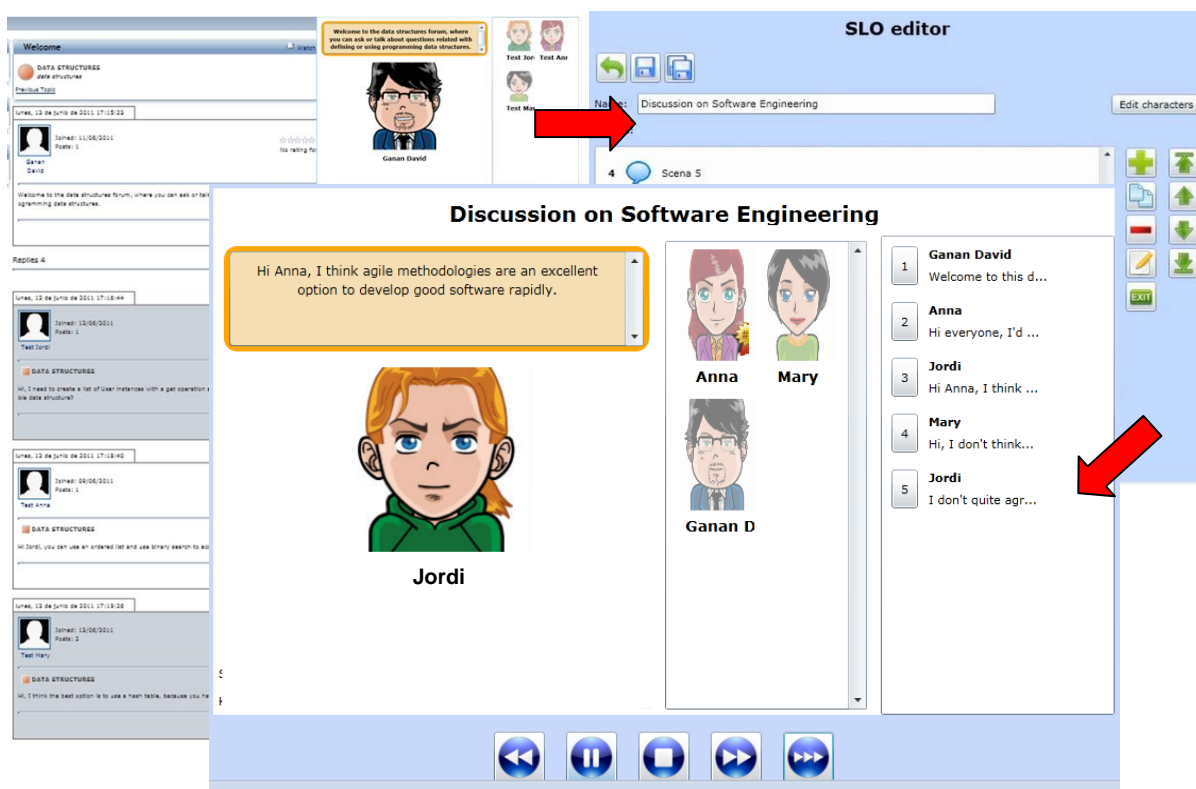


Figure 54: An SLO from a live discussion in IWT is edited by the VCS-SLO Editor to modify the involved characters, improve the text, cut non relevant scenes, etc. and eventually create a video-debate (CC-LR) as a new type of learning material that reuse live sessions.

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.

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 video-debates (CC-LR), which included a motivation test; (iv) test-based evaluation of the video-debates; (v) test-based evaluation on the usability of the VCS system; (vi) test-based evaluation on the emotional state; (vii) a test-based evaluation of the questionnaire.

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 experimental group and the control group.

For the section v (usability of the VCS player showing the video-debates) we used the System Usability Scale (SUS) developed by [8] 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 using the new system, which included 12 items of the Computer Emotion Scale (CES) [9]. The 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 the 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.

5.2.2.3 Procedure

A formal learning activity called “Practica 2” was scheduled during the last 3 weeks of May 2012 in the mentioned course of Organization Management and Computer Science Projects from the Bachelor in Engineering Computing degree at the UOC. The participants of the experimental group participated in this activity.

The activity was individual and mandatory for all students and consisted in developing a software project from the management perspective. A part from the usual didactical material of the course the students of the experimental group also received a new material to support specifically this activity in the form of a video-debate (CC-LR) called “Factors that lead a Computer Science project to failure” which contained a discussion about project management. The students entered IWT to find and watch this interactive video-debate.

Finally, since the participants of the experimental group were the same as the participants of the previous experiment (see Section 5.1), they had been already experimented with the VCS tools. Therefore, we will compare the evaluation and validation aspects with those students of the previous experiment forming the control group. All this data was already reported in Sections 5.1.3 and 5.1.4 and the feedback provided served to improve the prototypes for this new experiment.

5.2.3 Evaluation Results

Following the methodology described in Section 5.1.2, in this section we focus on activity levels, usability and emotional aspects of the video-debates (CC-LR) (H4.2.1) of both the control group (see Section 5.2) and experimental group. For this purpose we used metrics M4.2.1 and M4.2.2. 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.

5.2.3.1 Activity levels

The CC-LR activity is measured in terms of number of executions (i.e. start a reproduction) of the video-debates and raw interactivity (e.g. buttons to move forward, back, etc). From the log files that monitor the activity with the video-debate during the time scheduled (from May 12, 2012 to May 25, 2012), we count 166 video reproductions and 4929 interactions.

As $n=25$, each student reproduced the video-debate more than 6 times on average and interact with the video-debate 197 times on average. This is considered a good results since all students visualized the video-debate and all of them shown a good level of activity when watching the video-debate, which means that they were engaged in the learning material.

5.2.3.2 Usability of the video-debates (CC-LR)

To evaluate student’s satisfaction of the experimental group with the tool as for an efficient and user-friendly management (H4.2.1), we collected data from students’ ratings and open comments on the usability/functionality/integration of the tool.

To investigate the overall usability of the video-debates, we used the SUS (see Section 5.2.2.2) 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 **25 SUS scores of 68.20**, thus above the SUS mean and above the control group (40 SUS scores of 64.87) (see Section 5.1.2.2). Next, we present the most relevant results of the SUS score by providing several statistics: Mean (M), Standard Deviation (SD) and Median (Md).

The results of the experimental group and control group as for the usability of the VCS to watch the video-debates are the following:

- Students of the **experimental group** (n=25) thought they will use the video-debates often (M = M = 3.13, SD = 1.29, Md = 4) (See Figure 55). Students found the video-debates particularly easy to use (M = 3.47, SD = 0.96, Md = 4) (See Figure 56). In addition, students stated that they did not need the support of a technical person to be able to use the video-debates (M = 1.47, SD = 0.65, Md = 1) (Figure 57), they thought that most people would learn to use this system very quickly (M = 4.04, SD = 0.61, Md = 4) (See Figure 58), and they felt quite confident using the video-debates (M = 3.69, SD = 0.80, Md = 4) (See Figure 59).
- Students of the **control group** (n=40, see Section 5.1.3.2 for the graphical results) thought they will use the video-debates often (M = 2.92, SD = 0.95, Md = 3). Students found the video-debates particularly easy to use (M = 3.34, SD = 1.04, Md = 3.5). Students stated that they did not need the support of a technical person to be able to use the video-debates (M = 1.84, SD = 0.91, Md = 2), they thought that most people would learn to use this system very quickly (M = 3.73, SD = 0.89, Md = 4) and they felt confident using the video-debates (M = 3.42, SD = 1.00, Md = 3.5).

So far, the results of the experimental group are significantly better than the control group in line with the SUS score found. This confirms the students of the experimental group found the usability of the video-debates is satisfactory or very satisfactory as well as they realized the improvements made on usability of the video-debates after the previous experiments run a few weeks before.

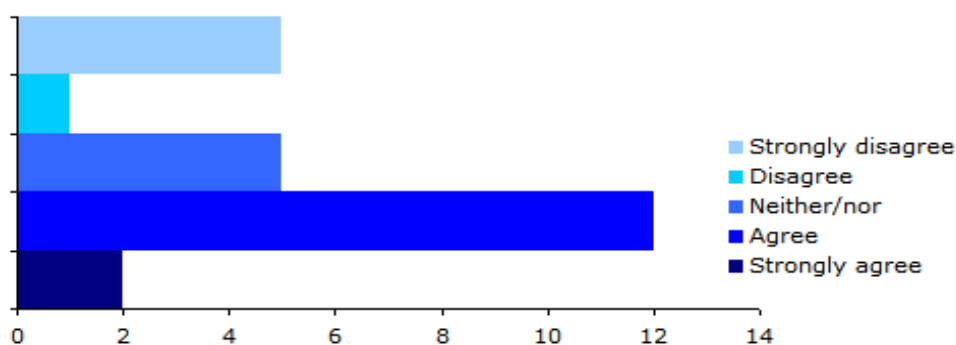


Figure 55: Results on the SUS item "I think I will use the video-debates often"

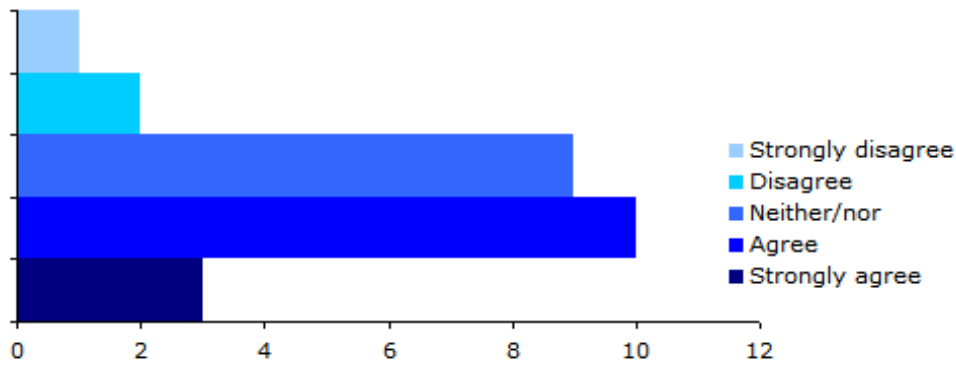


Figure 56: Results on the SUS item “I thought the video-debate was easy to use”.

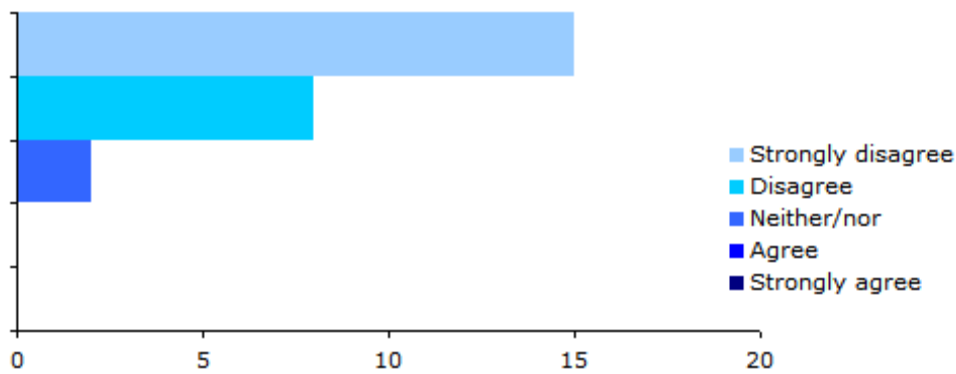


Figure 57: Results on the SUS item “I think that I would need the support of a technical person to be able to use the video-debate”.

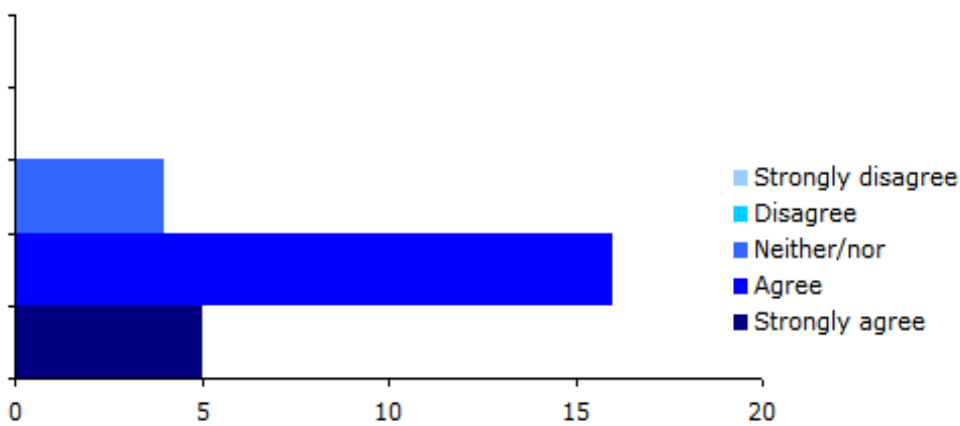


Figure 58: Results on the SUS item “I would imagine that most people would learn to use the video-debate system very quickly”.

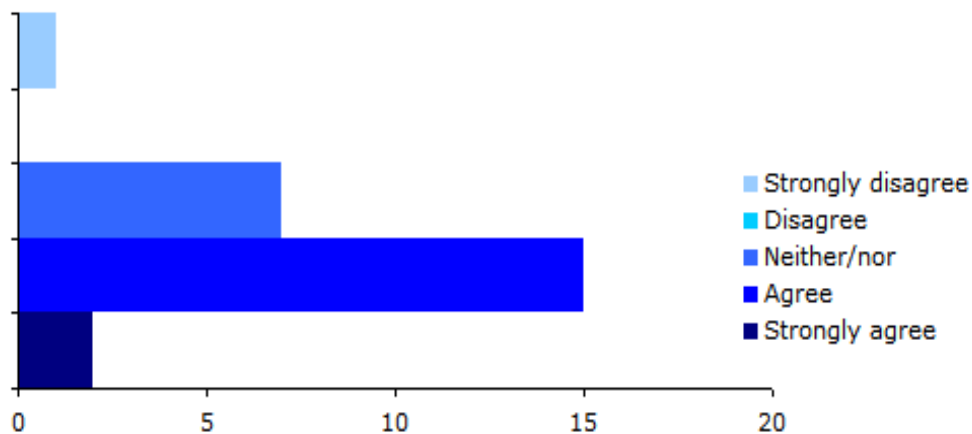


Figure 59: Results on the SUS item "I felt very confident using the video-debate".

Moreover, students of the experimental group stated that the video-debate functionality was well integrated ($M = 3.52$, $SD = 0.96$, $Md = 4$) (Figure 60) and the tool itself was adequately integrated in the UOC virtual campus and in turn in IWT (see Annex A). This result is better than the control group though very similar ($M = 3.47$, $SD = 1.00$, $Md = 4$), which confirms that students did not have problems to gain access to the video-debates in IWT.

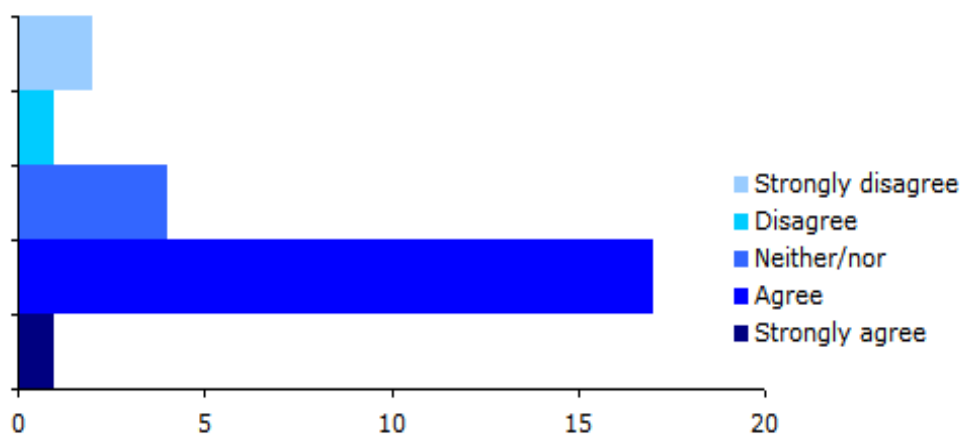


Figure 60: Results on the SUS item "I found the various functions in the video-debate were well integrated".

Finally, students indicated in a balanced way they found the video-debates unnecessarily complex ($M = 2.43$, $SD = 1.22$, $Md = 2$) (Figure 61). Despite this result is slightly lower than the control group ($M = 2.15$, $SD = 0.88$, $Md = 2$), it should be considered for this usability aspect the difference of purpose of the VCS for converting text-based forum into animated form (control group) and the use of the VCS to watch video-debates as learning material. The latter was compared by many students to written books they use normally as learning material to study.

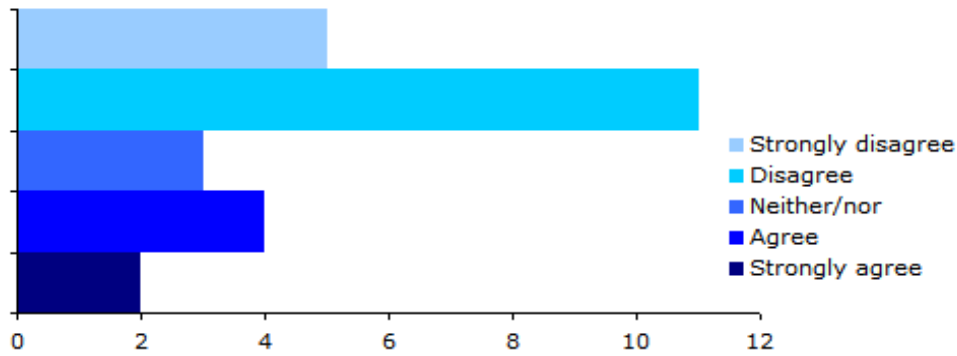


Figure 61: Results on the SUS item “I think the video-debate was unnecessarily complex”.

In summary, the improvements made to the VCS tool on usability from the control group were noticeable by the students of the experimental group using the VCS to watch video-debates who in general did not report the problems found by the control group.

5.2.3.3 Emotional aspects

Regarding the students’ emotions of the experimental group during the work with the video-debate tool to watch the video-debates (H4.2.1), the results from a 4-point rating scale (n=25) are presented next, and they are compared to the results of the control group (n=40). See Section 5.1.3.3 for the graphical results of the control group:

- Happiness (M=1.08, SD=0.90, Md=1) (Figure 62). This result is better than the control group (M=1.05, SD=0.81, Md=1) showing they were especially curious and satisfied on the video-debates as a new type of learning material despite they already knew from the previous experiment the video-debate tool that play them.

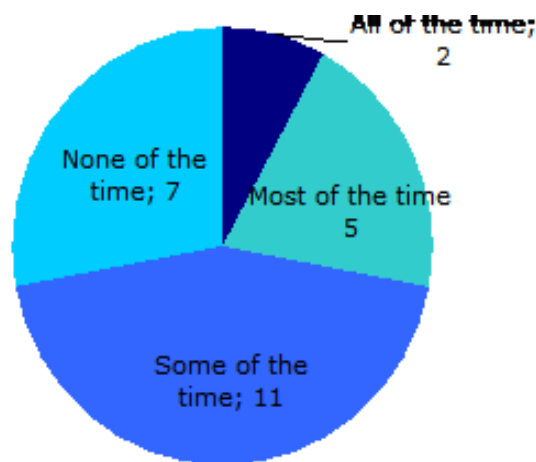


Figure 62: Results on the Happiness emotion

- Sadness ($M=0.48$, $SD=0.58$, $Md=0$) (Figure 63). This result is slightly worse than the control group ($M=0.32$, $SD=0.65$, $Md=0$). However, both results are very good with $Md=0$, which means that students of both groups did not experienced this bad feeling.



Figure 63: Results on the Sadness emotion

- Anxiety ($M=0.32$, $SD=0.69$, $Md=0$) (Figure 64). This result is slightly worse than the control group ($M=0.15$, $SD=0.36$, $Md=0$). However, both results are very good with $Md=0$, which means that students of both groups did not experienced this bad feeling.

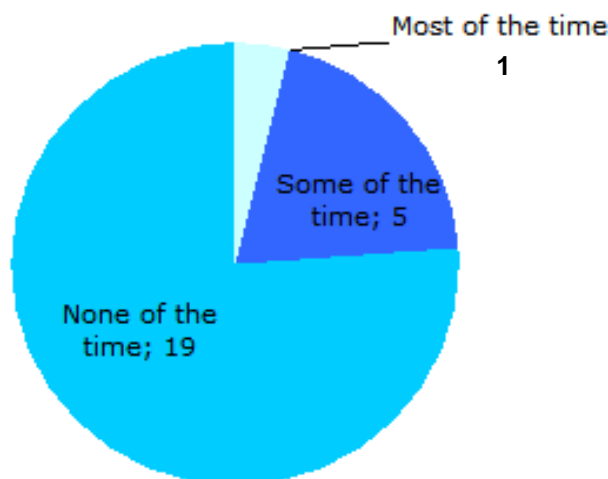


Figure 64: Results on the Anxiety emotion

- Anger ($M=0.40$, $SD=0.76$, $Md=0$) (*Figure 65*). This result is slightly worse than the control group ($M=0.22$, $SD=0.42$, $Md=0$). However, both results are very good with $Md=0$, which means that students of both groups did not experienced this bad feeling.



Figure 65: Results on the Anger emotion

In summary, students felt more often happiness than sadness, anxiety or anger when using the video-debates to study the new learning material (video-debates) The results in general are similar in the experimental group and the control group though, being the most noticeable result the highest value in happiness while the students felt the same level of sadness, anxiety and anger emotions, which were very low, almost inappreciable ($Md=0$), being the anxiety emotion the lowest.

In overall, this is a good result considering the students faced a new type of learning material and implicitly they assessed from both the pedagogical and technological perspective of the video-debates as a new type of learning material. Finally, this result is in line with the results presented above concerning the activity levels shown and the improvement of usability of the video-debate tool from the SUS mean (see Section 5.2.3.1 and 5.2.3.2).

5.2.3.4 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 an average time to fill out the questionnaire of about 30 minutes (*Figure 66*) and 72% of students found it appropriate to evaluate the experience (*Figure 67*) ($n=25$). This result is not far from the control group despite the questionnaire was more complex this time in order to evaluate a new type of learning material.

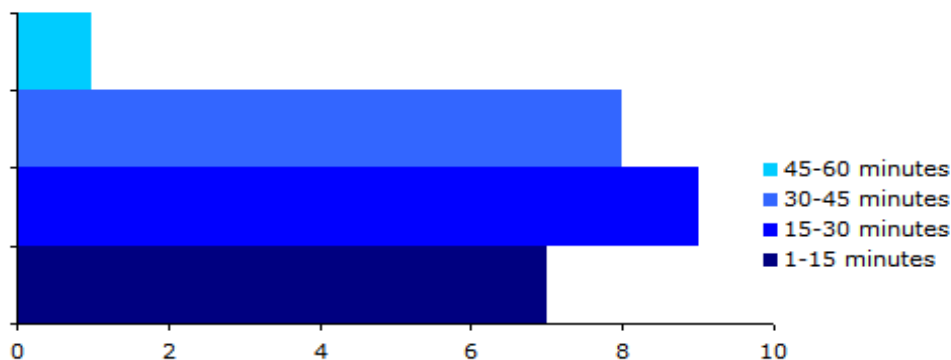


Figure 66: Time employed to fill the questionnaire

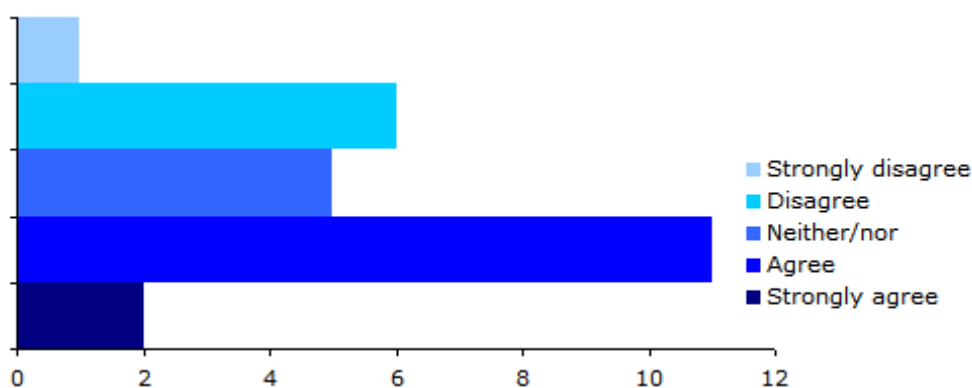


Figure 67: Appropriateness to evaluate the experience with the questionnaire

5.2.4 Validation Results

Following the methodology set out in Section 5.2.2 we will validate the improvement of motivation (H4.2.2), worthiness from the control group of the video-debates (CC-LR) as an educational tool (H4.2.3 and H4.2.6) as well as the acquisition of collaborative knowledge with this new type of learning material (H4.2.5). For these purposes we used the metrics M4.2.1, and M4.2.3 through M4.2.5.

5.2.4.1 The CC-LR as a valuable resource

In this section we evaluate the level of worthiness of the video-debates (CC-LR) as an educational tool (H4.2.6). To this end, we collected quantitative and qualitative data in order to know the user's satisfaction in the experimental group with the tool. Both quantitative and qualitative data were collected in section (iv) from 5 open questions of the questionnaire addressed to students. Finally, the lecturer in charge of the classroom also participated by providing his views of the new type of learning resource for teaching (H4.2.3). All this data was also collected with the same questionnaire and questions from students of the control group (Section 5.1.4.1). This will make it possible a fair comparison between both groups.

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 video-debates 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 under 5.0.

The following questions related to evaluate the video-debates were asked:

- 1- What did you like and what you did not like from the video-debates (assess the video-debates from this view in the scale 0-10).
- 2- Compare the video-debates with traditional learning material and tools (books, web pages, forums, etc) and indicate pros and cons of the video-debates (assess the video-debates from this view in the scale 0-10).
- 3- Do you think the video-debates have helped you acquire more knowledge about the discussion topics in comparison to the text-based forums? (assess the video-debates from this view in the scale 0-10)
- 4- Express your opinion about the video-debates in terms of efficiency and performance (assess the video-debates from this view in the scale 0-10)
- 5- Let us know your opinion about the potential of the video-debates to observe how people discuss and collaborate, and how knowledge is constructed (assess the video-debates from this view in the scale 0-10).

Despite both experimental and control groups use the same tool they assessed the tool with different purposes (i.e. the former evaluated the VCS to convert text-based discussion into storyboards or SLO in real time while the latter evaluated reusing the SLOs as a new learning material in the form of video-debates). Therefore, some questions are not the same while others have different purpose. Still the experimental and control groups can be compared by questions as students are not able to distinguish the purpose of the questions (i.e. technical vs. pedagogical), thus considering both aspects in the evaluation. These are the most noticeable differences in the questions of both groups:

Question 2 of the experimental was not included in the control group in the same terms (“Do you think the VCS tool has fostered your active participation in the discussion in comparison to the text-based IWT forum?”) though the purpose is similar by comparing technology-advanced resources with traditional resources. Question 3 of the control group (“Do you think the VCS tool has helped you follow the discussion in comparison to the text-based IWT forum?”) was excluded in the experimental group as the video-debates were not related to a discussion forum. The rest of questions were very similar and thus they were compared even though the purpose was not exactly the same.

All 25 students of the experimental group provided assessment marks. After calculating the 0-10 scale for all the questions of the experimental group we got a general mean score of 6.12 (SD=2.14 and Md=7). This result is significantly better than the control group (M=5.43, SD=2.20 and Md=5.60) (see Section 5.1.4.1) and in line with the previous results on usability and emotions, hence these results confirm the video-debates and the VCS supporting tool as valuable educational resource.

In particular, students of the experimental group liked the video-debates by the VCS (Question 1: M=6.24, SD=2.01, Md=7) and liked them more than the control group liked the

discussions with VCS if we consider the Md statistics (Question 1: $M=6.33$, $SD=1.79$, $Md=6$) (see Section 5.1.4.1). Similarly to the control group (see Section 5.1.4.1), they indicated to find this resource more attractive and pleasant to study from the video-debate rather than from a message or book in a traditional material. They noted that this new format of observing other to discuss invited to reflect on the topic more than read messages or books. Also students felt the video-debate simulated a “real” discussion, meaning that it was closer to a real discussion with the presence of the discussants.

In addition, students did not find problematic anymore to understand the VCS text-to-voice engine, which was solved thanks to the VCS Editor and the opportunity to correct the syntax of the original posts, which in turn improved the conversion text-to-voice. Finally, some students indicated again the benefits of the VCS tool for disable students.

Question 2 of the experimental group was compared to Question 2 of the control group, both related to compare the VCS system to previous system and the impact on the learning process. The experimental group ($M=5.88$, $SD=2.09$, $Md=6$) also achieved significant better scores than the control group ($M=4.87$, $SD=2.69$, $Md=5$).

Question 3 of the experimental group can be compared to Question 4 of the control group, both related to whether the VCS and the video-debates have helped acquire more knowledge on the topic. The experimental group achieved better scores ($M=5.48$, $SD=2.50$, $Md=6$) than the control group ($M=4.37$, $SD=2.31$, $Md=5$). This result confirms the didactical purpose of the video-debates as learning materials, beyond the use of the VCS to convert text-based discussion into animated storyboards. Students indicated in a balanced way that the video-debates helped them to acquire the knowledge more than reading books or other traditional sources of knowledge. Some students mentioned that the video-debates provided realism and thus they fostered knowledge retention more easily as they could memorize from listening instead of reading.

The previous pedagogical result is also confirmed by Question 5 of the experimental group, which can be compared to Question 6 of the control group, both related to the potential of the system VCS and the video-debates to observe how the knowledge is built. The experimental group ($M=6.20$, $SD=2.14$, $Md=7$) also achieved significant better scores than the control group ($M=5.98$, $SD=2.17$, $Md=6$). Most students mentioned that could observe the knowledge construction process by passing the different scenes of the video-debate. The opportunity to stop the video, go back, takes notes on interesting parts and resume the video help them to consolidate their knowledge. Some students mentioned to feel in a movie with the characters, which help them to form new ideas as the video went by.

Question 4 related to efficiency of the video debates also got better results ($M=6.78$, $SD=2.34$, $Md=7$) than Question 5 of the control group related to the efficiency of the VCS ($M=6.52$, $SD=1.97$, $Md=7$). Almost all the students indicated the video-debates were very easy to use, intuitive and fast, very convincing from the efficiency and performance perspective. Only a few students reported problems with performance due to their own computers and/or connection problems.

Finally, some students proposed to promote the use of video-debates in other courses. They also gave some hints for possible improvements of the tool, such as to export the video-

debates into different formats and devices (e.g., audio only) to have the opportunity to study them without a computer.

5.2.4.2 *Motivational aspects*

Students' motivation concerning the formal learning activity supported by the video-debates (CC-LR) was investigated by comparing the difference in motivation between the experimental and control groups (see Section 5.1.4.2 for the data on motivation in the previous experiments).

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 slightly higher ($M=3.11$, $SD=0.89$, $Md=3$) than the control group ($M=3.07$, $SD=0.75$, $Md=3$) (see Section 5.1.4.2). The results of the experimental group are in line with the results reported in Section 5.2.4.1. They were also in line with the control group (see Section 5.1.4.2), where the students of the experimental group found the VCS more attractive and pleasant to follow the discussion from a video rather than the traditional reading of the text-based messages in forums or books. Some of them mentioned to find the new type of learning resource very modern and thus they paid more attention. Moreover, the capability of the video-debates to revisit the information any time motivated them to spend more time to study this new and comfortable way. As a result they were more engaged and self-motivated in the material than reading books or from a discussion forum (control group). Finally, clear indications of amounts of motivation came from enthusiastic students who evaluated the video-debates as "I liked them!", "Good idea!", "very interesting", "surprising". On the other hand, a few students who did not understand the purpose of the video-debates or chose not use them due to lack of time or technical problems felt unmotivated. These comments are similar to the control group (see Section 5.1.4.2) though the experimental group put more emphasis in the good aspects and the potential applications in other courses, which is in line with the better results achieved.

5.2.4.3 *Tutor assessment and knowledge acquisition*

All students were evaluated on summarizing both the discussion in the control group and the learning activity "Practica 2" (see Section 5.2.2.3) in the experimental group. Both activities addressed the same topic of "Software project management". To this end, section (ii) of the questionnaire included 3 evaluative questions about this topic. In order to avoid repeating the same questions already asked to the control group in the previous experiment (see Section 5.1.4.3), we proposed to ask different questions to the experimental group though addressing the same topic: 2 first questions to evaluate the topic and the last question to evaluate the knowledge acquisition, as follows:

Control group (from the discussion with the VCS):

1. Indicate and justify what are the main factors seen during the discussion, which may lead a software project to fail.

2. Indicate and justify 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.

Experimental group (from the new learning material as video-debates):

1. Indicate and justify whether Human Resources are a key factor in management of Software project
2. Indicate and justify the responsibility of the company managers in a software project when it fails.
3. Comment what you learnt from the video-debates than can enrich your personal knowledge.

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

Evaluative questions	Experimental group Video-debates (n=25)	Control group (Sect. 5.1.4.3) Standard forum (+VCS) (n=40)
Question 1	M=8.12 SD=2.34 Md=8	M=6.43 SD=1.84 Md=7
Question 2	M=8.01 SD=1.42 Md=8	M=7.71 SD=1.52 Md=8
Question 3	M=7.98 SD=1.61 Md=8	M=7.26 SD=1.17 Md=7
Overall	M=8.03 SD=1.79 Md=8	M=7.13 SD=1.51 Md=7.30

Table 10: Results of the knowledge acquisition evaluation

From the results of *Table 10*, students from the experimental group scored significantly higher than the control group (see Section 5.1.4.3). Both groups got good marks on average and showed a good level of knowledge acquisition. These results are in line with the potential of knowledge acquisition and construction reported in 5.2.4.1, which are better than the control group.

The students obtained greater amounts of knowledge acquisition and retention by the capability to review the video-debates as many times as needed and having the video-debates been edited and improved, thus showing more quality material.

In summary, we can conclude that the inclusion of the video-debates had a significant impact in the quality of the student's knowledge acquisition and retention.

5.2.5 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 5.2.1). Then, based on the results summarized they are compared with the control group reported in Section 5.1.

In general the students liked the video-debates (CC-LR) and found them interesting to understand better the content of the live discussions supported by the VCS. They found a good idea to register and reuse the live sessions and create a new learning material (G4.2.1).

The majority of students could generate the video-debates efficiently (G4.2.6). In particular, during the study with the video-debates, the students found them very easy to use as no relevant technical problems were reported, also from the usability perspective (G4.2.4).

Complex aspects of the learning process, such as motivation and emotional were validated showing an impact of the use of the video-debates to make the learning process more effective. In particular, the new video-debates proved to become a useful educational resource (G4.2.5).

One of the most relevant results was found in the impact of the video-debates in knowledge retention and construction (G4.2.2), which was very significant in comparison to the use of the VCS to support the live discussion within the control group.

Finally, students provided some hints to improve the video-debates and CC-LR in general (G4.2.3) as well as they suggested to use this type of learning resources in more courses and programs of the UOC.

5.3 4-3. Live and Virtualized Collaboration: Experimenting with Collaborative Complex Learning Resources (CC-LR) enriched with authoring information from the student's viewpoint

5.3.1 Evaluation and Validation Procedure

In the previous experiment regarding live and virtualized collaboration (see Section 5.2), an experiment was conducted at UOC pilot site in order to test a new type of learning resources in the form of video-debates, formally called Collaborative Complex Learning Resource (CC-LR) from the virtualization of live sessions of collaborative learning to produce storyboard learning objects (CC-LO/SLO) embedded in a virtualized collaborative session system (VCS) to be experienced and played by learners. During the CC-LR execution, learners observed how avatars discuss and collaborate, how discussion threads grow, and how knowledge is constructed, refined and consolidated.

The experimentation results with the CC-LR were compared to a similar experiment with the VCS tool carried out by the same participants enrolled in the same course during the same academic term and just a few weeks later. The results validated the concept and notion of CC-LR as a new type of learning material produced by the transformation of a live

collaborative session and improved with an editor tool that provided lecturers and experts with edition capabilities of the CC-LRs, such as cutting scenes, modifying involved characters, selecting emotional states, dialogues and connected concepts. However, no new information at all was added in the CC-LR from the lecturer or expert side.

The purpose of this new experiment is to validate further the video-debates (CC-LR) by augmenting them with authoring information by means of the SLO Editor, thus showing the provision of complex aspects of the learning process in the CC-LRs, such as cognitive and emotional assessment, which make the new material highly interactive. See *Figure 68* below and [17] for more details.

The CC-LR tools (VCS Editor) were developed during the second phase [17] of the project and were experimented at UOC site (see Section 5.2) with real students who used the CC-LR to complement the official learning material of the course. From the experimentation data obtained, the CC-LR without authoring information was evaluated and validated from different aspects and perspectives related to impact on students and the learning process. It is worth mentioning that the IWT platform acted as a supportive infrastructure for the CC-LR during the experiment and students entered IWT to find the appropriate CC-LR and study with them.

To experiment with the CC-LR approach enriched with authoring information from the student's viewpoint and evaluate and validate it as for the usability, emotional, impact in the learning process, etc., we focused on the following scenario goals and hypotheses as well as criteria and metrics derived from [3]:

Scenario goals

- G4.3.1: The VCS Editor system that is able to build a CC-LR from a threaded discussion (coming from a forum).
- G4.3.2: To employ the CC-LR in online courses in order to enhance some aspects of the teaching/learning process.
- G4.3.3: To identify possible ways of improving further the utility of the CC-LR in online courses.
- G4.3.4: To create, store and playback the generated CC-LR through a user friendly interface.
- G4.3.5: To build (automatically) a draft CC-LR from a collaborative activity effectively
- G4.3.6: To build (automatically) a draft CC-LR from a collaborative activity efficiently

Scenario hypotheses

- H4.3.1: Use CC-LR by non-expert users (i.e., in a friendly way and efficiently).
- H4.3.2: Use of CC-LR contributes to significantly improve students' motivation.
- H4.3.4: Use of CC-LR to significantly increase students' activity levels, both in individual and collaborative activities.
- H4.3.5: Use of CC-LR contributes to significantly improve students' understanding of key concepts and students' results.
- H4.3.6: CC-LR is considered as a worthy educational resource by students.

Scenario criteria

- C4.3.1: Level of fulfillment of the VCS Editor features.
- C4.3.2: Potential increase in students' motivation caused by the use of CC-LR.
- C4.3.4: Potential increase in students' activity levels due to the incorporation of the CC-LR.
- C4.3.5: Potential increase in students' understanding of concepts and students' results.
- C4.3.6: Level of satisfaction of students with the inclusion of the CC-LR in their courses.

Scenario metrics

- M4.3.1: Number of students using the CC-LR.
- M4.3.2: Number of visits of the CC-LR.
- M4.3.3: Number of students passing the course and/or with high marks when the CC-LR with author information is used.
- M4.3.4: Number of students passing the course and/or with high marks when CC-LR without author information is not used.
- M4.3.5: Number of students that consider that the CC-LR is worthy.

5.3.2 Method

5.3.2.1 Participants

In the same way as in the previous experiment (see Section 5.2), the real context of this experience is the virtual learning environment of the Open University of Catalonia (UOC).

In order to evaluate the CC-LR enriched with authoring information and analyze its effects in the discussion process, the sample of the experiment consisted of 44 undergraduate students enrolled in the course Organization Management and Computer Science Projects of the Bachelor in Engineering Computing degree at the UOC were involved in this experience.

These same 44 students formed two groups and both participated during the Spring term of 2012 in the same course: the control group participated in the middle of the course (May 2012) while the experimental group participated at the end of the course (June 2012). All details about the experiment and results of the control group are found in Section 5.2.

Despite all 44 students participated in this experience, only 24 out of them (54.5%) submitted the final questionnaire, the rest of students (20) dropped out the discussion and the course for personal reasons. It is worth mentioning that the 45% dropout ratio found is considered normal in at the end of the academic term when the experience was run³.

Each group was supervised by the same tutor as the official lecturer in charge of the course.

³ 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.

5.3.2.2 Apparatus and Stimuli

Students of the experimental group were required to use standard forum IWT was equipped with the multimedia-based VCS tools, SLO Editor and Player (see *Figure 68* and [7]).

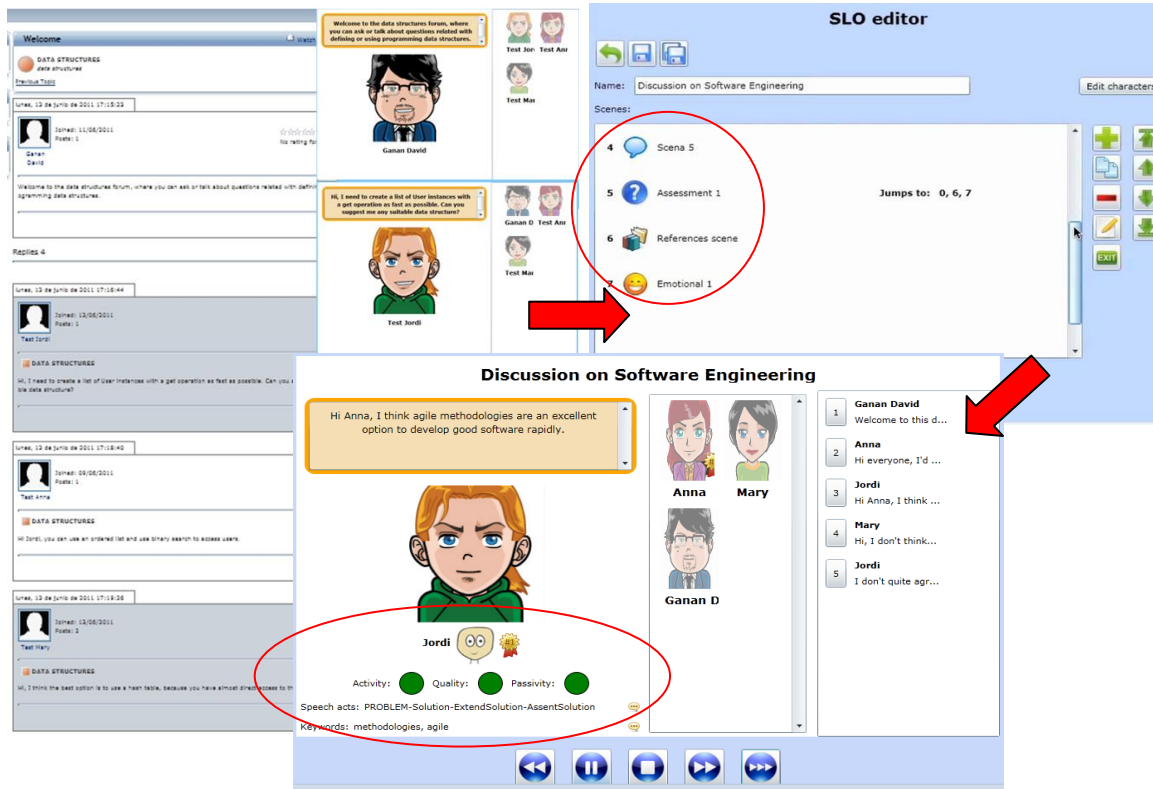


Figure 68: An SLO from a live discussion in IWT is edited by the VCS-SLO Editor to create a video-debate (CC-LR), which contains different types of scenes and author information, such as cognitive and emotional information of the original participant and contributions.

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 video-debates (CC-LR), which included a motivation test; (iv) test-based evaluation of the video-debates; (v) test-based evaluation on the usability of the VCS system; (vi) test-based evaluation on the emotional state; (vii) a test-based evaluation of the questionnaire.

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 experimental group and the control group.

For the section v (usability of the VCS player showing the video-debates) we used the System Usability Scale (SUS) developed by [8] 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 using the new system, which included 12 items of the Computer Emotion Scale (CES) [9]. The 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 the 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.

5.3.2.3 Procedure

A formal learning activity called “PAC 2” was scheduled during the first 2 weeks of June 2012 in the mentioned course of Organization Management and Computer Science Projects from the Bachelor in Engineering Computing degree at the UOC. The participants of the experimental group participated in this activity.

The activity was individual and mandatory for all students and consisted in filling a test about questions on Software projects management. A part from the usual didactical material of the course the students of the experimental group also received a new material to support specifically this activity in the form of a video-debate (CC-LR) called “Factors that lead a Computer Science project to failure” which contained a discussion about project management. This material was the same as in the previous experiment (Section 5.2) but it was greatly enriched with new types of scenes with emotional and cognitive information, which made the material highly interactive. The students entered IWT to find and watch this interactive video-debate.

Finally, since the participants of the experimental group were the same as the participants of the previous experiment (see Section 5.2), they had been already experimented with the VCS tools and in particular the CC-LR. Therefore, we will compare the evaluation and validation aspects with those results of the previous experiment forming the control group. All

this data was already reported in Sections 5.2.3 and 5.2.4 and the feedback provided served to improve the prototypes for this new experiment.

5.3.3 Evaluation Results

Following the methodology described in Section 5.3.2, in this section we focus on usability and emotional aspects of the video-debates (H4.3.1) of both the control group (see Section 5.2) and experimental group. For this purpose we used metrics M4.3.1 and M4.3.2. 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.

5.3.3.1 Activity levels

Due to a technical problem with the log system, we could not collect the interactivity of the experimental group with the video-debates during the learning activity.

5.3.3.2 Usability of the video-debates (CC-LR) enrich with author information

To evaluate student's satisfaction of the experimental group with the tool, enriched with cognitive and emotional information, as for an efficient and user-friendly management (H4.2.1), we collected data from students' ratings and open comments on the usability/functionality/integration of the tool.

To investigate the overall usability of the video-debates, we used the SUS (see Section 5.3.2.2) 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 **24 SUS scores of 69.27**, thus above the SUS mean and also above the control group (25 SUS scores of 68.20) (see Section 5.2.2.2). Next, we present the most relevant results of the SUS score by providing several statistics: Mean (M), Standard Deviation (SD) and Median (Md).

The results of the experimental group and control group as for the usability of the the video-debates are the following:

- Students of the **experimental group** (n=24) thought they will use the *video-debate* often (M = 3.36, SD = 1.09, Md = 3.5) (See *Figure 69*). Students did not find the tool unnecessarily complex (M = 2.18, SD = 0.96, Md = 2) (See *Figure 70*). In addition, students stated that they did not need the support of a technical person to be able to use the video-debate (M = 1.68, SD = 1.00, Md = 1) (*Figure 71*), they thought that most people would learn to use this system very quickly (M = 4.22, SD = 0.58, Md =

4) (See Figure 72), and they felt quite confident using the video-debate ($M = 3.86$, $SD = 0.70$, $Md = 4$) (See Figure 73).

- Students of the **control group** ($n=25$, see Section 5.2.3.2 for the graphical results) thought they will use the video-debate often ($M = 3.13$, $SD = 1.29$, $Md = 4$). Students did not find the tool unnecessarily complex ($M = 2.43$, $SD = 1.22$, $Md = 2$). Students stated that they did not need the support of a technical person to be able to use the video-debate ($M = 1.47$, $SD = 0.65$, $Md = 1$), they thought that most people would learn to use this system very quickly ($M = 4.04$, $SD = 0.61$, $Md = 4$) and they felt confident using the video-debate ($M = 3.69$, $SD = 0.80$, $Md = 4$).

So far, the results of the experimental group are above the average and better than the control group, though slightly. The results of both the experimental and control groups confirmed that the students found the usability in general satisfactory or very satisfactory in line with the SUS score found.

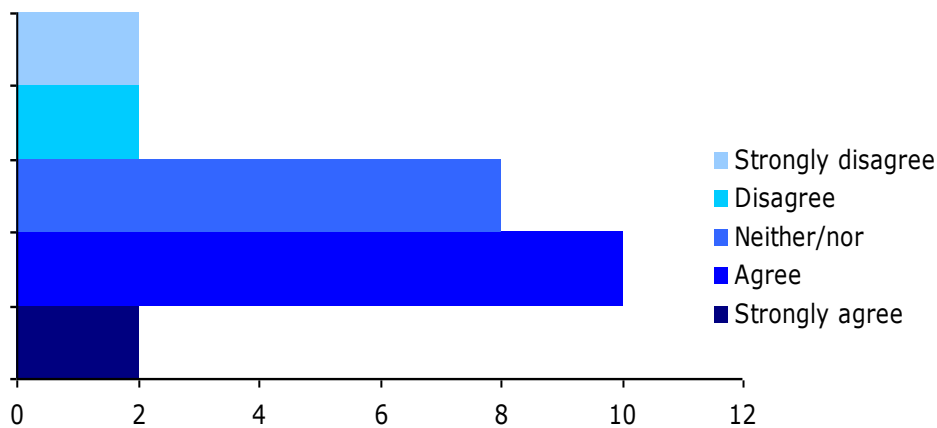


Figure 69: Results on the SUS item "I think I will use the video-debate often"

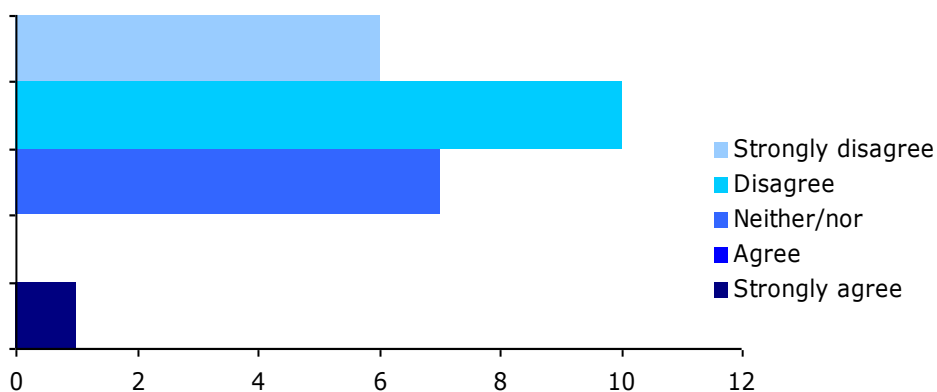


Figure 70: Results on the SUS item "I think the video-debate was unnecessarily complex".

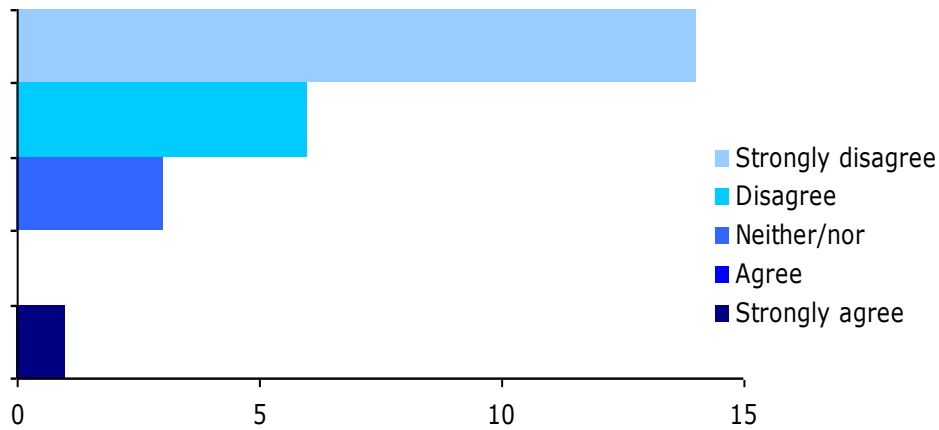


Figure 71: Results on the SUS item "I think that I would need the support of a technical person to be able to use the video-debate".

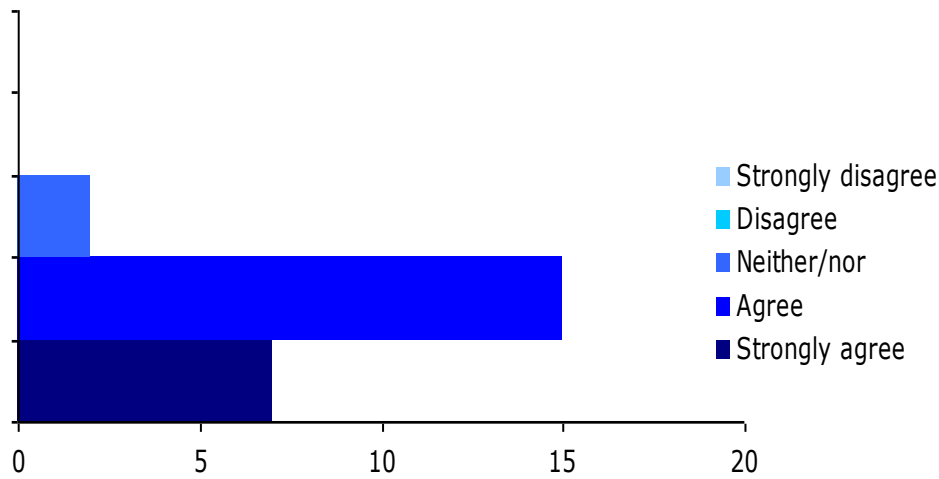


Figure 72: Results on the SUS item "I would imagine that most people would learn to use the video-debate very quickly".

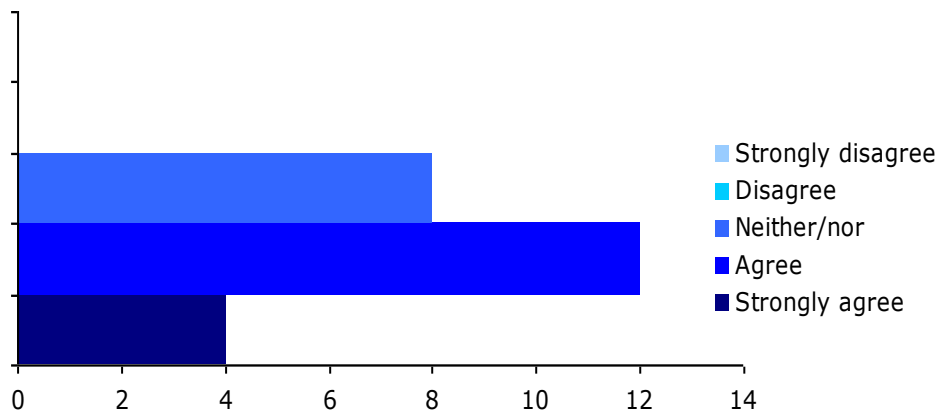


Figure 73: Results on the SUS item "I felt very confident using the video-debate".

Moreover, students of the experimental group stated that the VCS functionality was well integrated ($M = 3.95$, $SD = 0.50$, $Md = 4$) (Figure 74) and the tool itself was adequately integrated in the UOC virtual campus and in turn in IWT (see Annex A). This result is better than the control group though very similar ($M = 3.52$, $SD = 0.96$, $Md = 4$), which confirms that students did not have problems to gain access to the video-debates in IWT.

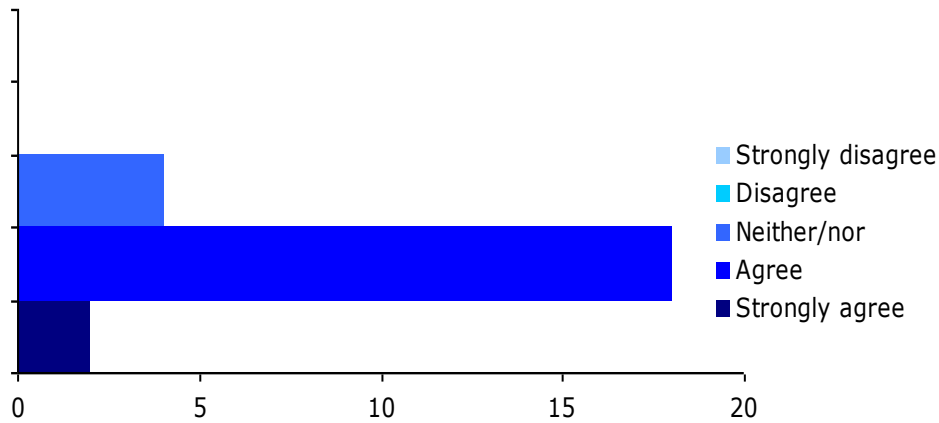


Figure 74: Results on the SUS item “I found the various functions in the video-debate were well integrated”.

Finally, students indicated in a balanced way they found the video debates easy to use ($M = 3.13$, $SD = 1.23$, $Md = 3$) (Figure 75). Despite this result is worse than the control group ($M = 3.47$, $SD = 0.96$, $Md = 4$), it should be considered for this usability aspect the inclusion of interactive scenes in the video-debates that the control group did not have (users just watch the video-debates as a regular video), making the study with this new material more complicated (e.g., they needed to interact with cognitive tests, select emotional state and interact with an affective agent, etc). Considering this, this result is quite satisfactory and well above the average (2.5).

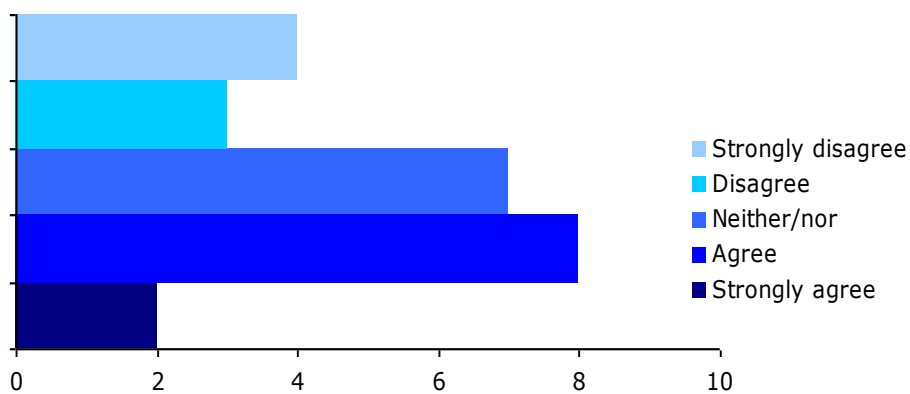


Figure 75: Results on the SUS item “I thought the video-debate was easy to use”.

In summary, the usability assessment of the video-debates enriched complex aspects, such as cognitive and emotional information, was quite satisfactory in comparison to both, the control group and the SUS scale as all SUS scores are above or well above the average (2.5). Finally, some improvements made to the VCS tool on usability since the previous experimentation (See Section 5.2.3.2) were also noticeable by the students of the experimental group who in general did not report some problems found by the control group.

5.3.3.3 Emotional aspects

Regarding the students’ emotions of the experimental group during the work with the video-debates (H4.3.1), the results from a 4-point rating scale (n=24) are presented next, and they are compared to the results of the control group (n=25). See Section 5.2.3.3 for the statistics and graphical results of the control group:

- Happiness (M=1.13, SD=0.67, Md=1) (Figure 76). This result is slightly better than the control group (M=1.08, SD=0.90, Md=1) showing they were curious with the new type of scenes incorporated in the video-debate (cognitive and emotional).

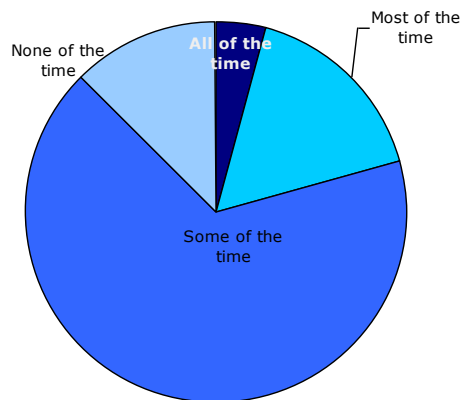


Figure 76: Results on the Happiness emotion

- Sadness (M=0.50, SD=0.78, Md=0) (Figure 77). This result is slightly worse than the control group (M=0.48, SD=0.58, Md=0). However, both results are very good with Md=0, which means that students of both groups did not experienced this bad feeling.

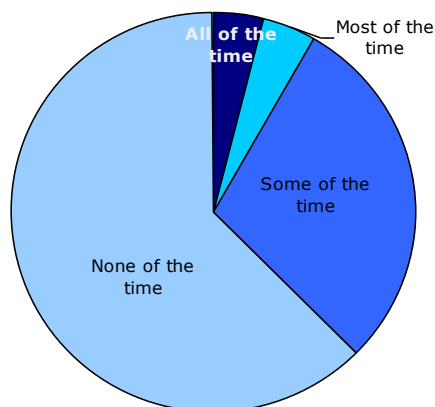


Figure 77: Results on the Sadness emotion

- Anxiety ($M=0.45$, $SD=0.72$, $Md=0$) (*Figure 78*). This result is slightly worse than the control group ($M=0.32$, $SD=0.69$, $Md=0$). However, both results are very good with $Md=0$, which means that students of both groups did not experienced this bad feeling.

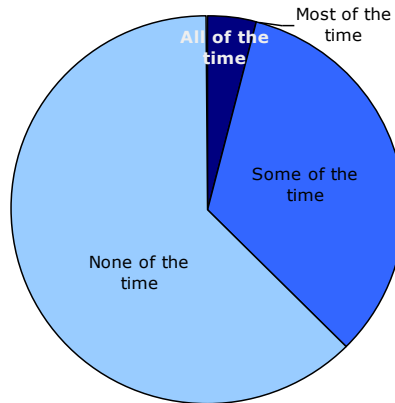


Figure 78: Results on the Anxiety emotion

- Anger ($M=0.54$, $SD=0.77$, $Md=0$) (*Figure 79*). This result is slightly worse than the control group ($M=0.40$, $SD=0.76$, $Md=0$). However, both results are very good with $Md=0$, which means that students of both groups did not experienced this bad feeling.

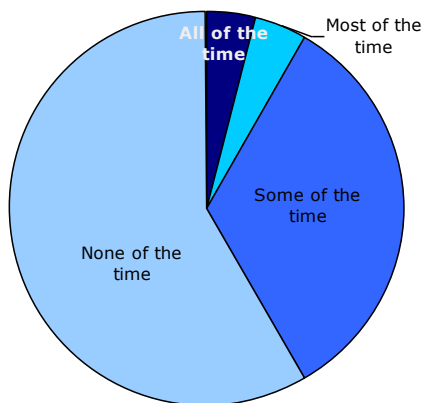


Figure 79: Results on the Anger emotion

In summary, students felt more often happiness than sadness, anxiety or anger when using the video-debates to study with the new learning material (video-debates) The results in general are similar in the experimental group than the control group though, being the most noticeable result the highest value in happiness while the students felt the same level of sadness, anxiety and anger emotions, which were very low, almost inappreciable ($Md=0$), being the anxiety emotion the lowest.

A particular increase of the bad feelings (especially anxiety and anger) from the control group is shown, which can be explained by the incorporation of complex information in the video-debates, such as cognitive tests, that the students had to pass to continue with the video. This might cause higher steps of anxiety and also anger, especially when they would fail the tests. However, as mentioned, bad emotions were assessed very low (Md=0).

In overall, this is a good result considering the students faced a complex type of learning material that was new for them and they had to learn how it worked and how to use it for their benefit. Finally, this result is in line with the results presented above concerning the usability (see Section 5.3.3.2).

5.3.3.4 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 an average time to fill out the questionnaire of about 40 minutes (Figure 80) and 75% of students found it appropriate to evaluate the experience (Figure 81) (n=24). These results are worse than the control group (n=24, 30 minutes on average to fill the questionnaire and 97% of appropriateness) (see Section 5.2.3.4) due to the questionnaire of the experimental group was more complex in order to evaluate more aspects of the video-debates.

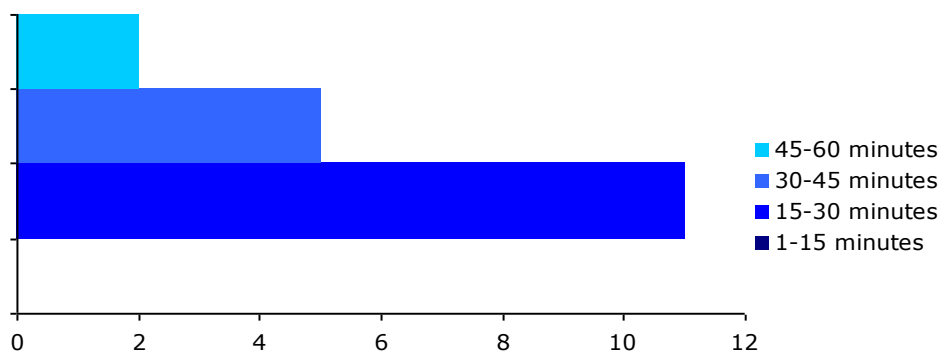


Figure 80: Time employed to fill the questionnaire

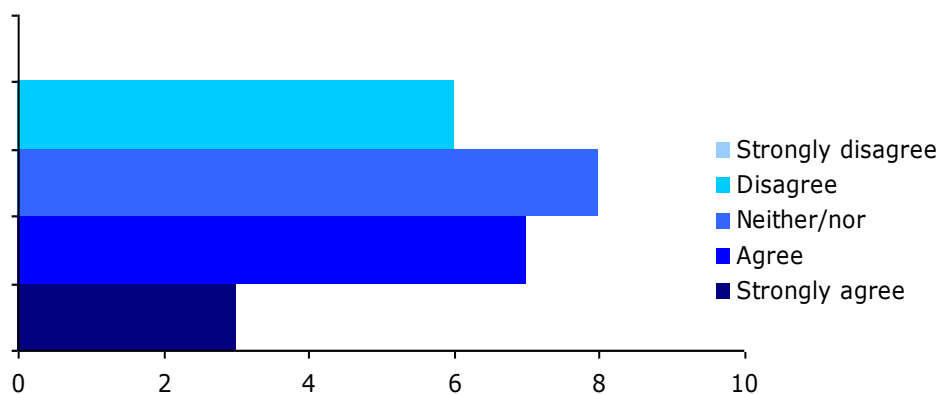


Figure 81: Appropriateness to evaluate the experience with the questionnaire

5.3.4 Validation Results

Following the methodology set out in Section 5.3.2 we will validate the improvement of motivation (H4.3.2), worthiness from the control group of the video-debates (CC-LR) enriched with author information as an educational tool (H4.3.3 and H4.3.6) as well as the acquisition of collaborative knowledge with this new type of learning material (H4.3.5). For these purposes we used the metrics M4.3.1, and M4.3.3 through M4.3.5.

5.3.4.1 The CC-LR as a valuable resource

In this section we evaluate the level of worthiness of the new version of the video-debates (CC-LR) enriched with complex information supported by the VCS as an educational tool (H4.3.6). To this end, we collected quantitative and qualitative data in order to know the user's satisfaction in the experimental group with the tool. Both quantitative and qualitative data were collected in section (iv) from 5 open questions of the questionnaire addressed to students. Finally, the lecturer in charge of the classroom also participated by providing his views of the new type of learning resource for teaching (H4.3.3). All this data was also collected with the same questionnaire and questions from students of the control group (Section 5.2.4.1). This will make it possible a fair comparison between both groups.

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 video-debates 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 under 5.0.

The following questions related to evaluate the video-debates were asked:

- 1- What did you like and what you did not like from the video-debates (assess the video-debates from this view in the scale 0-10).
- 2- Compare the video-debates with traditional learning material and tools (books, web pages, forums, etc) and indicate pros and cons of the video-debates (assess the video-debates from this view in the scale 0-10).
- 3- Do you think the video-debates have helped you acquire more knowledge about the discussion topics in comparison to the text-based forums? (assess the video-debates from this view in the scale 0-10)
- 4- Express your opinion about the video-debates in terms of efficiency and performance (assess the video-debates from this view in the scale 0-10)
- 5- Let us know your opinion about the potential of the video-debates to observe how people discuss and collaborate, and how knowledge is constructed (assess the video-debates from this view in the scale 0-10).
- 6- Do you think that the both the performance indicators on each character and test questions integrated in the video-debate allowed you to understand the contents of the video and acquire more knowledge? (assess the video-debates from this view in the scale 0-10)
- 7- Indicate if the consideration of both the character's emotional state and your own emotional state in the video-debate has had any impact in your learning experience? (assess the video-debates from this view in the scale 0-10)

Questions 1 through 5 are the same as those included in the control group's questionnaire (see Section 5.2.4.1). On the other hand, Questions 6 and 7 are new and particularized to the extensions of the video-debates with cognitive and emotional information. Therefore, Questions 1 through 5 will be compared between the experimental and control groups whilst Questions 6 and 7 will be analyzed individually to validate the inclusion of author information in the video-debates as a valuable resource.

All 24 students of the experimental group provided assessment marks. After calculating the 0-10 scale for all the questions of the experimental group we got a general mean score of 6.52 (SD=1.42 and Md=6.7). This result is significantly better than the control group (M=5.43, SD=2.20 and Md=5.60) (see Section 5.1.4.1) and in line with the previous results on usability and emotions, hence these results confirm the video-debates enriched with author information as valuable educational resource.

In particular, students of the experimental group liked the video-debates enriched with author information (Question 1: M=6.40, SD=1.44, Md=6.75) and liked them more than the control group who liked the video-debates without this information (Question 1: M=6.24, SD=2.01, Md=6) (see Section 5.2.4.1). Some students then found the video-debates quite impressive and original, and considered this as an innovation with respect to the current technology at UOC. On the other hand, others found the interface not very pleasant including the "robotic" voice. Also some students felt the video-debate simulated a "real" discussion, meaning that it was close to a real discussion with the presence of the discussants.

In addition, students did not find problematic to understand the VCS text-to-voice engine, which was solved thanks to the VCS Editor and the opportunity to correct the syntax of the original posts, which in turn improved the conversion text-to-voice. However, many of them did not like the "robotic" and monotonous voice of the video-debate as well as suggested to embed different voices for different characters. Finally, some students indicated the benefits of the video-debates for disabled people.

Question 2 of the experimental related to compare the video-debates to traditional learning systems and the impact on the learning process. The experimental group (M=6.33, SD=2.32, Md=7.00) achieved slightly worse scores than the control group (M=5.88, SD=2.09, Md=6). Similarly to the control group, students indicated in general to find this resource more attractive and pleasant to study from the video-debate rather than from a message or book in a traditional material. Some students indicated that with the video-debates they felt like studying in a collaborative fashion and as the characters in the video were based on real students they had the impression to form part of the collaborative activity registered in the video. On the other hand, others found the study with the video-debate a different way of individual study and "just yet another forum" where to share experiences in an only-read way. Some of them preferred the traditional way (i.e. read books) and missed a human moderator in the video-discussion. Finally, some students found the video-debates a "live" material that is interesting to dynamically share experiences and opinions about real life and when reliability is not a must but then it was necessary to get books and traditional material "made

by experts” to go deep in particular themes. From this view, they considered the video-debates as a “complement” rather than a “nuclear” material.

Question 3 of the experimental group related to whether the video-debates have helped acquire more knowledge on the topic. The experimental group achieved better scores ($M=5.92$, $SD=2.34$, $Md=6.50$) than the control group ($M=5.48$, $SD=2.50$, $Md=6$). This result confirms the didactical purpose of the video-debates as learning materials and the enrichment with author information reinforces their didactical purpose. Some students mentioned that by observing the different characters’ points of view helped them to understand better the main topics in the video as well as speeding up the understanding. However, they thought this material should “complement” the official material rather than replace it. Others also mentioned that the video-debates helped them to reflect and reason their ideas and because of the special format of the video it made easy to get into the topic easily. Finally, some students found the potential of the video-debates “huge” since it is more didactic and comfortable to learn from video with people sharing opinions rather than reading a book.

The previous pedagogical result is also confirmed by Question 5 of the experimental group, related to the potential of the system VCS and the video-debates to observe how the knowledge is built. The experimental group ($M=6.54$, $SD=1.93$, $Md=7$) also achieved significant better scores than the control group ($M=6.20$, $SD=2.14$, $Md=7$). Most students mentioned that could observe the knowledge construction process in a “natural” way and “progressive”. Some students found the video-debates a “simulative” way to build knowledge from real life, which was appreciated, and mentioned that it fostered to explore new views. The test scenes found in strategic points of the video-debate (see Question 6) also was mentioned to ease the process of knowledge construction by fostering knowledge retention and consolidation.

Question 4 related to efficiency of the video-debates got better results ($M=6.92$, $SD=2.41$, $Md=7.0$) in comparison to the control group ($M=6.78$, $SD=2.34$, $Md=7$). In line with the control group, almost all the students indicated the video-debates were very easy to use, intuitive and fast, very convincing from the efficiency and performance perspective. Only very few students reported problems with installing the application in Linux platforms.

Question 6 about the incorporation of cognitive aspects were very positive as the experimental group got a mean score as high as 6.94 ($SD=1.12$, $Md=7.0$), being the highest score of the all the questions. Students in general found very didactic to have the chance to self-evaluate the video-debate by several test scenes in certain points in the video. They reported to consolidate better the concepts and ideas (i.e. knowledge retention). Also the performance indicators showing up by the avatars were found useful to know the knowledge reliability of the avatar and incorporate certain opinions in their general knowledge. Most of students missed more test scenes to reflect, self-assess and check their knowledge. A few students found the test scenes broke the rhythm of the video reproduction.

The last Question 7 about emotional awareness scored low ($M=4.94$, $SD=1.99$, $Md=5.0$), being the lowest score of all the questions and the only under average, though the Median was on average (5). Many students mentioned that the emotional assessment was not

useful to understand better the concepts and they did not even understand the purpose of this feature. On the other hand, some students mentioned that they found interesting to know the emotional state of the character (i.e. participant) in the video when “talking” as the character transmitted confidence, cordiality and other positive feelings that cheered them up, though other students mentioned they could understand the contribution without knowing the emotional state of the character. A few students indicated that the feature that asked them to select their own emotional state even “bothered” them and interfered with the progress of the video while other indicated that this part helped them to keep concentrate on the work.

Finally, similarly to the control group, some students proposed to promote the use of video-debates in other courses. They also gave some hints for possible improvements of both pedagogical, such as increase the number of test scenes, and technical, such as distinguish between male and female voices in the video-debates.

5.3.4.2 *Motivational aspects*

Students’ motivation concerning the formal learning activity supported by the video-debates (CC-LR) was investigated by comparing the difference in motivation between the experimental and control groups (see Section 5.2.4.2 for the data on motivation in the previous experiments).

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 slightly higher ($M=3.27$, $SD=0.95$, $Md=3.5$) than the control group ($M=3.11$, $SD=0.89$, $Md=3$) (see Section 5.2.4.2). The results of the experimental group are in line with the results about the video-debates as valuable resources reported in the previous Section. The students in general found the new version of the video-debates enriched with authoring information on cognitive and emotional aspects more attractive and pleasant to follow rather than the video-debates without these features (control group). In particular, the students appreciated and felt very motivated by the test scenes incorporated in the video that allowed them to self-evaluate their leaning progress. This way, in overall, they found the video very intuitive and very good as learning material, which eventually engaged them in both the video-debate and also in the learning activity (“PAC 2”) supported by this material

Finally, clear indications of amounts of motivation came from enthusiastic students who evaluated the video-debates as “I liked it a lot!”, “Impressive!”, “Very surprising”, “Curious”. On the other hand, a few students who did not understand the purpose of the video-debates or chose not use them due to lack of time or technical problems felt unmotivated. These comments are similar to the control group (see Section 5.2.4.2) though the experimental group put more emphasis in the good aspects and the potential applications in other courses, which is in line with the better results achieved.

5.3.4.3 Tutor assessment and knowledge acquisition

All students were evaluated on summarizing both the discussion in the control group and the learning activity “PAC 2” (see Section 5.3.2.3) in the experimental group. Both activities addressed the same topic of “Software project management” and the students used the same video-debate as the control group but enriched with author information. To this end, section (ii) of the questionnaire included 3 evaluative questions about this topic. In order to avoid repeating the same questions already asked to the control group (formed by the same students as the experimental group) in the previous experiment (see Section 5.2.4.3), we proposed to ask different questions to the experimental group though addressing the same topic: 2 first questions to evaluate the topic and the last question to evaluate the knowledge acquisition, as follows:

Control group (from the video debates without author information):

1. Indicate and justify whether Human Resources are a key factor in management of Software project
2. Indicate and justify the responsibility of the company managers in a software project when it fails.
3. Comment what you learnt from the video-debates than can enrich your personal knowledge.

Experimental group (from the video-debates without author information):

1. Indicate and justify whether the IT budget is a key factor in management of Software project
2. Indicate and justify the degree of recycling of previous projects when it fails.
3. Comment what you learnt from the video-debates than can enrich your personal knowledge.

This part of each questionnaire was assessed by the lecturer who used the standard 10-point scale to score the students’ responses. *Table 11* shows the results.

Evaluative questions	Experimental group Video-debates (with author information) (n=24)	Control group (Sect. 5.2.4.3) Video-debates (without author information) (n=25)
Question 1	M=8.09 SD=1.91 Md=8	M=8.12 SD=2.34 Md=8
Question 2	M=8.16 SD=1.73 Md=8	M=8.01 SD=1.42 Md=8

Evaluative questions	Experimental group Video-debates (with author information) (n=24)	Control group (Sect. 5.2.4.3) Video-debates (without author information) (n=25)
Question 3	M=8.03 SD=1.89 Md=8	M=7.98 SD=1.61 Md=8
Overall	M=8.09 SD=1.84 Md=8	M=8.03 SD=1.79 Md=8

Table 11: Results of the knowledge acquisition evaluation

From the results of Table 9, students from the experimental group scored slightly higher than the control group (see Section 5.2.4.3). Both groups got very good marks on average and showed a good level of knowledge acquisition. These results are in line with the potential of knowledge acquisition and construction reported in 5.3.4.1, which are better than the control group.

The students obtained greater amounts of knowledge acquisition and retention by the capability to check and self-evaluate their knowledge about the topics of the video-debates in the own video.

In summary, we can conclude that the inclusion of author information in the video-debates had an impact in the quality of the student's knowledge acquisition and retention, though not very significantly.

5.3.5 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 5.3.1). Then, based on the results summarized they are compared with the control group reported in Section 5.2.

In general the students liked the extended features of the video-debates and found them interesting to understand better the content of the video and also for knowledge retention and construction (G4.3.2), even more than the control group. During the study with the new version of the video-debates, the students found them very easy to use as no relevant technical problems were reported and also from the usability perspective (G4.3.4).

The majority of students could generate the video-debates efficiently (G4.3.6). Complex aspects of the learning process, such as motivation and emotional were validated showing an impact of the use of the video-debates, and in particular, the extensions provided, in make the learning process more effective. In particular, the new video-debates proved to become a useful educational resource (G4.3.5).

The gain in knowledge acquisition by using the new features though it was increased from the control group (G4.3.2) were not significant, especially from the emotional features

incorporated that were not appreciated by the students and hid part of the real benefits provided by the incorporation of cognitive aspects (test scenes and performance indicators).

Finally, students provided some hints to improve the video-debates and CC-LR in general (G4.3.3) as well as they suggested to use this type of learning resources in more courses and programs of the UOC.

5.4 R4-4. Live and Virtualized Collaboration: Experimenting with Complex Learning Resources (CC-LR) from the instructor's viewpoint

5.4.1 Evaluation and validation procedure

The aim of this scenario is to build a Collaborative Complex Learning Resource (CC-LR) as a learning material to support the collaborative pedagogical model of academic courses from the instructor's viewpoint. The instructors use an Editor tool (see [17]) to edit, modify and augment the CC-LO obtained from live sessions of collaborative learning (see previous experiments (Sections 5.1 – 5.3)). To this end, an experiment was conducted on this scenario at UOC pilot site in order to test the Editor tool and collect feedback from the instructors when creating and managing CC-LRs from CC-LO to provide new learning resources to students.

The results of this study are analyzed to evaluate how the VCS-SLO Editor tool supports instructors and experts in order to create and manage CC-LR from CC-LO/SLO, the time spent in creating new CC-LRs as well as the problems and possible enhancements suggested. Therefore, in this study we were primarily interested in the functionality and usability of the tool as well as the time spent by lecturers and experts to create and manage CC-LR.

To experiment the live and virtualized collaboration from the instructor's viewpoint, we focused on the following goals and hypotheses as described in [3]:

Scenario goals

- G4.4.1: To build a VCS Editor system that is able to build a CC-LR from a threaded discussion (coming from a forum).
- G4.4.2: To ensure that the aforementioned tool allows efficient building of CC-LR even in the case of non-expert instructors (i.e., in a friendly way and without having to employ too much time).
- G4.4.3: To identify possible ways of improving further the utility of the CC-LR in online courses.
- G4.4.4: To create, edit, manage, store and playback the generated CC-LR through a user friendly interface.
- G4.4.5: To build (automatically) a draft storyboard from a collaborative activity effectively

- G4.4.6: To build (automatically) a draft storyboard from a collaborative activity efficiently

Scenario hypotheses

- H4.4.1: The VCS Editor prototype allows non-expert users to build and use CC-LR (i.e., in a friendly way and efficiently).
- H4.4.2: Use of CC-LR contributes to support instructors' task both in individual and collaborative activities.
- H4.4.3: CC-LRs are considered as a worthy educational resource by instructors.

Scenario criteria

- C4.4.1: To evaluate the level of fulfillment of the tool features.
- C4.4.2: To evaluate the level of satisfaction of the instructors with the tool for developing CC-LR.
- C4.4.3: To evaluate the level of satisfaction of the instructors with the inclusion of VCS in their courses.

Scenario metrics

- M4.4.1: Number of instructors using the VCS Editor tool.
- M4.4.2: Number of CC-LR created with the ASVCS tool.
- M4.4.3: Time employed in forming new instructors to use the Editor tool.
- M4.4.4: Time employed in creating each CC-LR.
- M4.4.5: Number of instructors that consider that the VCS is worthy.

5.4.2 Method

5.4.2.1 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 and they both are expert in e-Learning systems and applications. Lecturer A has 10 years of experience in teaching at UOC and he is a professional developer of software systems, especially e-learning systems, and is owner of a software company settled in Barcelona, whilst lecturer B has 8 years of experience of teaching and coordinating on-line courses at UOC as well as 9 years of performing research on e-learning. They currently teach on-line courses at UOC.

Hence they both have a strong background and advanced knowledge developing and using e-learning platforms, especially from the instructor's viewpoint.

5.4.2.2 Apparatus and Stimuli

First of all we asked the two lecturers to use the VCS-SLO Editor tool within IWT (Intelligent Web Teacher) to create a CC-LR from a CC-LO/SLO to provide new learning resources to students. In order to create the CC-LR they first selected an existing SLO in the SLO Repository [17].

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 annotated

for each step the time they spent on working with the VCS-SLO Editor in IWT. In addition, the lecturers listed all problems they had to face while working with the tool and wrote down advantages and disadvantages. For this task, the lecturers were provided with technical documentation on this scenario (see [17]).

In addition, both lecturers were asked to fill in the SUS (System Usability Scale [8]) after the end of the session in order to investigate the usability of the VCS-SLO Editor tool. 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 VCS-SLO Editor tool we used the Computer Emotion Scale (CES) [9]. 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.

5.4.2.3 Procedure

The experiment consisted of four sessions in a row conducted at the convenient time:

- **Work session 1:** Each lecturer selected a suitable CC-LO/SLO from a SLO Repository. Time spent in this work session was counted.
- **Work session 2:** Each lecturer checked the SLO selected with the VCS-SLO Editor and thought over the type of modifications and improvements to make in order to turn the SLO into a useful video-debate (CC-LR) as learning material to support a course. Time spent in this work session was counted.
- **Work session 3:** The 2 lecturers modified the SLO by using the VCS-SLO Editor as follows. Total time spent in this work session was counted.

Procedure:

- a. Edit each Dialog scene and scene part of the SLO and correct the syntax of the posts; select a suitable avatar for each original participant, select the appropriate performance indicators; select the emotional state; select the appropriate keywords; categories (speech

- acts) of each contribution. Use the semi-automatic capabilities to categorize posts and emotional states and report the experience. Time spent was counted.
- b. Create test assessment scenes and set appropriate assessment rules. Also create reference scenes to be connected with the assessment scene. Time spent was counted.
 - c. Create emotional scenes by selecting the most appropriate from the predefined emotional scenes given by the tool. Time spent was counted.
 - d. Sort out the scenes appropriately in the video timeline, connect them if necessary (by jumping from an assessment scene, etc.), test the new video-debate (CC-LR).
 - e. Check the video-debate created, perform final modifications and upload the video-debate in a CC-LR repository (server) for further use by students as learning material.

The lecturers were instructed to use the manual of the VCS-SLO Editor as provided in [17]. No training sessions on the tool were programmed given the strong background of the lecturers in developing and using e-learning systems.

After the task was finished, the lecturers were asked to fill out a questionnaire about their experiences with the Editor and the CC-LR, especially concerning the usability of the tool.

5.4.3 Evaluation and Validation Results

Following the methodology provided in Section 5.4.2, we will validate 3 aspects of the scenario: time to run the experience, the usability and emotions with the VCS-SLO Editor tool (H4.4.1) as well as this tool as a valuable resource (H4.2.2). For these purposes we used the metrics M4.4.1 through M4.4.5.

5.4.3.1 Time to run the experience

The experiment consisted of four sessions in a row conducted at the convenient time:

- **Work session 1:** Each lecturer selected a suitable SLO from the SLO Repository (*Figure 82*). Time spent in this work session:
Lecturer A: 5 minutes
Lecturer B: 7 minutes

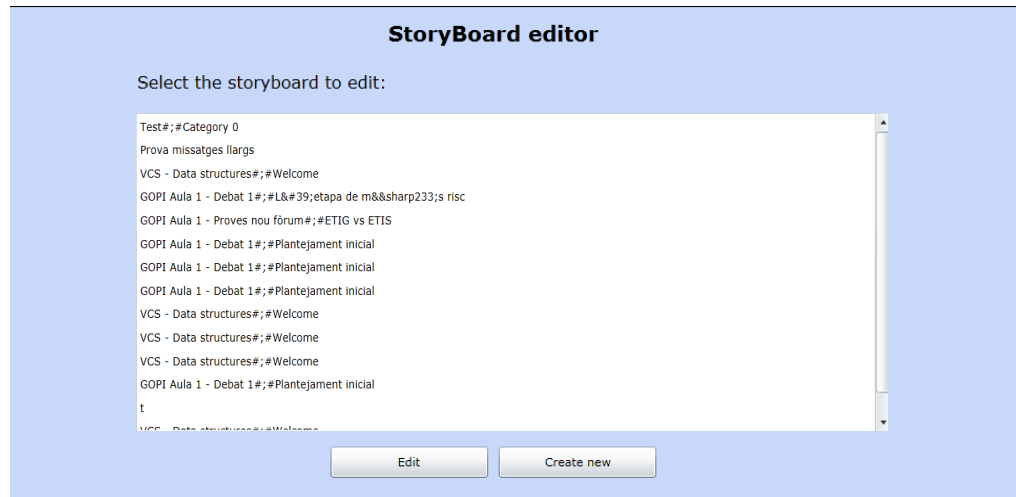


Figure 82: List of available SLO to be edited

- **Work session 2.** Each lecturer studied the selected SLO in the VCS-SLO Editor (Figure 83). Time spent in this work session:
 Lecturer A: 22 minutes
 Lecturer B: 30 minutes

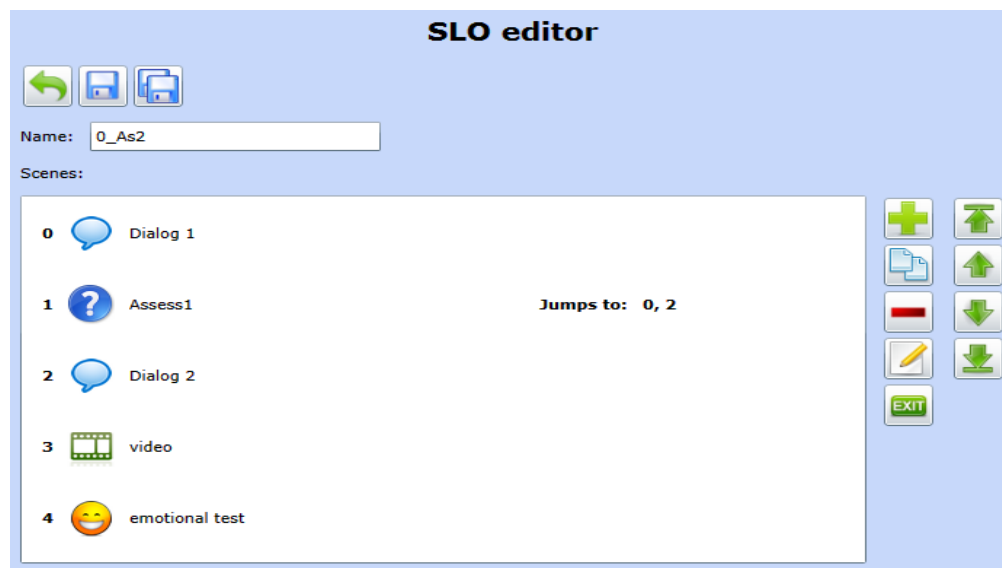


Figure 83: Structure of the SLO

- **Work session 3.** Each lecturer modifies the SLO with the Editor. Total time spent in this work session:
 Lecturer A: 1 hour and 9 minutes (for a 10-scene SLO).
 Lecturer B: 7 hours and 30 minutes (for a 65-scene SLO)
 - a. Edit and modify Dialog scenes (Figure 84). Time spent:
 Lecturer A: 15 minutes (for a 10-scene SLO).
 Lecturer B: 2 hours (for a 65-scene SLO).



Figure 84: Dialog scene editor

- b. Each lecturer created test assessment scenes. (see Figure 85). Time spent:
Lecturer A: 12 minutes (for a 10-scene SLO).
Lecturer B: 1h 30min (for a 65-scene SLO).

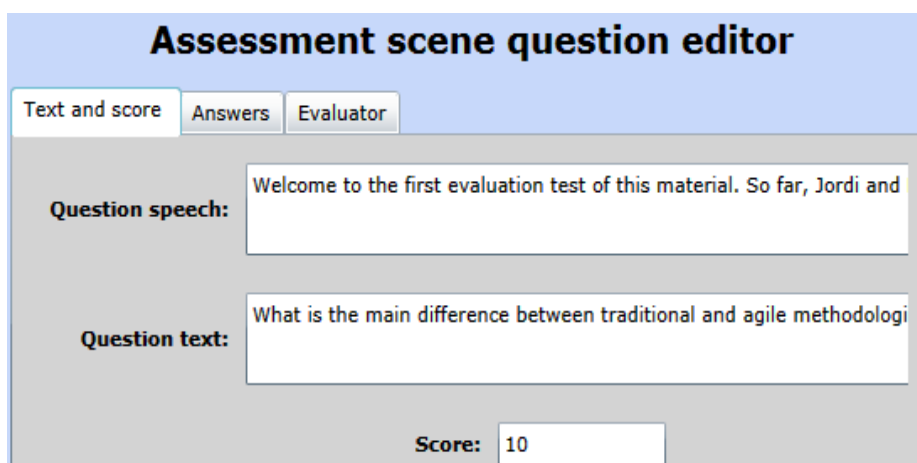
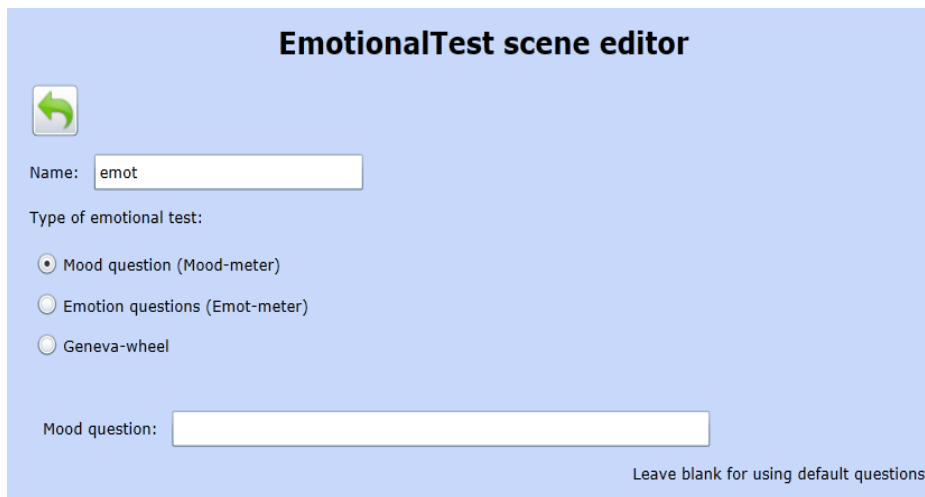



Figure 85: Assessment scene editor

- c. Each lecturer created emotional scenes (Figure 86). Time spent:
Lecturer A: 5 minutes (for a 10-scene SLO).
Lecturer B: 30 minutes (for a 65-scene SLO).



EmotionalTest scene editor



Name:

Type of emotional test:

Mood question (Mood-meter)

Emotion questions (Emot-meter)

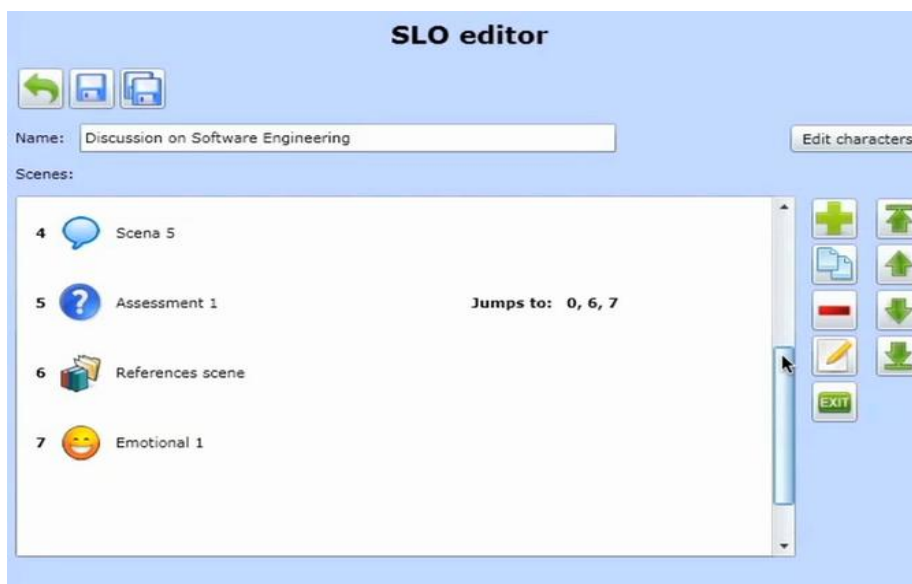
Geneva-wheel

Mood question:




Leave blank for using default questions

Figure 86: Emotional scene editor

- d. Each lecturer sorted out the list of scenes created in the video timeline. (Figure 87). Time spent:
 Lecturer A: 7 minutes (for a 10-scene SLO).
 Lecturer B: 1 hour (for a 65-scene SLO).







SLO editor

Name:

Scenes:

4		Scena 5	
5		Assessment 1	Jumps to: 0, 6, 7
6		References scene	
7		Emotional 1	









 
 
 
 

Figure 87: List of scenes created

- e. Each lecturer checked the newly created CC-LR, performed final modifications and uploaded in the CC-LR repository (Figure 88). Time spent:
 Lecturer A: 30 minutes (for a 10-scene SLO).
 Lecturer B: 2 hours and 30 minutes (for a 65-scene SLO).

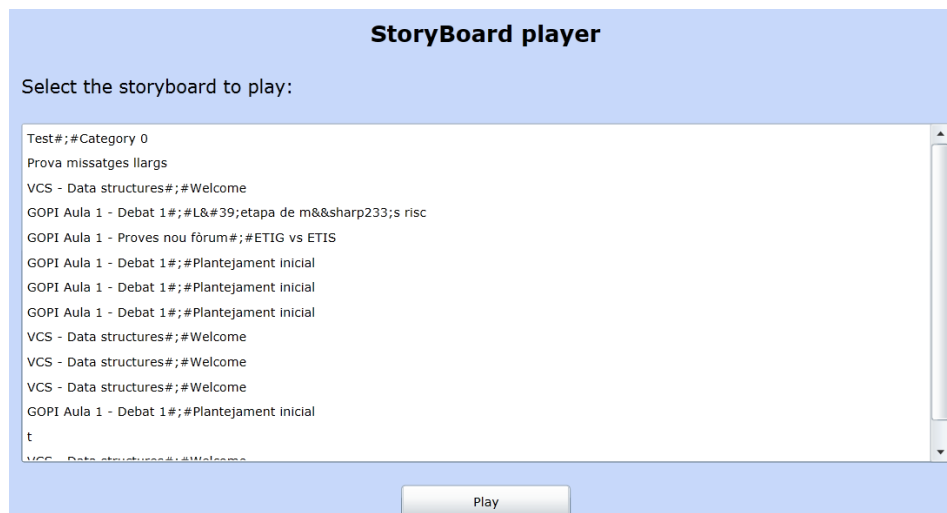


Figure 88: Repository of the CC-LR

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

Lecturer A: **2 hours and 57 minutes** (for a 10-scene SLO).

Lecturer B: **8 hour and 7 minutes** (for a 65-scene SLO).

5.4.3.2 Usability of the IWT

In this section, we analyzed the usability of the VCS-SLO Editor tool (H4.4.1). 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 85 and the score for lecturer B was 70 being both above the SUS mean score (68). **The 2 SUS score was 77.5.** The lecturers specially would like to use the VCS-SLO Editor tool often (M=5; SD=0; Md=5) and thought the tool was easy to use (M=4; SD=0; Md=4).

5.4.3.3 Emotion of the IWT

Regarding the lecturers' emotions during the work with the VCS-SLO Editor tool (H4.4.1), we used the mentioned CES scale. The results from the 4-point rating scale (n=2) are as follows:

- Happiness (M=2, SD=0, Md=2)
- Sadness (M=0.5, SD=0.7, Md=0.5)
- Anxiety (M=1, SD=0, Md=1)
- Anger (M=1.5, SD=0.7, Md=1.5)

As shown in the results, lecturers felt more happiness than the rest of bad feelings. Only some anxiety and anger was noticeable by Lecturer B, probably because the large SLO selected to work on (65 scenes) and the many hours employed to turn it into a CC-LR.

This is in line with the previous results on the IWT usability, where lecturer B found more problems than Lecturer A. and the procedure to create a contextualized course. In addition, the questionnaires showed in next section reflect some amounts of frustration and annoyance by Lecturer B, gain due to the large SLO selected and the time spent in it. On average though, both lecturers felt good or very good when using the Editor tool.

5.4.3.4 *The VCS-SLO Editor as a valuable resource*

In order to evaluate the tool and the resulting CC-LR as well as potential enhancements for the tool (H4.4.3), we asked the lecturers to answer six open questions:

1. Please describe what you liked regarding the SLO Editor.

Both lecturers liked the SLO Editor. In particular, Lecturer A reported that he liked “the ability of editing and personalizing each SLO in order to meet the specific requirements of the course”.

2. Please describe what you did not like regarding the SLO Editor

Both lecturers highlighted the great efforts to edit and customize an SLO. Lecturer A reported that “I had to make a great effort to customize each element of the SLO. It is inherent to the task itself, but the SLO Editor tools should be improved in order to facilitate more the work.” Lecturer B also reported that “I spent more than 8 hours to customize one single SLO...I can’t imagine customizing dozens of SLOs!” However, Lecturer B added that “...considering the resulting CC-LR is a real learning material if we compare it to the workload to create a regular learning module or activity, the effort to create a CC-LR may be reasonable”

3. Do you have any suggestions for improvements?

Lecturers’ recommendations were in line with the comments provided in the previous questions. In particular, even if usability was not considered an issue, they suggested improving it in order to speed up their work. In particular, Lecturer A suggested “add functionalities to enable ‘mass modifications’”. Lecturer B suggested improving many small functional details, such as getting the scene/part/character selected in the list when going back after working on this element, instead of always getting the first element of the list...this can avoid wasting a lot of time when the SLO is very large”.

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

Both lecturers agreed that the user’s manual was complete and detailed enough. However, Lecturer B missed more information about how to set up the assessment rules as he had to ask a technician about this.

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

Lecturer A thought that the tool would be useful for teachers “because it let them to personalize the SLO content in order to adapt it to the course requirements.” However, he mentioned that “the main disadvantage is the large effort needed to edit an entire SLO so, as already said, some improvements to help on that should be great.” Lecturer B found “it may be quite problematic to introduce this tool to lecturers at UOC as nobody would spend hours customizing a single SLO”. However, he thought “...it is however a matter of mentality as the resulting CC-LR can replace part of the official material of the course, which also needs spending a lot of time to create every course; this argument can be very convincing to create CC-LR with the Editor”

6. Do you think that your students would benefit from the course with CC-LRs?

Both lecturers were very positive with respect to having customizing SLO based on students’ contributions as regular material to support the courses. In particular, Lecturer A mentioned that “students will be able to access contents about the course in a more attractive and intuitive way with interaction functionalities.” Lecturer B indicated that “I think students would appreciate having such a resources and will very beneficial for them as they can learn by seeing other students’ performance in a social way”.

5.4.4 Conclusion

This experiment at UOC was conducted by real experts in developing complex e-Learning 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. Considering this strong background in web applications as developers and users, it was outstanding not to report relevant technical problems when using the tool, which means that the tool performed very well, even if non-expert users (G4.4.2). Since no trial was designed for instructors in the initial experimentation we cannot provide here comparison results.

From the analysis of the usability of the tool it was shown that both lecturers considered usability was satisfactory and referred this aspect above the SUS mean score meaning that this tool has a high perceived usability (G4.4.4). The lecturers’ emotions when using the tool were in line with the usability results, feeling more satisfied (happiness emotion) than the rest of bad feelings. However, some improvements on usability, even if minors, were suggested by the lecturers (G4.4.3).

Even if one lecturer reported to spend a lot of time to customize a single SLO he also mentioned that it is comparable to the time needed to create any regular material. Hence both lecturers were quite satisfied of creating new learning material efficiently (G4.4.6) and especially in an effective way (G4.4.5) as the material was based on students’ contributions, thus having an important impact in the learning process (See also Section 5.3.4.3).

To sum up, the lecturers liked the idea of editing and personalizing each SLO in order to meet the specific requirements of the course and found it very beneficial for students. This achieved the main goal (G4.4.1) as for creating learning material (CC-LR) from a threaded discussion.

6 R5. Storytelling

The goal of this scenario is to allow an efficient learning about knowledge and behavior 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 learner's needs.

6.1 Evaluation and Validation Procedure

The purpose of the second experimentation phase is to satisfy all the scenario goals and criteria that are not completely covered in the first phase.

Following we report, as already exposed in [3] the evaluation hypotheses in correspondence of the scenario goals and the metrics for fulfilling specific criteria.

Scenario 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.

Scenario 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.

Scenario 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.

Scenario 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.

6.2 Method

6.2.1 Participants

In order to evaluate the storytelling scenario and validate it through the effects in the learning process, 4 schools have participated in the experience. In the specific, 4 tutors and 58 students have been enrolled.

For each school the students were allocated into one classroom composed by two groups: experimental and control. The groups use IWT platform, in two different way: the experimental group delivers a learning course by using complex learning resources (as the SLO); the control group delivers a learning course by using traditional learning materials as power point presentations; pdf. file...That in order to compare, through qualitative and quantitative data, the learning process for each group.

All students were supervised by one tutor during the experiment.

6.2.2 Apparatus and Stimuli

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

The total students belonging to experimental groups are 29.

On completion of the session they have filled a Post-Questionnaire, which includes the following sections: demographic data, storytelling learning object activity, 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. 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).

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

- During the use of the storytelling resource you've got to navigate autonomously and explore the various ways that the story presented?
- Have the reflection moments give you the possibility to find the key of lecture wrt the objective?
- The assessment events, distributed within the various key situations has allowed you to understand what you are learned and understand concepts about which you needed to enforcement before moving forward?
- To a specific point of the history, you had the opportunity to evaluate your emotional state. Do you think that it's useful for more understanding if your emotional state are influencing your interaction with the story?
- The storytelling's interface is easily usable?
- The visual and auditory stimuli were defined so that you can follow the events of history and allow you to pay attention to the most important events?

The answer categories for this section are: “Not at all”, “A little”, “Moderately”, “Very”, “Completely” and correspond to 5 points of the Likert's scale. The rating scale ranged from “Not at all” (1), “A little” (2), “Moderately” (3) to “Very” (4), “Completely” (5).

For usability of the storytelling, we used the SUS (System Usability Scale), which contains 10 items and a 5 points Likert's scale to state the level of agreement or disagreement, for instance “I think that I would like to use this system frequently”. 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).

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”.

Finally, quantitative data was also collected from IWT database and log files.

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 within IWT. The course has been delivered by two groups of users having different learning styles: experimental and control.

The experimental group has been composed by **analytical** students: a kind of student that likes testing and in a second time to match if the correct solution to a specific problem is correct or not with respect to its hypotheses.

The control group has been composed by **holistic** students: a kind of student that likes to analyze the problem and the associated information before to start a specific activity.

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 by 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 sees his class.

Within the class, each student delivers a personalization of the course about the big emergency. In the specific, for the experimental student has been created a personalized learning path by using 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.

At the end of the course, a qualitative survey has been given both the students and the tutor in order to test the knowledge acquired through the storytelling with respect to a passive learning resource.

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. We also include in this section the evaluation of the questionnaire. For these purposes we used metrics M5.5, M5.6 and M5.7.

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 four schools belonging to the experimental group and the quantitative data obtained from the IWT Database.

6.3.1 The storytelling activity

The Figure 89 shows the average of the students' evaluation with respect to the storytelling activities.

The Figure includes the average's answers to the six questions exposed in the Section 6.2.2 and obtained by using a 5-points Likert's scale for analyzing the answers.

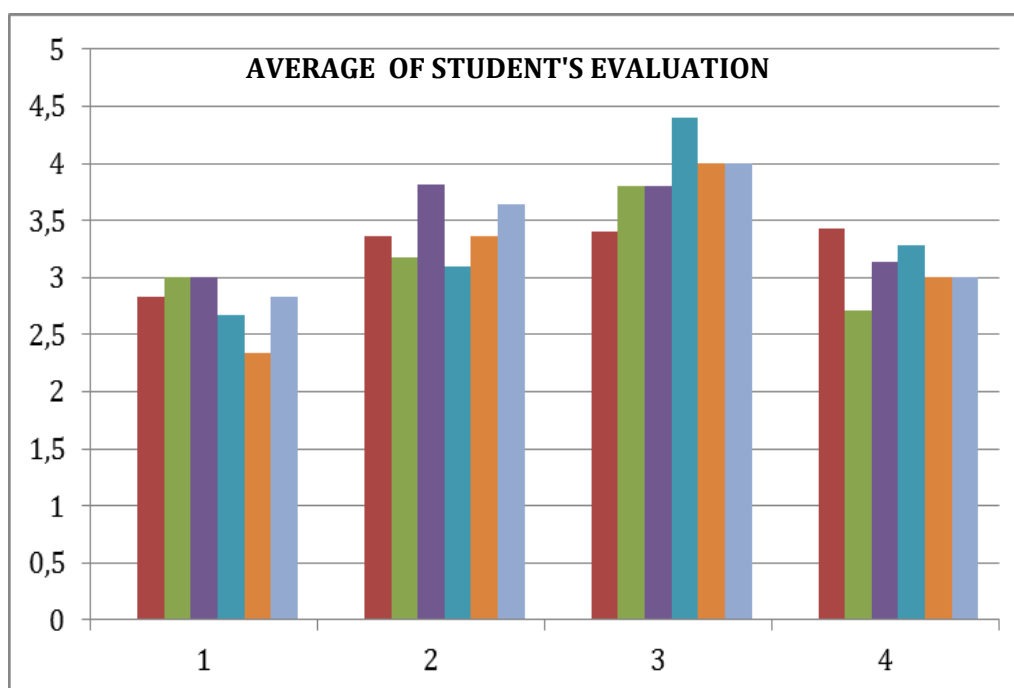


Figure 89: Average's answers to opened questions

A lot of students have answered in a successful way to the Q2: that shows how the students have appreciated the combination of activity and reflection for each situation. This value has reached its maximum score within the third school.

A good percentage has been obtained also for the third item Q3 related to the assessment events: the assessment events have contributed to mature in their a judgment with respect to what they had to acquire before that history presented new difficulties and new challenges in terms of knowledge and skills to be deployed.

Very appreciated has been the opportunity to evaluate the emotional state in a specific point of the history (Q4) that indirectly influences the continuation of the history.

The general analysis of the averages for each question confirms the data previously discussed and partially valid the SLO as a possibility to change and renew the educational experience of the digital natives; the experience takes into account an instructive architecture for which the assessment and reflective events have obtained the most successful.

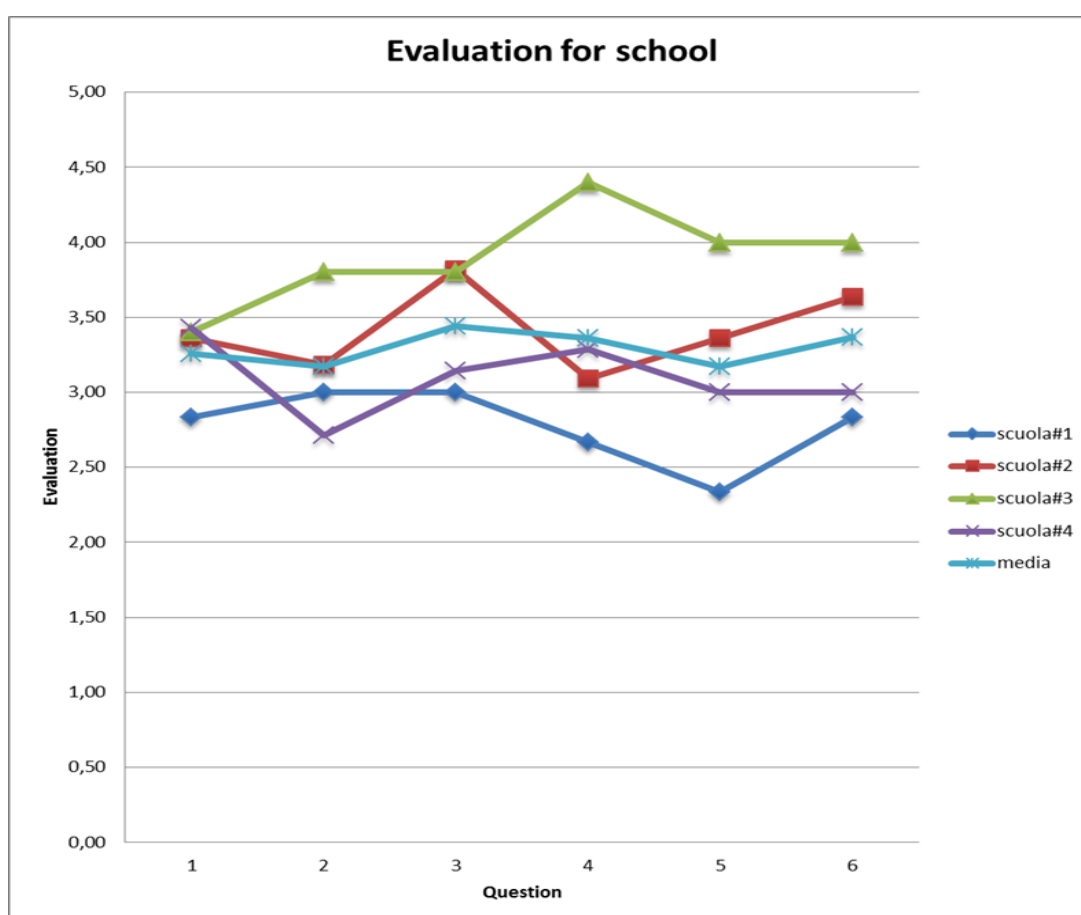


Figure 90: Evaluation for school

6.3.2 Emotional Aspects

Regarding the students' emotions during the work with the IWT tool, 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=29) were as follows:

- Happiness (M=1.72, SD=1, Md=2) (Figure 91)

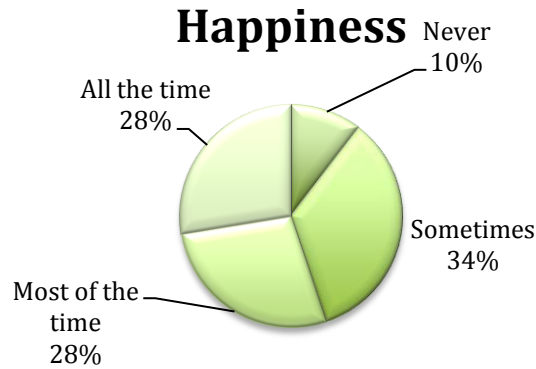


Figure 91: Results on the Happiness emotion

By analyzing the Figure 91 from another point of view (see Figure 92), we have obtained the more high values for the first two schools, that have appreciated in particular the new structure of the instructional events and the opportunity of auto-evaluate the specific knowledge.

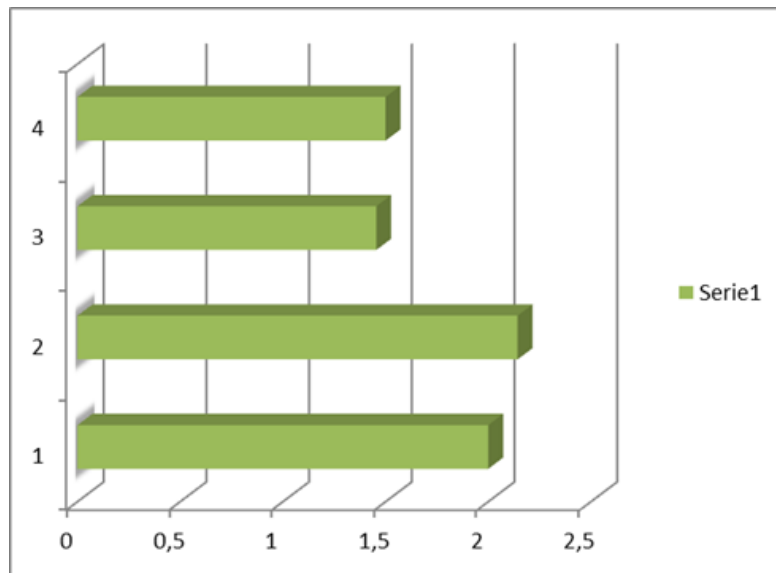


Figure 92: Happiness emotion from a different point of view

The percentages of Sadness (M=0.62, SD=0.62, Md=1) and Anxiety (M=0.55, SD=0.63, Md=0) (see Figure 93) are significant: by considering the topic of the storytelling (earthquake in an Amusement Park) we can confirm that the construction of the story was able to empathize with the student in overcoming of the various trials and tribulations.

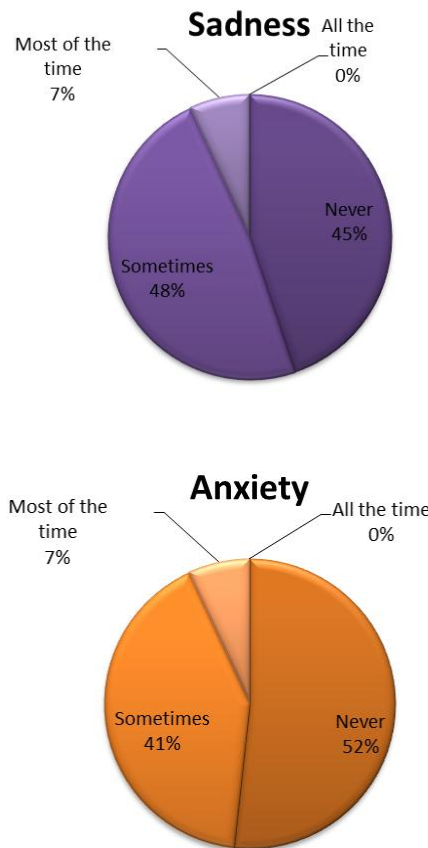


Figure 93: Results on the Sadness and Anxiety emotions

The partial view show a good average with respect to the Q3 and Q6 questions: that indicates a total involvement of the students in the different situations, where they have registered a constant level of performance anxiety and a functional tension to do better for helping the others (see Figure 94).

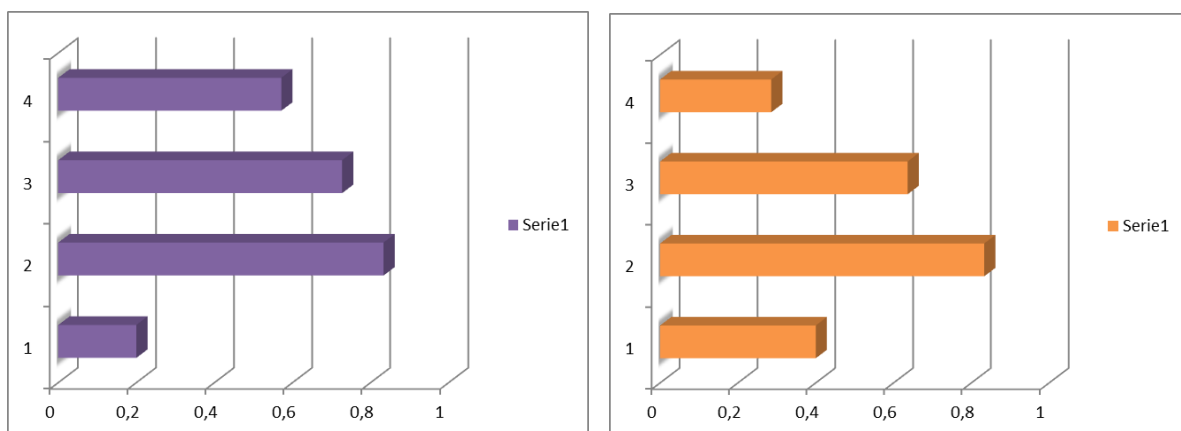


Figure 94: Sadness and Anxiety emotions from another point of view

A few students have registers an anger sentiment ($M=0.28$, $SD=0.53$, $Md=0$) (see Figure 95)



Figure 95: Results on the Anger emotion

6.3.3 Usability of the Storytelling

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

To investigate the overall usability of the tool, we used the SUS and included it a specific section of the qualitative 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).

The 10 items that composed the SUS questions are:

1. I would use this tool regularly
2. I found it unnecessarily complex
3. It was easy to use
4. I'd need help to use it
5. The various part of the tool worked well together
6. Too much inconsistency
7. I think others would find it easy to use
8. I found it very cumbersome to use
9. I felt very confident using the tool
10. I needed to understand how it worked in order to get going

This questionnaire has been submitted immediately after the interaction of the student with the system: in such a way the students must materialize their perception with respect to the system.

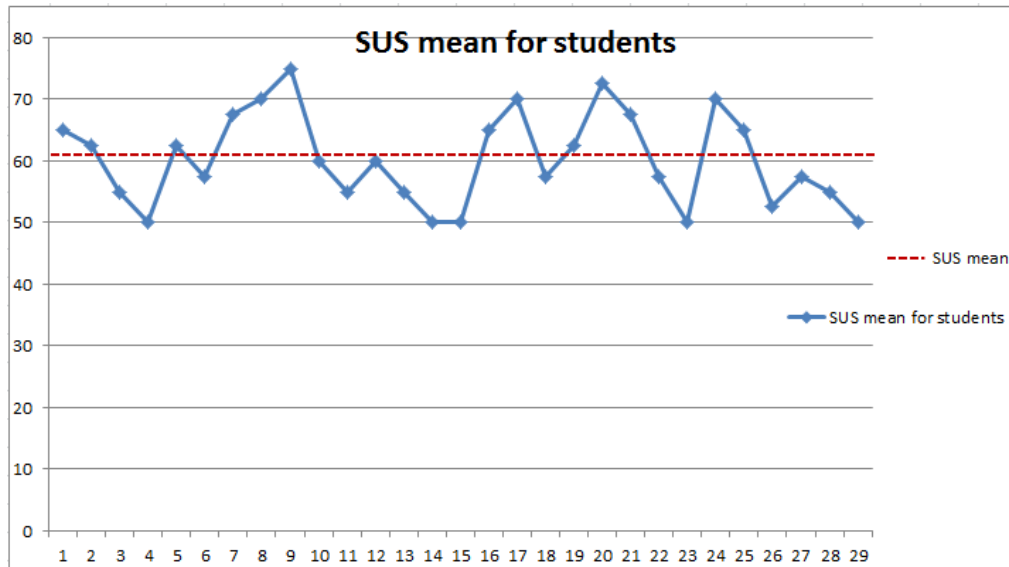


Figure 96: Storytelling's usability

The SUS mean score is 60.25. The minimum score is 50 the maximum score is 75 (see Figure 96). In particular, some questions (like Q3 “ I thought the system was easy to use“ and Q5 “I found the various function in the tool were well integrated”) show a predominance of score 4 of the Likert scale (I Agree).

Some questions (like Q7 “I would imagine that most people would learn to use the tool very quickly” or Q9 “I felt very confident using the tool”) show a good presence of score 4 e 5 (Agree/Strongly Agree) of the Likert scale. These data show as the SLO has been appreciated by the students both for its usability and for the integration in IWT.

Considering the answers to the items Q4 “I think that I would need the support of a technical person to be able to use the tool” a predominance of the score 4 has been registered; probably a more use in terms of permanence time and SLO exploration could improve students' sense of confidence.

6.4 Validation Results

Following the methodology described in Section 6.2, in this section we will analyze the metrics M5.5-M5.11, related to the interaction of the students/instructors with the Storytelling Learning Object.

6.4.1 The Storytelling as a valuable resource

In order to analyze the interaction with the Storytelling educational resource, we investigate for each classroom the number of accesses to the resource and the time spent for it (M5.5 and M5.6): all the experimental students (29) have interacted with the SLO also by using multiple accesses.

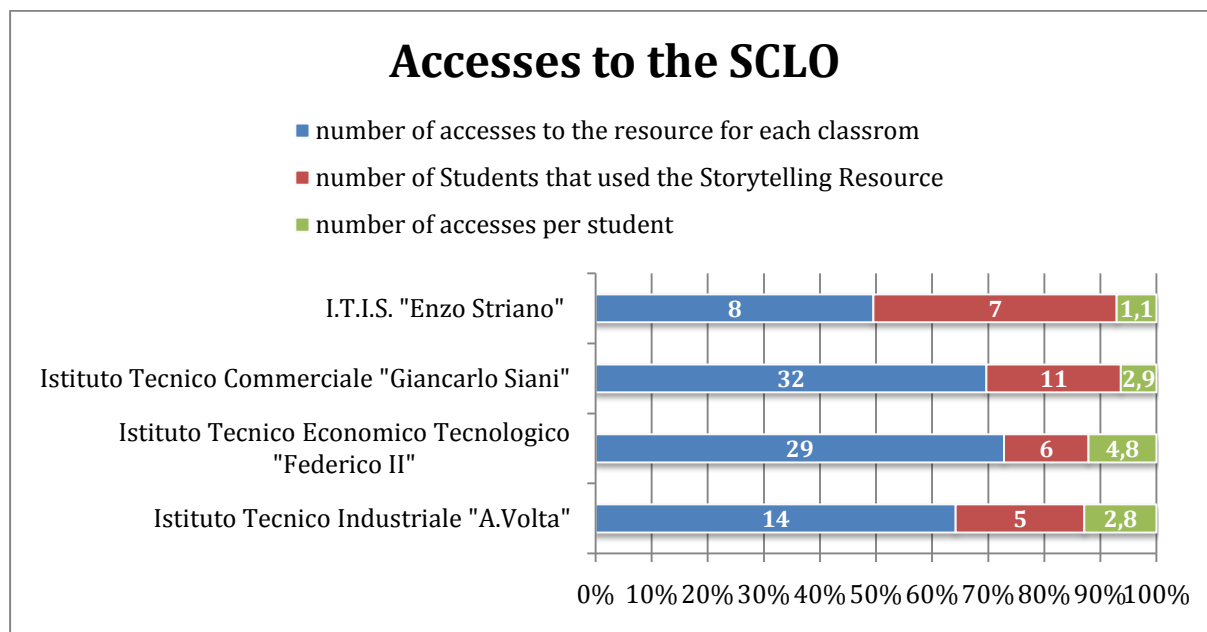


Figure 97: Access to the Storytelling LO

The Figure 97 shows as the first experimental class students “ITIS Striano” have had an approach continued in the delivery of the resource storytelling. In fact the students, after the first access to the resource, have continued in the use without interrupting the path. This aspect could be linked to the engagement of the student in the history.

On the contrary, a different behaviour has been obtained by the third school “Federico II”, where the students have stopped and restarted several times the complex resource so that to have an average access of about 4. This could be justified by a more reflective style of learning with the need of focus and reflect on specific topics: indeed these students are the same that have taken notes and tagging of particular situations, additional functionalities of the implemented tool.

So we have had two different styles of resource use: on the one hand the tendency to the discovery and to the progressive approximation to the learning; on the other hand the tendency to multitasking and the preference to a cognitive moment.

In order to also validate the didactic structure of the six situations that compose the storytelling resource, we compare the navigation time (expressed in seconds) to the different knowledge types to be acquired (M5.10 and M5.11):

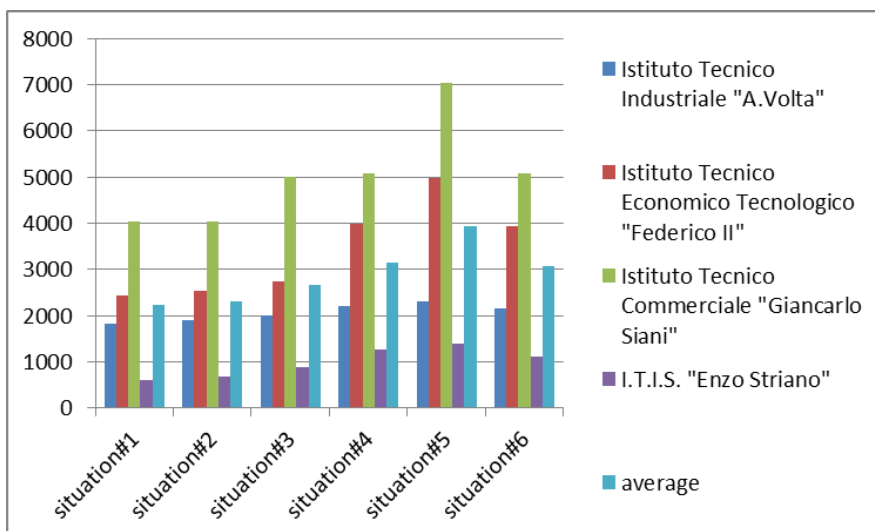


Figure 98: Navigation time with respect to different knowledge types to be acquired

The Figure 98 shows a satisfactory linear progression: indeed from the situation #1 to situation #6, the students' attention increases taking into account a more involvement compliant with the correspondent level of the Bloom's taxonomy.

As we can see, the first three situations (*Beginning, Call Adventure e Problem*) are quite introductory, so the average fruition time is between 20 and 30 minutes; while for the other three situations (*Middle, Solution, Closure*), that required a more cognitive involvement of the student, the average fruition time is between 30 and 40 minutes.

In order to investigate the added value of the microadaptivity, we focus on the situation #5 and on how the assessment results change after the students have taken a different role:

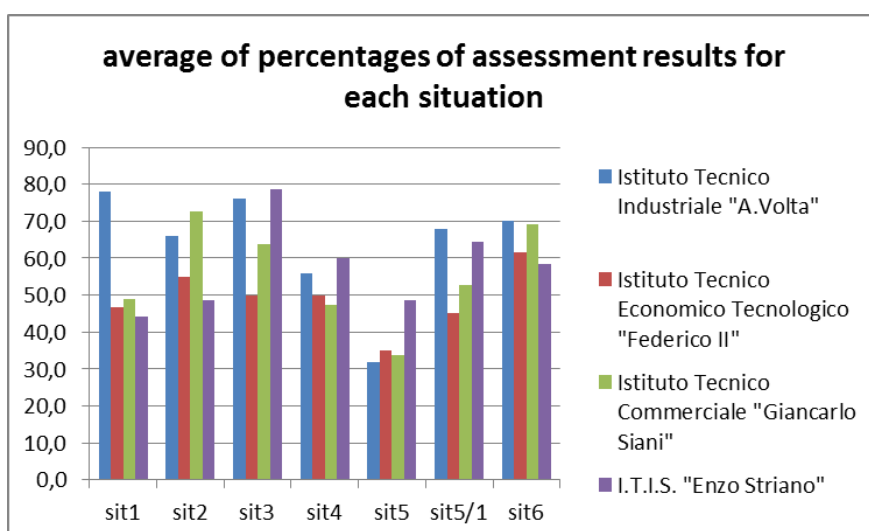


Figure 99: Focus on the micro-adaptivity

The *Figure 99* shows as some analytical students (#12) have taken a new role that has allowed for filling the competence gaps obtained with the previous role. Thanks to an emotional test, combined with a cognitive assessment, the system has individuated the correct role to associate to the specific student. In such a way, the student can see the history from another point of view, registering, as shown in the figure, an assessment results compatible with the rest of the experimental group.

6.4.2 Acquired competences and didactic efficiency of the Storytelling

In this section we investigate on the results obtained from competences and assessment point of view in order to validate the remaining metrics (M5.7- M5.8–M5.9).

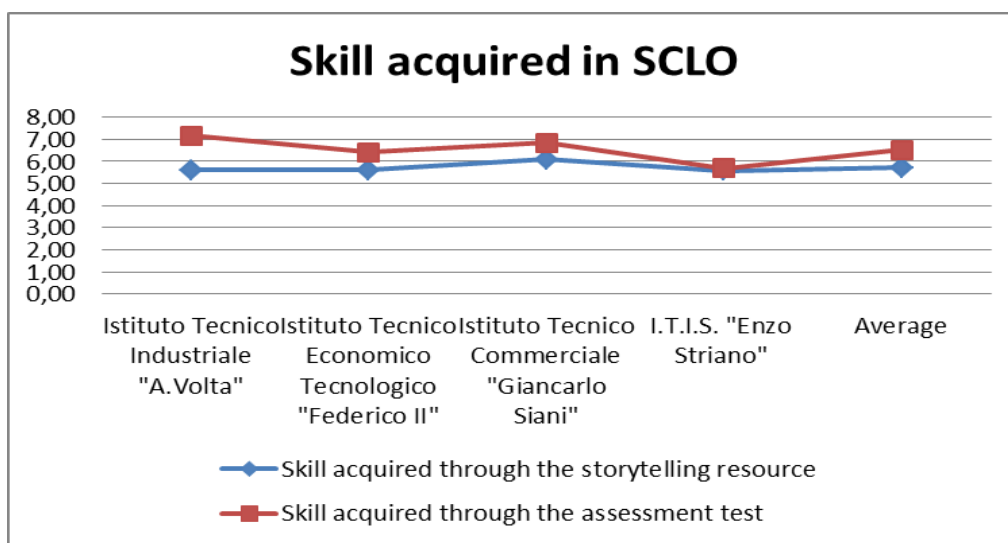


Figure 100: Skill acquired through the Storytelling

The Figure 100 allows for comparing the skill acquired through the storytelling resource respect to them acquired through the assessment test.

The first significant datum that could be brought out is that the gap between the two modalities is not relevant: this means that the storytelling resource was able to guide the student in the acquisition of specific concepts by also resorting to alternative paths. That it is also confirmed by the average competences obtained only through the delivery of resource (M=5.72): level over the minimal threshold fixed by the teachers (equal to 5).

In a second time, the competences average has fixed through the final assessment that has given to the students the institutionalization of the implicit knowledge acquired by delivering the six situations of the resource. The final average competence's level, reached by the students, is equal to 6.53.

Comparing the results obtained in the Figure 100 with the results derived by alternative LO, having the same learning objective, we can see how the storytelling resource give the added

value in a learning context: the Figure 101 indeed shows the competences achieved by delivering traditional Learning Objects.

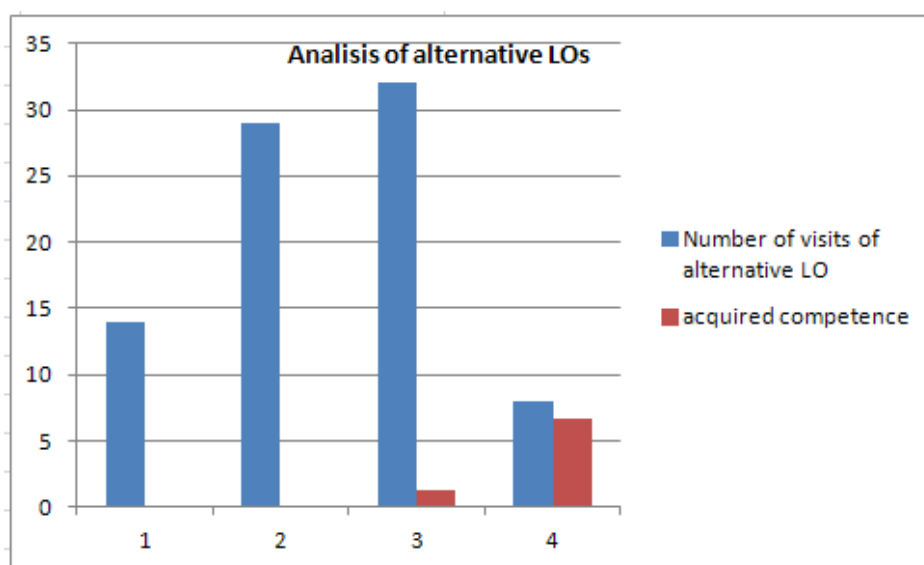


Figure 101: Analysis of alternative LOs

As exposed in the Figure 101, though the number of visits of alternative Learning Object is quite high, very low are the competences linked to them. In particular, they have been achieved only for the third and fourth schools.

Another relevant aspect could be the mapping between the accesses number of the experimental groups to the SLO and the skill's level acquired.

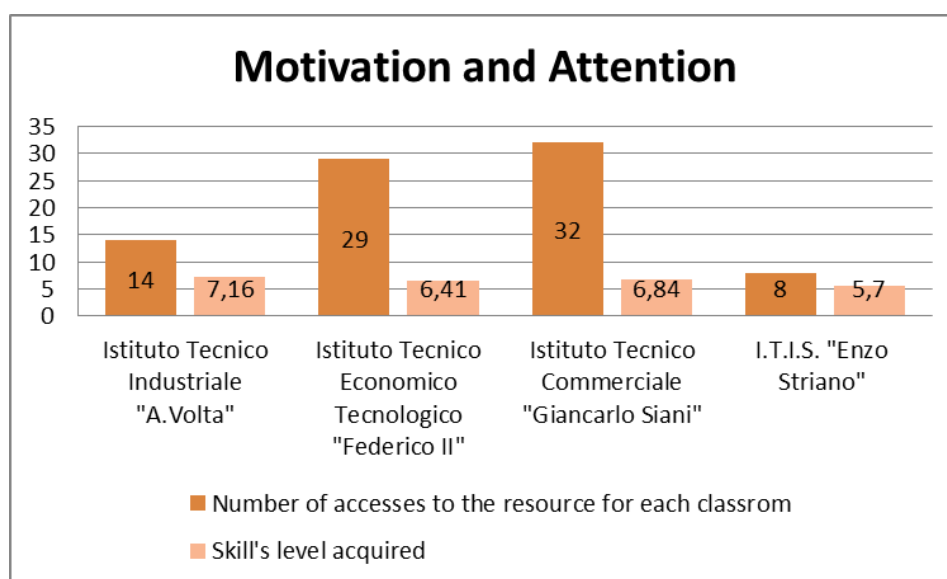


Figure 102: Motivation and Attention

This datum is relevant since allows us to understand if the student has used the time out to explore some concepts individually and make a self assessment. As shown in the Figure 102, though the accesses number is high (see second and third classroom), the reached competences are above the average: that confirms the didactic validity of the storytelling resource that has allowed students to build some conceptual links recalling even in the case of stops and starts.

6.5 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 6.1). The results are summarized by taking also into account the survey given to the teachers. The answers were given on the 5-point Likert scale, so that teachers 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 teachers (M= 4.25) all agree that the storytelling resource provides to the students the opportunity to express their native style characterized by a progressive exploration of knowledge in a guided and structured context

With respect to the Q2 item “*Is, from your point of view, the role taking a good strategy for filling some gaps through a different perspective?*” the teachers strongly agree. They suggest the application of this methodology to a more large domain by taking into account also humanistic and artistic topics.

Very appreciated is also the average with respect to the item Q4 “*Do you think that the role making is an innovative aspect for reviewing the history built in a collaborative way by the students?*”. Indeed the teachers think that the story-wiki is a good space that gives the added value to a co-narration of the history.

From the student’s point of view, the output of the 1st experimentation phase, as the need to be able to stop the flow of the storytelling for having brainstorming with the tutor and their peers, has been taken into account, by putting this additional functionality in the development of the storytelling prototype (see D6.2.2 deliverable)

7 R6. A Serious Game for Civil Defense

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

7.1 Evaluation and Validation Procedure

The purpose of the second experimentation phase is to satisfy all the scenario goals and criteria that are not completely covered in the first phase.

Following we report, as already exposed in [3] deliverable the evaluation hypotheses in correspondence of the scenario goals and the metrics for fulfilling specific criteria.

Scenario 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.

Scenario hypotheses

- 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 contributes 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.

Scenario 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.

Scenario 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.

7.2 Method

7.2.1 Participants

In order to evaluate the Serious Game scenario and validate it through the effects in the learning process, 4 schools have participated in the experience. In the specific, 4 tutors and 58 students have been enrolled.

For each school the students were allocated into one classroom composed by two groups: experimental and control. The groups use IWT platform, in two different way: the experimental group delivers a learning course by using complex learning resources (as the Serious Game); the control group delivers a learning course by using traditional learning

materials as power point presentations; pdf. file...That in order to compare, through qualitative and quantitative data, the learning process for each group.

All students were supervised by one tutor during the experiment.

7.2.2 Apparatus and Stimuli

We asked to the experimental group of each school to interact with the Serious Game Learning Object related to the risk management in a complex context as the evacuation from an afire virtual building.

The total students belonging to experimental groups are 29.

On completion of the session they have filled a Post-Questionnaire, which includes the following sections: demographic data, Serious Game learning object activity, 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 Serious Game) 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. 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).

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

1. How responsive was the game to actions that you initiated (or performed)?
2. How much did the visual aspects of the game involve you?
3. How compelling was your sense of the game moving through space?
4. How much did the visual display quality interfere or distract you from performing the assigned tasks?
5. How completely were your senses engaged in this experience?
6. To what extent did external events distract from your experience of the game?
7. How easily did you adjust to the control devices used to interact with the game?
8. Was the information provided through different senses in the game (e.g., vision, hearing, touch) consistent?

The answer categories for this section are: “Not at all”, “A little” , “Moderately” , to “Very”, “Completely” and correspond to 5 points of the Likert’s scale. The rating scale ranged from “Not at all” (1), “A little” (2), “Moderately” (3) to “Very” (4), “Completely” (5).

For usability of the storytelling, we used the SUS (System Usability Scale) which contains 10 items and a 5 points Likert’s scale to state the level of agreement or disagreement, for instance “I think that I would like to use this system frequently”. 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).

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/dispirted.”)
- 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”.

Finally, quantitative data was also collected from IWT database and log files.

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 within IWT. The course has been delivered by two groups of users having different learning styles: experimental and control.

The experimental group has been composed by **analytical** students: a kind of student that likes testing and in a second time to match if the correct solution to a specific problem is correct or not with respect to its hypotheses.

The control group has been composed by **holistic** students: a kind of student that likes to analyze the problem and the associated information before to start a specific activity

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 by a Serious Game, for supporting intuitive learning processes in case of fire in school.

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

After that each student logged in IWT platform, he sees his class.

Within the class, each student deliveries a personalization of the course about the big emergency. In the specific, for the experimental student has been created a personalized learning path by using 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.

At the end of the course, a qualitative survey has been given both the students and the tutor in order to test the knowledge acquired trough the storytelling with respect to a passive learning resource.

7.3 Evaluation Results

In this section we focus on the activity level, usability and emotional aspects of the Serious Game Learning Object delivered by IWT platform. We also include in this section the evaluation of the questionnaire. For these purposes we used metrics M6.2, M6.3 and M6.4.

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 four schools belonging to the experimental group and the quantitative data obtained from the IWT Database.

7.3.1 The Serious Game Activity

The Figure 103 shows the average of the students' evaluation with respect to the Serious Game activity.

The Figure includes the average's answers to the eight questions exposed in the Section 7.2.2. and obtained by using a 5 points Likert's scale for analyzing the answers.

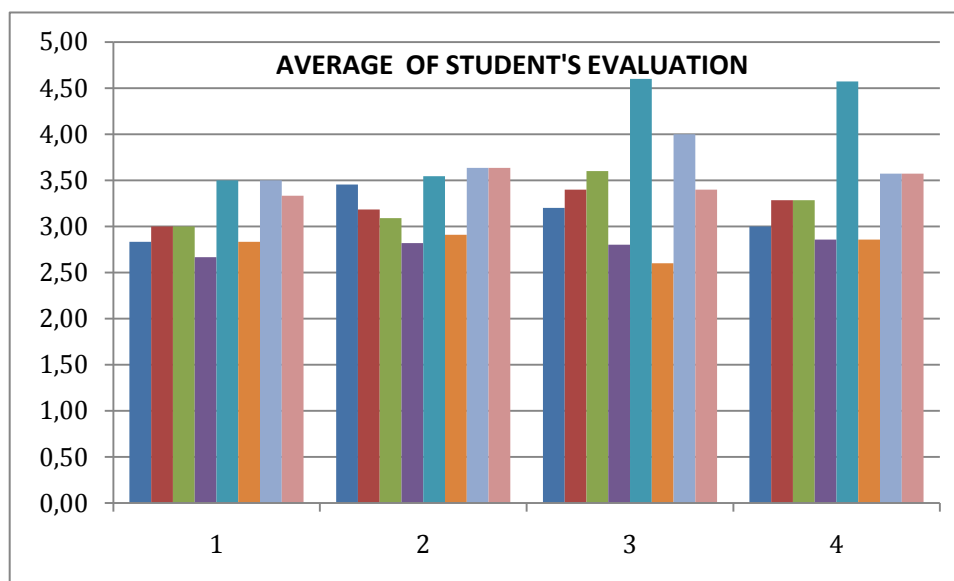


Figure 103: Evaluation of the SG activity

A lot of students have answered in a successful way to the item Q5 (M.3.9), related to the sense of engagement in the experience: that shows how the students have appreciated the immersive reality of the game. This value has reached its maximum score within the third and fourth schools. On the contrary the percentage obtained with respect to the item Q6 (M:2.8) is very low and that confirms the previous results.

Also the results obtained in correspondence of items Q7 (M.3.6) and Q8 (M.3.5) are interesting: the first one, related to the interaction with the game by using the control devices,

shows as the students have not found particular problems with the new device. The second one, related to the information obtained through different senses of the game, as vision, hearing, touch, have caught the students attention.

A quite low percentage has been obtained in correspondence of item Q4 (M:2.7), related to the quality of the visual display. That could be justified taking into account that the PC used by each classroom had hardware performance not very high.

7.3.2 Emotional Aspects

Regarding the students' emotions during the work with the IWT tool, 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=29) were as follows:

- Happiness (M=2.07, SD=0.65, Md=2) (Figure 104)

Happiness

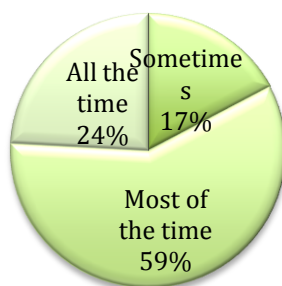


Figure 104: Results on the Happiness emotion

- Sadness (M=0.34, SD=0.48, Md=0) (Figure 105)

Sadness

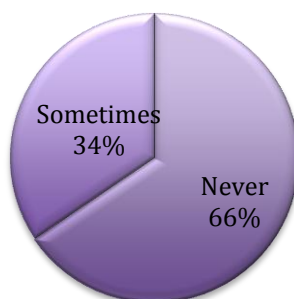


Figure 105: Results on the Sadness emotion

- Anxiety ($M=0.38$, $SD=0.63$, $Md=0$) (*Figure 106*)

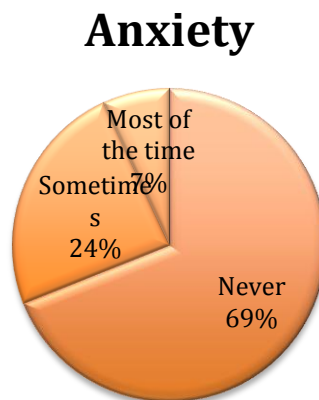


Figure 106: Results on the Anxiety emotion

- Anger ($M=0.36$, $SD=0.77$, $Md=0$) (*Figure 107*)

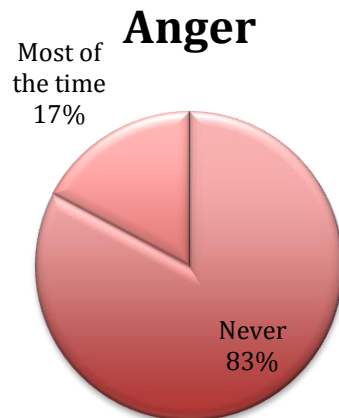


Figure 107: Results on the Anger emotion

7.3.3 Usability of the Serious Game

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

To investigate the overall usability of the tool, we used the SUS and included it a specific section of the qualitative 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).

The 10 items that composed the SUS questions are:

1. I would use this tool regularly
2. I found it unnecessarily complex
3. It was easy to use
4. I'd need help to use it
5. The various part of the tool worked well together
6. Too much inconsistency
7. I think others would find it easy to use
8. I found it very cumbersome to use
9. I felt very confident using the tool
10. I needed to understand how it worked in order to get going.

The results are summarized in the following picture:

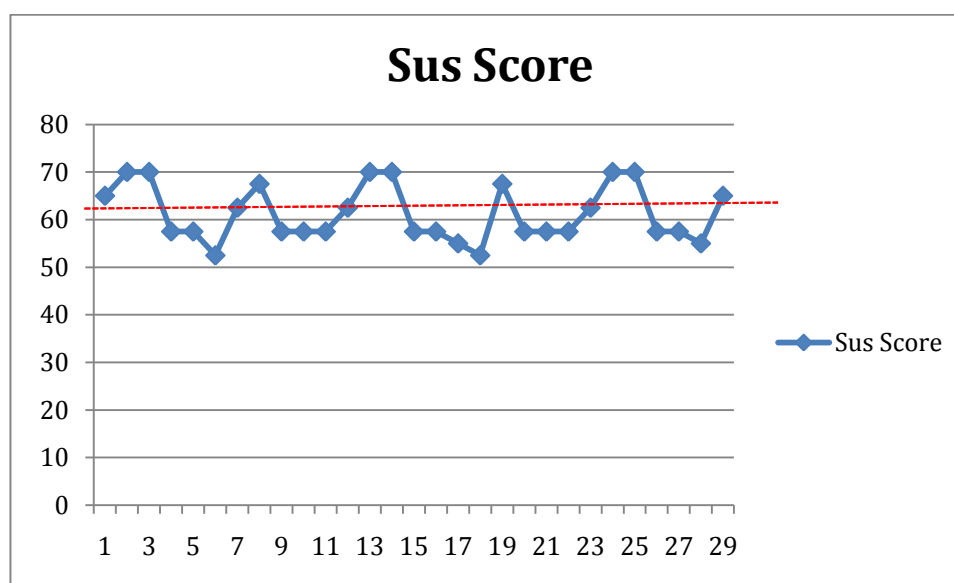


Figure 108: Usability of the SG

The SUS mean score is 61.29. The minimum score is 52.5 (achieved twice) the maximum score is 70 (see Figure 108). It is worth mentioning that due to the specimen nature (students) some questions (like Q1 “I would use this tool regularly” and Q7 “I think other would find the tool easy to use”) show a predominance of score 3 of the Likert scale (Neither/Nor). A lot of score 5 of the Likert scale (Strongly Agree) has been registered in correspondence of Q9; that indicates that the students have interacted enough easily with the tool.

7.4 Validation Results

Following the methodology described in Section 7.2, in this section we will analyze the metrics M6.1 –M6.9, related to the interaction of the students/instructors with the Serious Game Learning Object.

7.4.1 The Serious Game as a valuable resource

In order to analyze the interaction with the Serious Game educational resource, we investigate for each classroom the use spent for the resource (M6.1-M6.3):

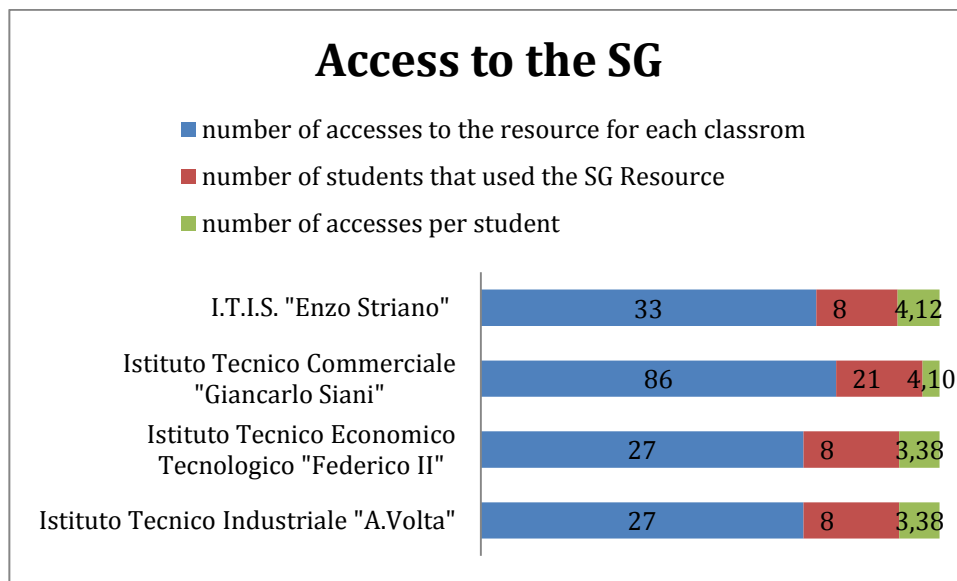


Figure 109: Access to the SG

As we can observe, the average value for which each student has interacted with the SG resource is 3.74. This is an appreciable result, if we consider that the SG is a very interactive game that needs to be investigated different time before to discover the end of it by respecting the fixed time.

7.4.2 Acquired competences and didactic efficiency of the Serious Game

In this section we investigate on the results obtained from competences and assessment point of view in order to validate the metrics M6.4 to M6.7.

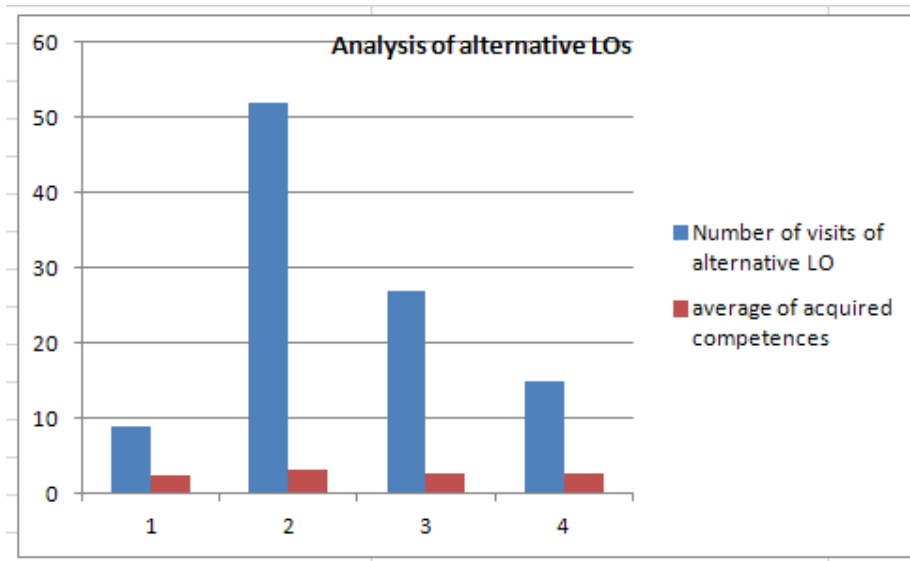


Figure 110: Analysis of alternative LOs

The Figure 110 shows the relationship between the number of visits of alternative LOs (M 6.4) and the average of the relative acquired competences (M6.6). As we can observe this latter aspect (M= 2.71) is very low respect to the average of the visits of alternative LOs (M= 25.75). So we can affirm as the traditional LOs have not much advantages in a learning course related to the management risk.

In order to evaluate the level of worthiness of a SG as a educational resource, we compare the skill obtained by delivering the SG resource and the skill obtained through the assessment test

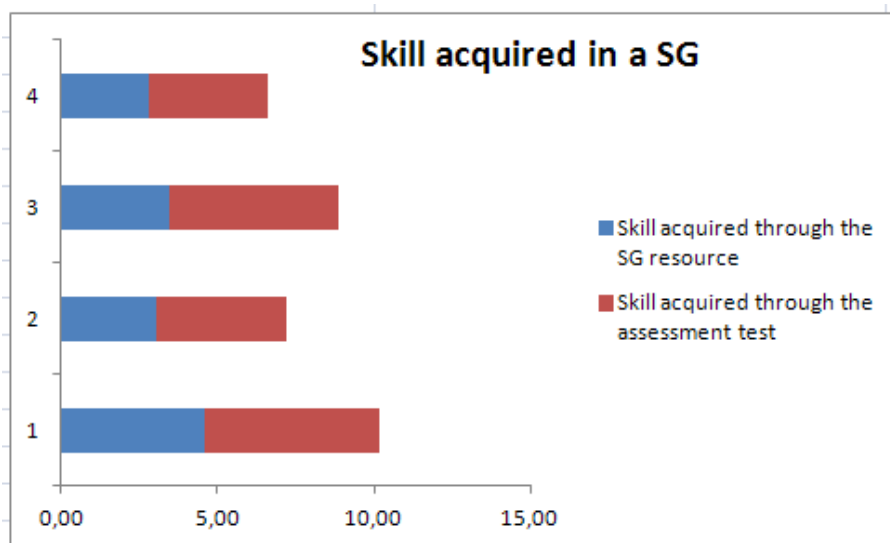


Figure 111: Skill acquired through a SG

As we can see in *Figure 111*, the SG resource includes an appreciable auto-assessment ($M= 3.47$) that allows to the student to obtain specific skill related to the management risk.

The knowledge is institutionalized through the assessment test that validates the intrinsic competence previously obtained.

7.5 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 7.1).

In general the students liked the SG Learning Object and found it interesting to acquire specific concepts as the management of big risks in an emergency situation (G6.5).

The SG has been considered a very educational resource by both instructors and students (G6.3). Indeed as reported by the student's evaluation to the open questionnaire, a lot of students have appreciated the immersive reality of the game (G6.2).

On the contrary, by analyzing the survey given to the four teachers, all of them agree on the use of SG learning resource as improvement of students' understanding of key concepts (G6.4).

They have found very interesting the topic of the game since it is considerable for the curricula profile of the students and helps them by teaching particular topic in a way very near to the students' motivation and emotional status (G6.2).

Despite all, some teachers show difficulties in the monitoring of students' performance and progress. Probably it is due to the difficulty noted in the creation and use of a so complex resource: respect to the first experimentation, the usability and interaction with the game have been ameliorated. In any case for a better experimentation and to have an improved feeling with the game a more powerful PC is required, not usually used in all learning environments (as school, etc.).

All in all, the SG learning resource could be deployed and used successfully within the schools, thus achieving the main goal (G6.1).

8 R7. Affective and Emotional Approaches

The goal of this scenario is to monitor the particular emotion taken by the student during his interaction with the complex learning resources. That in order to modify the learning experience if the emotional state is altered or not compliant with the assessment results.

8.1 Evaluation and Validation Procedure

The purpose of the second experimentation phase is to satisfy all the scenario goals and criteria that are not completely covered in the first phase.

Following we report, as already exposed in [3] the evaluation hypotheses in correspondence of the scenario goals and the metrics for fulfilling specific criteria.

Scenario 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.
- G7.5: to ensure that the system is able to modify the learning experiences according with the detected affective/emotional state.
- G7.6: to ensure that the components of the modified learning experience are relevant to the type of emotion/affection identified.
- G7.7: to identify possible ways to improve the evaluation of the emotional state of the learner and its exploitation to modify a learning experience.

Scenario 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.
- H7.6: the use of system for emotional/affective management contributes to significantly increase students' activity levels.

Scenario criteria

- C7.1: To evaluate the level of fulfilment 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.
- C7.4: To evaluate the level of satisfaction of the instructors with the inclusion of the affective and emotional support in their courses.
- C7.5: To evaluate the increase in students' activity levels due to the affective and emotional support.
- C7.6: To evaluate the increase in students' understanding of concepts and students' results due to the affective and emotional support

Scenario metrics

- M7.1: Number of students requiring affective/emotional support.
- M7.2: Number of courses in which it is required the affective/emotional support.
- M7.3: Number of interventions by the system to provide 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.

8.2 Method

8.2.1 Participants

In order to evaluate the emotional approach and validate it through the effects in the learning process, 4 schools have participated in the experience. In the specific, 4 tutors and 58 students have been enrolled.

For each school the students were allocated into one classroom composed by two groups: experimental and control. The groups use IWT platform, in two different way: the experimental group delivers a learning course by using complex learning resources (as the SLO and SG); the control group delivers a learning course by using traditional learning materials as power point presentations; pdf. file...That in order to compare, through qualitative and quantitative data, the learning process for each group.

All students were supervised by one tutor during the experiment.

8.2.2 Apparatus and Stimuli

The experimental group of each school interact with emotional tool only when his emotional status is not enough to have a positive assessment feedback.

The total students belonging to experimental groups are 29.

On completion of the session they have filled a Post-Questionnaire, which includes the following sections: demographic data, emotional tool facilities, 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. 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).

Regarding the section “Affective/Emotional facilities”, 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. Do you think that the emotional test is very clear and easily understandable?
3. Do you think that the emotional test is representative of your emotional state?
4. Do you think that the emotional/affective state impact greatly on the results of your educational experience?
5. The display of your emotional and affective state leads you to improve your performance levels?
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?
8. Would you like to share your status with only a small group of students selected by you?

The answer categories for this section are: “Not at all”, “A little” , “Moderately” , to “Very”, “Completely” and correspond to 5 points of the Likert’s scale. The rating scale ranged from “Not at all” (1), “A little” (2), “Moderately” (3) to “Very” (4), “Completely” (5).

For usability of the storytelling, we used the SUS (System Usability Scale) which contains 10 items and a 5 points Likert's scale to state the level of agreement or disagreement, for instance "I think that I would like to use this system frequently". 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).

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".

Finally, quantitative data was also collected from IWT database and log files.

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 within IWT. The course has been delivered by two groups of users having different learning styles: experimental and control.

The experimental group has been composed by **analytical** students: a kind of student that likes testing and in a second time to match if the correct solution to a specific problem is correct or not with respect to its hypotheses.

The control group has been composed by **holistic** students: a kind of student that likes to analyze the problem and the associated information before to start a specific activity.

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

After that each student logged in IWT platform, he sees his class.

Within the class, each student deliveries a personalization of the course about the big emergency. In the specific, for the experimental student has been created a personalized learning path by using 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.

At the end of the course, a qualitative survey has been given both the students and the tutor in order to test the knowledge acquired through the storytelling with respect to a passive learning resource.

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). For these purposes we used metrics M7.1 and M7.2.

The survey is a study of the qualitative questionnaires submitted to all the experimental students.

8.3.1 The Emotional Tool Activity

The *Figure 112* shows the average of the students' evaluation with respect to the emotional tool activity.

The Figure includes the average's answers to the eight questions exposed in the Section 8.2.2. and obtained by using a 5 points Likert's scale for analyzing the answers.

The four classrooms involved in the experimentation have been noted with a different colour

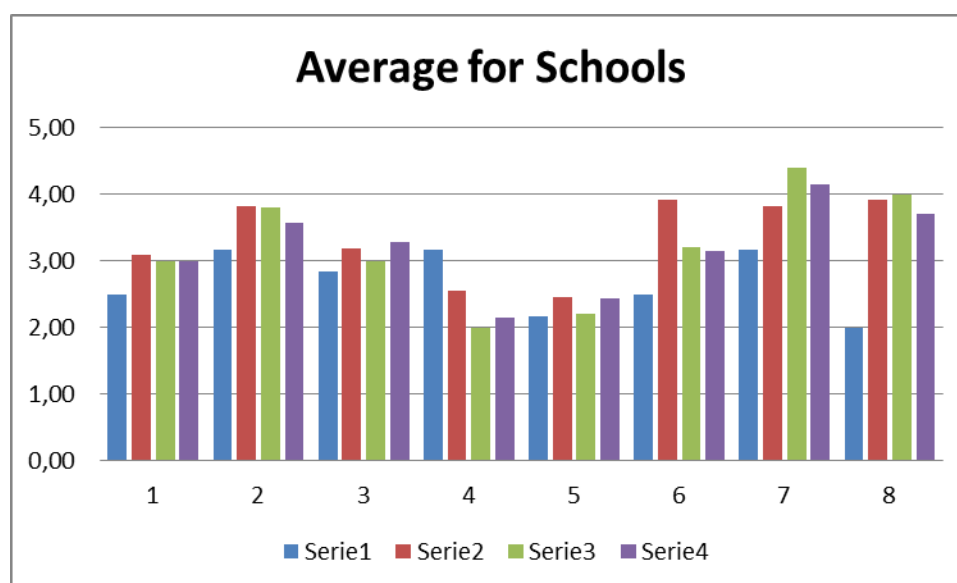


Figure 112: The emotional tool's activity

For the item Q1 "The recognition of your emotional state feels you at the centre of the attention during the learning path?", the average value of a lot of students is settled at about the item 3 of the Likert's Scale, while the student belonging to the first school have registered a lowered average. This means that the student do not consider the emotional approach strictly learner center learning. Students in fact, although appreciated the consideration of

their emotional state do not think that this factor alone could affect the results of experience teaching. This analysis is also confirmed by the answers to the item Q5 “*The display of your emotional and affective state leads you to improve your performance levels?*” that has registered score about the item 2 of the Likert’s Scale.

On the contrary, though the students don’t consider the tool very interesting for their learning path, they think that the functionalities of the tool are more useful for the teacher, for creating personalized learning path. Indeed, the answers to the item Q6” *Do you think that the data collected can be used to provide additional activities useful for recovering the emotional balance?*” register a value equal to the item 2 of the Likert’s Scale. This datum confirms that the tool is more useful for the teacher’ s activity rather than the students one since it allows to arrange corrective activities in order to bring the students into learning functional equilibrium conditions.

Regarding the sharing of own emotional state respect to this one of the other students (see item Q7 “*Do you think that the emotional test should be made visible to the peers in order to trigger a social support?*” and Q8 “*Would you like to share your status with only a small group of students selected by you?*”) all students agree about the possibility to share their emotions with the others because it’s could help them to overcome critical situations.

8.3.2 Usability of the Emotional Tool

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

To investigate the overall usability of the tool, we used the SUS and included it a specific section of the qualitative 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).

The 10 items that composed the SUS questions are:

11. I would use this tool regularly
12. I found it unnecessarily complex
13. It was easy to use
14. I’d need help to use it
15. The various part of the tool worked well together
16. Too much inconsistency
17. I think others would find it easy to use
18. I found it very cumbersome to use
19. I felt very confident using the tool
20. I needed to understand how it worked in order to get going.

The results are summarized in the following picture:

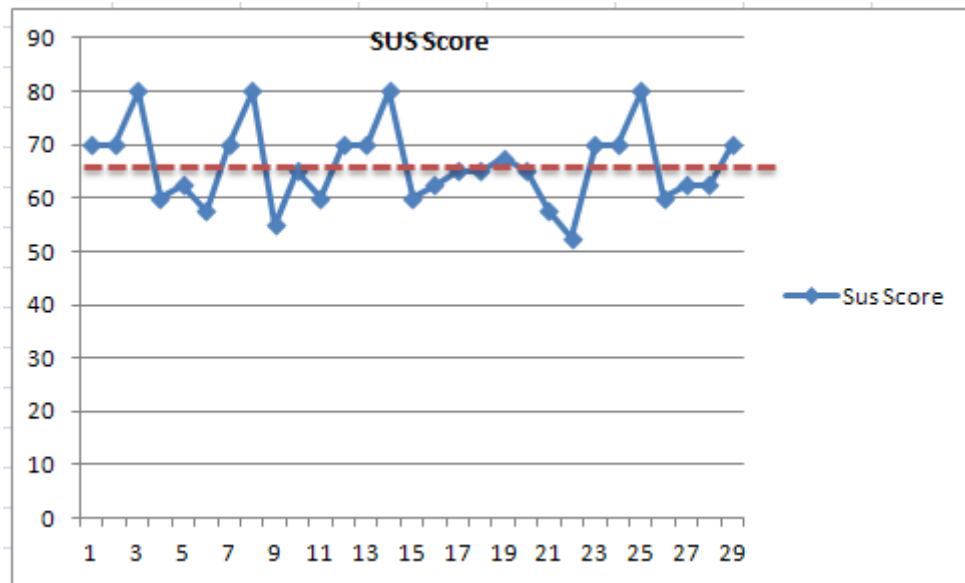


Figure 113: Usability of emotional tool

The **SUS mean score is 66.21**. The minimum score is 52.5 (achieved only 1 time) the maximum score is 80 (see Figure 113).

This is a considerable result that denotes how easily the students have interacted with the emotional tool.

8.3.3 Emotional Aspects

Regarding the students’ emotions during the work with the IWT tool, 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=29) were as follows:

- Happiness (M=1.34, SD=0.55, Md=1) (Figure 114)

Happiness

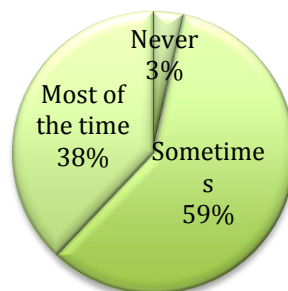


Figure 114: Results on the Happiness emotion

- Sadness (M=1.10, SD=0.62, Md=0) (Figure 115)

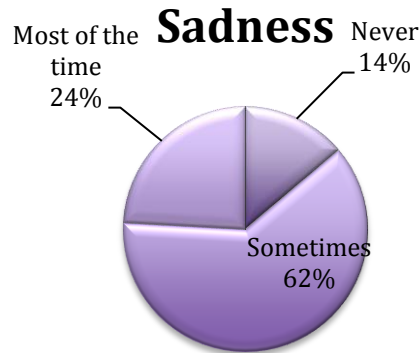


Figure 115: Results on the Sadness emotion

- Anxiety (M=0.48, SD=0.51, Md=0) (Figure 116)

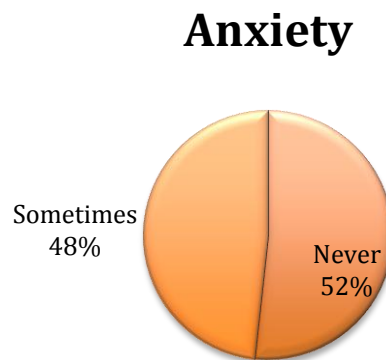


Figure 116: Results on the Anxiety emotion

- Anger (M=0.55, SD=0.57, Md=1) (Figure 117)

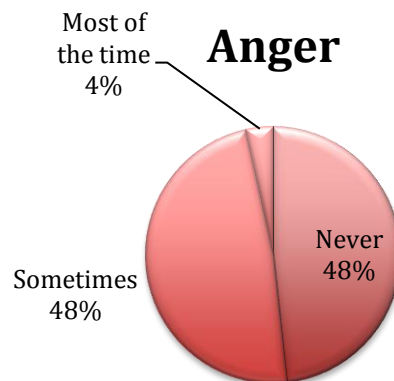


Figure 117: Results on the Anger emotion

Considering the SUS and CES results, obtained in this section and section 8.3.2 respectively, we can affirm that further works should be done in order to ameliorate the qualitative structure of pre and post quantification questionnaires.

8.4 Validation Results

The Figure 118 shows a mapping between the number of students and the access times for each classroom. In general all students have accessed to the emotional tool twice at least.

That confirms the interest of the students respect to the emotional tool in a context of a learning experience. For these purposes we used metrics M7.3 – M7.8.

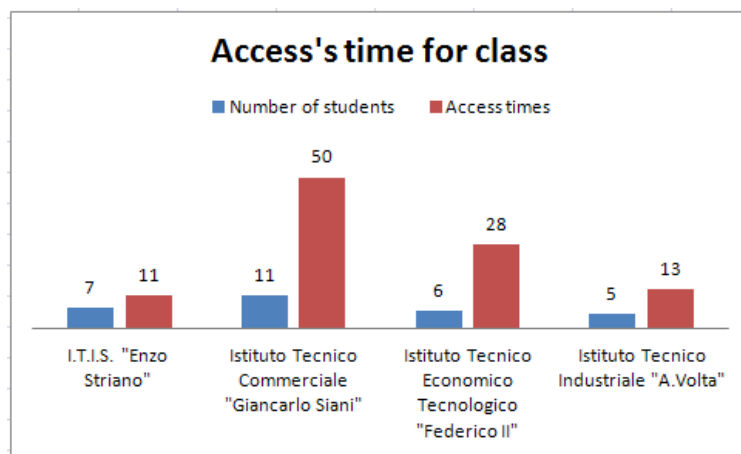


Figure 118: Access's time for class

This result is also confirmed by the Figure 119 that denotes the average number of accesses for each student.

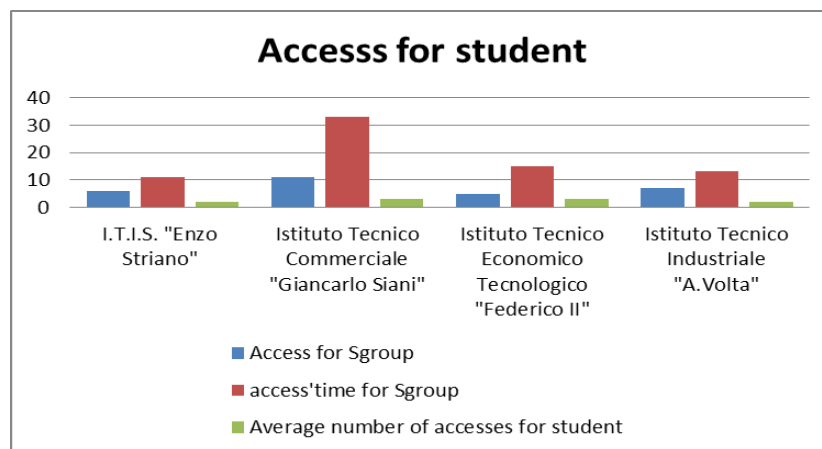


Figure 119: Access for student

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.

The qualitative data showed 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.

With respect to the results obtained from the 1st experimentation phase, in this phase the students have had the possibility to choose if testing or not their affective/emotive status.

From teacher point of view, they have had the possibility to enable or not different parameters in order to ameliorate the personalization of the learning path.

9 R8. 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.

9.1 Evaluation and Validation Procedure

Scenario 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 give 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.

Scenario goals

- H8.1: The tool allows an efficient and user-friendly management.
- H8.2: Using the tool supports the 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 give 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.

Scenario criteria

- C8.1: To evaluate the level of fulfilment of the tool features.
- C8.2: To evaluate the level of satisfaction of the students with the tool regarding functionality.
- C8.3: To evaluate the level of satisfaction of the students with the tool regarding self-, peer, and group assessment activities.
- C8.4: To evaluate the level of satisfaction of the instructors with the tool.
- C8.5: To evaluate the learning outcomes of the students when using the tool.
- C8.6: To evaluate the potential increase in students’ motivation when using the tool.

Scenario metrics

- M8.1: Ratings of students’ satisfaction with the tool.
- M8.2: Ratings of instructors’ satisfaction with the tool.
- M8.3: Ratings of students’ self-assessment activities.
- M8.4: Ratings of students’ peer-assessment activities.
- M8.5: Ratings of students’ motivation while/after using the tool.
- M8.6: Comparison between results from self- and peer assessment.
- M8.7: Ratings of students regarding their learning outcome due to the tool

9.2 General Methodology

In order to test the above listed hypothesis, three experiments were conducted in this second phase of experimentation. After each study, the functionality of the co-Wiki was improved in accordance with the results from the respective experiment. Additionally, we tested the tool in three different settings. Table 12 gives an overview of the studies conducted in phases 1 and 2 of the Alice project, their specific goals, settings, and the achieved improvements.

-

Exp. Phase	Title	Participants	Setting and Goals	Improvements afterwards
Ph I	TUG: HCI Course	$N_{students} = 18$ $N_{instructors} = 3$	<ul style="list-style-type: none"> • Setting: regular course on HCI in mobile learning, Computer Science students • Stand alone system • First test of functionality and usability 	<ul style="list-style-type: none"> • Performance increase for homepage • Performance increase for edit page
Ph II	Curtin University	$N_{students} = 15$ $N_{instructors} = 1$	<ul style="list-style-type: none"> • Setting: controlled environment, business students 	<ul style="list-style-type: none"> • Show authors of peer-group assessments on home page

		<ul style="list-style-type: none"> • Stand alone system • Improved functionality 	<ul style="list-style-type: none"> • Implementation of motivation charts page • Implementation of tagged teacher feedback • Enhancement of rubric control to support criteria weights • Enhancement of revision player to show internal peer-reviews • Implementation of usage pattern recording
TUG: ISR Course	$N_{students} = 23$ $N_{instructors} = 3$	<ul style="list-style-type: none"> • Setting: regular course, home assignments, Computer Science students; • Test improved functions • Record and analyse log data (number of accesses to different pages, access paths, etc.) • Setting: regular course, home assignments, Psychology students • Test improved functions • Test Co-WIKI integrated into IWT 	<ul style="list-style-type: none"> • Implementation of final grades • Implementation of SSO integration for the Wiki • Implementation of page-structure page
KF-University	$N_{students} = 30$ $N_{instructors} = 2$		

Table 12: Overview of the studies testing the co-writing Wiki (R8)

In the following, we give a short overview of the Wiki’s main functionality. For a better readability, the evaluation and validation instruments that are common to all three experiments are also explained in this section. Additional questionnaires are described where applicable.

9.2.1 Co-writing Wiki system

Generally, wikis are websites with an easy-to-use group and knowledge management system to support online collaboration. Functions include the possibility to add, edit, delete, and comment current as well as previous versions of a site, receive and give feedback, and interact with peers. The enhanced co-writing Wiki system for collaborative writing and peer-review has the following features (for a detailed description, see D5.2.2, ALICE, 2012):

- Enhancement of ScrewTurn wiki⁴ to maintain *task and social awareness*.
- *Self-, peer-, and group-assessment* with use of assessment rubrics for grading and feedback.
- *Continuous feedback provision* for learner scaffolding and for teachers to follow the collaboration progress.
- *Visualization tools* to support both students and teachers in knowing who did what and when.

⁴ ScrewTurn Wiki - Free ASP.NET Wiki Software, <http://www.screwturn.eu/> (accessed 14 April 2012)

- *Motivational charts* to motivate peers (groups) to contribute and work in comparison with other group members (other groups).

Figures Figure 120 through Figure 122 show examples of the enhanced functionality provided by the co-writing Wiki.

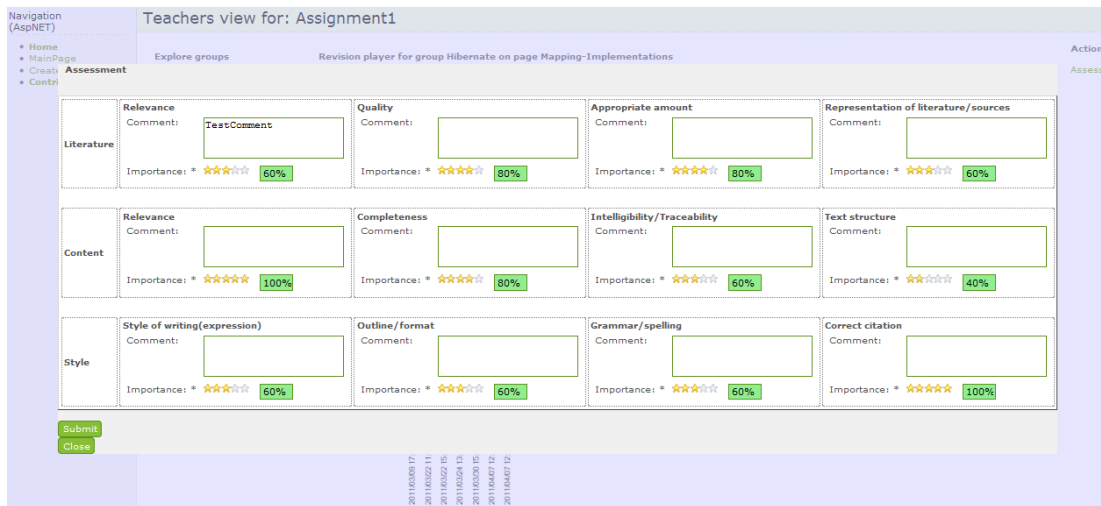


Figure 120. Assessment rubric for grading and feedback

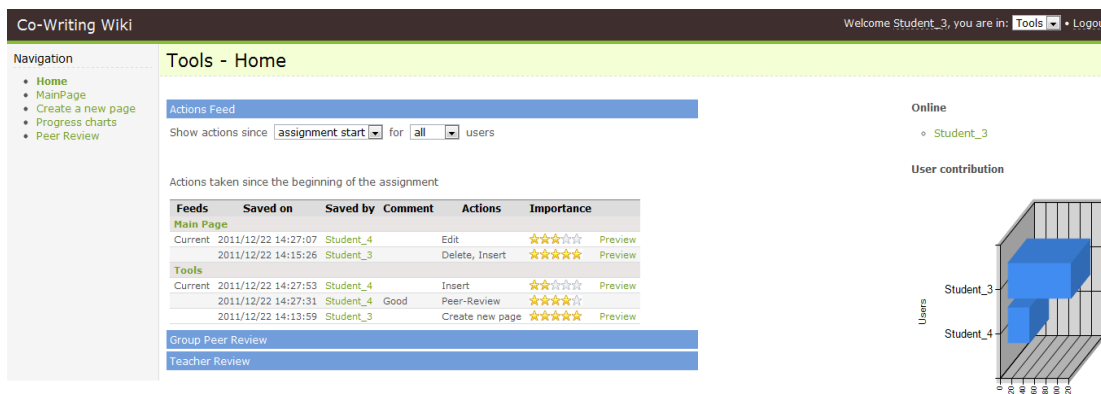


Figure 121. Actions feed and contribution chart in the co-WIKI's assignment homepage

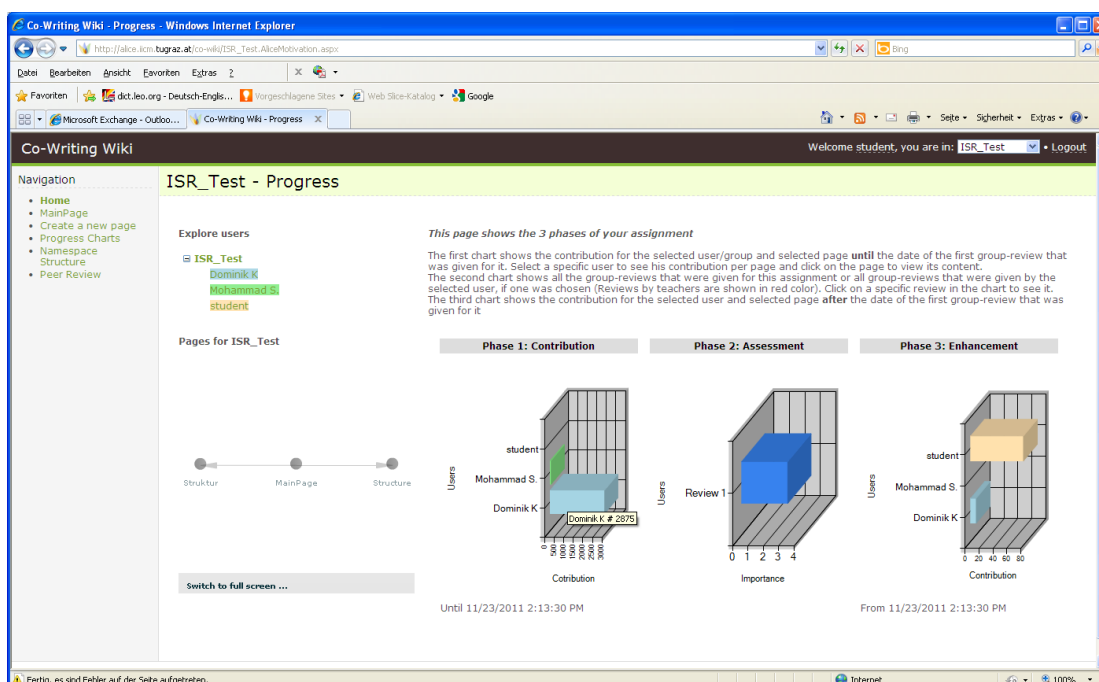


Figure 122. Motivations Charts Page with contribution and assessment graphs

9.2.2 Questionnaires used for evaluation and validation

9.2.2.1 Demographic data

In all three studies the inquired demographic data covered participants' age, gender, and highest level of education. Furthermore, they had to indicate whether the course they were enrolled in was mandatory and whether they knew all of their group members.

9.2.2.2 Previous experience in group working and working with wiki-tools

Participants had to answer 7 (Curtin) and 9 (ISR and KFU) questions regarding their familiarity, previous experience, and attitude with/to collaborative work and wiki-tools. Example items are "How much experience do you have in working with a group face-to-face", "What do you think are the (dis)advantages of collaborative work?", "I have already worked with wiki tools" or "What did you (not) like concerning these tools?"

9.2.2.3 System Usability Scale (SUS; Brooke, 1996)

We used the System Usability Scale (SUS) by Brooke (1996) [8] to investigate the usability of the co-writing Wiki and its integration into the IWT. The SUS is a simple, ten-item attitude scale giving a global view of subjective assessments of usability. Responses are given on a 5-point Likert scale by stating the level of agreement or disagreement (e.g. "I think that I would like to use this system frequently", "I find the system unnecessarily complex", "I feel very confident using the system" or "I needed to learn a lot of things before I could get going with this system"). The SUS 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%).

Additionally three open questions asked what participants liked, disliked, and what they would improve about the system.

9.2.2.4 Task Awareness

To assess task awareness while working with the co-writing Wiki, we developed 10 questions asking how well the different features (actions feed, coloured difference tool, contribution graphs) of the co-Wiki support and motivate students. The items are listed in Table 13. Answers were given on a 5-point Likert scale ranging from “I strongly disagree” to “I strongly agree”.

Items	
The actions feed in the assignment homepage...	
	supports me in tracking the activities of my peers effectively.
	supports me in getting an overview about the actual state of the paper.
	supports me in coordinating the tasks with my group members.
	supports me in directing my effort towards the group product.
Knowing what others are doing motivates me to effectively contribute towards the group product.	
The enhanced colored difference tool in the Co-writing wiki gives me a good overview about the latest changes on the contribution.	
The contribution graphs in the assignment homepage give me a good overview about...	
	who of my colleagues had contributed to the task.
	the amount to which my colleagues had contributed to the task.
	the progress of the other groups.
The contribution graphs motivate me to contribute more to the paper.	

Table 13: Items used for assessing task awareness while working with the co-writing Wiki

9.2.2.5 Computer Emotion scale

To investigate in which emotional mood the students were while they worked with the co-writing Wiki, we used the Computer Emotion scale (CES), which includes 12 items. Kay and Loverock [9] 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”)

Answers are given on a 4-pt. rating scale with the categories “None of the time”, “Some of the time”, “Most of the time” and “All of the time”.

9.2.2.6 Attitudes towards self-and peer assessment

To measure participants’ attitudes regarding the self- and peer-assessments, we used four subscales developed by Tseng and Tsai (2010) [11] to assess the self-efficacy and motivation of learning in online peer-assessment environments. Two scales are taken from the online peer assessment self-efficacy scale (OPASS), two from the motivations in online peer assessment scale (MOPAS). The entire section contains 2 items on self-assessment (“In a self-assessment activity, I can find the strength/weaknesses of my work”) and 17 items on peer-assessment, which are distributed as follows:

MOPAS (motivation):

- The intrinsic motivation scale (5 items) 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 will still learn something even if I get an unsatisfied score on my work”
- The extrinsic motivation scale (4 items) 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 teachers’ opinions are more important than my peers.”

OPASS (self-efficacy):

- The evaluating scale (4 items) measures the confidence of the students in evaluating the peer’s work, e.g. “In a peer-assessment activity I can find the weaknesses of my peers’ work when I review it”
- The receiving scale (4 items) measures students’ ability to constructively use peer assessments, i.e. how well students can handle the peer-assessments in order to recognize their own weaknesses. An example item is “In a peer-assessment activity I can examine the problems in my own work when I get 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).

9.2.2.7 Group assessment

For the group-assessment, we provided assessment rubrics (see Figure 120) 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)

For each category, students should assess the work by means of a rating and a short comment explaining their assessment.

Participants' experiences with the group assessment were measured by asking students to rate the supportiveness and usability of the group assessment rubric on a 5-pt. rating scale as well as by open comments regarding what students (dis)liked about this function.

Table 14 summarizes, which aspects were covered by the different questionnaires presented in each of the three studies.

Study	Questionnaire	Sections						
		Demo-graphic data	group work and wiki exp.	MOPAS/OPASS	Task awareness	SUS	CES	Group assessment
Curtain	Post	√	√	√	√	√		
ISR	Pre	√	√	√				
	Inter				√	√	√	
	Post		√	√	√	√	√	√
KFU	Pre-Qu.	√	√	√				
	Intermediate Qu.				√	√	√	
	Post-Qu.		√	√	√	√	√	√

Table 14. Overview of student-questionnaires provided during the co-WIKI studies

9.3 Study R8.2: Business Course at Curtin University

9.3.1 Method

9.3.1.1 Participants

Fifteen students participated in the course, nine of which are male and six are female. 60% of the students are between 20 and 29 years old, 27% belong to the category 30 to 39 years and just 2 of the students are between 40 and 49 years old. Regarding the main field of study, the majority of the participants stated Master of Commerce (IS) as their main field. One participant stated Postgraduate Diploma in Commerce (IS) and 1 student answered that his main field is Master of ISS. In addition, 9 of them stated that the course was mandatory for them. Almost all of the students said that their group members were known or partially known to them.

9.3.1.2 Apparatus and Stimuli

After working with Co-writing wiki, students were asked to fill in a Post-Questionnaire concerning (i) their demographic data, (ii) attitudes regarding collaborative working and previous experience in working with wiki-tools, (iii) usability of Co-writing wiki, (iv) task awareness, (v) attitudes concerning self- and peer-assessment (based on their experiences), and (vi) their attitudes concerning internal peer-review. For a detailed description of (i) through (v) see section 9.2.2.1 through 9.2.2.4 and 9.2.2.6 respectively. For the assessment of (vi), students rated their agreement (5-pt. scale) to three questions concerning the internal peer-review. The following items were used: “The internal peer-review allowed me to effectively ... (a) rate the importance of my peer’s contribution, (b) comment on my peer’s contribution, and (c) track the latest changes in the paper”.

9.3.1.3 Procedure

In cooperation with a lecturer from Curtin University in Perth, Australia, it was possible to run through this study and to test the collaborative Co-writing wiki as part of a course. As part of a postgraduate unit in the field of Information Technology, students were asked to work with Co-writing wiki in order to write an assignment collaboratively. The experiment took place in a controlled environment, because they were using Co-writing wiki within the lecture, supervised by their professor. Co-writing wiki was available all over this time. Additionally, technical support has been available online via Skype during the lecture.

Additionally to the Post-Questionnaire students were also asked to give informal feedback regarding the usability of the Co-writing wiki. Furthermore, the lecturer provided a feedback about her experiences with Co-writing wiki after the course.

9.3.2 Evaluation Results

In this Section Hypotheses H8.1, H1.2, H8.3, and H8.4 are evaluated by using the following metrics from [3].

- M8.1: Ratings of students’ satisfaction with the tool.
- M8.2: Ratings of instructors’ satisfaction with the tool.

- M8.7: Ratings of students regarding their learning outcome due to the tool.

9.3.2.1 Usability of the Co-writing wiki

For each item, we computed the mean and its standard deviation as an exact measure of central tendency. However, in this section 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 data of this section 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.

According to Knussen and McQueen (2006) [13] 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.

After working with Co-writing wiki, students were asked about its usability. Students stated that Co-writing wiki was written in a clear and simple language ($M = 3.54$, $SD = 0.97$, $Md = 4$) and that it was easy to learn the functions ($M = 3.23$, $SD = 1.17$, $Md = 4$). In addition, students agreed that Co-writing wiki was very effective ($M = 3.31$, $SD = 1.11$, $Md = 4$). Nevertheless students disagreed that Co-writing wiki was very friendly ($M = 2.62$, $SD = 1.12$, $Md = 2$). Besides, it can be assumed that students were not in favor of the presentation, because they denied that the tool was presented in an attractive way ($M = 2.46$, $SD = 0.88$, $Md = 2$).

Furthermore, a student suggested enabling both, adding a new page and uploading an attachment in one session. Students also complained that Co-writing wiki is not compatible with all search engines and that there are too many bugs. In addition, a student suggested adding parallel input function. Another student mentioned that in his/her point of view the background is too simple. One student described Co-writing wiki as unstable and that its interface is not intuitive, especially when trying to import pictures, chart or other graphical information. Besides, a student stated that it was easy for him/her to follow the explanation. However, another student suggested an online manual which explains all functions.

9.3.2.2 Task Awareness

The answers of the students imply that the actions feed in assignment homepage supported them in tracking the activities of their peers effectively ($M = 3.85$, $SD = 0.69$) and in orientating about the actual state of the paper ($M = 4$, $SD = 0.71$).

Moreover, the students stated that these actions also supported them to coordinate with their group members ($M = 4.08$, $SD = 0.76$) and direct their effort towards the group product ($M = 4$, $SD = 0.58$). Additionally, the actions feed in assignment homepage motivated them to

effectively contributed towards the group product and recognizing their peer's contribution ($M = 3.85$, $SD = 0.69$).

The students agreed that the enhanced colored difference tool gave them a good overview about the latest changes on their peer's contributions ($M = 4$, $SD = 0.58$). According to the contribution graphs in the assignment homepage, the results show that these graphs gave almost all students a good overview about who of their peers ($M = 3.46$, $SD = 1.20$) and to which amount their colleagues ($M = 3.54$, $SD = 1.27$) had contributed to their task. Half of the students also agreed that these contribution graphs gave them a good overview about the progress of the other groups ($M = 3.38$, $SD = 0.96$). Besides, they stated that the contribution graphs motivated them to contribute more to the assignment ($M = 3.69$, $SD = 1.03$).

According to additional comments, students mentioned that the contribution graphs don't give accurate information, because even if somebody submits irrelevant contents, it will be considered as a contribution in the graphs. Moreover, the contribution graphs will 100% represent the activities if all contributions which are posted online (also the discussions) are considered. For one of them, it was also unclear how the information for the graphs was calculated. Another student stated that the contribution graphs motivated him/her to improve the speed to state his/her opinions. One student suggested that it would be helpful to have a better organization of pages created by users because the unit required them to create a page for each student, so it was confusing and disorganized to see many pages in the homepage. Finally one student stated that the software could be more user-friendly.

9.3.2.3 Students' informal feedback regarding Co-writing wiki

Students stated that Co-writing wiki enables posting opinions, constructive comments and sharing ideas with group members. Additionally they described Co-writing wiki as a very handy tool that encourages and supports teamwork. Furthermore a student mentioned that the peer and teacher review allowed him/her to understand more. On the one hand students described Co-writing wiki as excellent, clear, brief and precise. On the other hand they also noted several problems such as loss of contents, slowness and crashes. Besides, students complained that Co-writing wiki cannot be used with all browsers. A student also reported difficulties in navigation through the pages and another student suggested a proper manual which explains each feature in detail.

9.3.2.4 Lecturer's feedback regarding Co-writing wiki

The lecturer stated that she was very pleased working with Co-writing wiki and that she liked using the tool within her units. During working with Co-writing wiki she also noted several difficulties and provided some suggestions for improvements. So she suggested for example to release some information regarding the rating, respect to the color and the number of stars. In her point of view the tool should also be able to display the teacher review for a specific page. Adding different color or highlighting pages would support both, teacher and student to know whether a page was rated by the teacher or another student. Besides, she mentioned that under the discussion section spelling is missing and also the format and style should be carried out from one line to another. According to the discussion, it would also be helpful to get a message when a new discussion was added. Moreover, the lecturer argued that more information regarding access and access page should be available. She also faced

difficulties with the compatibility under various browsers, renaming files and uploading images. Additionally she complained about the fact that Co-writing wiki shut down without any warning message.

9.3.3 Validation Results

The following metrics as they are specified in [3] were used for validation:

- M8.3: Ratings of students' self-assessment activities.
- M8.4: Ratings of students' peer-assessment activities.
- M8.5: Ratings of students' motivation while/after using the tool.
- M8.6: Comparison between results from self- and peer assessment.

9.3.3.1 Attitudes concerning working collaboratively

Half of the students like working collaboratively, 36% have a neutral attitude concerning this term and 2 of them stated that they do not like working collaboratively.

First, the students were asked about disadvantages regarding working collaboratively. Some students mentioned that coordinating a group could be time consuming, because of different point of views and the need of tracking and managing changes from peers. Students stated that it could also be difficult to get involvement from all group members, especially when group members do not want to work collaboratively or act in a stubborn way.

Second, the students were asked about advantages concerning collaborative working. Almost all of the students stated that sharing various ideas and different perspectives enrich the work final result. Additionally, the students also mentioned the advantage that collaborative working enhances teamwork.

9.3.3.2 Previous experience with Co-writing wiki

The students stated that they are more or less familiar with the tool and two participants worked with the wiki tool previously in the context of other units at the University. According to their experiences, the students mentioned that they liked using the wiki in order to work collaboratively, but they also complained about the slowness of the tool.

9.3.3.3 Experiences concerning Internal Peer Review

The students stated that the internal peer review allowed them to effectively rate the importance of their peers' contribution ($M = 4.71$, $SD = 0.47$). They also agreed on the fact that the internal peer review allowed them to effectively comment on their peers' contribution ($M = 4.07$, $SD = 0.27$) and track the latest changes in the paper ($M = 4.07$, $SD = 0.47$).

Moreover, students explained that they had learned how and what aspects should be included while assessing their peer's work. Additionally they felt inspired by new ideas or how to see a single topic from the group point of view.

9.3.3.4 Experiences concerning peer- and self-assessment

According to students' motivation during working on the self- and peer-assessment, a comparison of the mean values showed that they were more intrinsically motivated ($M = 3.98$, $SD = 0.37$) than extrinsically ($M = 2.90$, $SD = 0.79$; $t(12) = 4.12$, $p < .01$). So they agreed for instance on gaining more knowledge by discussing their work with peers ($M = 4.46$, $SD = 0.52$). Additionally the students were convinced that they still learn something even if they get an unsatisfied score on their work ($M = 4.15$, $SD = 0.55$).

Regarding students' experiences receiving feedback, the students stated that comments from peers supported them in examining problems in their work ($M = 4.15$, $SD = 0.69$). Moreover, the students could decide whether or not to revise their work after they got peers' feedback ($M = 4$, $SD = 0.71$).

Concerning the term of evaluating, students answered that they could share their opinions or suggestions during reviewing their peers' work ($M = 4.15$, $SD = 0.80$). Furthermore, the students were convinced that they could recognize the strengths ($M = 4$, $SD = 0.82$) and the weaknesses of their peers' work ($M = 3.85$, $SD = 0.38$).

Regarding the self-assessment, students stated that they could find weaknesses ($M = 3.85$, $SD = 0.69$) and strengths ($M = 3.85$, $SD = 0.55$) of their own work.

In additional comments, students emphasized that the self- and peer-assessment supported them in recognizing the plus and minus points from their personal work. According the peer-assessment, students stated that peers tended to be "kind" while evaluating their work.

Regarding their experiences concerning self- and peer-assessment, students also mentioned that Co-writing wiki is a strong tool for sharing and gaining knowledge and that spending more time on working with the tool would have been more useful.

9.3.4 Conclusion

In this Section the main results are summarized and discussed with respect to the goals defined at the beginning of this section. The specific goal of this first study in the second round of experimentation was to test the improved functionality of the co-Wiki in a controlled environment. Main technical improvements of the tool after the study in Phase 1 of experimentation concerned a performance increase for the homepage and for the edit page. Furthermore, the WIKI was tested with a different group of users. Whereas participants in the first round were all computer science students from TUG Graz, this second test of the Wiki was performed with business students from an Australian University.

The first goal G8.1 states, that the developed tool should allow an efficient and user-friendly management. Although the students rather disagreed on statements regarding the user-friendliness and attractiveness of the wiki, they agreed on questions concerning the clarity of the used language and the easiness to learn the functions. Furthermore, the results yield high task awareness scores, indicating that the single components like the actions feed, the enhanced colored difference tool, or the contribution graphs supported them in tracking their own and their peers' activities, in coordinating the work with their group members, or in getting an overview about the latest changes and the amount of contributions per person.

Thus, the data from the questionnaire imply that the features provided in the wiki support students to work collaboratively (G8.2) and that the actions feed and the contribution graphs are perceived as useful functions by the students (G8.4). Students also stated that the actions feed as well as the contribution groups motivated them to contribute to the group product. Thus, also G8.5, the provision of the tool that motivates students in their learning activity could be met.

Open comments from students and teacher also point to several aspects of the tool, which still need improvement (G8.3). As a consequence the group-assessment was changed to be more transparent by showing the authors, a motivation charts page showing the contributions before and after an assessment was added, and the revision player showing the internal peer-reviews was enhanced. Furthermore, a tagged teacher feedback and the recording of usage patterns were implemented.

Regarding students' experiences with the peer assessment, the two MOPAS scales (Tseng & Tsai, 2010) [11] showed that students' intrinsic motivation was higher than their extrinsic motivation after working with the co-WIKI. From the rather high scores on intrinsic motivation and students' open comments we can also infer that the provided peer-assessments motivated the students (G8.5). The achieved scores from the two OPASS scales indicate that students are able to receive feedback and to evaluate peer's work rather well and that the peer-assessment helped them to improve their work, and thus supported their learning process (G8.6).

Finally, with regards to goal G8.7, namely whether the tool facilitates the work for the instructors, the teacher's open comments reported in Section 9.3.2.4 show, that she liked working with the co-Wiki, but also faced various difficulties, part of which could already be improved (see above).

9.4 Study R8.3: Computer Science Course at TUG (ISR)

9.4.1 Method

Students enrolled in the course "Information Search and Retrieval" were asked to use the Co-writing Wiki to communicate and collaborate in writing a scientific paper. Researchers (e.g. Morris, 2005) [14] have emphasized the usefulness of computer log analyses to examine students' online behaviors. Thus, we used log data to get information on students' usage patterns, such as overall working time, number/duration of edits, number of access to different pages, and number of self-, peer-, and group-assessments. For the assessment of satisfaction, motivation, and emotional state online questionnaires were presented. To evaluate the tool's usability, we differentiated between satisfaction, efficiency, and effectiveness as it was suggested by Frøkjær et al. (2000) [15]. The peer- and teacher assessments of the group work were taken as indicator for performance.

9.4.1.1 *Participants*

From 26 students enrolled in the course, 23 gave their consent to participate in the study (18 male, 5 female) by filling out at least one out of three presented questionnaires. Participants' age ranged between 21 and 39 years with an average of 26.46 (SD = 4.52) years. 79% had a Bachelor degree, the remaining participants had either a high-school diploma (8.33%) or a Master degree (12.5%). The course was mandatory for 54.17%, but all students gave their consent to participate in the study. For the collaborative writing assignments, students were assigned to 7 groups with three and one group with four members. One student worked alone. At the beginning of the assignment 62% of the students already knew their group members, 25% knew part of them, and 12.5% did not know any of their group members. Regarding their experience in working collaboratively, 79.16% indicated to have quite much or a lot of experience with face-to-face group work, 45.84% with group work in an online environment. Furthermore, 79.16% like working collaboratively, whereas the remaining 20.84% are not decided. 33.34% of the participant agreed or strongly agreed to having experience with Wiki-tools, whereas 37.34% (strongly) disagreed. The tools most often listed by the students are TWiki and MediaWiki.

9.4.1.2 *Apparatus and Stimuli*

Three online questionnaires were presented to the students. The pre-questionnaire (PreQ) at the beginning of the study contained the sections demographic data, previous experience with group work and wiki-tools, and attitudes towards self- and peer-assessment. The intermediate questionnaire (IntQ) was delivered after students had turned in the first version of their papers. It contained questions on task awareness, usability (SUS), and emotional aspects (CES). The post-questionnaire (PostQ), presented after students had finished their assignments, contained task awareness items, SUS, CES, the self-and peer-assessment scale (reformulated to refer to current experiences made during working with the co-WIKI), experiences with working collaboratively, and experience with the group assessment. See Section 9.2.2 for a detailed description of the used scales.

9.4.1.3 *Procedure*

As part of the course, each student group had to select a topic on information search and retrieval and collaboratively work out a short paper using the co-WIKI. The assignment was divided into four phases, starting with preparing a paper structure and performing a literature search in Phase 1. In Phase 2 students had to work out a first version of the paper and peer-review the work of one other group (group-assessment). After receiving additional feedback from the instructors, the first draft was revised and re-submitted (Phase 3). These final papers were again reviewed by the instructors and after the final feedback presented in class (Phase 4). The three questionnaires were presented via lime-survey over the course of the study. The pre-questionnaire was delivered after the selection of topics at the beginning of Phase 1, the intermediate questionnaire was delivered in Phase 2 after students had turned in the first version of their papers, and at the end of Phase 4, students' were asked to fill out the post-questionnaire.

While working with the co-Wiki, students could rate the importance of their peers' latest contribution (peer-assessment) as well as of their own contribution (self-assessment activity).

While the self-assessment was mandatory after creating or editing a page, the peer-assessment was optional. Peer- and self-assessments were given as short comments and ratings on a five-star scale. The group-assessment at the end of Phase 2 was based on the three assessment rubrics references, content, and formal aspects with two, six, and four subcategories respectively. For each sub-category a short comment-field and a 5-star rating scale was provided (see Figure 120). The same rubrics were used by the instructor to provide the final grade.

Additionally, students had the possibility to continuously monitor the actions of all group members via the actions feeds and the contribution charts on the assignment homepage (Figure 121), check changes from the latest version of the assignment via the difference page, and call the progress or motivation charts page (Figure 122) to view the contributions of each student over the course of an assignment (i.e. contributions per page before/after an assessment and assessment results).

9.4.2 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, H8.2, H8.3, and H8.4 with the corresponding metrics M8.1 as they are specified in [3].

- M8.1: Ratings of students' satisfaction with the tool.
- M8.2: Ratings of instructors' satisfaction with the tool.
- M8.7: Ratings of students regarding their learning outcome due to the tool.

From the 23 participating students, 20 filled out the pre- and 19 the intermediate questionnaire. Regarding the post-questionnaire items on motivational and emotional aspects were answered by 17 participants, SUS items by 18, and task awareness items by 19 participants. The presented log data is based on the behaviour of those 22 students who worked collaboratively. Thus, in the analyses below, sample sizes and degrees of freedom vary.

9.4.2.1 Usability

Corresponding to the findings by Frøkjær et al. (2000) [15] we evaluated usability with regard to three different aspects, namely satisfaction, efficiency, and effectiveness. The measures used to assess the three aspects are also in accordance with Frøkjær et al [15]. Task awareness ratings and SUS scores were taken as indicators for satisfaction, quality of solution in terms of peer- and teacher grades as indicator for effectiveness, and completion time in terms of working and editing time as indicator for efficiency. With respect to the hypotheses stated in Section 9.1, SUS scores and mean task awareness scores refer to H8.1 (the tool allows and efficient and user-friendly management), open comments to the tool's usability to H8.3 (suggestions for improvements), and H8.4 (support of actions and contribution graphs to get overview of learning process) can be answered by sub-questions of the task awareness scale.

9.4.2.2 Usability in terms of satisfaction

Mean task awareness ratings for the 10 presented questions ranged between $M=2.42$ ($SD_{inter} = 1.346$, $SD_{post} = 1.216$) and $M= 3,26$ ($SD_{inter/post} = 1,327$) per item for both the intermediate and the post-questionnaire. The overall mean score for task awareness (i.e. across all items) was $M = 2.88$ ($SD = .963$) for IntQ and $M = 2.85$ ($SD = .931$) for PostQ. The mean SUS scores of 44.11 ($SD = 20.696$) for the intermediate and 41.17 ($SD = 23.595$) for the post-questionnaire also indicate an average usability of the tool (SUS scores have a range between 0 and 100; Brooke, 1996) as far as the aspect of satisfaction is concerned. Related t-tests showed that both task awareness and SUS-scores did not change significantly over the course of the study. See Table 15 for the details. Thus, for H8.1, it can be stated that the tool allows an efficient and user-friendly management on a medium level. Further improvements of the tool from a students' point of view (see H8.3) should mainly concern the following aspects: more intuitive navigation menu, performance increase, skip/shorten mandatory self-assessments, implementation of bibtex, integration of latex, html-code instead of wiki-markup. Regarding H8.4, a closer look at the respective items in the task awareness scale of the PostQ (IntQ) showed that 54.54 (31.58)% agreed or strongly agreed that the actions feed supported them in getting an overview about the actual state of the paper, whereas 40.9/27.27% (52.63/42,11)% agreed or strongly agreed that the contribution graphs gave them a good overview about who/how much was contributed.

9.4.2.3 Usability in terms of efficiency and effectiveness

Regarding effectiveness, the collaborative work with the co-Wiki lead to good grades given by peers as well as teachers. Out of possible 100% students graded their peers' first version of the assignment with a mean of 84.45 ($SD = 7.29$) and teachers gave an average of 90.91 ($SD = 9.03$) after students revised the first version according to the received feedback. As far as efficiency is concerned, the working and editing times and their variability indicate that students used the co-Wiki in very different ways. Average working time (WT) was 1040 minutes ($SD = 752$) and the average editing time (ET) was 246 min with $SD = 305$ (here and in the following, ET refers to all successful edits; mean and SD for successful and unsuccessful edits are 288.61 and 376.53 min respectively). When relating the three usability indicators satisfaction, efficiency, and effectiveness we found a significant correlation between effectiveness (as indicated by peer grades) and efficiency (as indicated by working time with $\rho = .547$, $p < .01$ and editing time with $\rho = .515$, $p < .05$), whereas the correlations with satisfaction (as indicated by SUS and task awareness) range between $\rho = -.394$ and $\rho = .327$, with all $p > .05$. Grades given by the teacher were also unrelated to all other variables. See Table 16 for the exact results derived from correlating the three usability aspects.

9.4.2.4 Usage patterns

For the analysis of students' usage patterns, we logged number and time of the typical actions performed when working with the co-WIKI. *Figure 123* shows the mean number of logins, how often students edited or created a page (edits and creates), how often they accessed the assignment home page, difference page, and motivational charts page, as well as the mean number of self-, peer-, and group-reviews. The results indicate that students mainly used the Wiki to edit a page, i.e. to work on the assignment. The strong differences in

the number of self-, peer-, and group-reviews (in the given order $M_{SR} = 45.86$, $SD_{SR} = 26.36$; $M_{PR} = .09$, $SD_{PR} = .294$; $M_{GR} = 5.36$, $SD_{GR} = 5.9$) can be explained by the fact that self reviews were mandatory after saving a page and students were explicitly asked to review one of their peers' work (group-review), whereas peer-reviews were optional.

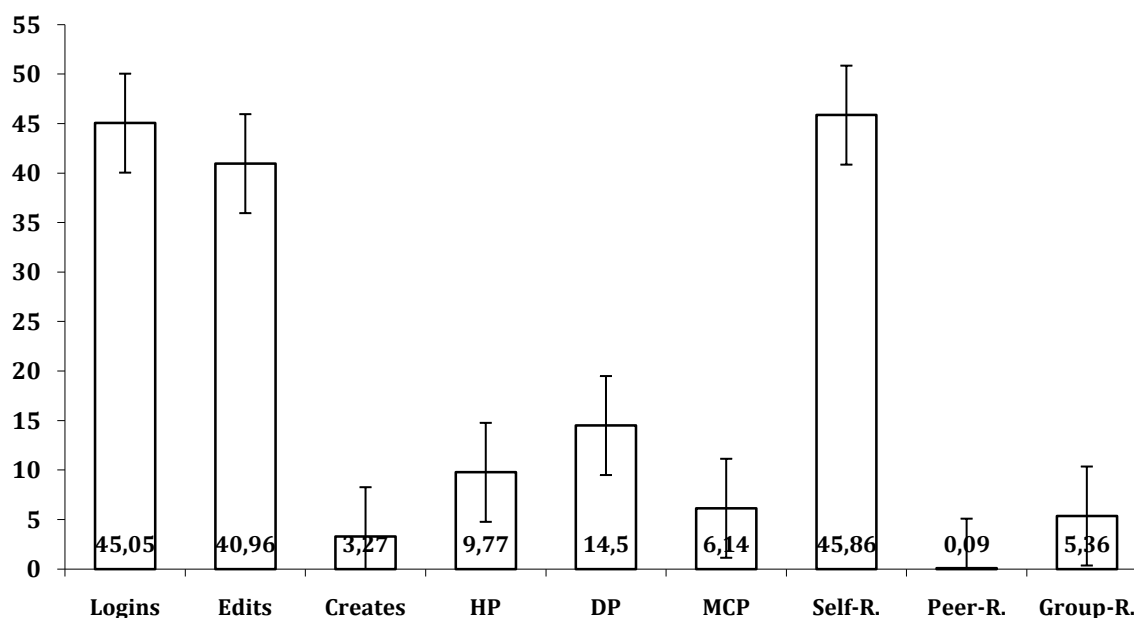


Figure 123. Mean number of various actions performed during the assignment; HP = assignment home page, DP= difference page, MCP=Motivation Charts page, R.=Review.

Regarding the number of times students used the different features of the WIKI, the log data show that besides editing, the difference page was visited most often ($M = 14.5$, $SD = 14.72$), followed by the assignment homepage ($M = 9.77$, $SD = 9.85$). Generally it should be noted that the standard deviations are extremely high indicating a wide range of usage behavior. For example, the number of accesses to the difference page ranged from 1 to 59, that for the assignment homepage from 0 to 46. Also the motivation charts page with a mean of $M = 6.14$ ($SD=4.39$) was viewed between 0 and 16 times. Similar data were found for the number and time of edits ($M = 40.96$, $SD = 25.88$ and $M = 244.36$ min, $SD = 312.07$ min) as well as the overall working time ($M = 1049.53$ min, $SD = 768.07$ min). Thus, some students spent only 2.86 hours in the WIKI (editing text for only 23.42 min), whereas others spent up to 63.36 hours (with a maximum edit time of 24.69 hours).

Another aspect concerning the students' usage patterns is how and in which order students entered the different pages. After logging in to the WIKI, on average students most often started to edit a page (37.31%, $SD = 14.57$), followed by starting a group review (17.76%, $SD = 10.63$) or going to the homepage (13.47%, $SD = 14.09$). Sometimes, they also went to the motivation charts page (9.73%, $SD = 9.64$) or created a new page (5.39%, $SD = 5.82$) right after logging in (the reported percentages are corrected for logins that were directly

followed by a logout, which happened in 46.53% of all logins). Having a closer look at the editing function, which was used most often by students (on average they spent 21% [$SD = 12.74$] of the overall working time on editing), the usage patterns show that 80.72% ($SD = 15.17$) of the cases students completed their edit successfully by going to the end edit page function. Before editing, students usually visited either the difference page ($M = 30.42\%$, $SD = 19.8$) or just logged in to the WIKI ($M = 22.7\%$ $SD = 8.9$). Regarding the paths right before students left the WIKI, the data show that students mostly logged out after editing a page ($M = 40.61\%$ $SD = 14.6$), starting a group review ($M = 10.93\%$ $SD = 6.38$), or visiting the homepage ($M = 10.13\%$ $SD = 8.04$).

9.4.2.5 Group work and group assessment

Results concerning the collaborative working aspect and group assessment of the assignment are taken as indicators for H8.2 and H8.6.

Generally, students stated, that they liked working collaboratively ($M = 3.59$, $SD = 0.94$; $Md = 4$ on 5-pt. scale). With respect to H8.2, namely that the tool supports the students in working collaboratively, they stated that they liked that the work is shared, that it is possible to work on different aspects at the same time, to receive more inputs on the topic and to get faster feedback. Disadvantages were only seen, when one of the group member did not collaborate fairly. This aspect was mentioned by two members of the same groups.

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 effectively supported the students to learn more about other groups' topics ($M = 3.24$, $SD = 0.97$, $Md = 3$). However the students neither agreed nor disagreed on the statements, that the provided assessment rubric supported them in reviewing the product of other groups ($M = 2.71$, $SD = 1.16$, $Md = 3$) and that it was easy to use ($M = 2.71$, $SD = 1.26$, $Md = 3$).

Open comments on the group assessment revealed, that the students liked getting in touch with another topic, seeing how other groups solved the assignment and to learn from that, and they liked using new technologies. Thus, for H8.6, students confirmed, that the group assessment supported them in their learning process. However, some students did not like the categories used in the rubric, some disliked the interface and the pre-structuring of the peer review form and that sometimes information was incomplete and could therefore not be assessed properly.

9.4.2.6 Emotional aspects

Students emotional status, indicated on a 4-pt. rating scale, ranged from $M = 1.58$ ($SD = .69$) for anxiety to $M = 2.32$ ($SD = 1.06$) for anger, both in the IntQ. To check for differences regarding participants' emotional states, one-way ANOVAs with repeated measures were performed for the four types of emotions covered in the intermediate and post-questionnaire. With $F_{(1.469, 23.509)} = 1.209$ and $p = .303$ we found no effect for the post-questionnaire, but a small effect for the intermediate questionnaire, $F_{(1.734, 31.215)} = 3.719$, $p = .041$, $\eta^2 = .171$. Related t-tests performed with Bonferroni correction revealed significant differences for sadness vs. anger ($t = -3.284$, $df = 18$, $p = .004$) and anxiety vs. anger ($t = -3.986$, $df = 18$, $p = .001$).

Regarding eventual changes over the course of the study, we found no differences for motivational and emotional aspects except for the increase in extrinsic motivation. Table 15 summarizes the results from related t-tests for the different questionnaire sections.

Scale	Questionnaire	M	SD	t	df	p
Intrinsic Motivation	Pre-Qu.	3.57	.514	2.11	13	0.055
	Post-Qu.	3.21	.426			
Extrinsic Motivation	Pre-Qu.	2.71	.825	-2.28	13	0.040*
	Post-Qu.	3.29	.611			
Receiving	Pre-Qu.	3.64	.497	0.00	13	1.000
	Post-Qu.	3.64	.497			
Evaluating	Pre-Qu.	3.71	.469	1.39	13	.189
	Post-Qu.	3.50	.519			
Happiness	Interm. Qu.	2.00	.707	.898	12	.387
	Post-Qu.	1.77	.725			
Sadness	Interm. Qu.	1.69	.947	-.56	12	.584
	Post-Qu.	1.85	1.068			
Anxiety	Interm. Qu.	1.54	.660	-1.0	12	.337
	Post-Qu.	1.85	.899			
Anger	Interm. Qu.	2.15	1.068	-.185	12	.856
	Post-Qu.	2.23	1.166			
Task awareness	Interm. Qu.	2.77	1.006	-.194	14	.849
	Post-Qu.	2.85	.960			
SUS	Interm. Qu.	46,14	22.156	.349	13	.732
	Post-Qu.	43.07	23.734			

Table 15. Results of paired samples t-tests (pre/intermediate vs. post-questionnaire) for motivational and emotional aspects, task awareness and usability

9.4.2.7 Feedback from tutors

In order to evaluate, whether the tool facilitates the work of instructors (H8.7), we also asked the two tutors (and the instructor) of the course to evaluate the assessment forms and functions for tracking students' contributions provided in the WIKI. The following results are based on the ratings given by the two tutors on 5-pt. Likert scales ranging from (1) I strongly disagree to (5) I strongly agree. Regarding the assessment forms, both tutors strongly agreed (5) that the group-assessment was helpful for evaluating students contributions and agreed (4) with regard to the self- and peer assessment. They also found the provided rubric for the group-assessment to be appropriate (5) and the rate control stars to be helpful (4) in assessing the students' contributions.

With regard to the various functions provided in the WIKI, they found the actions feed in the assignment homepage supportive for knowing about the state of the paper (4), knowing about the progress (4), and providing feedback for the groups (3, 4). The contribution graphs were rated to give a good overview about who had contributed to the task (4, 5) and the progress of the groups (3, 4). The revision player supported the tutors in knowing about the progress of the final product (3, 5), in knowing who and what was contributed (both 4 and 5), and in assessing the groups' final product (3, 4). Finally, the tutors stated that the charts in the contribution tool supported them in knowing about the works progress (4), knowing who contributed (3, 4), and when a student contributed (2, 4).

Open comments by the tutors just pointed to one problem, namely the performance of the group-contribution chart on the home page. Thus, with respect to H8.7 the results clearly indicate that the tool supports the work of instructors.

9.4.3 Validation Results

In this Section the main results regarding the pedagogical and psychological perspective of the tool are reported, namely data gathered with respect to the motivational status of the users. Furthermore, relationships between the logged usage patterns and the data received from the questionnaires are analysed. From the metrics specified in [3], the following are relevant for validation:

- M8.3: Ratings of students' self-assessment activities.
- M8.4: Ratings of students' peer-assessment activities.
- M8.5: Ratings of students' motivation while/after using the tool.
- M8.6: Comparison between results from self- and peer assessment.

9.4.3.1 Attitudes and experiences concerning self- and peer-assessment

Figure 124 shows the mean ratings from the three questionnaires for motivational and emotional aspects. Ratings from the post-questionnaire are compared to those from either pre- or intermediate one to show eventual changes during the learning process. Results from the MOPAS and OPASS (attitudes towards peer-assessment, Tseng & Tsai, 2010 [11]) are also indicators for H8.5 and H8.6 (motivation and support from peer-assessments).

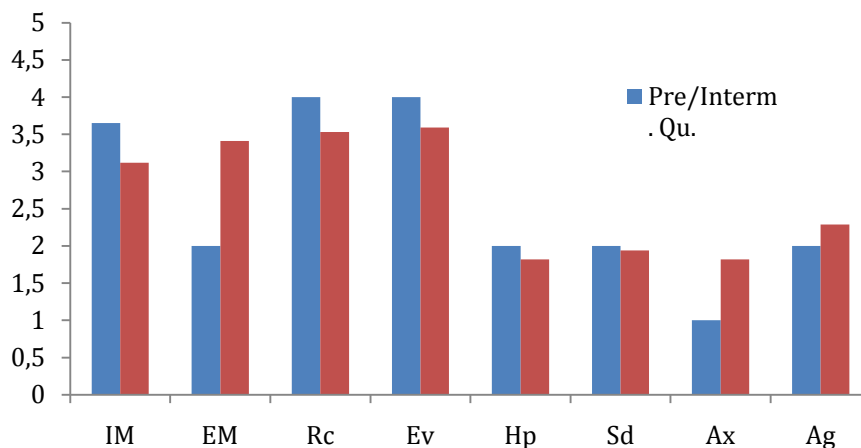


Figure 124. Mean ratings for motivation (5pt. Likert scales), and emotion (4pt. scale); IM-intrinsic motivation, EM-extrinsic motivation, Rc-receiving, Ev-evaluating, Hp-Happiness, Sd-Sadness, Ax-Anxiety, Ag-Anger.

Students' ratings on 5-pt. Likert scales regarding their attitudes towards self- and peer-assessments (Tseng & Tsai, 2010 [11]) showed that at the beginning of the study peer-assessments were motivated more intrinsically than extrinsically ($M_{intr} = 3.65$, $SD_{intr} = .489$, $M_{extr} = 2.65$, $SD_{extr} = .813$, $t=4.156$, $df=19$, $p=.001$), whereas we found no difference in motivation after the course ($M_{intr} = 3.12$, $SD_{intr} = .485$, $M_{extr} = 3.41$, $SD_{extr} = .618$, $t=-1.429$, $df=16$, $p=.172$). Also the scores on the nine single questions concerning motivation revealed only one median below 3 ($M = 2.41$, $SD = 0.71$; "I felt that I have learned nothing, if I got a low peer score on my work"), which indicates that the peer-assessments rather motivated the students concerning their learning activity (H8.5).

Results from the scales for receiving and evaluating indicate that students are able to handle peer-assessments to recognize their own weaknesses (PreQ: $M_{Rc} = 3.7$, $SD_{Rc} = .47$) and that they are confident in evaluating their peer's work (PreQ: $M_{Ev} = 3.8$, $SD_{Ev} = .52$). These findings did not change significantly over the course of the study (see Table 15). Thus, the feedback provided by the peer- and group-assessments support the students in their learning process (H8.6) by helping them to recognize their own strengths and weaknesses. For H8.5 (peer-assessments motivate students concerning their learning activity), the respective items in the postQ show that

9.4.3.2 Relationships between motivational-emotional aspects and usage patterns

Considering the increase in extrinsic motivation over the course of the study and the highly varying usage patterns we also looked for possible relationships among behavioural and questionnaire data. Table 16 summarizes the correlations between measures derived from the questionnaires as well as from the log data. Because of the 4 and 5 pt. rating scales used to measure motivational and emotional aspects and the great variability in the behavioural data, we applied Spearman's Rho coefficient. Considering only log data, Pearson's product-moment correlation coefficient reveals similar results.

	WT	ET	MC	SR	PG	Sus	Hp	Sd	Ax	Ag	IM
<i>N</i>	22	22	22	22	22	19	19	19	19	19	17
<i>M</i>	1050	244	6.14	45.86	84.45	44.11	1.84	1.74	1.58	2.32	3.12
<i>SD</i>	768	312	4.39	26.36	7.29	20.7	.688	.872	.692	1.06	.485
WT	-	.627** (22)	.376 (22)	.697** (22)	.547** (22)	.147 (19)	.359 (19)	.030 (19)	.305 (19)	-.045 (19)	.140 (16)
ET		-	-.075 (22)	.724** (22)	.515* (22)	-.083 (19)	.093 (19)	-.088 (19)	.098 (19)	.016 (19)	-.049 (16)
MC			-	.097 (22)	.223 (22)	-.360 (19)	-.086 (19)	.533* (19)	.362 (19)	.226 (19)	.528* (16)
SR				-	.404 (22)	.098 (19)	.336 (19)	-.043 (19)	.182 (19)	-.090 (19)	-.118 (16)
PG					-	-.394 (19)	-.016 (19)	.024 (19)	.209 (19)	.160 (19)	-.164 (16)
SUS						-	.554* (19)	-.512* (19)	-.426 (19)	-.612** (19)	-.418 (13)
Hp							-	-.172 (19)	.166 (19)	-.308 (19)	-.537 (13)
Sd								-	.639** (19)	.677** (19)	.728** (13)
Ax									-	.678** (19)	.494 (13)
Ag										-	.622* (13)

Table 16. Means, SD, and Spearman’s ρ (*N*) for behavioral and self-report measures

Note. * $p < 0.05$; ** $p < 0.01$. WT-working time, ET-edit time (minutes); MC-motivation charts page, SR-self-reviews (frequencies); PG-peer-grade (out of 100); Hp-happiness, Sd-sadness, Ax-anxiety, Ag-anger (4pt. scales), IM-intrinsic motivation (5pt.scales); Emotional and SUS ratings are from intermediate, motivational ratings from post-questionnaire. Note. Correlations were also calculated for group reviews, teacher grades, task awareness, and extrinsic motivation. For all $p > 0.05$.

With respect to motivation in self- and peer-assessment activities, we found that higher scores on the intrinsic motivation (IM) scale in the post-questionnaire are associated with higher numbers of calls of the motivation charts page (which shows the amount of contributions and assessment results), as well as with higher ratings for sadness and anger. Having a closer look at the emotions, results show that anger, anxiety, and sadness are positively interrelated, whereas happiness is independent. Significant correlations of SUS-scores with happiness ($\rho = .554$, $p < .05$), sadness ($\rho = .512$, $p < .05$), and anger ($\rho = .612$, $p < .01$) indicate that participants’ emotional state is also closely related to their satisfaction with the tool.

9.4.4 Conclusion

All students successfully finished their assignment using the co-writing WIKI, thus the goal to provide a WIKI system for collaborative writing assignment (G8.2) could definitely be met. The goal to provide a tool for an efficient and user-friendly management (G8.1) could only be reached in part. The two measures for usability in terms of satisfaction, mean task awareness scores and SUS scores are both below average. However, more than half of the students agreed that the tools' features supported them in getting an overview of their learning progress (G8.4). With respect to G8.3 (identification of possible improvements), a closer look at the single questions and open comments shows that students found the system to respond too slowly, and some features to be too complex. Also a redesign of some graphs was suggested. Since also the log data show that most students didn't access the assignment homepage and the motivation charts page very often, improvements of the tool should primarily concern these aspects. However, considering that the tool is still in developmental status, the overall results are very promising.

With respect to self- and peer-assessment attitudes, emotional aspects, and usability perception over the course of the collaborative writing assignment, the results can be summarized and interpreted as follows. For all three aspects, we compared ratings from two different phases of the study, and except for extrinsic motivation students' data did not change. Extrinsic motivation increased during the collaborative work, which means that external rewards, such as grades became more important at the end of the course. Considering that the teacher assessed the work and gave his final grade at the end of phase 4, this result is not surprising. Overall, the motivational scales revealed values towards agreement in both questionnaires, indicating that students had positive attitudes towards self- and peer-assessment at the beginning and end of their assignment (G8.5, G8.6). However, in spite of their positive attitudes towards peer-assessments, students did not use the respective function very often. Results from the Computer Emotions Scale (means range between 1.63 and 2.29) on the other hand indicate that students' emotions were not very strong during work with the co-Wiki.

Another goal of this work was to provide a tool that facilitates the work of teachers (G8.7). The results of the questionnaires filled in by the two participating tutors clearly show that the functions of the WIKI support teachers in tracking and assessing students' contributions.

Another goal of this study was to examine the relationships between the before mentioned variables, usage patterns, and performance. Morris et al. (2005) [14] found participation to be a significant factor for achievement and pointed out that persistence is important for motivation. In our study, students with higher intrinsic motivation used the provided feature of the motivation charts page more often and those with higher participation (longer working and editing time) achieved higher peer-grades. Thus, finding further ways to foster intrinsic motivation and to motivate students to increase their participation (in form of contributions, but also discussions and assessments) seems to be one important factor for future developments. Our findings also show that high SUS scores go along with happiness, whereas low scores are related to sadness and anger. Thus, another focus of future research needs to be on the enhancement of usability in terms of satisfaction.

9.5 Study R8.4: Psychology Course at Graz University (KFU)

9.5.1 Method

This section deals with the evaluation of the wiki KFU study. The systematic approach to the evaluation of the WIKI tool was the same as in the ISR study. The tests that were used in the ISR study were also used in this study. In addition to that, some new tests and questions were added. Additionally, this is the first study with the Wiki fully integrated into the IWT. The integration in the IWT was the subject of some newly added questions.

9.5.1.1 Participants

Subjects were 30 Psychology students of the Karl-Franzens-University in Graz, Austria. They used the Wiki to complete an assignment within a course. While the course itself was not mandatory, the course was for a mandatory module, i.e. students can choose among several courses, but have to complete one. Participation in the Wiki experiment was voluntary.

The following data was raised mainly during the Pre-Questionnaire, with the Inventory for Learning Strategies in academic Studies (LIST) stemming from the Intermediate Questionnaire. Of the 30 students, 5 were male and 25 were female. They were between 19 and 31 years old ($M = 22.0$, $SD = 2.74$). For 26 of them, Matura, the Austrian high-school diploma (qualification for university entrance), was the highest level of education; two had acquired a Bachelor degree and another two a Master's degree. At the point of the Pre-Questionnaire, the students had already formed groups. 13 of them stated that they knew all the members of their respective groups, 13 knew the members of their group partially and four did not know them at all. When asked, 19 of the students stated that they had "some" experience in working with a group face-to-face. Three had "not much" and "a lot", respectively, while eight had "quite much" experience. The students were less experienced when it came to working with a group in an online environment. 14 chose "not much", six "some" and only one "quite much" while nine students said that they had no experience at all ("none"). The next question concerned whether the students liked working collaboratively. The answers were given on a 5-point-Likert-scale from "I strongly disagree" (1) to "I strongly agree" (5). 18 students agreed, one agreed strongly and 11 opted for the middle category "neither/nor" ($M = 3.67$, $SD = .55$). The students were then asked what the disadvantages and advantages of collaborative work are. As disadvantages, the students listed the difficulties that come with coordinating with a group, potential malicious or lazy peers, longer and more difficult decision processes and negative group dynamics like holding back ones opinion for the sake of acceptance by peers. As advantages, the students listed the coming together of different perspectives and ideas, shared work effort, social aspects (meeting people, working together) which according to some students increase motivation and fun, practicing soft skills and that working together with others can be motivating in itself. Next up, the students were asked regarding their previous experience with Wiki tools. Only three students had worked with Wikis before, those wikis being Wikipedia, wikidot.com and an unspecified regional Wiki. What the students liked regarding these Wiki tools was the sharing of information with others regardless of time and place, the structuring of the information on

Wiki pages and working and creating together with others. The students stated they did not like the “boring layout” or design in general of those tools.

With the LIST in the Intermediate Questionnaire, we investigated the learning strategies regarding the collaborative writing task. On average, the students showed values around the middle category in both the Learning with Peers subscale ($M = 3.1$, $SD = .62$) and the Meta-Cognitive Strategies subscale ($M = 3.5$, $SD = .56$).

9.5.1.2 Apparatus and Stimuli

Three online questionnaires were presented to the students via LimeSurvey over the course of the study. See Section 9.2.2 for a detailed description of the used scales. The pre-questionnaire (PreQ) was presented at the beginning of the study, before the students started working with the Wiki. It contained the sections demographic data, previous experience with group work and wiki-tools, attitudes towards self- and peer-assessment and the Motivated Strategies for Learning Questionnaire (MSLQ, see below). The intermediate questionnaire (IntQ) was delivered after students had turned in the first version of their papers and had received feedback from the course instructor. It contained questions on task awareness, usability (SUS), emotional aspects (CES), and the Inventory for Learning Strategies in academic Studies (LIST, see below). The post-questionnaire (PostQ), presented after students had finished their assignments and the Group-Assessment, contained task awareness items, SUS, CES, the self-and peer-assessment scale (reformulated to refer to current experiences made during working with the co-WIKI), experiences with working collaboratively, and experience with the group assessment, the students’ self-reported usage behaviour with the Wiki, questions regarding the usability of IWT and motivational aspects (MSLQ).

In addition to the MOPAS and OPASS (see Section 9.2.2.6), participants’ motivation regarding goal orientation and task values as well as their learning styles were inquired. For motivational aspects, we used three subscales from the *Motivated Strategies for Learning Questionnaire* (MSLQ) by Pintrich et al. (1991) [10], for learning styles two scales from the “Inventory for Learning Strategies in academic Studies” (LIST) by Wild et al. (1994) [16] were presented.

Motivated Strategies for Learning Questionnaire (MSLQ). Pintrich et al. (1991) [10]

To investigate motivational aspects, three subscales of the MSLQ were presented to the students in the pre- and post-questionnaire:

- 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 ranging from (1) I strongly disagree to (5) I strongly agree.

Inventory for Learning Strategies in academic Studies (LIST), Wild et al. (1994) [16]

The LIST is a standardized questionnaire containing 11 subscales, two of which we used to investigate students' learning strategies regarding the collaborative writing task. Since the original LIST is in German, the following two scales were translated into English:

- Meta-cognitive strategies:

The scale has 11 items concerning the three aspects planning, monitoring and regulating which focus on the self-regulation of current learning processes. Examples are "I think in detail about which parts of a topic I have to learn and which ones I don't" or "I do additional exercises to test if I really understood the topic".

- Learning with peers:

This scale measures the degree of collaborative learning. It includes seven items regarding different forms of collaborative tasks as well as forms of one-sided demands on peers. An example item is "I take time to discuss course matter with my fellow students" or "I compare the notes I took in class with those of my fellow students".

9.5.1.3 Procedure

The students were tasked with writing a scientific paper. In this assignment they had to write an "exercise paper", consisting of the two parts method and results. For this purpose the 30 students were split into nine groups consisting of three to four people each. These groups were formed before the Pre-Questionnaire was answered and before any writing happened in or outside the Co-writing Wiki. After answering the Pre-Questionnaire, the students started working on a first version of their assignment in the Wiki, while being supervised and familiarised with the Wiki by their tutor. Two weeks later, they received feedback from the teacher on their completed first version. Immediately after this, we asked them to fill out the Intermediate Questionnaire. Now the students had to complete the final version of their work and submit it. After that, they were asked to perform the Group-Assessment. Each student had to assess two groups which he was not a part of. After they had finished the Group-Assessment and their assignment, we sent them the Post-Questionnaire.

Analogous to the ISR-study, the Self-Assessment when editing a page was mandatory, while the Peer-Assessment was not. In this study, the rubrics for the Group-Assessment were

defined by the instructor of the course. The instructor defined the three rubrics method, results, and formal aspects with four subcategories each.

This is also the first study with the Wiki fully integrated into the IWT system. This allowed the tutor to set up the student accounts and assign students into groups via the IWT at the beginning of the study. To evaluate this process, he was asked to complete a Post Authoring Questionnaire.

9.5.2 Evaluation Results

Following the Study R8.3 (see section 9.4) 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, H8.2, H8.3, H8.4, H8.5 and H8.7 with the corresponding criteria and metric as they are specified in [3].

- M8.1: Ratings of students' satisfaction with the tool
- M8.2: Ratings of instructors' satisfaction with the tool.
- M8.7: Ratings of students regarding their learning outcome due to the tool.

All 30 students completed the Pre-Questionnaire, 26 also completed the Intermediate Questionnaire and 23 students completed all three questionnaires. The presented log data is based on the behaviour of all 30 students. Thus, in the analyses below, sample sizes and degrees of freedom vary. Since the intermediate questionnaire was erroneously not personalized, the results cannot be traced back to individual students and thus comparisons between intermediate and post-questionnaire are based on unrelated samples.

9.5.2.1 Usability of the IWT integrated WIKI-tool

In contrast to the previous Wiki-studies, in this study participants connected to the WIKI via the IWT system. The following results refer to the IWT integrated WIKI-tool. Sixteen of 26 participants agreed or strongly agreed that the access to IWT always worked (mean= 3.62, SD= 1.20, median= 4). Less than one third stated that they have faced problems with the IWT-system. Participants, who answered positively on the question if they faced any problems, could specify their problems: 4 of the participants reported they had been disconnected from the system. Two of the participants criticised the layout of the IWT display as confusing. To evaluate student's satisfaction with the WIKI-tool regarding an efficient and user-friendly management (H8.1), we analyzed students' ratings and open comments on the usability/functionality of the tool. Therefore we collected the mean SUS-scores and task awareness ratings for all participants for intermediate and post questionnaire.

9.5.2.2 System Usability scale (SUS)

Mean SUS scores were 42.31 (SD= 18.93) for the intermediate questionnaire and 37.39 (SD= 14.47) for the post-questionnaire. However there was no significant change in usability rating over the course of the study ($t_{(47)} = 1.011, p = .317$). Eighteen of 26 usability ratings in

the intermediate questionnaire where below 51 (bottom 15% of norm sample) whereas in the post-questionnaire 20 of 23 students reported a usability value lower than 51.

Reasons for this low usability score seem to be grounded in technical problems, as 17 out of 21 participants agreed or strongly agreed on facing technical problems while working with the WIKI. Eleven complained about missing clarity of the layout and display errors of various WIKI page elements, four of the students described the wiki as slow and two were facing problems while saving their contributions. Six participants claimed to be disconnected from the system out of no reason. Comparing this answer to the behavioural data we tracked for each subject while using the wiki, we assume, that most of the system kick outs happened due to auto-logout. This happens every 20 minutes a user does not executes a traceable action in the WIKI-system. On average students were automatically logged out 15.57 times ($SD=9.74$. Range 34) during their activity in the WIKI-system.

To get additional details about the implications of the auto logout function we conducted relative auto logout per hour for each participant and defined working consistency as overall working time divided by the number of logins. We found a correlation between relative auto logout per hour and overall edit times ($r= -.423, p= .044$), working consistency ($r= -.436, p= .038$) and happiness ($r= -.508, p= .013$). This again reflects participants troubling with the auto logout function. Students which were facing many auto logouts per working time did less edits, were more inconsistent while editing and less happy (see Table 17).

Apart from the technical problems, students liked the features of the WIKI. Five students commented that they liked the online collaborative aspect and eight the features that allowed tracking the various changes of their group members. Five reported liking being able to see the progress of the collaborative work and the amount of contribution by their group members.

Regarding H8.1 the reported SUS-scale ratings and participant’s feedback indicate a level of usability below average for the WIKI-system. In comparison to the ISR-usability results it can be concluded that there has been no improvement of the usability of the WIKI-system measured by the SUS-Scale. This could partly be an effect of the integration in the IWT-system and the new functionalities that were tested for the first time (see Table 17).

In terms of H8.3 students were asked to give suggestions of improvement of the WIKI-tool. They recommended making the interface clearer and more user-friendly. They mentioned there should be a better explanation to the functions of the system and some wished to have more formatting options. As mentioned above students were facing system logouts. They stated there should be no auto-logout and recommended to make the system faster.

	Hp	Sd	Ax	Ag	Wc	SUS	OET	CET	A/h	TA
<i>N</i>	23	23	23	23	23	23	23	23	23	23
<i>M</i>	1.80	1.80	1.70	2.20	58.87	44.1	78.93	77.01	3.15	3.47
<i>SD</i>	1.70	2.20	2.00	2.60	101.37	20.7	82.44	81.17	1.91	.82
Hp	-	-.278	.003	-.208	.169	.261	.187	.202	-.508*	.400
Sd		-	.637**	.780**	-.053	-.631**	-.003	.003	-.017	-.497*

Ax	-	.671**	.003	-.630**	-.172	-.162	-.159	-.376
Ag		-	-.070	-.741**	.214	.209	-.304	-.287
Wc			-	.010	.257	.250	-.436*	.288
SUS				-	.098	.096	.139	.362
OET					-	.997**	-.423*	.406
CET						-	-.409	.421*
A/h							-	-.360
TA								-

Table 17. Means, SD, and Spearman’s ρ for behavioral and self-report measures

Note. * $p < 0.05$; ** $p < 0.01$. Hp-happiness, Sd-sadness, Ax-anxiety, Ag-anger (4pt. scales), Wc-working consistency (time per login) in minutes; SUS-System usability scale; OET-Overall edit time in minutes; CET-Correct edit time in minutes; A/h-relative Auto logout (per hour); TA-Task awareness. All metrics are taken from post-questionnaire.

9.5.2.3 Task awareness

Another indicator of usability (H8.1) in terms of satisfaction is task awareness, which was used to measure satisfaction with the single functions provided within the Wiki. The overall mean score for task awareness measured by 10 questions were $M=3.47$ ($SD_{post}= 0.82$) for the post questionnaire and $M= 3.46$ ($SD_{inter} = 0.61$) for the intermediate questionnaire. Thus, there was no significant change over the course of the study (see also Table 18).

Nine of the 10 task awareness questions had a median of 4 which indicates a high average agreement. For example stated the students that the enhanced coloured difference tool in the Co-writing wiki gave them a good overview about the latest changes on the contribution (mean rating of 4.00 with $SD= 1.00$ on a scale from 1-5). They further found the actions feeds in the assignment homepage supported them in getting an overview about the actual state of the paper (mean rating of 3.96 with $SD= 1.11$). Regarding the contribution graphs students reported they did not get an overview about the progress of the other groups (mean rating of 2.48 with $SD= 1.24$). This low rating can partly be explained by technical problems as some students reported the contribution graphs could not be displayed correctly. However 12 students reported contribution graphs gave them an overview about the colleague’s amount of contribution (mean rating of 3.22, $SD= 1.28$, median= 4) and 13 agreed or strongly agreed that they motivated them to contribute more (mean 3.35, $SD= 1.27$, median= 4). It can be stated that in terms of H8.4 the contribution graphs for the most part fulfil useful functions. For H8.5 results show students are motivated through the features of the WIKI-system.

Whereas participants’ SUS scores did not improve from previous experiments, resulting task awareness scores are higher in this study, which indicates an improvement of the tool’s functionality.

9.5.2.4 Usability in terms of efficiency

Average overall working time per student was 6:07:21 hours ($SD=4:00:30$). Participants’ working times ranged from 31 minutes to 14 hours. The average amount of edit time was

1:16:05 hours ($SD= 1:18:27$, Range= 5:46:31 hours) that is 18.52% ($SD= 12.82\%$) of overall time spent in the WIKI. To get an idea if students used the WIKI editing function in a correct way we had a look at the relative amount of finished edits on overall edits. On average participants saved their edits by going to the end edit page in 95.40% ($SD= 9.09\%$) of their overall editing cases.

Further analysis of the behavioural data showed a correlation between time spent editing and usability ratings in terms of task awareness: Students who gave good ratings at task awareness also had more overall edit times ($r=.406$, $p=.055$). This correlation was not significant on the .05 level, but tends towards significance. There was also a correlation between task awareness and amount of correct editing times ($r=.42$, $p<.05$). No correlation was found between task awareness and overall working time. Hence it can be concluded if participants used the WIKI-system in a correct way (editing via path *edit* -> *End-edit*, see descriptive for usage patterns below), they gave better task awareness ratings on average. With respect to H8.1, these results indicate that the WIKI-tool provides efficient user management, if the users interact with the WIKI in the appropriate manner.

9.5.2.5 General usage patterns

To get an idea of how participants used the functions of the WIKI we collected behavioural data for all possible actions in the WIKI as we did for ISR-study. See Figure 125 for the average amount of actions in terms of page views per student ($N=30$).

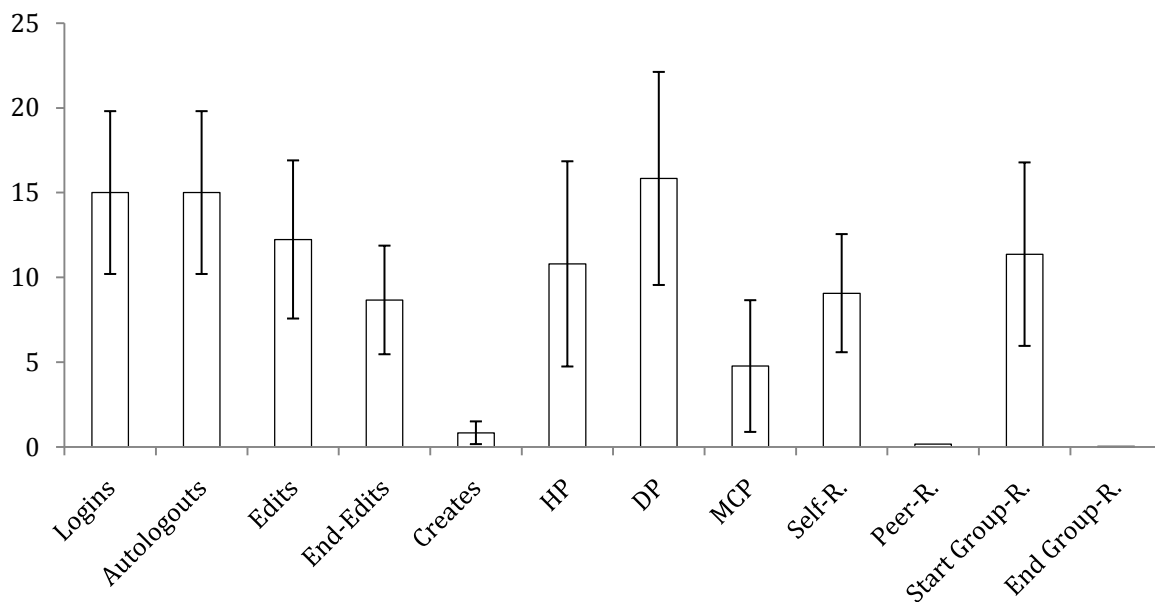


Figure 125. Mean number of various actions performed during the assignment; HP = assignment home page, DP= difference page, MCP=Motivation Charts page, R.=Review.

Logins and auto-logout both average at 9.61 per participant ($SD=4.80$, Range= 34). Thus, participants almost never used the manual logout. On average a student went to the Edit

page 12.23 times ($SD= 9.33$) whereas only 8.67 ($SD= 6.41$) edits were successfully ended via the End-Edit page. With an average 14.85% ($SD=7.05\%$) of total webpage views the difference page made up the highest portion of overall actions. As group review being mandatory every participant visited the Start Group-Review page. On average students started 11.37 group reviews ($SD= 10.82$, Range= 53). Apart from one person, nobody finished their group review by going to the End Group-Review page. This reflects the technical problems students mentioned. Students were being logged out of the system during group assessment if it took more than 20 minutes due to the auto logout function. Because of missing feedback or introduction participants were not aware of this mechanism. Additionally they mentioned the display of the group assessment page was unclear. Taking these reasons into account it can be explained why just one person was able to successfully finish the group assessment. Participants on average spend 27.38% ($SD= 14.34\%$) on group reviews although their effort was not being saved. This may be one major reason for the high amount of frustration participants experienced.

9.5.2.6 *Experience during the Group-Assessment*

In this experiment, the group assessment was mandatory for all students. Each student had to assess two groups he or she was not a part of. To investigate H8.6 (whether the feedback provided by the group-assessment supports the students in their learning process) we asked the students for feedback regarding the group-assessment. The students generally found the assessment of their group by peers helpful ($M = 3.9$, $SD = .6$) and agreed that reading other students' work helped them in understanding more about the topic ($M = 3.9$, $SD = .79$). When asked what they liked about the group-assessment, the students listed getting an insight into the work of other groups at all, seeing how they solved the assignment, and – via the feedback of peers - getting new ideas and being able to learn from one's mistakes to improve oneself. However, as is highlighted by the generally high agreement to “I faced technical problems while doing the group assessment” ($M = 4.2$, $SD = .94$), the students had problems with the Group-Assessment. Some students stated that they had to do the Group-Assessment multiple times because they were forcibly logged out when they tried to submit it. In the free text feedback, the students also reported display errors like an overlapping graphic hindering the group-assessment, the submit button being invisible due to the assessment window being too small, problems when opening the files of other groups, not receiving feedback whether or not the Group-Assessment had been successfully submitted, as well as not being able to find the peers' assessment of one's own work. These problems could be solved before most students finished their group-feedback, however it still influenced students' perception of the tool's functionality. Consequently, most students found the assessment rubrics for the group-assessment “neither/nor” easy to use ($M = 3.1$, $SD = 1.39$). Most students, however, agreed that these rubrics supported them in effectively reviewing the products of the other groups ($M = 3.7$, $SD = 1.03$) and in learning more about the other groups' work ($M = 4.0$, $SD = .93$).

9.5.2.7 *Emotional Aspects*

To investigate another aspect of H8.1, i.e. student' satisfaction with the tool, we analyzed student's emotions during working with the wiki-tool. For this we used the Computer Emotion

Scale (see Section 9.2.2.5). For getting an idea about differences between amount of experienced emotions and potential change of emotions over the course of the study we conducted an ANOVA with the within factor *emotion* (containing the 4 factor levels *happiness*, *sadness*, *anxiety* and *anger*) and the between subject factor *questionnaire* with the two levels *intermediate* and *post*. Number of participants for the intermediate questionnaire was 26, whereas the post questionnaire contained 23 students in total.

As the ratings on the Computer Emotion Scale can be interpreted as nonparametric, we also conducted equivalent nonparametric analysis for emotion. Specifically the Friedman-Test was used to replace the ANOVA and Wilcoxon-Tests to replace the related *t*-tests (for pairwise comparisons of the four emotions). Due to comparable results we do not report them here.

The within subjects effect emotions was significant on the .001 level ($F_{(2,12, 98.92)} = 14.588$). On average students felt more often anger than happiness ($p < 0.001$), sadness ($p < 0.001$), and anxiety ($p < 0.001$). These findings are partly in line with the ISR-results. These differences in experienced emotions during work with the WIKI-system may be caused by technical problems the students reported (see usability section).

Moreover anger changed significantly over the course of the study (see also Table 18) as participants on average rated their anger at 2.2 ($SD = .57$) in the intermediate questionnaire and at 2.6 ($SD = .90$) in the post questionnaire ($t_{(47)} = -2.184$, $p = .034$). Main ratings on the sadness subscale also increased significantly ($t_{(47)} = -2.168$, $p = .035$). These results are in line with the decrease of usability rating reported above.

Scale	Questionnaire	M	SD	t	df	p
Intrinsic Motivation	Pre-Qu.	3.77	.57	1.202	22	.242
	Post-Qu.	3.53	.58			
Extrinsic Motivation	Pre-Qu.	2.76	.44	-1.239	22	.228
	Post-Qu.	2.88	.60			
Receiving	Pre-Qu.	3.6	.63	.629	22	.536
	Post-Qu.	3.42	.47			
Evaluating	Pre-Qu.	3.6	.54	-.895	22	.380
	Post-Qu.	3.75	.35			
Happiness	Interm. Qu.	1.80	.70	.659	47	.513
	Post-Qu.	1.70	.45			
Sadness	Interm. Qu.	1.80	.61	-2.168	47	.035
	Post-Qu.	2.20	.81			
Anxiety	Interm. Qu.	1.70	.55	-1.521	47	.135
	Post-Qu.	2.00	.67			

Anger	Interm. Qu.	2.2	.57	-2.184	47	.034
	Post-Qu.	2.6	.90			
Task awareness	Interm. Qu.	3.47	.82	-.037	47	.971
	Post-Qu.	3.46	.61			
SUS	Interm. Qu.	42.31	18.93	1.01	47	.317
	Post-Qu.	37.39	14.47			

Table 18. Results of paired samples t-tests (pre/intermediate vs. post-questionnaire) for motivational and emotional aspects, task awareness and usability

A closer look at the subscales of the Computer Emotion Scale shows that 10 of 23 participants stated they were frustrated most of the time and 4 of 23 all of the time while using the WIKI-system. 17 of 23 and 21 of 23 reported they felt helpless/ angry at least sometimes. These results are in line with low usability ratings discussed earlier. With regards to H8.1 it can be concluded that participants experienced more negative than positive emotions while working with the WIKI-tool. One main reason for that seems to be the instable system mentioned by the students (see section usability). As their work on the wiki was to be graded afterwards the problems they encountered led to frustration and distress.

9.5.2.8 Instructors' feedback

In this experiment, the setting up of student accounts for the Wiki/IWT and the assignment of students into groups was done by the tutor (a psychology student from a higher semester) using the IWT. After these tasks were performed, the tutor was asked to complete a Post Authoring Questionnaire. He agreed that he would like to use the tool frequently, that the tool was easy to use, that the functions in the tool were well integrated and that he felt very confident using the tool. He could further imagine that most people would learn to use the tool very quickly. He disagreed with the notion, that he would need the support of a technical person to be able to use the tool and disagreed with the statement, that he needed to learn a lot before he could get going with the tool. He further strongly disagreed that the tool was unnecessarily complex and that there was too much inconsistency in the tool. In the Computer Emotion Scale, the tutor scored 1 out of 4 on the Sadness, Anxiety and Anger subscales, and a 3 on the Happiness subscale.

After the end of the experiment, the teacher and the tutor were asked to answer a Post Questionnaire, very similar to that of the ISR study (see Section 9.4.2.4) but with added IWT questions. The teacher and the tutor both received the same Questionnaire, but were asked not to answer questions which they could not answer (for instance the tutor had no part in grading the students). The teacher reported that she did not consult the self- and peer-assessment before doing the group review. Consequently, she found the self- and peer-assessment forms neither helpful nor unhelpful for grading. She did however strongly agree that the Group Assessment form was helpful for grading purposes. The teacher found the rubrics for Group Assessment to be a helpful feature for assessing students' contributions ("I strongly agree"). Note that the rubrics in this study were for the first time defined by the course instructor herself. In this experiment, the tutor performed the task of setting up the

rubrics. He agreed that the corresponding function in the IWT was easy to use. The teacher strongly agreed that using the rate control was very helpful to assess the student's level of mastery based on the rubric criteria. When asked for possible improvements concerning the rubrics, she asked for a zoom-in function for the students' free text fields to be able to see the whole comment at once. The teacher strongly agreed that the concepts were helpful to assess the students' contributions while the tutor was undecided ("neither/nor") on whether the tool for the extraction of the concepts was easy to use. Following the results of the Post Authoring Questionnaire, the tutor strongly agreed that the tool for assigning students to groups was easy to use. Both the tutor and the teacher agreed strongly, that the assistance they received while working with the IWT was adequate. The teacher found the information received through the group assessment helpful for grading ("I strongly agree") and was certain, that the Wiki-system can support students in doing exercises in groups. The teacher agreed strongly, that the actions feed in assignment homepage supported her in tracking the activities of the students effectively, in knowing about the progress and in providing feedback for the group. She further agreed, that the actions feed supported her in knowing more about the state of the paper. The tutor and the teacher both stated, that they tracked the students' activities using the Wiki "some of the time". The teacher gave the following suggestions for improvements: an "are you sure?" popup to prevent students from accidentally deleting content and a way to find student content easily if the students did not link additional pages to the main page. The teacher reported display problems with the contribution graphs. Consequently, she only used the contribution graphs "some of the time" and opted for "neither/nor" regarding whether the contribution graphs gave her a good overview about who of the group members had contributed to the task and the progress of the other groups. The teacher further reported, that the Revision Player did not work at all and thus disagreed with the usefulness of this feature to know about the progress of the final product, know who contributed, know what a student contributed and to assess the group final product. The teacher agreed that the charts in the Contribution tool had supported her in knowing about the progress of the final product and knowing when a student contributed, however disagreed that they allowed her to know who contributed. She reported that the charts contribution only worked sometimes and was rather slow and shared her observation that the speed seemed to be browser-dependant. Next, the tutor and the teacher were asked to evaluate the Co-writing Wiki and the IWT separately on the System Usability Scale. While the Wiki scored 57.5 points from the tutor, the result for the Wiki by the teacher is 60 points. The IWT got 62.5 points from the tutor and 60 points from the teacher. At the end of the questionnaire, the teacher and the tutor were asked to state what they liked and disliked about the Wiki, as well as their suggestions for improvements. The teacher highlighted especially the possibility for the students to perform a group assessment in an easy way and tracking the status of the documents and providing feedback. She stated that she did not like that some features didn't work or worked unreliably, that the Wiki window was too small and the general slowness of the system. Her suggestion for improvement was adding a "request for peer-review" function to the Wiki that would allow students to explicitly ask their peers for a review of their work, for instance when they have made an important contribution and want their group members to look over it.

So regarding H8.7 it can be said that in spite of some technical problems, the tool did facilitate the work for the instructor. We have also identified possible improvements to the tool from the instructor's point of view (see H8.3).

9.5.3 Validation Results

Following the methodology in Section 9.1 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. From the metrics specified in [3], the following are relevant for validation:

- M8.3: Ratings of students' self-assessment activities.
- M8.4: Ratings of students' peer-assessment activities.
- M8.5: Ratings of students' motivation while/after using the tool.
- M8.6: Comparison between results from self- and peer assessment.
- M8.7: Ratings of students regarding their learning outcome due to the tool.

As discussed at the beginning of Section 9.3.2, 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.

9.5.3.1 Students' self-reported usage of the wiki

To better understand how the WIKI can support students learning process and what parts of the wiki can influence student' motivation, in the Post-Questionnaire we asked the students to make some statements or estimates regarding their usage behavior with the Wiki. These estimates were then compared to actual behavioral data, whenever both were available. The 23 students that completed the Post-Questionnaires were asked whether they liked working collaboratively during the Wiki assignment. The responses ($M = 3.7$, $SD = .7$) were on average not statistically significantly different from the same question in the Pre-Questionnaire, suggesting that the students' expectations towards the collaborative work were neither undercut nor exceeded ($t_{(22)} = -.624$, $p = .539$). On average, the students agreed that they did work collaboratively with their group using the wiki ($M = 3.7$, $SD = 1.07$, $Md = 4$) and outside of the Wiki ($M = 4.1$, $SD = 1.14$), like via e-mails or meeting face-to-face. The statement "The Wiki made my work easier" was met with little agreement. Only 3 people agreed while 17 people disagreed or disagreed strongly ($M = 2.1$, $SD = .97$). This result correlates negatively with the Anger subscale of the Computer Emotion Scale ($r_{(21)} = -.5935$, $p < .01$), suggesting a pronounced emotional component in answering this question. This result is also reflected in the negative undertone of the free text feedback we received in the Post-Questionnaire. The students were also asked to estimate the time they spent working on the assignment within the wiki and outside of the wiki. The estimates ranged between 1 and 10 hours for time spent inside the Wiki ($M = 3.67$, $SD = 2.26$) and 0 to 6 hours spent

outside of the Co-writing Wiki ($M = 2.41$, $SD = 1.7$). Compared with actual behavioural log data, it shows that, on average, the students underestimate the time they spent in the Wiki by 31 minutes, with the highest overestimation at 3 hours, 10 minutes (time estimated more than actually spent) and the highest underestimation at almost 6 hours less than actual time spent. This general underestimation-trend still holds true if we eliminate the three people who agreed that the Wiki made their work easier, from the equation, in which case the estimation is on average 29 minutes lower than actual time spent. Next, the students were asked if they wrote their texts inside the wiki or if they wrote them outside (for instance in Word) and copied them into the Wiki later. Except for two students, a pattern could be observed: students who agreed to having written their texts inside the wiki disagreed with having written them outside, and the other way around. This way, 10 students stated that they wrote their texts inside the wiki, while 11 stated that they wrote them outside and copied them inside. The remaining two students agreed to both question-statements.

9.5.3.2 Attitudes and experiences concerning self- and peer-assessment

The general attitude of the students regarding self- and peer-assessment was tested in the Pre-Questionnaire. Their experience concerning self- and peer-assessment while working with the co-wiki was tested in the Post-Questionnaire. In the Pre-Questionnaire we also asked the students, how experienced they were with both self- and peer-assessment. In response to “I already have a lot of experience in peer-assessment” 12 students chose the middle category “neither/nor” while 11 disagreed on the 5-point-Likert-scale ($M = 2.7$, $SD = .92$). When asked the same question but regarding experience in self-assessment, the answers were more evenly distributed amongst “I agree” (11), “I disagree” (10) and “neither/nor” (8) ($M = 3.1$, $SD = .92$). Generally speaking, the students were more experienced in self-assessment than in peer-assessment ($t_{(29)} = -3.525$, $p < .05$). The students also agreed that a self-assessment activity can help them find the strengths ($M = 3.83$, $SD = .59$, $Md = 4$) and weaknesses ($M = 3.6$, $SD = .81$, $Md = 4$) of their work. In the MOPAS (for description see Section 9.2.2) the 30 students scored an average of 3.79 ($SD = .57$) on the Intrinsic scale, while they scored an average of 2.76 on the Extrinsic subscale ($SD = .44$). This makes them significantly more intrinsically than extrinsically motivated ($t_{(29)} = 7.323$, $p < .05$). On the OPASS Evaluating subscale the students scored an average of $M = 3.6$ ($SD = .54$) and on the Receiving subscale $M = 3.53$ ($SD = .63$).

To investigate if the feedback provided by the peer-assessment supports the students in their learning process, the same questions were asked again in the Post-Questionnaire, concerning their experience with the (self- and) peer-assessment, instead of general attitudes. Note that in this experiment the self-review was optional and was generally used very little. The results for the different subscales of the OPASS and MOPAS are fairly similar to the Pre-Questionnaire. After the experiment, the students were still significantly more intrinsically ($M = 3.53$, $SD = 0.58$) than extrinsically ($M = 2.88$, $SD = .6$) motivated ($t(22) = 3.424$, $p < .05$). The OPASS Evaluating and Receiving subscales attained similar results as in the Pre-Questionnaire ($M_{Evaluating} = 3.75$, $SD_{Evaluating} = .35$; $M_{Receiving} = 3.42$, $SD_{Receiving} = .47$). There were no significant changes when comparing the same subscales in the Pre- and Post-Questionnaire (all $p \geq .05$). We asked the students again if a self-reflection activity support them in findings the strengths ($M = 3.35$, $SD = 1.19$, $Md = 4$) and weaknesses ($M =$

3.26, $SD = 1.21$, $Md = 4$) of their work. The change in the strength-item from the Pre- to the Post-Questionnaire was significant ($t_{(22)} = 2.228$, $p < .05$). Students were statistically less likely to agree to it in the Post-Questionnaire. We also asked the students regarding the self-reflection activity. They neither disagreed nor agreed that the self-reflection activity support them in effectively reflecting on their work ($M = 3.22$, $SD = 1$, $Md = 3$) and in indicating the importance of their contribution ($M = 3$, $SD = 1$, $Md = 3$). They did however on average agree that it supported them in providing feedback about their contribution intentions ($M = 3.42$, $SD = 1.08$, $Md = 4$). Regarding the internal peer-review, the students agreed that it allowed them to effectively rate the importance of their peers' contributions ($M = 3.43$, $SD = 1.08$, $Md = 4$), comment on their peers' contributions ($M = 3.91$, $SD = .9$, $Md = 4$) and track the latest changes of the paper ($M = 3.65$, $SD = .88$, $Md = 4$). An item that especially stood out during the statistical analysis was "In a peer-assessment activity I understood some ideas better by discussing them with my peers" ($M = 3.52$, $SD = 1.2$, $Md = 4$). While this item is part of a subscale and interpretation of the item alone is thus limited, this was the response with the highest standard deviation of all responses in this test and highly statistically significant compared to the equivalent item in the Pre-Questionnaire ($t(22) = 3.023$, $p < .01$). These finding will be discussed in the conclusion.

9.5.3.3 Motivational Aspects

With the MSLQ, we looked into whether the students were more intrinsically or extrinsically motivated regarding this course, and how high they valued the course matter. To investigate if the students' motivation changed during the course, the MSLQ was presented unchanged in the Pre- and in the Post-Questionnaire. The results in the Pre-Questionnaire were in accordance with the MOPAS findings above: On average, the students' motivation on the intrinsic subscale was rather high ($M = 3.89$, $SD = .57$), while they were averagely extrinsically motivated ($M = 3.03$, $SD = .89$) and valued the course matter ($M = 3.96$, $SD = .48$). The difference between the extrinsic and the intrinsic score was statistically significant: the students were more intrinsically than extrinsically motivated ($t_{(29)} = 4.176$, $p < .05$). Virtually the same results were attained in the Post-Questionnaire: the students scored equally high or low on the Extrinsic ($M = 2.98$, $SD = 1.05$) and Intrinsic ($M = 3.69$, $SD = .91$), as well as Task Value ($M = 3.85$, $SD = .93$) subscales. Results from independent samples t -tests show that all $p \geq .05$. As before, the students were again significantly more intrinsically than extrinsically motivated ($t_{(22)} = 3.55$, $p < .05$).

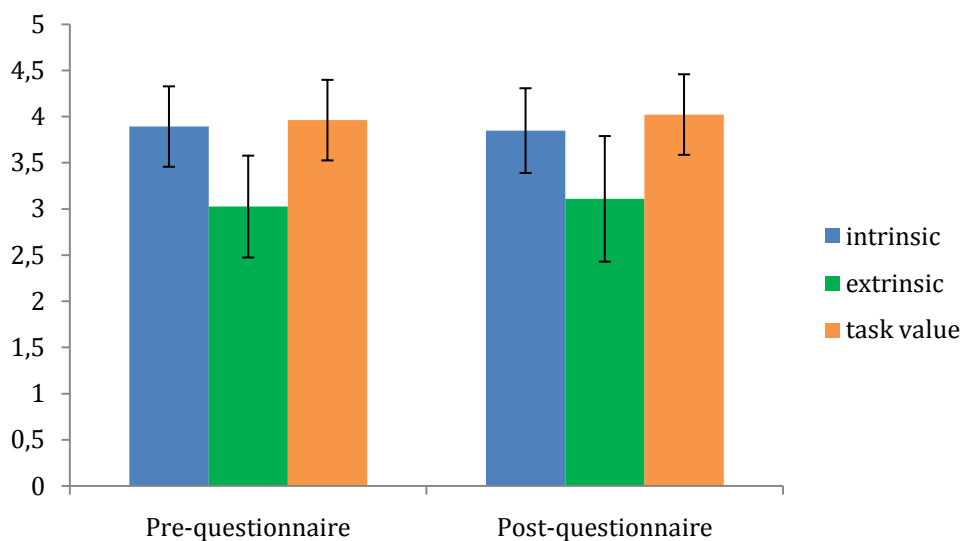


Figure 126. Comparison of intrinsic, extrinsic motivation and task value for preQ and postQ

However, it has to be noted that seven students that completed the Pre-Questionnaire did not complete the Post-Questionnaire. It is possible that students whose motivation was affected dropped out of the experiment.

9.5.4 Conclusion

This study was conducted to evaluate the IWT-integrated co-WIKI-system using a different sample of students without technical background. Most of the data yield very similar results to the ISR-experiment. The students spent on average a little under four hours in the Wiki ($M = 3.82$, $SD = 3$). The time spent by individual students fluctuated widely. Although students had close to no experience working with a collaborative online tool, they liked the collaborative aspects of the WIKI. Connecting to the system worked for the majority of the participants, however, most students complained about being logged out automatically and having problems displaying parts of the content. Hence the usability ratings were below average and students experienced negative emotions specifically anger and frustration while working with the WIKI. Although the structure of the WIKI has been improved in the latest version the interface still needs revision. In contrast to the low usability ratings, task-awareness ratings were mostly positive. Especially the contribution graphs were perceived as useful by the students. So towards G8.1 it can be stated that on the technical side the system still needs some more improvement, whereas the build in functions of the WIKI are mostly perceived as useful and enable students an overview of their learning progress (see G8.4).

To improve usability of the system changing the functioning of the auto logout and revising the interface is an inexpensive solution. We suggest to view these results in context of the time spent in the Wiki, as well as the wealth of feedback we received regarding possible improvements for the tool (G8.3), including reports of technical difficulties. We asked the students to leave their comfort zone of the programs which they usually use to write

assignments and use the Co-writing Wiki instead. It is likely that they spent much less time with the Wiki in the course of this experiment than they did with the programs we asked them not to use. The feedback we received regarding the ideas incorporated in the Wiki was positive. The biggest part of the negative free text feedback we received concerned technical problems, like misbehaviour of the auto-logout function. Students who spent more than 20 minutes on any one page were forcibly logged out of the Co-writing Wiki, because it was assumed that they were idling. This also happened, but was not meant to happen, when the students were editing a page or doing the group-assessment for more than 20 minutes. As can be seen by the distinct negative correlation of the item “The Wiki made my work easier” to the Anger subscale of the Computer Emotion Scale, students who disagreed to this item were more angry, frustrated and irritated while working with the Wiki. We suggest that this pronounced emotional component in judging the Wiki was caused in large parts by the frustration caused by the students encountering technical problems while working with the Wiki. In particular, the statistically significant changes in the subscales of the Computer Emotion Scale between the Intermediate and the Post-Questionnaire are likely related to the frustration encountered when performing the Group-Assessment. At the point of the Intermediate Questionnaire, the students had already spent time working with the Wiki and getting to know its functions. The only new thing to do between the Intermediate and the Post-Questionnaire was indeed the Group-Assessment. Despite the technical problems, we received a lot of positive feedback. The feedback we received from the students was that they enjoyed the possibility to get to know the work of the other groups (both their mistakes and what they did better), to be able to assess them anonymously, testing one’s own knowledge by assessing others and getting new ideas and perspectives. The students agreed to the usefulness of the assessment of their own group’s work by peers and that reading other students’ work helped them in understanding more about the topic. We conclude that the Group-Assessment supported students in their learning process (G8.6).

Regarding G8.2, the students stated that they did work collaboratively using the Wiki, and on average estimated that they spent more time working on their assignment in the Wiki than they did outside. In spite of the rather high frustration levels reported above, students, on average, still underestimated the time they spent in the Wiki, when compared to behavioural log data. The motivation of the students did not change in the course of the study, with the caveat that students whose motivation changed may have been among the seven who did not complete the Post-Questionnaire. The students were significantly more intrinsically than extrinsically motivated before and after the study. Working with the Wiki did not change their motivation, for better or worse. Regarding the Peer-Assessment, the students’ attitudes before the start of the experiment and the students’ experiences in the course of the experiment did not differ significantly. The internal peer-review was generally rated well, however it was used almost none of the time. We suspect that the students rated the idea of the peer-review instead of judging the actual peer-review itself. When we asked for feedback regarding the Peer-Review, some students wrote feedback concerning the Group-Assessment, suggesting that they did not even know what the Peer-Review was. We got feedback saying that the input by peers was welcome; however it is unclear if this feedback was referring to the Peer- or Group-Assessment. The response to the item “I understand some ideas better by discussing them with my peers” is worth pointing out: the response is

the one with the highest standard deviation of all self- and peer-assessment items, and the item was agreed upon significantly less in the Post-Questionnaire compared to the Pre-Questionnaire. We suspect that this result is closely related to the free text feedback we received regarding the peer-assessment: some students asked for a forum- or chat-like discussion feature to be able to discuss back and forth instead of sending feedback one way. Thus this point is another one to consider for future improvements (G8.3).

Concerning goal G8.5, we believe that there is simply not enough evidence to reach a final conclusion. On the one hand, the students stated that they liked the Peer-Review, on the other hand they almost never did it. However, as mentioned above, students' intrinsic motivation was high throughout the study, and feedback on the group-assessment was positive. Group-assessments were done at the end of the assignment, though. Thus, it can be assumed, that obligatory peer-assessments at an earlier point in time would also support students learning process. Since the Peer-Review was performed so few times, there is not enough evidence to back up the first part of G8.6 either. The Self-Review was also optional and was used only slightly more. It was rated at around the middle mark of the scale, worse than the Peer-Review. The idea of a self-reflection activity was apparently not appealing to the students. We got a single feedback on this, saying that if a student has already decided to change something, then it is important. Ergo rating the importance was viewed as superfluous. As already stated above, the group assessments can definitely be viewed as valuable support for students' learning process.

Regarding G8.7 we received positive feedback from the instructor. The feedback we received from the teacher was generally very positive. Her criticism regarded technical problems of the Wiki like the not working Revision Player. She strongly agreed with the usefulness of the Wiki for grading purposes and the tutor agreed that the setting up of the course via the tools in the IWT was easy to use.

10 R9. 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.

10.1 Evaluation and Validation Procedure

In this Section, we report a study which aimed at further investigating the quality of the automatic question creation tool (AQC) developed within the ALICE project. Generally, the AQC generates four types of questions out of a given text, namely open end, fill-in-the-blank, multiple choice, and true/false questions. The generation of questions can be divided into two processes. In a first step, the AQC extracts concepts out of the text, which can be viewed by the user. In a second step, questions are generated based on the extracted concepts. Therefore, the user can choose which concepts are to be used for the generation of questions, and which type of questions are to be generated (e.g. two multiple choice, three true/false questions, etc.). The generated questions are presented as test to the user and right after taking the test, students receive feedback on their performance. Thereby, the AQC lists again all questions but this time with the answer given by the student, the correct answer, and the received points (please refer to D5.2.2, ALICE 2010, for a detailed description of the tool).

After the studies conducted during the first round of experimentation, the AQC was integrated into the IWT and furthermore improved by two features, namely the implementation of a function for adding concepts and a function for tagging concepts. Adding concepts can be done by calling the list of concepts the AQC extracted from a text and then choosing further concepts from a list of words and phrases contained in the text. Furthermore, the user can order the final list of concepts by relevance and choose those concepts for which he or she wants to generate questions. To tag concepts, students simply have to highlight a concept within the text and then save it to their concept list. Afterwards, they can generate questions out of the self-extracted concepts.

The main goals of this study conducted in the second round of experimentation were (a) to test the IWT-integration and the improved functions of the AQC (b) to test the quality of concepts and questions with a well-balanced study design, and (c) to compare the concepts and questions generated by the AQC to concepts and questions generated by a real teacher (with course material used in a actual learning setting). In the following the goals and hypotheses are presented in detail.

Scenario 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.

Scenario 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.
- H9.7: 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.

Scenario criteria

- C9.1: To evaluate the different question types provided by the AQC.
- C9.2: To evaluate the quality (i.e., pertinency and terminology) of the questions provided by the AQC.

- C9.3: To evaluate the level (i.e., difficulty) of the questions provided by the AQC.
- C9.4: To evaluate the relevance of answers provided by the AQC.
- C9.5: To evaluate the distractors provided by the AQC for multiple-choice questions.
- C9.6: To evaluate the concepts provided by the AQC.
- C9.7: To evaluate questions generated by the AQC, using concepts created from users.
- C9.9: To evaluate the level of satisfaction of the users with the tool.
- C9.10: To evaluate the potential increase in students' motivation caused by the use of the tool.

Scenario metrics

- M9.1: Ratings regarding the pertinence of the questions provided by the tool
- M9.2: Ratings regarding the terminology of the questions provided by the tool.
- M9.3: Ratings regarding the level (i.e., difficulty) of the questions provided by the tool
- M9.4: Ratings regarding the relevance of the answers provided by the tool.
- M9.5: Ratings regarding the quality of the distractors provided by the tool.
- M9.6: Ratings for the quality of questions and answers generated by humans.
- M9.7: Ratings for concepts extracted by humans.
- M9.8: Ratings regarding the relevance of the concepts extracted by the tool.
- M9.9: Difference in relevance between human-extracted concepts and concepts extracted from AQC.
- M9.10: Ratings for questions when the tool uses human-extracted concepts.
- M9.11: Ratings for functionality/usability of the tool itself.
- M9.12: Ratings for the opinion of the users whether the tool supports them in self-generated learning.

10.2 Method

10.2.1 Participants

Thirty students enrolled in a full online course on “Learning models and processes for e-Learning” participated in the experiment. Twelve of the students are male and 18 of them are female. Participants are between 22 and 48 years old, on average they are 36 years old ($SD = 10.45$). Concerning the highest level of education, 11 students finished their Bachelor, 16 of them reached a Master degree and 3 of them have already a PhD. Participants' native language was Spanish, but they indicated having good writing ($M = 3.57$, $SD = 0.84$) and

reading ($M = 4.13$, $SD = 0.72$) skills in English. Twenty students stated that the course was mandatory for them. All participants are familiar with e-learning environments ($M = 4.43$, $SD = 0.63$) and prefer online-courses over face-to-face courses ($M = 3.57$, $SD = 0.94$). Students gave their consent to participate in the study by filling out the first questionnaire.

10.2.2 Design

One main goal of the study was to test the tool's quality with a balanced design covering the following aspects: The first factor concerns the question type and comprises the three factor levels multiple choice (MC), true or false (TF) and fill-in-the-blank (FiB) questions. Open ended questions were not included, because the automatic assessment cannot account for different wordings. The second factor refers to the creator of the concepts on which the questions are based on and has two levels, instructor and AQC. Hence, the concept is either extracted by the instructor or automatically by the AQC. The third independent variable refers to the creator of the questions. As the question is either generated by the instructor or by the AQC, there are also the two factor levels, instructor and AQC. Hence, for the questions, we've got a $3 \times 2 \times 2$ design. All questions were presented for two learning contexts, namely problem based learning and project based learning.

The dependent variables concern the quality and difficulty of the questions and their respective answers. The quality of the questions is measured by the four aspects pertinence, level, terminology, and difficulty, out of which a mean score is calculated. Pertinence means the relevancy of the question with respect to the topic. Level concerns whether the question is trivial or expresses a significant meaning. Terminology focuses on the appropriateness of the words chosen. Difficulty means the perceived difficulty. Quality of answers is measured for FiB and MC questions by the aspects terminology of answer, ambiguity of answer, and additionally for MC questions by the quality of distracters. Out of these aspects a mean score for the answer quality was calculated. All aspects were evaluated by the participants on a 5-pt. rating scale with 1 indicating a low quality and 5 indicating a high quality (except for ambiguity, where 5 indicates high ambiguity and therefore low answer quality).

According to the design of the experimentation, we calculated a multivariate ANOVA with three factors.

Regarding the quality of the concepts, we differentiate between teacher concepts, AQC concepts, and student concepts. Concepts were evaluated regarding their relevancy on a 5pt. rating scale ranging from (1) not relevant at all to (5) very relevant.

10.2.3 Apparatus and Stimuli

During the self-regulated learning experiment, the participants had to read two texts, take knowledge tests, which were provided on IWT, and received three Questionnaires, which were presented via lime-survey.

The two texts on problem-based and project based learning were provided by the instructor of the course. They had 1307 and 1002 words respectively and dealt with basic knowledge (definition, history, theoretical foundations, etc.) on the two topics.

Before the students started working on IWT, they were asked to fill in a Pre-Questionnaire (Questionnaire 1). This Questionnaire included the following sections: demographic data, previous knowledge regarding e-learning, general questions about learning preferences, evaluation of question types and students' English skills.

After the first unit, the students received an Evaluation Questionnaire (Questionnaire 2), where they were asked to evaluate 20 concepts on a 5-point rating scale regarding their relevancy ("not relevant at all" (1), "not relevant" (2), "neither/nor" (3), "relevant" (4), "very relevant" (5)). 10 concepts were automatically extracted by the AQC, whereas the other 10 concepts were extracted by the instructor. In the Evaluation Questionnaire the concepts were randomised and students didn't know which concepts were extracted by the AQC and which ones by the instructor. The Questionnaire also included a second section about students' experiences with extracting concepts.

After a second session in the IWT, students were asked to fill in another Questionnaire (Questionnaire 3) with the following sections: evaluation of 10 student concepts, evaluation of questions, and usability of IWT.

Finally, after the experiment finished, students were presented a post-questionnaire concerning their motivation during the study. It contained the subscale "Task Value" from the MSLQ by Pintrich et al. (1991) [10]. This scale measures students' perception of the course material in terms of interest, importance, and utility. High task value should lead to more involvement in one's learning outcome and Pintrich et al. found a high correlation between task value and intrinsic goal orientation ($r = .68$). More specifically, students have to indicate their (dis)agreement to six questions regarding the task value, i.e. how interesting, important, and useful the task was perceived. Example items are "I think I will be able to use what I've learned in this course in other courses" or "It was important for me to learn the course material in this class". The rating scale is a 5pt. Likert scale ranging from (1) I strongly disagree to (5) I strongly agree. Furthermore, the Post-Questionnaire asked for students' motivation to do the different tasks involved in the study, e.g. reading the texts, extracting concepts, or working with the IWT. Answers were given on 5pt. rating scale ranging from (1) not motivated at all to (5) very motivated. Three more questions concerned the impact on their learning activities, their understanding of domain concepts, and their general perception of the course.

Additionally, a Post-Questionnaire for the instructor was provided, which included the following sections: usability of IWT, functions on IWT, emotional aspects, and some open questions.

We used the SUS (System Usability Scale) by Brooke (1996) [8] 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%).

To investigate in which emotional mood the instructor was when he used IWT, we added the section “emotional aspects”, which includes 12 items. Kay and Loverock (2008) [9] 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”.

10.2.4 Procedure

The experiment consisted of four phases.

The **first phase** had taken place before the students were working on IWT. Two learning resources were provided for the students in order to get an overview about the topics “Problem-based Learning” and “Project-based Learning”. For these two texts, we let the AQC and the instructor extract concepts and make an order of relevance. These concepts were collected in an Evaluation Questionnaire (Questionnaire 2), which was given to the students in the second phase in order to evaluate their relevance, whereas students didn’t know which concepts were extracted by the AQC and which ones by the instructor. Hence, we could check whether the quality of the AQC concepts is as good as the quality of the instructor’s concepts. Later on we also used these concepts in order to let the instructor and the AQC generate questions out of them. In the meanwhile the students had received a Pre-Questionnaire (Questionnaire 1) before they started working on IWT.

The **second phase** was the first unit with the students. In this phase students were assigned to two groups, Group 1 and Group 2. First, the students were asked to read the learning resource which was presented on IWT. Group 1 read the text regarding the topic “Problem-based Learning” and Group 2 got the text about “Project-based Learning”. Second, the students had to extract at least 6 concepts from the text by tagging and highlighting keywords. Additionally the students were asked to put their concepts in an order of relevance. Third, the students answered the Evaluation Questionnaire (Questionnaire 2), which was prepared in Phase I. Here the students evaluated the quality of the concepts extracted by the AQC and the instructor from the learning resource they had read before.

As the first phase, the **third phase** also took place without the students. In this phase, we generated questions based on the extracted concepts. Questions were generated by the AQC and by the instructor. Hence the instructor as well as the AQC had to generate questions based on AQC and instructor concepts. For each learning resource, 24 questions (12 instructor / 12 AQC questions) based on 12 concepts (6 instructor / 6 AQC concepts) were generated. Thus, for each concept, the instructor as well as the AQC generated a question. So finally we’ve got 4 different variants of questions: an instructor question based on an instructor concept, an instructor question based on an AQC concept, an AQC question based on an instructor concept and an AQC question based on an AQC concept. Besides,

the question types (Multiple Choice, Fill-in-the-blank, and True/False) also were balanced. More exactly, two questions of each type were generated for the six instructor as well the six AQC concepts. The three questions types combined with the four different variants described above (e.g. teacher concept - AQC questions or teacher concept – teacher questions) yield 12 different questions.

From these questions the knowledge tests for Phase IV were constructed and provided on IWT. To achieve a fair comparison of AQC and teacher questions, we provided for each concept (teacher or AQC) one teacher and one AQC questions of the same type (TF, MC, FiB). Out of this, two parallel test forms for each topic were created. At the same time, students were assigned to 4 groups, Group A, B, C and D. Students from Group 1 were assigned to Groups A and B, Group 2 was divided into Group C and D. Each pair of groups received the parallel test forms regarding their topic. Group A got for example an instructor question based on the first concept, whereas Group B got an AQC question based on the first concept. Then, Group A got an AQC question based on the second concept, whereas Group B got an instructor question based on the second concept and so on.

Clearly, students weren't aware of the origin of the questions. So they didn't know whether the questions were generated by humans or automatically and whether the underlying concepts were from the instructor or the AQC.

The **fourth phase** was the second unit with the students. First, the students learned the text on IWT, which they had read in Phase II. Second, they were asked to extract six concepts as in Phase II in order to compare consistency of the extracted concepts. Then, the students should generate “on-the-fly” questions from the AQC based on their concepts. After that, the students got a test on IWT about the text, they had learned before. So Group A and B got a test about “Problem-based Learning” and Group C and D received a test about “Project-based Learning” (see above for the detailed construction of the tests).

Afterwards the students read through another learning resource. Group A and B read about “Project-based Learning” and Group C and D read through the text regarding “Problem-based Learning”. In Questionnaire 3, the students were asked to evaluate students' concepts, which were extracted in the first unit. Additionally they evaluated the questions generated by the instructor and the AQC. Group A and B got the concepts and questions regarding the text ““Project-based Learning”, they had read before and Group C and D evaluated the concepts and questions which were related to the text “Problem-based Learning”. So Group A and B evaluated the concepts from students which were in Group C and D and the questions which Group C and D had received in their tests and vice versa.

The Post-questionnaire regarding students' motivation was sent after all students had finished their tasks in phase four.

10.3 Evaluation Results

In this section we focus on the hypotheses concerning the quality of the tool, namely H.9.1 through H.9.7 (see Section 10.2 for details). From the metrics M9.1 – M9.11 specified in Section 10.1 and in [3], the following are relevant for the evaluation:

From the 30 participants filling out the pre-questionnaire, 25 took part in phase 2, in which they evaluated the relevancy of 20 concepts extracted by the teacher and AQC. In phase 4, we could collect data from 20 participants, who all evaluated 10 student concepts and the 12 questions created by the teacher and AQC. The knowledge tests provided within the IWT, which also contained the AQC and teacher questions (but were presented cross-wise as described in Section 10.2.4), were taken by only 13 students. The data of one student was not included in the analysis below, because she called and saved the test, but did not answer a single question. Thus, for each of the two topics, six students took the prepared knowledge test. Furthermore, four students took an on-the-fly test with a total of 21 generated questions (seven for each question type).

Regarding H.9.1, it could already be shown in the first round of experimentation that the AQC is able to generate four types of questions from a given text, namely open ended questions, fill-in-the-blank questions, multiple choice questions and true/false questions. For the three latter question types this was again proven with two new texts (as discussed above, open ended questions were not included for this study). In contrast to phase 1, the two texts were not chosen by the experimenters, but by the teacher of the course herself. Independent samples *t*-tests performed for concept relevancy (RelConc), mean question quality (QualQu), and mean answer quality (QualAns) yielded no differences between the two topics problem-based and project based learning (RelConc: $t_{(68)}=.155$, $p=.877$; QualQu: $t_{(238)}=.715$, $p=.475$; QualAns: $t_{(158)}=.243$, $p=.808$). Thus, for the following statistical analysis the data were aggregated across the two topics.

10.3.1 Relevancy of extracted concepts

One important requirement for the generation of high quality questions is the extraction of concepts that are relevant within a given context. Thus, in Phase 2 of the experiment (see Section 10.2.4) the 10 most relevant concepts extracted by instructor and AQC were given to the students to be evaluated (Questionnaire 2). For each topic (problem vs. project base learning) the concepts were presented in a random order. In the case of equivalent concepts, the next one in the order of relevancy was chosen. Ten of the concepts extracted by the students in Phase 1 were presented to the other half of students in Phase 4 (Questionnaire 3) and also evaluated with regard to relevancy. Thus for the comparison of AQC and teacher concepts paired samples are available, whereas comparisons with student concept relevancy involve independent samples. Table 19 gives examples for concepts as they were extracted by AQC, teacher, and students for the two different texts.

Problem based learning		
AQC	Teacher	Students
Collaboration and Learning	Problem-based learning	Working in Groups
Learning activity	Prob. BL curriculum	Activation of prior knowledge
The instructor	Prob. BL environment	Similarity of contexts
The use	Instructional strategy	Stimulate the learner
Problem	Prob. BL process	Constructivist perspective
Future roles	Knowledge Construction	Self-directed
Project based learning		
AQC	Teacher	Students
Project	Project-based Learning	Planning
Learners	Components of Proj. BL	Creating
Knowledge construction	Phases of Proj. BL	Projects
The planning phase	Teaching and Learning strategy	Active builders of knowledge
Complex activities	Projects focus	Share their artifacts
Learning	End product	Processing the project

Table 19. Example concepts extracted by AQC, teacher, and students for two texts

Figure 127 shows the average ratings for the three concept types for each topic. Across the two topics mean ratings from 25 (for teacher and AQC concepts) and 20 (for students) participants ranged from 3.72 ($SD = .7$) for student concepts over 3.92 ($SD = .57$) for AQC concepts to 4.12 ($SD = .44$) for teacher concepts. A one-way ANOVA for the three concept extractors revealed a significant effect ($F_{(2,67)} = 3.66, p = .031$). Post-hoc tests after Scheffé revealed a difference between teacher and student concepts ($p = .032$) but not between AQC and teacher or student concepts ($p = .353$ and $.42$ respectively). Since teacher and AQC concept evaluations are based on the same sample, we also performed a repeated measures t -test for a stricter comparison of these ratings; however, with $t_{(24)} = 2.0$ and $p = .057$ AQC concepts do still not differ significantly from the concepts extracted by the teacher. Thus, with regard to H9.4, it can be stated that concepts extracted from the AQC are as relevant as concepts extracted by humans.

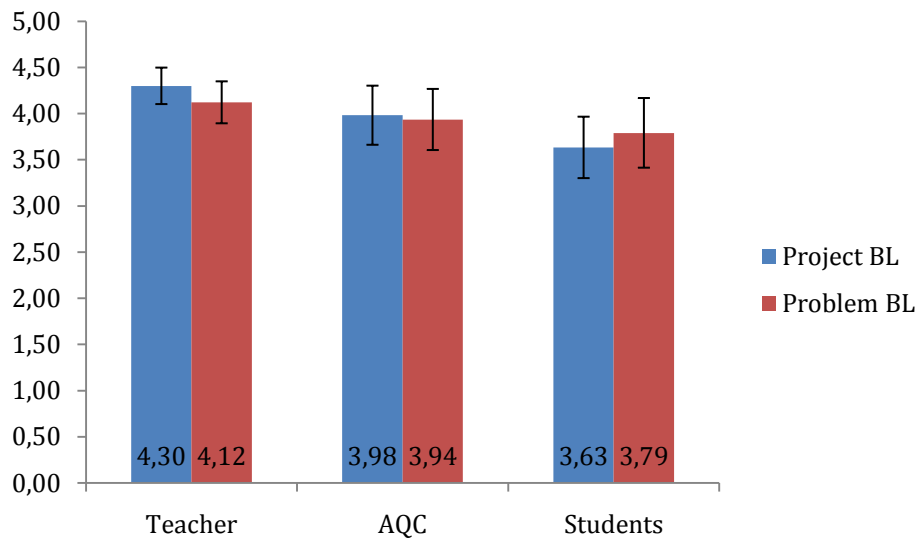


Figure 127. Relevancy ratings for concepts extracted by AQC, teacher, and students

10.3.2 Quality of questions

To evaluate the quality of questions 3x2x2 MANOVAs were performed by aggregating the data from both topics (problem based and project based learning). With 20 students performing the evaluation and 12 questions per student, 240 answers were collected for each criterion. These are divided equally over the question types (80 data points for TF, MC and Fib questions each), concept extractors (120 from teacher and AQC each), as well as questions creators (also 120 from teacher and AQC each). The evaluation metrics consisted of four measures for the quality of the questions themselves (pertinence, level, terminology, and difficulty) and two and three measures for the quality of the answers of FiB and MC questions (terminology of answer and ambiguity of answer for both, plus quality of distractors for MC). Since FiB questions are not included in this analysis, the number of data points for answer quality decreases to 160 (80 for distractor quality). Figure 128 shows the mean ratings for question and answer quality per question variant (concept extractor x question creator) and question type (TF, MC, FiB). Because of the different numbers of questions types (TF, MC, Fib) involved, we performed two 3x2x2 ANOVAS for mean questions and mean answer quality.

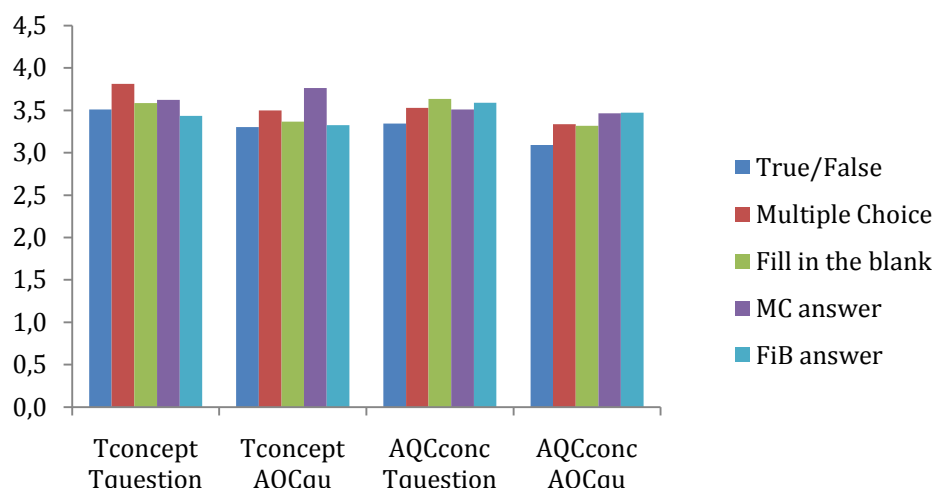


Figure 128. Mean ratings for question and answer quality of T/F, MC, and FiB question (Tconcept = concept extracted by teacher, AQCqu = question created by AQC, etc.)

The results are summarized in Table 20. Whereas none of the three factors had an effect on mean answer quality, we found one significant effect on mean questions quality. More specifically, with $M_{teacher} = 3.57$ and $M_{AQC} = 3.32$ ($SE = .063$) questions created by the teacher were evaluated significantly higher, than those created by the AQC. However, with a partial η^2 value of .034 the effect size is rather small. Interactions are also non-significant. To investigate, in which aspect the questions differ, we had a closer look at the different aspects contributing to question (and answer) quality. Figure 129 gives a detailed description of each dependent variable separate for the three questions types (TF, MC, FiB).

	Factor	df	F	p	η_p^2
Mean question quality (N = 240)	Question Type	2,228	2.432	.09	.021
	Concept extractor	1,228	2.399	.123	.01
	Questions creator	1,228	8.025	.005	.034
Mean answer quality (N=160)	Question type	1,152	1.084	.299	.007
	Concept extractor	1,152	.044	.834	.000
	Question creator	1,152	.070	.792	.000

Table 20. Effects from three-way ANOVAS on mean question and answer quality

A three-way MANOVA including the four aspects of question quality yielded the expected effect of question creator for the multivariate results ($F_{(4,225)} = 2.51$, $p = .043$, $\eta_p^2 = .043$) and no other main effects or interactions. Univariate results showed that the effect is due to higher evaluations of teacher questions' pertinence and level. For pertinence teachers questions reached a mean rating of $M = 3.95$ as compared to $M = 3.58$ for AQC questions with $SE = .092$ ($F_{(1,228)} = 7.82$, $p = .006$, $\eta_p^2 = .033$). Ratings for level imply that teacher questions are more meaningful ($M = 3.62$) than AQC questions ($M = 3.27$, $SE = .096$), with $F_{(1,228)} = 6.56$, $p = .011$, $\eta_p^2 = .028$. However, for both aspects the effects are only small in size. Ratings for questions' terminology and perceived difficulty did not differ significantly, and there were no significant interactions among the three factors.

With respect to the quality of answers, a closer look at the three aspects revealed no significant main effects and no interactions for the multivariate results of the performed MANOVA (for MC and FiB questions), for the aspect terminology of the answer, and for the quality of distracters (ANOVA for only MC questions). However, we did find an effect of questions type on the ambiguity of answers. With $M = 2.35$ ($SD = 1.17$) for MC and $M = 2.8$ ($SD = 1.26$) for FiB questions, participants evaluated answers of the latter questions type to be more ambiguous.

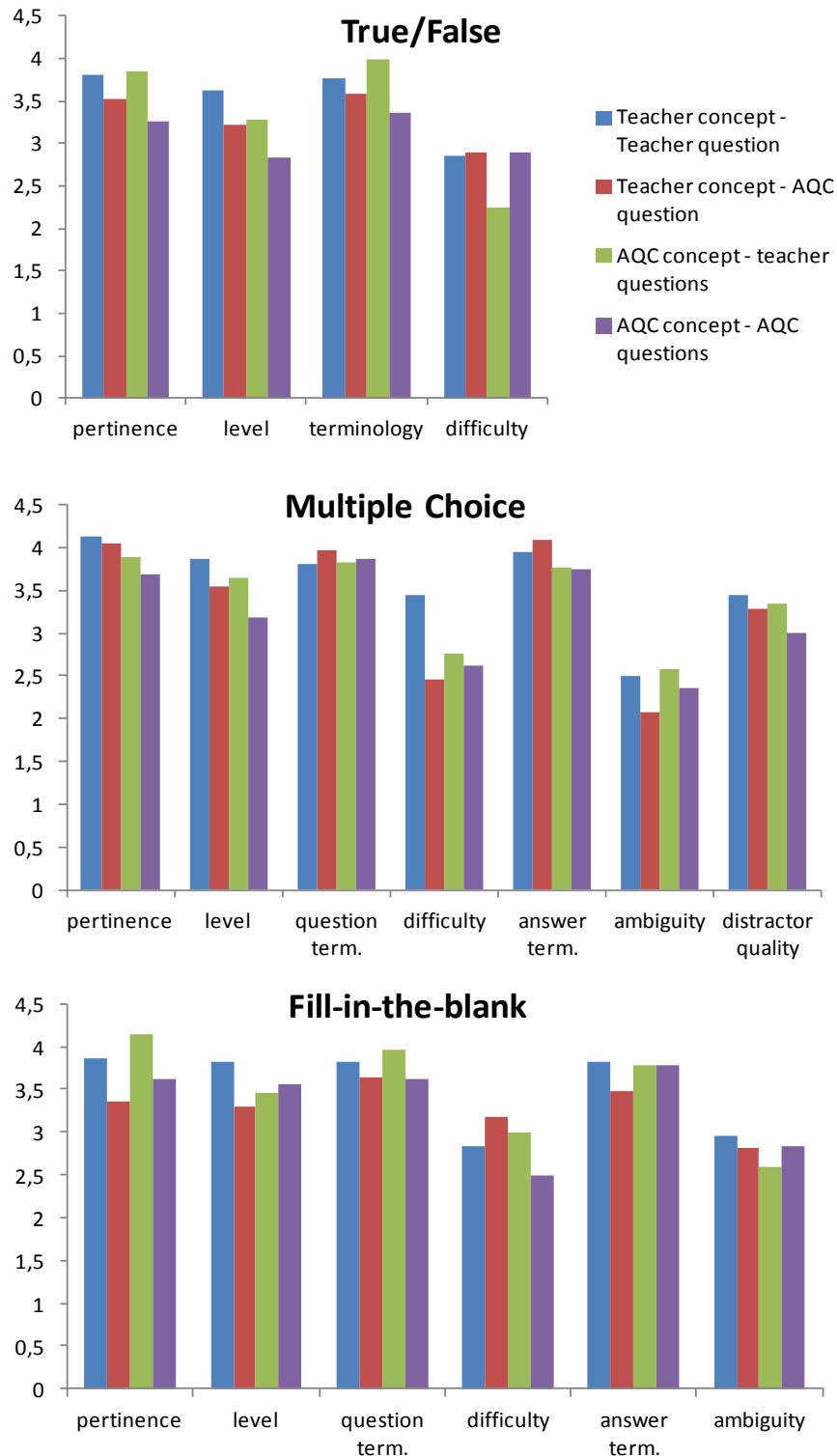


Figure 129. Aspects of question and answer quality for different question variants

Coming back to the hypotheses listed at the beginning of this section, the results allow a clear verification of hypothesis H9.3 and H9.5. For H9.3, the answers to the questions provided from the tool are relevant. Moreover, participants evaluated the quality of answers

generated by the AQC as equally high as the quality of answers generated by the teacher. By using teacher concepts for half of the questions created by the AQC, we could also show that the tool is able to generate questions from concepts entered by users (H9.5). Regarding H9.2, which states that all types of questions generated from the tool are as high in quality as questions generated by humans, the answer is two-fold. Whereas the quality of answers, the terminology, and the perceived difficulty of AQC questions are evaluated equally high as those of teacher questions, the level and pertinence of questions received lower ratings. Thus, teacher questions seem to be less trivial and address the topic in a more meaningful way.

10.3.3 Difficulty of questions

To further investigate H9.2, we also collected data concerning the real difficulty of questions (as compared to the perceived difficulty described above). Therefore, the same questions which were presented for evaluation were also prepared as knowledge test and uploaded to the courses in the IWT. To avoid an influence of the evaluation process on the test taking or vice versa, each test was given to half of the students as test and to the other half for evaluation purposes. Whereas 20 students did the evaluation of questions only 13 took the knowledge test. Data of 12 students who each answered 12 questions could be analyzed (see also above).

All together 45.83% of the questions were answered correctly, which equals 66 out of 144 questions. Table 21 gives an overview on how many items per questions variant have been answered correctly. Since there is no difference between the topics problem based and project based learning (32 vs. 34 correct responses), the data are aggregated across the two topics. We calculated χ^2 tests to compare the frequencies for questions that (a) are based on AQC vs. Teacher concepts, (b) are generated by AQC or teacher, and (c) are designed as either TF, MC, or FiB question. Except for the question type, the critical χ^2 values exceeded the empirical ones. With $\chi^2 = 22.46$, the differences between the three question types are statistical significant, which can clearly be attributed to the very low solution rate for fill-in-the blank questions (8% correct solutions compared to 69% for MC and 60% for TF). With regard to H9.2, the results indicate that the difficulty of AQC questions does not differ from that of teacher questions.

To investigate the relationship between perceived difficulty of actual difficulty, we correlated the mean ratings and number of correctly solved items per questions variant (i.e. for 12 different item types, as e.g. TF with teacher concept and teacher questions or MC with AQC concept and teacher questions, etc.). The resulting correlation of $r_{(12)} = -0.71$ ($p = .009$) indicates that questions which are perceived as more difficult are also solved by less participants.

We also looked at the difficulty of the on-the-fly questions. From seven questions per type, five TF, six MC and 1 FiB have been answered correctly, which is in line with the findings reported above (high difficulty of FiB, no difference between TF and MC).

	Teacher concept		AQC concept		
	Teacher questions	AQC question	Teacher question	AQC questions	
TF	9	6	11	3	29
MC	0	15	6	12	33
FiB	2	0	0	2	4
Sum	11	21	17	17	66
Sum concept	-	32	-	34	
Sum question	-	-	28	38	

Table 21. Number of correctly answered questions per question type

10.3.4 Usability of the AQC integrated into IWT

The basis for the validity of usability measure is the time participants spent in the system, i.e. the more time users spent with the system the more valid are their judgments. Log data show that students accessed the IWT on average 3.31 times (SD = .72, MIN = 1, MAX = 7) and spent M = 103.78 min (SD = 89.3) within the system. The teacher (and her assistant) spent together 39 hours 28 minutes in the IWT, accessing it 102 times.

To evaluate the usability of the integrated AQC (H9.6), the SUS scale (Brooke, 1996, see Section 9.2.2.3 for a detailed description) was presented to students as well as the teacher. Students' mean SUS scores amount to 57.59 (SD = 16.99) with a rather large range from 23.75 up to 87.5. The SUS score provided by the teacher was 48.13. Thus students perceived the usability of the AQC integrated in the IWT very differently, but on average higher than the teacher. However, both student and teacher scores are below the average of 68, which is the reference value suggested by Brooke. Despite the low SUS score and the great amount of time the teacher spent in the IWT, ratings from the Computer Emotion Scale (see Section 10.2 3) show very positive emotions while working with the system (mean scores for happiness/sadness/anxiety/anger in the given order are 3/1/1.25/1).

A closer look at what participants liked and disliked about the system can be taken from the corresponding open comments. The teacher stated that he liked the "Automatic Test Generator", i.e. the AQC and disliked that he had sometimes troubles to select concepts and that for loading a text clicking on the "Explain Section" is necessary. Students stated that they liked: highlighting concepts to make auto-evaluations and to save them in a list for later reference, the easy tools to create an evaluation and generate questions, the generation of questions from concepts, the simple but functional interface, the interactivity and the easy use of the tool. On the other hand they disliked: that they could see only small portions of the text when reading, the awkward and difficult to use interface, that the test generation did not work right away, design and color, that it was only in English, that errors are not commented

(e.g. whether it is a system error or a user error), to little help and instructions, that the home page is confusing with regard to the difference between actual and available courses.

With respect to H9.7, namely the derivation of improvements from user's feedback, the teacher stated that he would improve the back functions ("when you press the back button or just return the functionality changed"), make the messages more clear (e.g. does 100% completed mean that an activity or only the loading of the activity is completed), and enable to instructor to do more course preparations by himself (e.g. loading a text). Students noted that: the system should allow a person to see the whole reading at once, to improve the interface, to give clear information on the item scores, to make the system fast and improve the graphics, to give clearer information on the functionality of the icons/buttons, and to make the system multilingual.

10.4 Validation Results

For the validation of the scenario, H9.8 and H9.9 are relevant, as well as the criterion C9.10 and metric M9.12 specified in Section 10.1.

For the following analysis, the results from 14 students filling out the post-questionnaire are aggregated. Furthermore, one teacher filled out the instructor-post-questionnaire.

10.4.1 Motivational aspects and task value

With respect to H9.8, a requirement for having a positive impact on students' motivation is that they are generally comfortable with self-regulated learning settings. Participants of this study indicated that they like self-regulated learning environments ($M = 3.77$, $SD = .76$) and that they prefer learning on their own over being supervised all the time ($M = 3.8$, $SD = 1.01$). Furthermore, they agreed on the statements that testing themselves helps when they learn something ($M = 3.87$, $SD = .72$) and that they need clear instructions when they learn something ($M = 3.93$, $SD = 1.03$). These results are in line with the comments on the tool itself, namely that they liked the automatic question generation and automatic assessment as well as the possibility to highlight and save important concepts to support their learning process.

The task value scale by Pintrich et al., (1991) [10], which was presented in the post-questionnaire showed that students were highly interested in the task and also perceived it as being important and useful. Mean ratings to the six single questions ranged between 4.5 ($SD = .65$) and 4.71 ($SD = .47$) resulting in a mean task value of 4.58 ($SD = .48$). Due to the high correlation of task value and intrinsic goal orientation reported above, it can be assumed, that students were also intrinsically motivated and involved in their learning activities. This result is also supported by the high motivation ratings for the single tasks required during the study, which ranged between 3.77 ($SD = 1.01$) and 4.31 ($SD = .85$). Figure 130 shows the mean ratings for task value and motivation for doing different tasks.

Summarized, H9.8 can be viewed as confirmed, since the students indicated to be highly motivated for their learning activities.

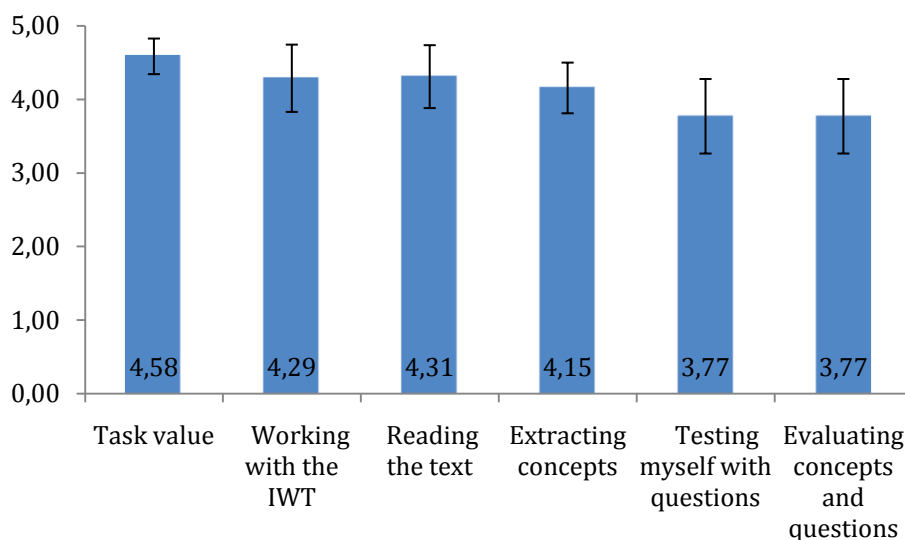


Figure 130. Mean task value (MSLQ) and level of motivation for various tasks

10.4.2 Support of self-regulated learning

To investigate the pedagogical and psychological impact of the tool, we checked, whether the tool supports self-regulated learning and students can thus benefit from using the tool (H9.9). According to the teacher the tool constitutes a support for students in the self-study process. From a teacher's point of view, the functions tested by teacher were too few to judge the worth of the tool for teachers. Especially the fact that creating the courses on IWT, providing the tests, and providing the learning material could not be used easily was seen as drawback. Furthermore, the teacher stated that observing the progress of students and monitoring the course as a whole without technical support was difficult.

From student's point of view, testing themselves with questions had a positive impact on their learning activities ($M = 4.43$, $SD = .76$), taking the course improved their understanding of domain concepts ($M = 4.5$, $SD = .65$), and the course was a worthy educational resource ($M = 4.43$, $SD = .76$). In their open comments, all 14 students stated that they would benefit from self-assessments (self-generated tests) when learning in general. More specifically, they said, that self-assessments are a good preparation for real assessments, that they help to know one's level of knowledge or progress, and that it helps to improve the understanding and retention of concepts. They also thought that automatic test generation is a good support for teachers. Only one student indicated that he wouldn't go through a self-generated test, because for him reflection on his knowledge is more important. Another student stated that the self-assessments help to study a text, but that the questions were not good.

10.5 Conclusion

The aim of the conducted study was to evaluate a tool for automatic question creation (AQC) and its application within the IWT. We now want to address the specific goals developed in [3]. As already concluded after the first phase of experimentation, it was possible to provide a tool that generates different types of questions from a text (G9.1). The four offered question types include open ended, fill-in-the-blank, multiple choice, and true/false questions, the latter three of which were re-evaluated in this second round of experimentation. All questions generated by the AQC are based on concepts, which are in a first step automatically extracted from a given text. In a second step the AQC creates for each concept the required types of questions (up to four different questions per concept). Thus, the quality of the extracted concepts is an important factor for the achieved quality of the generated questions. Before considering G9.2, the quality of questions, and G9.3, the quality of the provided answers, it is therefore necessary to have a look at the underlying concepts. In this study the teacher of an online course on e-learning provided two texts and extracted the 10 most relevant concepts out of each text. Simultaneously, the AQC extracted concepts out the same texts and put them into an order of relevance. For each text, 10 AQC and 10 teacher concepts were then randomly presented to a group of students, who had read the respective text and extracted their own concepts beforehand. Students rated the concepts with respect to relevancy. Later on in the study, students were given their colleagues' concepts for evaluation. Hence we could compare the relevancy of concepts extracted by the AQC, by the teacher, and by students. The obtained results show no difference between concepts extracted by the AQC and by humans (teacher or student). Thus for G9.4, we can conclude that the tool is able to extract relevant concepts from a text, which form a suitable basis for knowledge questions.

Coming back to goals G9.2 and G9.3 the quality of questions and their respective answers was evaluated by comparing it to the quality of questions created by teachers (both AQC and teacher questions were based on a equal number of AQC and teacher concepts). The quality of questions was measured by means of four different aspects (pertinence, level, terminology, and difficulty), that of answers by two or three different aspects (terminology and ambiguity for FiB and MC, plus distractor quality for MC). A three-way MANOVA including the factors questions type (TF, MC, FiB), concept extractor (teacher, AQC), and questions creator (teacher, AQC) revealed an effect of questions creator for the dependent variables pertinence and level. Thus, AQC questions are equally well formulated as teacher questions (no effect on terminology) and are perceived as equally difficult, but are evaluated as more trivial and less relevant than teacher questions. However, considering that the factor "questions creator" accounts for only about 3% of the overall (effect and error) variance ($\eta_p^2 = .033$ and $.028$ for the two measures) and that the AQC is mainly meant as tool to support self-regulated learning, the outcome of the evaluation is definitely positive. Additionally, the results from the evaluation of answers revealed no difference between teacher and AQC terminology, ambiguity, or distractor quality. The same is true for the actual difficulty of questions, indicated by the number of correctly solved items.

With respect to G9.5, namely to provide a tool that creates questions using concepts entered by users, we asked students to extract concepts and take an on-the-fly-test using these concepts. Four students successfully generated and took the test with a total of 21 questions. Furthermore, half of the questions created by the AQC were based on teacher concepts, which were also entered manually. Thus, G9.5 could definitely be met.

Goal G9.6 and G9.7 concern the user-friendliness of the tool and the identification of possible improvements. Regarding the tool's usability, SUS scores from both teacher and students were below average, but the teacher was still in a very positive emotional state and open comments from both sides show that they appreciate the tool and its functions. Thus, the results suggest that the tool is a very useful and valuable resource for self-regulated learning (see also G9.9.), but basic functions, such as clarity, insightfulness, and also performance still need to be improved. However, students indicated that they would benefit from automatic self-assessments and that the course is a worthy educational resource. Regarding G9.8, the results show high task values and high motivational ratings for the different tasks performed during the study. Thus, we can conclude that the tool is able to motivate students in their learning activities.

11 Final conclusions

The aim of ALICE is to build an adaptive and innovative environment for e-learning that extends the IWT platform [1]. 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 has described the results of the second round and final experimentation, evaluation and validation activities of the project ALICE within Work package 8. The final objective of the WP8 was to experiment the developed tools (delivered as independent working packages) and resources in order to provide feedback to theoretical and technological activities.

More specifically, this report has presented the results of the execution of the experimentation and validation plan of the research and technology developed in ALICE reported in [3]. 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 was followed (see Table 1). Therefore, the purpose of this report was to collect information about the experience of performing the different tasks where the experimentation and validation are based on in the different pilot sites.

In summary, the objectives and research goals of the ALICE project reported in this document were achieved by experimentation and validation and provided 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 [3]. Moreover, by implementing the developed scenarios of use, as well as enhancing the learning experiences of the various users could contribute to more effective and efficient learning activities, more motivation and inspiration for learners and teachers in various formal and informal learning circumstances.

11.1 Overview

The aim of this section is to make an overview and summary of the results of the entire experimentation, evaluation, and validation activities 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.

The experimentation and validation activities reported in this document and summarized in this section consider six work packages, which investigate the scientific aspects of the project:

- 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 work packages set the theoretical and technological grounds towards the evaluation and validation of the impacts of the innovative features offered by ALICE inside the selected learning and training environments. Two different contexts and two e-learning modalities were considered to evaluate ALICE prototypes making three evaluation contexts in three pilot sites:

- Full e-learning for science teaching at university: UOC
- Blended learning for science teaching at university: TUG
- Blended learning for civil defence and emergency at secondary school: MOMA

In all, nine scenarios were experimented in the above pilot sites:

- R1. Upper Level Learning Goals
- R2. Knowledge Model contextualization
- R3. Semantic Connections between Learning Resources
- R4. Live and Virtualized Collaboration
- R5. Storytelling
- R6. A Serious Game for Civil Defense
- R7. Affective and Emotional Approaches
- R8. Enhanced Wiki-Test and Peer-review for writing assignments
- R9. Assessment in Self-Regulated Learning

Two rounds of experiments were scheduled for each of the above scenario, some of the scenarios needed several studies or trials to evaluate and validate all the hypotheses, making a total of **26 experiments**. The total number of participants in these experiments was **949 students and 36 tutors** in the three pilot sites.

11.2 Scenarios

In this section we present an overview and summary of each of the nine scenarios. The most relevant results are shown towards achieving the scenario goals according to [3] as well as a comparison when relevant between first and second rounds of experiments. Moreover, the results are interpreted and justified from different perspectives, such as type of participants and pilot sites.

11.2.1 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 goals 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. As a result, the ULLG recommendation algorithm provides suitable learning resources that meet the learners' needs [7].

To evaluate and validate the above aim, two rounds of the experiments were carried out at the UOC pilot site involving 231 students and 2 lecturers in all following the same methodology and reflecting on the scenario goals as reported in [3]. Moreover, the second experiment used an improved version of the same prototype and ULLG recommendation algorithm as that used in the first experiments [5]. Therefore the two rounds of the experiments had the same evaluation and validation objectives and the second round of experiments (Section 2) served also to validate the improvements made in the prototype from the first experiments.

Considering the results of both rounds of experiments, in general the students liked the IWT tool and found it interesting to have a personalized system to study. IWT was able to generate a course from the ULLG recommender system from a need expressed in natural language by the learner (G1.1) and these courses generated had been fulfilled the expectations of the learners (G1.2). However, these courses could not provide students with significant amount of new knowledge though it managed to satisfy the needs expressed by them (i.e. students met specific knowledge needs by using the ULLG recommended system), thus G1.5 could only be partially satisfied.

In addition, at usability level, in the second round the students noticed the improvements (e.g., new navigational panels, automatic searching suggestions, etc.) made in the second phase of the project and students did not report any particular usability aspects that had influenced negatively their emotions during the first experiments. However, still IWT was considered a technical barrier by students and important amounts of resilience to change the e-learning platform from UOC to IWT were found partially due to the usual learning curve when facing a new system. Numerically, the SUS scores in the first experiments was 60.78 and it was decreased up to 53.97 in the final experiments whilst Happiness emotion was 1.51 in the second round (versus 1.39 in the first experiments) and all bad feelings were ameliorated in the second experiments. Hence, goal G1.4 was achieved partially.

Finally, motivation was found to be quite high and increasing from the first to the second experiments ($M=2.79$ to $M=3.02$ in the scale 0-5). As for the gaining of knowledge, the first round of experiments found that the group with IWT achieved significant better marks than the group without IWT (7.11 vs. 6.22 in the scale 0-10). Moreover, the improvements made in the prototype in the second phase of the project also impacted in the gain of knowledge during the final experiments but not significantly (6.96 vs. 6.58 in the scale 0-10).

Finally, possible ways of improving further the utility of the ULLG and a larger extend of IWT were suggested by the students, especially related to usability. Those suggestions made in the first round of experiments were addressed and, as mentioned previously, students realized these improvements by showing more positive emotions in the second round of

experiments. More comments and suggestions were collected from students after the final experiments related to solve new usability issues found, but especially they gave hints to be addressed as future work with IWT, such as improve even more the system performance and compatibility with mobile devices. All in all, goal G1.6 was fully achieved.

11.2.2 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. TUG and UOC pilot sites run trials on this scenario: from the students' and instructor's viewpoint. The scenario is underpinned by a Visual Ontology Editor (VOE) coming out from WP7 development and research [7]. Three experiments were run at TUG and UOC involving 8 students and 3 lecturers in all. Next, all the experiments are summarized along with the results.

11.2.2.1 TUG Site

In order to stand on the level of goals achievement with respect to [3], the scenario was experimented twice on the TUG site. The first experiment was reported in [5] while the second in Section 3.1. This section discusses the findings from the two experiments and reflects on the scenario goals as reported in [3].

The first study was more teacher oriented and only investigated aspects related to teacher's perception on the scenario and its prototype, whereas the second experiments investigated the teachers and students perception on the scenario. Concerning tools usability and user-friendly interfaces (G2.1.1), the first experiment shows that the teachers did not like tool (VOE) complexity and stressed the need for user support. The same findings came from the second experiment as the teachers indicated that the VOE tool is too complex and they would need technical support to manage and administer contextualized courses. Nevertheless, the teachers high curiosity to use the tool was faced with moments of feeling helpless and frustrated as they needed technical support to manage permissions and to author learning material using IWT, which obviously influenced their satisfaction on the VOE tool to have a SUS value of 45 for lecturer A and 37.5 for lecturer B below their votes in the first experiment of 65 and 57.5 respectively.

Regarding the ability to generate automatically contextualized courses (G2.1.2), despite the complexity of managing the scenario requirements the teachers in both experiments mentioned that their students could benefit from the features provided by this scenario. However, having a closer look on the finding from the student third phase in the second experiment, some of the student mentioned that the provided course fitted their preferences and context - i.e. beginner of advanced (G2.1.3). Some also mentioned that they already knew the provided knowledge and the system does not perfectly considered their knowledge state.

In order to focus on the student learning support and better learning via contextualized courses (G2.1.4), the second experiment was designed to provide a static course and a contextualized one. The student were asked whether the contextualized course improved their domain understanding, and thereby their votes on a 5-point Likert scale question indicates that the dynamic course supported them to improve their learning and to have a

better understanding of the domain concepts - covered by the contexts ontology. Nevertheless, the students argued that the contextualized course provided worthy educational materials. However, there was no significant difference in their votes on this question from both static and contextualized courses.

Focusing on further improvements (G2.1.5), the teachers from the two experiments conducted in TUG site stressed the importance of elaborating the system difficulty through more user friendly interfaces and further information concerning the VOE provided services. More focus on easy-of-use aspects should be considered and try to avoid having technical support to manage and administer contextualized courses.

11.2.2.2 UOC Site

Similarly to the previous scenario at TUG, the aim of this scenario at UOC site was also to build an ontological description of a teaching domain that is able to automatically adapt to a context [7]. To this end, two rounds of experiments were conducted at UOC pilot site in order to test the tool from the instructors' viewpoint. The results of this study provided relevant feedback of how the Visual Ontology Editor (VOE) tool of IWT supports instructors in order to create online courses with the tool.

The experiments were primarily interested in the functionality and usability of the tool. Moreover, second round of experiments aimed at validating the improvements made from the feedback collected in first round of experimentation and developed during the second phase of the project. The scenario goals were achieved according to [3].

In contrast to the previous experiments conducted at TUG, all the experiments at UOC were conducted by 3 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.

In the first experimentations, the VOE tool experimented serious technical problems that impeded to complete the experiments (i.e., the 2 participants had to give up before completing the experiment). From the feedback collected in the first phase of the project, key improvements were developed and incorporated in the tool. Then, the second round of experiments the participant running the experiment confirmed that the use of the tool was satisfactory and the VOE tool was validated for defining domain ontologies and contexts with a user friendly interface (G2.2.1).

Therefore, the tool did not experience any technical problem during the second round of experiments and could be completed, thus achieving the main goal (G2.2.2). This is in line with TUG site that could finalize the experience with success. This confirms that the technical problems faced in the initial experiments were sporadic and exceptional as no relevant technical problem was reported at the final experiments.

The analysis of the usability and emotions of the tool confirmed and were in line with the previous achievements in the tool improvements as in the first experiments the usability scored as low as 15 in the SUS scale (Grade F, in the bottom of 15%) while the second

experiment it scored up to 55, though still under the SUS mean score (68). As for emotions, in the initial experiments great amounts of anxiety and anger were found (i.e. the participants felt these emotions all the time) as well as low levels of happiness (none of the time). In contrast, the second experiment, the participant was happy most of the time, and feeling anxious and angry only some of the time.

All in all, the participants in both rounds of experiments 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 when using the tool and the learning curve to exploit its potential efficiently is rather high.

Finally, the participants were very helpful and active, and provided many hints and suggestions for improvements at different levels, being the most productive the technical level. This led to achieve the last goal of this scenario (G2.2.3).

11.2.3 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 (see [7]). Two rounds of experiments were on this scenario at UOC pilot site involving 231 students and 3 lecturers. The experiments were run from both student's and instructors' viewpoints.

11.2.3.1 Students' viewpoint

Two rounds of experiments were run at UOC by using the CLR with semantic connections [7] as an alternative course run in IWT to support a specific lesson of the official course. The following scenario goals were achieved according to [3].

In general the students liked the CLR tool with semantic connections and found it interesting to extend and go deeper in certain learning concepts by means of the semantic connections. Students got better marks when assessed of these concepts (G3.1.2). The recurrent comment on finding too few semantic connections confirmed positively that students liked this approach for their study. Also, the CLR were reported to be reproduced efficiently by students who could use them to find further information about these concepts (G3.1.3).

One of the most relevant results in both experiments was that most of students (70%) in both experiments 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 reinforced the achievement of G3.1.2 by providing students with the appropriate links and target information. 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 implicitly achieves also G3.1.3.

From the usability point of view, the goals were also achieved by providing CLRs with a friendly user interface (G3.1.1) and also it was noticeable the improvements made in the second stage of the project. In particular, a specific technical issue reported by many

students in the first round of experiments related to navigation problems, was not reported in the second round, thus considering the overall usability of the system satisfactory. However, the general view on usability of both rounds of experiments was the same and keep the SUS mean Grade the same or even lower in absolute terms (53.97 in the final experiments versus 60.78 in the initial experiments). The students participating in the second round of experiments indeed still reported a sense of disorientation when using external connections. To this direction, students commented on possible ways of improving further the utility of the CLR and semantic connections (G3.1.4).

Finally, motivation was found to be quite high and increasing from the first to the second experiments ($M=3.01$ to $M=3.83$ in the scale 0-5). As for the gaining of knowledge, the first round of experiments found that the group with IWT achieved significant better marks than the group without IWT (7.83 vs. 6.33 in the scale 0-10). Moreover, the improvements made in the prototype also impacted in the gain of knowledge but not significantly (7.81 vs. 7.32).

11.2.3.2 Instructors' viewpoint

One single experiment was conducted at the UOC pilot site on the instructor's viewpoint in order to test the CLR editor tool of IWT. The experiment was addressed mainly towards the functionality and usability of the tool and eventually to achieve the scenario goals according to [3]. A real expert in developing complex computer systems participated in the experience.

He lecturer liked the idea of personalizing a course by semantic connections that link different learning resources to fit the specific students' needs and different contexts. Depending on the topic to be taught, the lecturer thought that the lecturers and students could be benefitted from incorporating semantic connections between learning resources.

Despite having a strong background in web applications as developer and user, the participant did not find strong technical inconveniences with the editor tool nor in the IWT tools in general in R3 scenario. A more accurate analysis on the usability, the participant considered the CLR Editor tool was satisfactory (SUS score was 70, just above the SUS mean) and confirms from this perspective the tool is working very well. His emotions when using the tool were also in line with the level of satisfaction and usability, as he felt happy most of the time while none of the time he felt bad emotions. Therefore, the tool did not experience any major technical problem during the experiment and could be completed, thus achieving the main goal (G3.2.1).

Finally, the lecturer was very helpful and active, and provided some hints and suggestions for improvements at different levels. This leads to achieve the second goal of this scenario (G3.2.2).

11.2.4 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 to be experienced and played by learners (VCS). During the resource execution, learners observe how avatars discuss and collaborate, how discussion threads grow, and how knowledge is constructed, refined and consolidated.

In the first phase of the project, despite the VCS at that stage was fully functional and the development goals had been achieved, it was still far from offering the actual potential to be provided at the of the current stage. 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 are now available.

In the second phase of the project, a new type of technological resource called CC-LR [17] was introduced as learning material that leverages CC-LO and shows animated storyboards such that learners can interact with the CC-LR and observe the consequences of their decisions. To this end, an editor tool [17] was developed as a component of the VCS to augment the CC-LRs with author-generated information, thus showing the provision of complex aspects of the learning process in the CC-LRs. Specifically, cognitive assessment and emotional awareness were addressed and developed in the CC-LR. In addition, lecturers and tutors are provided with edition capabilities of the CC-LRs, such as cutting scenes, modifying involved characters, selecting emotional states, dialogues and connected concepts.

The notion and nature of the CC-LR was validated by running extensive tests on a proof of concept of the VCS system that embeds a CC-LR [4]. These validation activities were carried out following the same methodological procedures in the UOC pilot site with different perspectives and expectations. In general, the tests allowed for exploring how better to convey the underlying process and principles to novices, supporting them in developing their understanding of the use and application of CC-LO/SLO in the form of new learning resources (CC-LRs). All the feedback received by the testers was used to improve both the VCS prototype and in turn enhance the features of the sample of the CC-LR used for the later experimentation reported in this deliverable.

Therefore, the goals and hypotheses formulated for the above scenarios were experimented and validated at UOC in two rounds of experiments involving 5 trials, 257 students and 6 lecturers in all. The experiments were run from both student's and instructors' viewpoints and achieved the scenario goals according to [3].

- **From the student's viewpoint**, we followed an iterative approach driven by empirical phases. First phase set out a solid basis for the next experiments by evaluating the VCS with a CC-LO embedded evaluated in the first round of experiments whose improvements were evaluated in the final experiments. Second and third phases were run in the final experiments to evaluate the usability and functionality of the VCS tool to edit and play the current text-based discussion in a multimedia attractive format (CC-LR). To this end, within the second phase, an experiment was run to pilot this scenario from both the student' view in support for a formal in-class assignment of collaborative learning based on a discussion. Finally, in the third phase an experiment was run focused purposely on the cognitive and emotional aspects of the CC-LR as complex aspects.

- **From the instructor's viewpoint**, an experiment was run to validate the CC-LR approach and especially the Editor tool to create CC-LR from the usability and functionality as well as a valuable resource to support the teaching processes at UOC.

11.2.4.1 Students' viewpoint

In the first phase of experimentation regarding live and virtualized collaboration, an experiment was conducted at UOC pilot site in order to test the virtualization of live sessions of collaborative learning to produce storyboard learning objects (CC-LO) embedded in a virtualized collaborative session system (VCS) to be experienced and played by learners. During the resource execution, learners observed how avatars discuss and collaborate, how discussion threads grow, and how knowledge is constructed, refined and consolidated.

In the second phase, the experimentation was repeated in order to validate improvements of the VCS tools. In particular, the usability and functionality of the VCS tool to play and observe the 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).

In general, the students confirmed they liked the VCS tool and found it interesting to have another option to follow the in-class discussion assignments (G4.1.3). During this specific assignment, students could generate the storyboard from the VCS (G4.1.1) and it was effective to support the discussion for review and summary purposes (G4.1.5).

Most of students in both rounds of experiments could generate the storyboard (SLO) efficiently (G4.1.6) and create, store and playback it as many times as needed. During the experiments it was noticeable an increase in usability (SUS scores went up to 64.87) and emotions (being more happy and less anxious, sad and angry) when using the tools (G4.1.4).

Complex aspects of the learning process, such as motivation ($M=3.07$ in the scale 0-5) and the gain of knowledge were validated by showing an impact of the use of the VCS tool on these aspects. However, the gain in knowledge acquisition obtained by using the VCS in the two rounds of experiments was not significant. The VCS was proved to become a valuable educational resource by assessing several aspects of the learning process, such as knowledge construction and participation (G4.1.2).

Next experiments at UOC went one step further and validated the use of CC-LO as complex learning resources (CC-LR) [17] to be provided to students as regular learning material. A first trial in this context was experimented with the VCS-SLO Editor with limited capabilities to create simple CC-LRs by cutting scenes, modifying involved characters, etc. The second trial validated the incorporation of complex author information into the CC-LR, such as cognitive and emotional aspects.

From the results of the last trial on the last update of the prototypes, in general the students liked the use of CC-LR as learning material and its extended complex features since they helped them to understand better the contents (G4.3.2). In particular, the students

appreciated the test scenes (cognitive assessment) of the CC-LR for knowledge retention and construction. During the study with the CC-LR the students found them very easy to use and no relevant technical problems were reported, also from the usability perspective (SUS score was 69.27, above the SUS mean score and higher than the CC-LR without extended features). This achieved G4.3.4.

The majority of students could generate the video-debates efficiently (G4.4.6). Complex aspects of the learning process, such as motivation (M=3.21 in the scale 0-10) and emotional (feeling more times happy than bad emotions) were validated in absolute terms and also relatively from the CC-LR without complex extensions. These results impacted in the use of the CC-LR and made the learning process more effective. In particular, the CC-LR proved to become a useful educational resource (G4.3.5).

The gain in knowledge acquisition by using the new features though it was increased from the previous experiments (G4.3.2) were not significant, especially from the emotional features incorporated that were not appreciated by the students and hid part of the real benefits provided by the incorporation of cognitive aspects (test scenes and performance indicators).

Finally, students provided some hints to improve the video-debates and CC-LR in general (G4.3.3) as well as they suggested to use this type of learning resources in more courses and programs of the UOC.

11.2.4.2 Instructors' viewpoint

The aim of this scenario was to build a CC-LR as a learning material to support the collaborative pedagogical model of academic courses from the instructor's viewpoint. To this end, an experiment was conducted on this scenario at UOC pilot site in order to test the Editor tool and collect feedback from the instructors when creating and managing CC-LRs from CC-LO to provide new learning resources to students. This experiment at UOC was conducted by real experts in developing complex e-Learning systems.

The results of this study were analyzed to evaluate how the VCS-SLO Editor tool supported instructors and experts in order to create and manage CC-LR from CC-LO/SLO, the time spent in creating new CC-LRs as well as the problems and possible enhancements suggested.

The participants liked the idea of editing and personalizing each SLO in order to meet the specific requirements of the course and found it very beneficial for students. This achieved the main goal (G4.4.1) as for creating learning material (CC-LR) from a threaded discussion

From the analysis of the usability and emotions, the use of the Editor tool was considered very satisfactory (SUS score was as high as 77.5) and the participants felt happy most of the time while feeling sad and anxious none of the time, only anger was observed some of the time (G4.4.4).

As a result of the above, the tool performed very well, even if non-expert users (G4.4.2) and the participants could create new CC-LR efficiently (G4.4.6) and especially in an effective way (G4.4.5) as the material was based on students' contributions, thus having an important

impact in the learning process. However, some improvements on usability, even if minors were suggested (G4.4.3).

11.2.5 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 was created for providing suitable learning resources that meet the learners' needs.

Two rounds of experiments were performed at MOMA's school network, involving 6 schools, 100 students and 6 teachers in all. The following scenario goals were achieved according to [3].

From the usability and emotional perspective the results of the experiments validated the use of the Storytelling tool (SUS score was 60.25, this nearby the SUS mean score), feeling happy most of the time and the bad emotions were felt none or some of the times, thus achieving G5.3. Feedback from the initial experiments was considered for improvements (G5.7). In particular, some comments on usability as for the need to stop the flow of the storytelling for having brainstorming with the tutor.

As a valuable resource, it was found some steps of engagement in the Storytelling educational resource (G5.6). Interestingly, from these results it was found two different styles of resource use: on the one hand the tendency to the discovery and to the progressive approximation to the learning; on the other hand the tendency to multitasking and the preference to a cognitive moment.

Moreover, it was found that the storytelling learning resource can offer more variation than the traditional practicing methods. The experimentations confirm 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 easier the reflection. This achieves G5.4.

Finally, the teachers participated in a survey that helped validate the Storytelling resources from the instructor's view (G5.1 and G5.2). In particular, they agreed that the resource provide to the students with the opportunity to express their native style characterized by a progressive exploration of knowledge in a guided and structured context.

11.2.6 R6. A Serious Game for Civil Defense

The goal of this scenario is to allow an efficient learning about knowledge and behavior to be adopted in civil emergency situation (like seismic event in Amusement Park) through the use of complex and innovative learning resource (Serious Game or SG for short). As a result, an Emergency Course was created for providing suitable learning resources that meet the learner's needs.

To evaluate and validate the above aim, 2 rounds of experiments were performed at MOMA's school network, involving 6 schools, 100 students and 6 teachers in all. The following scenario goals were achieved according to [3].

From the usability and emotional perspective the results of the experiments validated the use of SG resources (SUS score was 61.29, this nearby the SUS mean score), feeling happy most of the time and the bad emotions were felt none or some of the times, thus achieving. These results confirmed the motivation of the students with the use of serious games as learning resource (G6.2) Feedback from the initial experiments was considered for improvements (G6.4).

As a valuable resource, a lot of students confirmed the sense of engagement in the experience with SG showing that students appreciated the immersive reality of the game. In particular, the interaction with the game by using the control devices shows as the students did not find particular problems with the new device and from the information obtained through different senses of the game, as vision, hearing, touch, have caught the students attention (G6.3).

A relevant drawback was found in the quality of the visual display, which could be justified taking into account that the PC used by each classroom had hardware performance not very high. Moreover, as a traditional LOs, the SG resources have not much advantages in a learning course related to the management risk.

Finally, by analyzing the survey given to the teachers, all of them agreed on the use of SG learning resource as improvement of students' understanding of key concepts, thus validating G6.5.

All in all, the main goal related to deploy SG resources within schools was validated (G6.1).

11.2.7 R7. Affective and Emotional Approaches

The goal of this scenario is to monitor the particular emotion taken by the student during his interaction with the complex learning resources. That in order to modify the learning experience if the emotional state is altered or not compliant with the assessment results.

Two rounds of experiments were performed at MOMA's school network, involving 6 schools, 100 students and 6 teachers in all. The following scenario goals were achieved according to [3].

From the usability and emotional perspective the results of the experiments validated the use of the Emotional tool (SUS score was 66.21, this nearby the SUS mean score), feeling happy most of the time and the bad emotions were felt none or some of the times, thus achieving G7.4. Feedback from the initial experiments was considered for improvements (G7.7).

As a valuable educational resource with the Emotional tool, the results showed that the student do not consider the emotional approach strictly learner center learning. Students in fact, although appreciated the consideration of their emotional state do not think that this factor alone could affect the results of experience teaching. On the other hand, the teachers indicated the tool gave them the possibility to enable or not different parameters in order to ameliorate the personalization of the learning pat. Hence G7.5 it was partially achieved.

Following, this, a relevant result was that the use of the Emotional tool was more useful for teachers than for students as it may help the instructional designer or the teacher to differentiate the learning path taking into account the different learning styles of the students (G7.2 and G7.3).

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

Over the course of the Alice Project, four studies were conducted to evaluate and improve the enhanced Wiki-tool for writing assignments. An overview of the studies is given in Table 9.1., which also lists the main goals of each study as well as the improvements done to the Wiki-tool after each study. Two rounds of experiments with several trials for each round were performed involving 82 students and 3 tutors in all.

Summarized, the first test of the co-Wiki took place in the first phase of experimentation with a sample of 18 computer science students. The tool was used as a stand-alone system. Comments from students and teachers led to an increase of performance for the assignment homepage and the edit page. Thereafter the Wiki was used again in the context of a self-regulated learning study (R9) reported in Section 10 of this report's first version [5]. In the second phase of experimentation three studies were conducted, which differed in their settings (business vs. computer science vs. psychology course, home assignments vs. controlled environment, stand-alone vs. integrated system) and individual goals. The latter started with a first test of the improved functionality, followed by the recording and analysis of behavioural data (log data), and the full integration into the IWT. After each study, various functions of the Wiki were changed or added, and thus tested thereafter for the first time. Due to the extensive amount of collected data, in this Section we will focus on the goals outlined in [3] and in which way they could be reached during this project.

One main goal was to provide a Wiki system that allows an efficient and user-friendly management (G8.1). Over the course of the studies, usability was measured in various ways, namely in terms of satisfaction, efficiency, and effectiveness. Regarding the latter two, we found that students who spent more time in the Wiki (working time as well as editing time) got better grades from their peers and that grades (from teacher and peers) were generally high when students did their contributions using the co-Wiki. However, we did not find a correlation with satisfaction. Data about satisfaction with the tool was collected in all four studies by means of SUS [6] and task awareness. In all studies the obtained SUS scores were below the reference value of 68 suggested by Brooke and even decreased over the course of the studies, whereas most task awareness ratings were above average and increased over time. From the ratings to single questions and participants' open comments we can infer that the low SUS scores are mainly due to technical problems such as slow performance of the system, too many auto-logouts, or suboptimal presentation of some graphs. The single features of the Wiki, on the other, such as the actions feed, the contribution graphs, or the enhanced coloured difference tool were rated as useful functions supporting the students in their learning activities. Feedback from instructors and tutors is on the same line – low SUS scores due to technical problems, but good evaluation of the co-Wiki's features. Thus, future work should focus on eliminating the performance problems, further enhancing the interface, and optimizing some of the graphs (G8.3). An increase of

usability as it is measured by SUS is also important in the light of the findings, that low SUS scores are highly correlated with high anger, anxiety, and sadness as well as lower happiness scores. The usefulness of action and contribution graphs was confirmed by teacher and students, and also that these features can give students a good overview of their learning progress. In all studies, students mentioned that it was helpful for them to see, who had contributed and how much (G8.4).

Since the Wiki was designed for supporting collaborative writing, it is important to have a closer look at this aspect (G9.2). Log data from the last two studies showed that students in the ISR study worked on average 1040 minutes in the Wiki (editing text about 246 minutes), whereas the assignment in the psychology course lead to an average of 367 minutes working and 76 minutes editing time. Hence, students were definitely using the wiki and thus working collaboratively. Questionnaire data also implies that the features provided in the wiki supported them in working together and also motivated them to contribute more to the group product. With respect to motivation, it was also investigated, whether the peer assignment function motivates students concerning their learning activity (G8.5). Whereas peer-reviews were mandatory in the first study, students were not forced to review their peer's work after each change in the remaining studies. This change of function was done because of student complains regarding the permanent task of reviewing even little changes such as format edits. However, making the peer-review voluntarily, lead to a strong neglect of this function. Most students never performed a peer-review, although the function was rated to be a helpful resource. We assume that the peer-review function was confused with the group-review function or group-assessment which was performed by all participants. The group-assessments were rated to be a good support to learn more about other groups' contributions and to get new ideas for one's own work. Thus, this feature turned out to be a very valuable function not only for the persons receiving the group-review, but also for those performing it. Additionally, group-assessment ratings for the last study increased as compared to the previous one. Summarized, results from across the studies indicate that peer- as well as group-reviews support students in their learning process (G8.6). The technical problems faced in the last study could for the most part be resolved before the end of the study.

Finally, the developed tool was also meant to facilitate the work of teachers (G8.7). Feedback from the instructors and their tutors could only be collected in phase two of experimentation, but the instructors of those three studies all stated that they liked the tool. Furthermore, their comments confirmed that the co-writing Wiki supports teachers in tracking and assessing students' contributions and that the group assessment form was very helpful for grading the students.

Summarized, it can be concluded that the features of the co-writing Wiki, such as the actions feed, the contribution graphs, or the assessment rubrics are a very valuable instrument for students working on a collaborative assignment as well as teachers supervising their work. However, the functionality in terms of power and at some points design has to be further improved in order to be applied in broader contexts.

11.2.9 R9. Assessment in Self-Regulated Learning

One aim of the ALICE project was to provide a tool that supports students in self-regulating learning environments by allowing them to assess their knowledge state at any time and to any topic they want to. Thus, the automatic questions creator AQC was developed, which automatically generates knowledge questions from a given text. The provided questions types include open end (OE), fill-in-the blank (FiB), multiple choice (MC), and true/false (TF) questions. In order to create a question, the AQC first extracts concepts out of the text. The questions are then build around the concepts, which constitute in most cases the correct answer to the questions. Besides extracting relevant concepts and creating meaningful questions around them, another challenge is to find alternative answers/statements for multiple choice and TF questions, which do not allow a simple elimination strategy (e.g. by grammatical errors) to get the correct answer. With respect to the basic goal G9.1 to provide a tool that generates different types of questions, we can clearly say that this goal could be reached.

Over the course of the project, we conducted four studies to evaluate and improve the quality of the AQC involving 71 students in all (including 12 students participating in a pre-study). In the first round of experimentation, the quality of concepts, questions, and answers was tested in two pre-studies (R9.0a and R9.0b), whereas the tool's ability to support students in their learning process was investigated together with the co-Wiki with PhD students enrolled in a life-long learning course as sample (R9.1). The samples for the pre-studies were 29 Computer Science and 8 PhD students and post docs. Finally, in phase two of experimentation, another sample of 30 students enrolled in an online course for an e-learning master evaluated the AQC after its integration into the IWT system (R9.2). Between the two phases of experimentation the tool was improved by implementing functions to add concepts (out of a comprehensive list of words and phrases from the text) and to tag concepts by highlighting and saving them. Additionally, the integration into the IWT was done between the two phases.

The first quality tests were performed by means of the two pre-studies R.0a and R.0b, in which participants had the tasks to extract concepts out of a text on NLP (Natural Language Processing), create questions out these concepts, answer knowledge questions created by the AQC and the experimenters, and to evaluate concepts and questions created by the AQC (based on either AQC or student concepts) and the experimenters. Since high quality concepts are a natural prerequisite for high quality questions, we first want to look at G9.4 which concerns the quality of concepts. One main result of the pre-studies is that the mean relevancy of 7 concepts extracted by the experimenters is higher than that of 49 AQC concepts. However, comparisons with only the seven most relevant concepts extracted by the AQC lead to inconsistent results. Whereas the 29 students from the first pre-study rated them as equally relevant as human concepts, the 8 participants of the second pre-study evaluated human concepts better. Thus, in the last study (R9.2), students were again asked to evaluate the relevancy of concepts, but this time we extracted concepts from two different texts (problem based and project based learning) provided an equal number of AQC and human concepts (per text 10 each), and had the actual instructor of the course select the texts and extract the concepts. Furthermore, concepts were extracted by students and

evaluated by their peers. For both texts, results showed no difference between AQC and human concepts (neither for teacher nor for student concepts). Thus, concepts extracted by the AQC are as relevant as human concepts, which is a clear indication for the achievement of G9.4.

With respect to the goals G9.2, G9.3, and G9.5 a closer look at the quality of questions is necessary. The evaluation of questions in R9.0a was based on 20 AQC questions and 5 questions created by the experimenters for each questions type (OE, TF, MC, FiB). In R9.0b participants evaluated 10 human questions and 20 AQC questions, of which one half was based on AQC and the other half on human concepts. Scenario criteria in both studies concerned the pertinence, terminology, and level for questions (all types), the quality of the answers (OE, MC, FiB) and the distractors (MC). The overall results show that students perceive the quality of the questions as rather good (on 5 pt. rating scales, all means are higher than 3). Comparisons with the human questions give a mixed picture, as there is no effect for Fib question, while for TF questions the AQC ones got higher ratings and for OE and MC questions the ones created by human have been evaluated better. For the pre-studies, human questions have not been based on concepts extracted by the AQC, thus it is not possible to judge in which step of the question creation process the AQC performs worse than humans. In R9.2 we therefore used a balanced design and also added some evaluation criteria (perceived difficulty for questions, terminology and ambiguity for answers) to get a more differentiated picture. Furthermore, the actual difficulty of questions was measured by providing knowledge tests to the students. Results showed no effect of who extracted the concepts (AQC vs. teacher) and for which type a question was created (TF, MC, FiB), but for the question creator regarding the pertinence and level of questions. However, partial η^2 values of .033 and below indicate only small effect sizes. Terminology and perceived difficulty are rated equally to teacher questions. There is also no difference between AQC and teacher questions regarding the terminology, ambiguity, and distractor quality of answers. Regarding the actual difficulty of questions, the number of correctly solved questions did not differ between AQC and teacher questions. Summarized, the results indicate that quality of questions as it is addressed in G9.2 and G9.3 is high and for sure sufficient for self-assessments. Although the difficulty does not differ from that of questions created by teachers, the lower ratings of pertinence and level (i.e. relevancy and meaningfulness with respect to the tested topic) suggest that some more improvements are desirable before applying the AQC for real teacher assessments with grading. Considering G9.5 it has also been shown, that the AQC is able to create questions based on concepts that are entered by users.

Goal G9.6 addresses the usability of the tool, which was measured by means of the System Usability Scale (SUS) by [6]. The average SUS score obtained in study R9.1 amounts to 66.88, the average student SUS score from study R9.2 is 57.59 and the teacher score is 48.13. Thus, the user-friendliness of the tool decreased from R9.1 to R9.2 and was even worse from a teacher's point of view. The decrease might be attributable to the fact, that the latter study was the first one after integration of the tool into the IWT. Thus, all encountered bugs had to be solved while students have already started to work with the tool. Suggestions for possible improvements (G9.7) concern for the main part the tool's usability (see Section

10.3.4 for more details) whereas the main functions and idea behind the tool are perceived as a very valuable resource by students and teachers.

Finally, the studies aimed at providing a tool that motivates students in their learning activities (G9.8) and that can be used for assessments for self-regulating learning (G9.9). Participants from study R9.1 were intrinsically motivated and stated that testing themselves with the AQC also motivated and supported them in their learning activities. R9.2 participants also commented that they liked self-regulated learning environments, automatic questions generation and assessment, and that testing themselves supports their learning process. Furthermore, they showed high task value ratings and were highly motivated to perform different tasks. Additionally, students and the teacher indicated that the tool constitutes a support for students and that they can benefit from the self-assessments. Thus, the AQC seems to be a worthwhile instrument for assessments in self-regulated learning environments.

References

- [1] Intelligent Web Teacher (IWT) web site; <http://iwtalice.cmpa.unisa.it/IWT>
- [2] 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.
- [3] 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.
- [4] ALICE (Adaptive Learning Via Intuitive/Interactive, Collaborative And Emotional Systems) project, Deliverable D3.4.2 “Sample Collaborative Complex Learning Object v2”, project co-funded by the European Commission within the 7th Framework Programme (2007-2013), n. 257639, 2010.
- [5] Bortz, J. & Döring, N. (2006). *Forschungsmethoden und Evaluation*. Heidelberg: Springer.
- [6] ALICE (Adaptive Learning Via Intuitive/Interactive, Collaborative And Emotional Systems) project, Deliverable D8.1.1 “Initial Experimentation and Evaluation Results”, project co-funded by the European Commission within the 7th Framework Programme (2007-2013), n. 257639, 2010.
- [7] ALICE (Adaptive Learning Via Intuitive/Interactive, Collaborative And Emotional Systems) project, Deliverable D7.4.2 “Prototype Components for Adaptive e-Learning v2”, project co-funded by the European Commission within the 7th Framework Programme (2007-2013), n. 257639, 2010.
- [8] Brooke, J. (1996). SUS: A “quick and dirty” usability scale. In *Usability evaluation in industry*. London: Taylor & Francis
- [9] Kay, R.H., & Loverock, S. (2008). Assessing emotions related to learning new software: The computer emotion scale. *Computers in Human Behavior*. 24, 1605-1623.
- [10] 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.

- [11] 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.
- [12] Wild, K.-P. (2000). *Lernstrategien im Studium. Strukturen und Bedingungen*. Münster: Waxmann.
- [13] Knussen and McQueen (2006): *Introduction to Research Methods and Statistics in Psychology*. Harlow: Pearson Education Limited.
- [14] Morris, L.V., Finnegan, C., & Wu, S. (2005). Tracking student behavior, persistence, and achievement in online courses. *Internet and Higher Education* 8, 221-231.
- [15] Frøkjær, E., Hertzum, M. & Hornbæk, K. (2000). Measuring Usability: Are Effectiveness, Efficiency, and Satisfaction Really Correlated? *Proceedings of the CHI 2000 Conference on Human factors in computing systems*, The Hague, The Netherlands, April 1-6, 2000, 345-352.
- [16] Wild, K.-P., Schiefele, U., & Winteler, A. (1994). Das Inventar zur Erfassung von Lernstrategien im Studium (LIST) [Inventory for learning Strategies in academic Studies]. In A Krapp (Ed.), *Arbeiten zur empirischen Pädagogik und pädagogischen Psychologie* (Bd. 20). [Studies on empirical Pedagogy and pedagogical Psychology] Neubiberg: Gelbe Reihe.
- [17] ALICE (Adaptive Learning Via Intuitive/Interactive, Collaborative And Emotional Systems) project, Deliverable D3.3.2 "D3.3.2: Prototype Components for Creation and Execution of Collaborative Complex Learning Objects v2", project co-funded by the European Commission within the 7th Framework Programme (2007-2013), n. 257639, 2010.

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 sections, 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 131: 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:

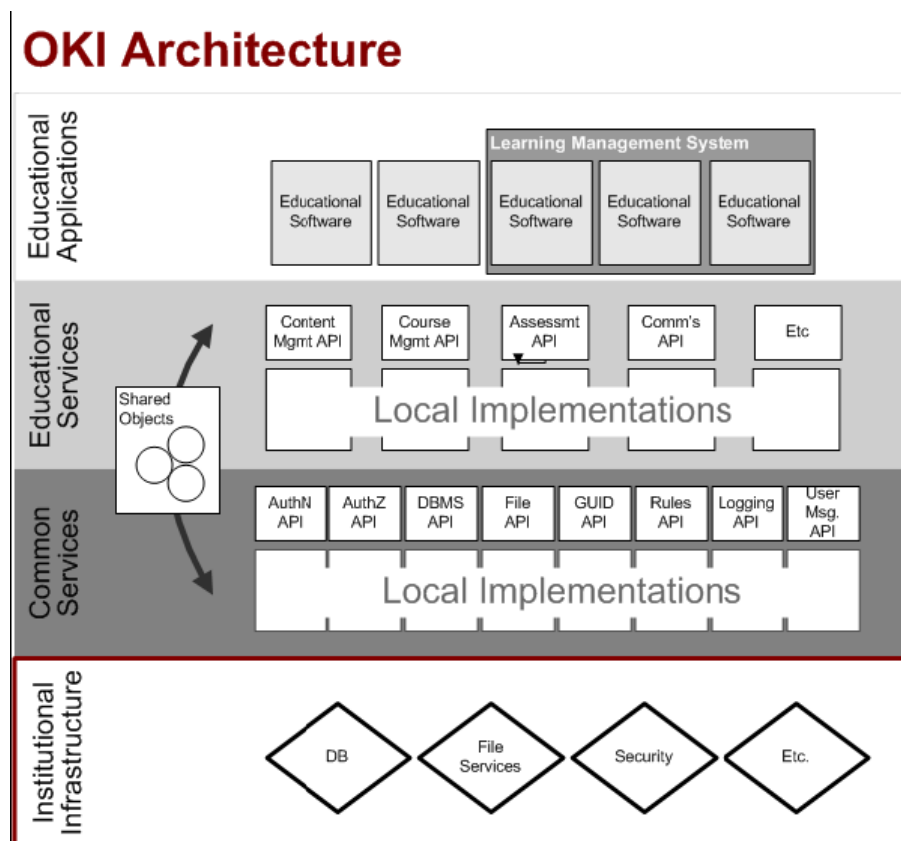


Figure 132: Tiers of the OKI architecture

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:

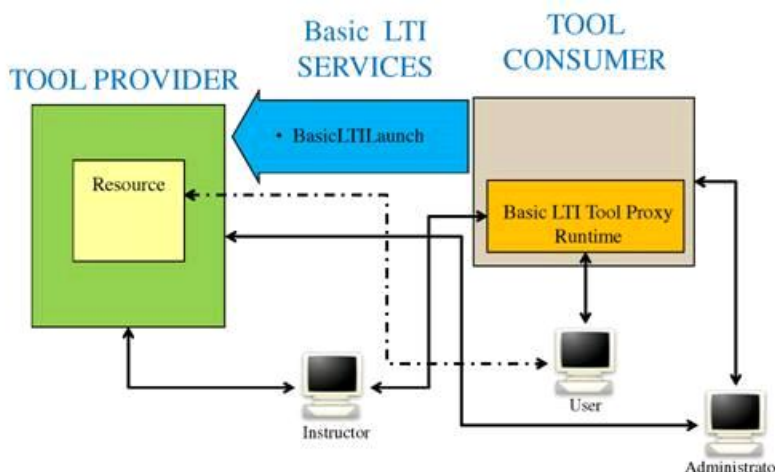


Figure 133: Overview of Basic LTI

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 section 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 134 and Figure 135 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

[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 134: The customized web portal before the login

The screenshot shows the ALICE web portal interface after a user has logged in. The layout is organized into several sections:

- Personalise your home page:** A callout pointing to the top right corner where users can customize their view.
- Experimentation Welcome Message:** A callout pointing to the central banner area that welcomes users to the project's experimental portal.
- Collaboration tools:** A callout pointing to the 'Comunica/collabora' section on the right, which includes options for communication and collaboration.
- My Groups:** A callout pointing to the 'I miei gruppi' section on the left, which lists the user's groups.
- My Classes:** A callout pointing to the 'Le mie classi' section at the bottom, which displays the user's current classes.
- FAQ, Links and Download:** A callout pointing to the 'Utilità' section on the right, which provides quick access to frequently asked questions, links, and downloadable resources.

The main content area features a central message: "Benvenuti nel Portale per la Sperimentazione del Progetto di ricerca ALICE". Below this, it outlines the project's objectives and lists several scenarios for experimentation, such as fire safety training and emotional state assessment during online courses.

Figure 135: The customized web portal after the login

A3 Integration at TUG site

A3.1 Co-Writing Wiki

Login to Co-Writing Wiki

Teacher and students can access to the co-wiki pages directly through the IWT Co-Wiki Service instantiated for each Class or group in a Class. Through a SSO mechanism the user is logged automatically in the Co-wiki system and redirected to the right page taking into account the user id, the group he entered and his role.

Screw Turn system will maintain and use the users and group information through the AliceSSOProvider which save the data on the IWT DB. As depicted in Figure 1, users can access the Wiki via a Web application which opens a special SSOLogin page in the Wiki. This page just redirects the request with all necessary parameters like the corresponding username, the role and the user group to the auto login mechanism of the Wiki. This mechanism uses the implemented user provider to fetch the user data from the calling Web application which also logs the user into the Wiki. It is important to say that all further accesses to the user data is also done via this provider because the Wiki should only save Wiki related data such as namespaces, pages and permissions and no user related data such as accounts and user groups. Another important thing happens on the SSOLogin page in step 2 where the Authenticate method is called. This method checks if the Web application that accesses the Wiki is allowed to do so using the OAuth standard.

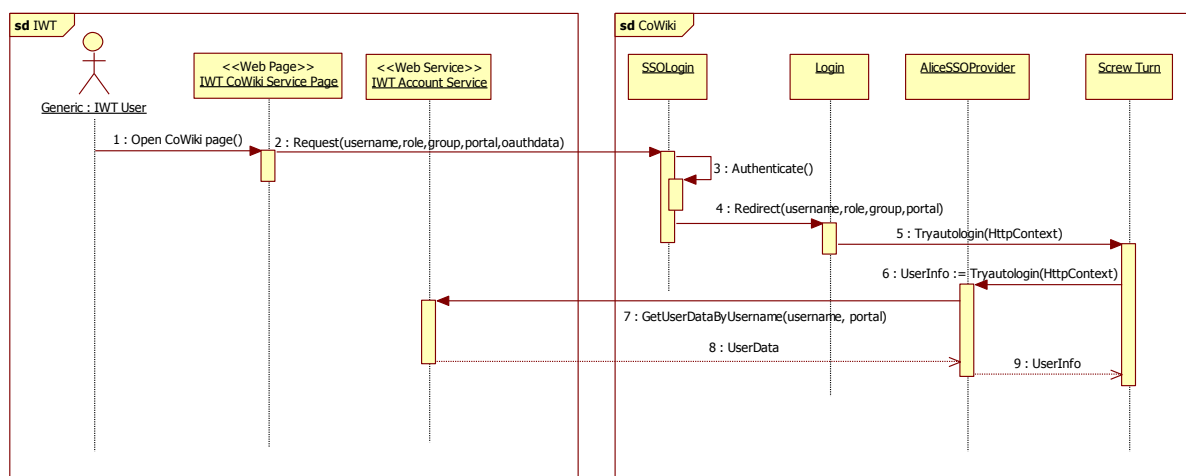


Figure 136. Sequence diagram of the SSO login process.

The ACL (Access Control List) engine handles the whole permission system of ScrewTurn (see Figure 136). It is used to save an ACL entry to the database and to check if a *subject* has the necessary permissions to perform an *operation* on a *resource*. In ScrewTurn terms, a subject can be a user group or a single user, an operation can be an activity such as read or write and a resource can be a page, a namespace, an upload directory or a global resource.

Figure 137 shows the possible operations that can be granted or denied for a namespace. This is similar to the ones of pages and global resources. (ScrewTurn, 2012)

Action	Description
Full Control	Full control on the namespace
Read Pages	Read pages
Modify Pages	Edit pages
Create Pages	Create new pages
Delete Pages	Delete, Rename pages
Manage Pages	Create, Edit, Delete, Rename pages
Read Page Discussions	Read page discussions
Post Messages in Page Discussions	Post messages in page discussions
Manage Page Discussions	Edit, Delete other users' messages in page discussions
Manage Categories	Modify category bindings of pages, create and delete categories
Download Attachments	Download page attachments
Upload Attachments	Upload page attachments
Delete Attachments	Delete, Rename page attachments

Figure 137. ACL operations of ScrewTurn Wiki for namespaces (ScrewTurn, 2012).

Teacher Use Case

Within IWT, teachers can create a new assignment or edit an existing one and enter in the Co-Wiki in order to check the activities of the students and give a feedback (see Figure 138). The teacher can also manage groups and assign them to the assignment running on Co-writing Wiki as depicted in Figure 139.

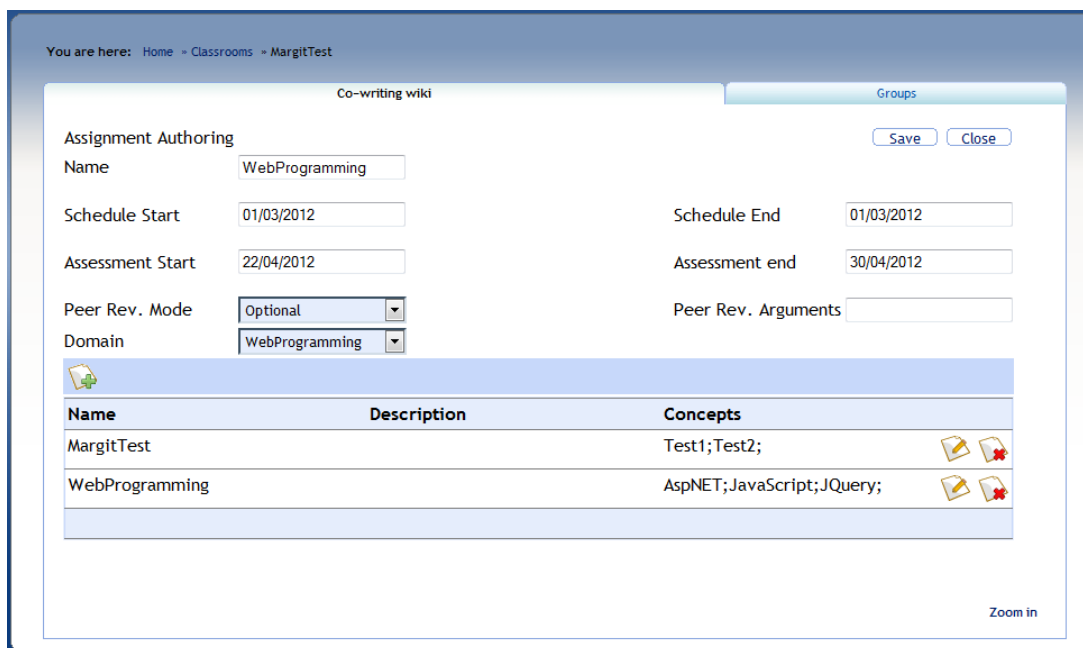


Figure 138. Authoring an assignment using Co-writing Wiki from IWT (SSO based)

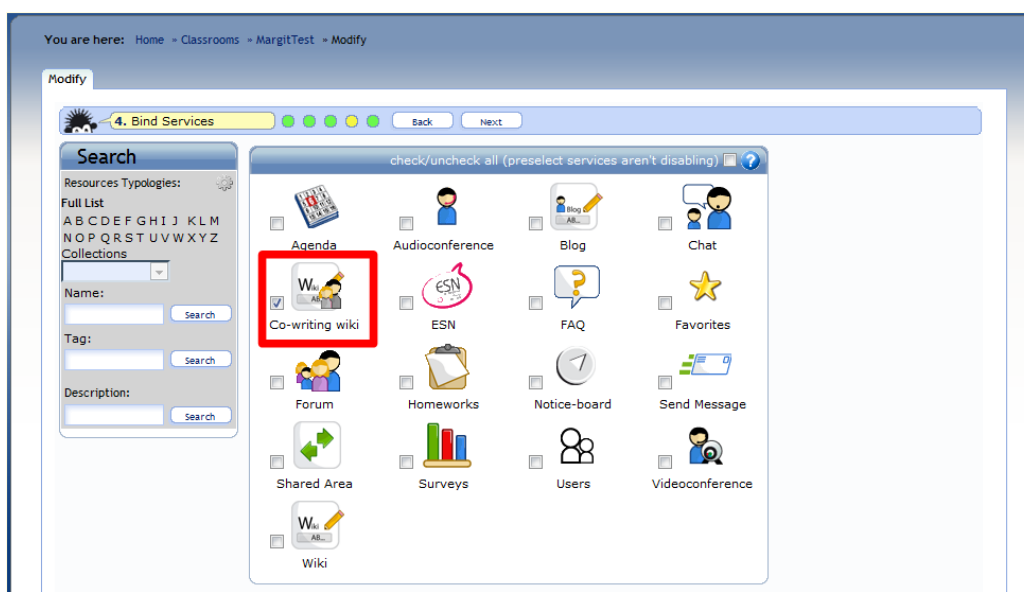


Figure 139. Assigning groups of students to Co-writing Wiki

Create Assignment

Teachers can create new assignments and assign a topic to each group in an assignment directly from the IWT Co-Wiki Service page. In order to achieve this goal, a Web service interface has been implemented and published that offers CRUD operations for assignments, namespaces and ACL permissions. To ensure consistency, it is essential that user group names in the LMS are directly mapped to namespace names in the Wiki and that teachers

are part of a user group called “*Teachers*”. The sequence of this communication can be seen in Figure 140.

In IWT the teacher can divide the class group into a number of subgroups, entering a subgroup area and the Co-wiki service page associated, he can set (create or edit) the assignment for that group. The Web application first checks if an assignment for the currently selected user group exists and creates a new one if not. The *AssignmentDetails* structure holds all assignment related data like the scheduling dates and an optional competence dictionary which will be further explained in the next subsection. After an assignment has been created, user groups can be added to it using the *AddUserGroupToAssignment* method. This method is the most complex one because it needs to create the namespace for the corresponding group, its main page and all permissions. Furthermore, its method signature also contains the username of a teacher because it could be possible that each group should be assessed by another teacher or tutor. The Web service interface also contains edit and delete-methods for all the data that is not shown in Figure 5.

Through the ScrewTurn API and the IWT Co-Wiki Service page, IWT can create assignments, configure them and add or remove groups inside them. For each group a namespace and a Main page will be created in the Co-writing Wiki DB, the name of the group in practice is the topic associated by the teacher.

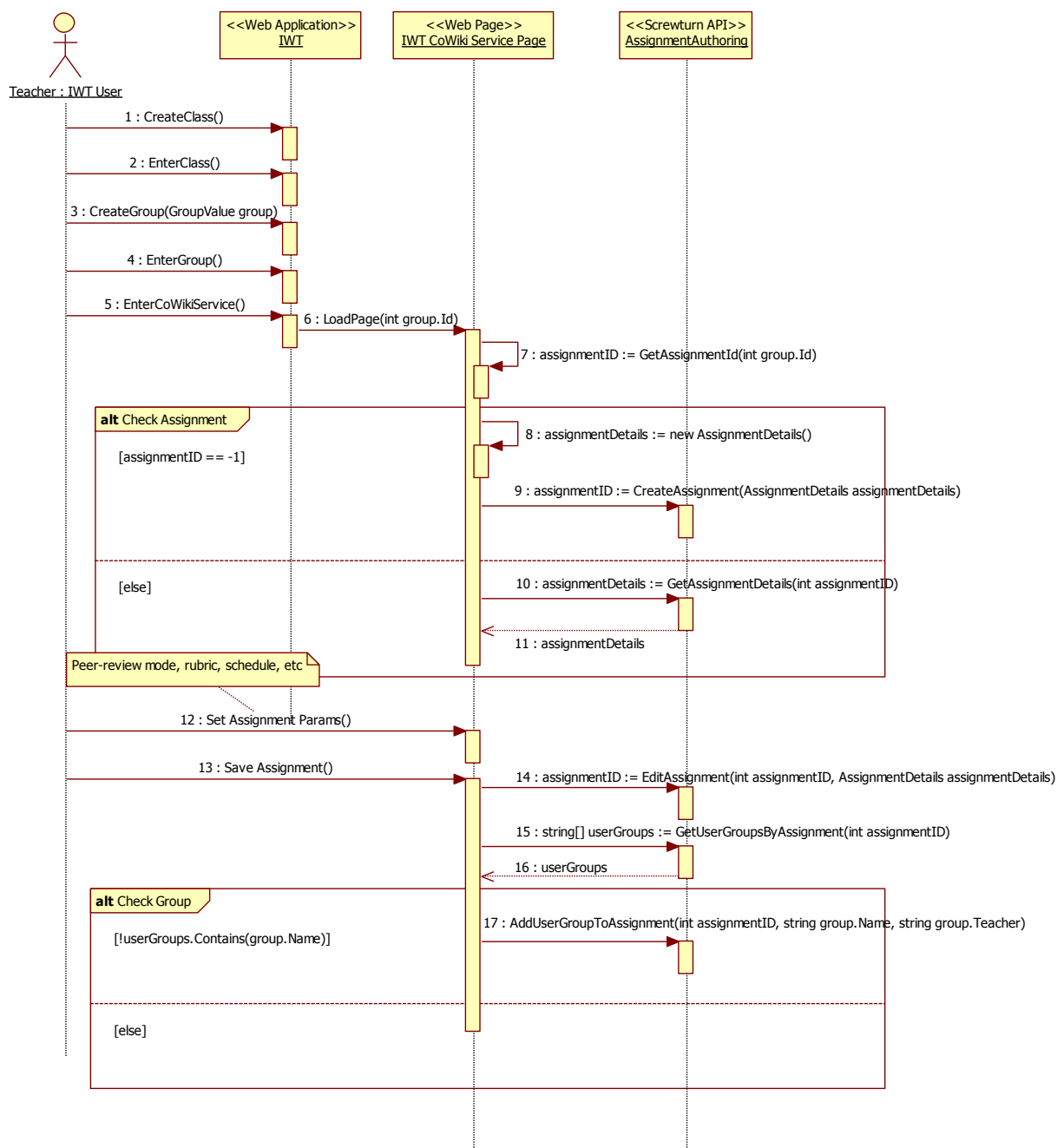


Figure 140. Sequence diagram of the assignment authoring process.

Edit Assignment

In edit mode a Teacher can change the configuration of an assignment, changing so the peer-review mode, the rubric or the assessment scheduling. All the groups in one assignment share the same configuration (see Figure 141).

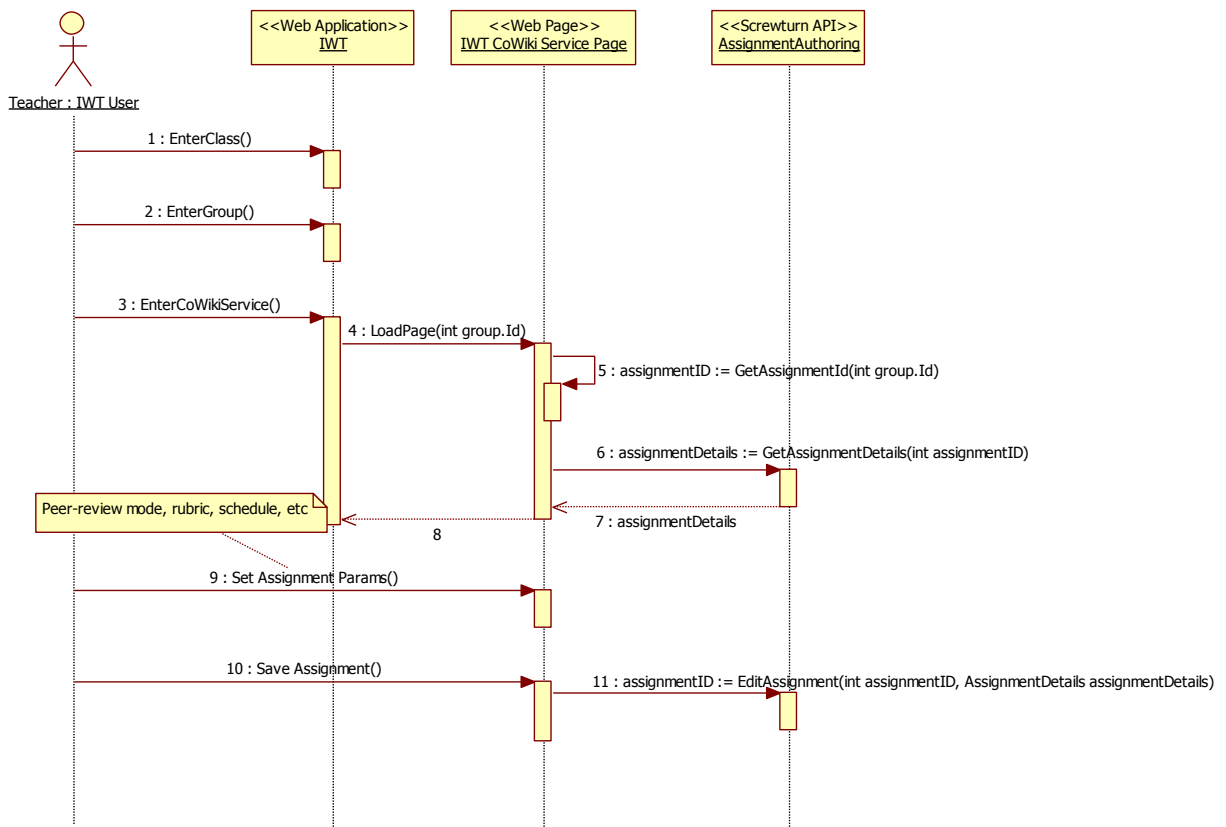


Figure 141. Sequence diagram of the assignment editing process.

Delete Assignment

When a Teacher enters in a class group and chooses to delete the binding to the assignment the UserGroup is removed from the Assignment in the Co-writing Wiki system (see Figure 142). If explicitly specified by Teacher and the Assignment has no group associated the entire Assignment can be deleted.

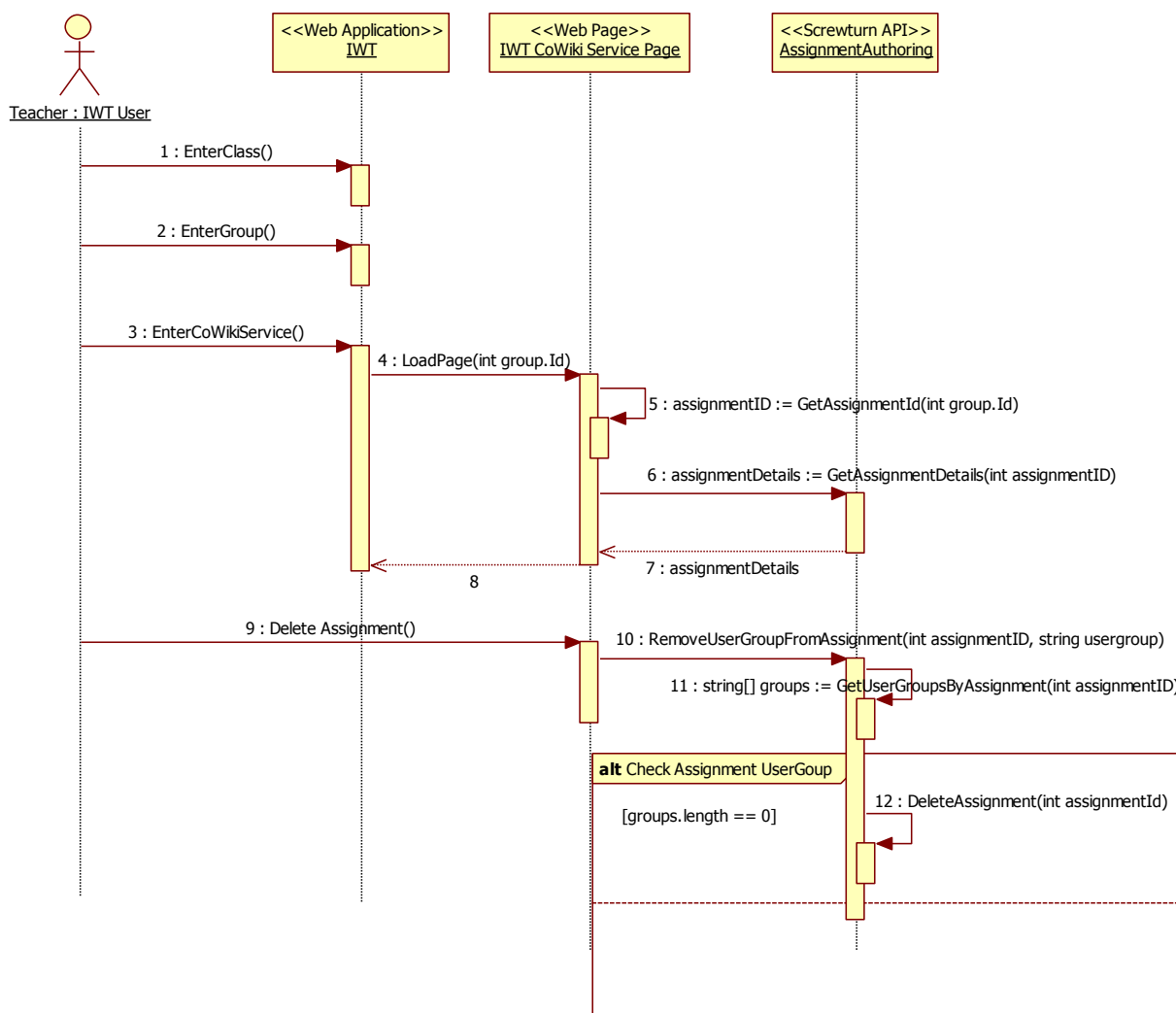


Figure 142. Sequence diagram of the assignment deletion process.

Give a grade to Students

In this use case a Teacher opens the Co-writing Wiki page for giving a grade to the UserGroup students. Assuming that for each group a set of concepts have been previously bound, the Teacher, after having checked the group contributions, have to update for each student the Cognitive State on those concepts accordingly (see Figure 143).

In this use case for the IWT integration, a teacher opens the grading schema page for giving summative assessments based on the assigned concepts. Thus, this page interacts with two Web services to receive the current Knowledge state - i.e. *IWT_KnowledgeModelService* - and the learners' preferences via *IWT_LearnerModelService*. The *GetUserCognitiveState(string learnerUserName, int portalId)* method is used to get the actual cognitive state of a Student; the groups' usernames are known after the has been obtained through the previous *GetUserGroups()* invocation and the portalId has been hold in Session during the SSO (see the SSO sequence). The Teacher set a grade for the concepts bound to

that group and the method `SetUserCognitiveState(int learnerID, TaxonLevelValue[] txLevList)` is invoked to update the entire user cognitive state.

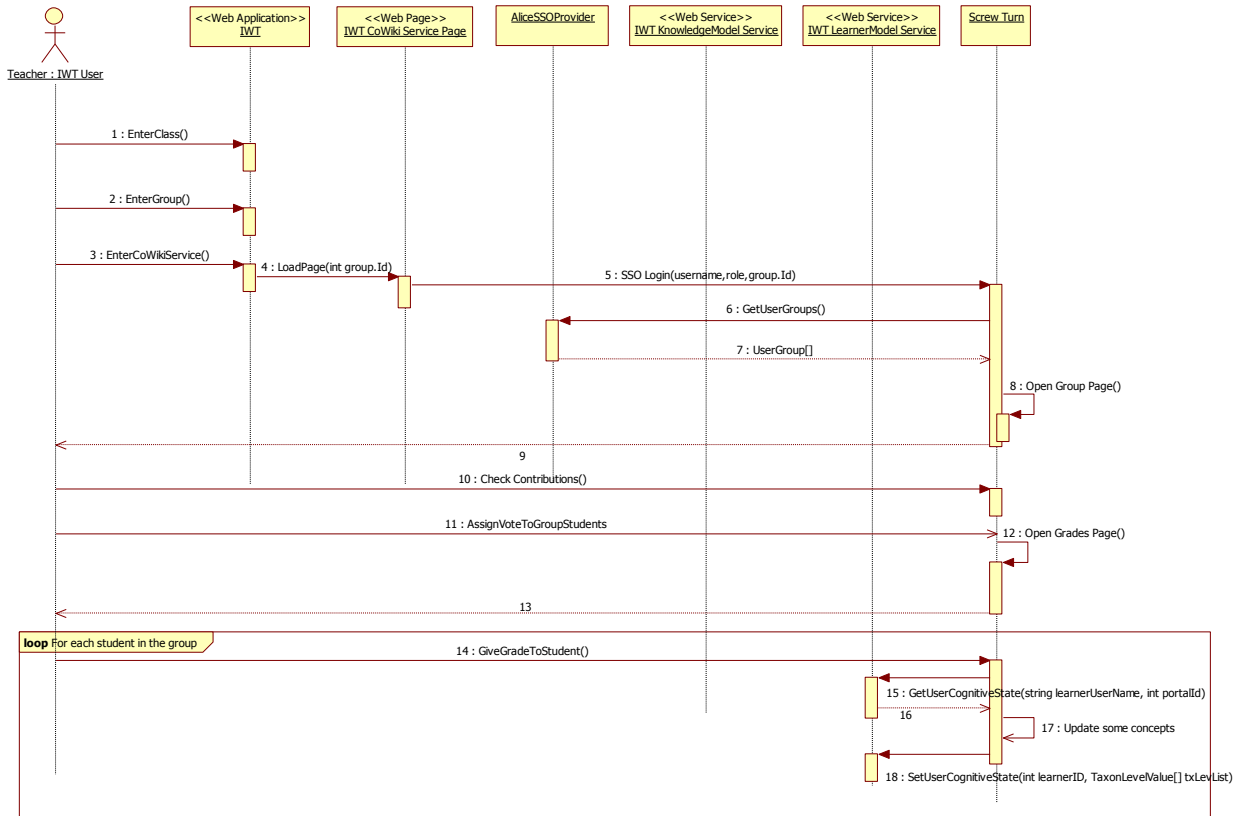


Figure 143. Sequence diagram of the grading process

Student Use Case

Students can enter in the co-wiki environment directly from their Class group area in IWT and participate to the assignment task as usual.

Groups and Usage of Co-writing Wiki

Once a Student enters in the Co-Wiki service page of his group, is automatically logged and redirected to the main page of his group within the assignment (see Figure 144).

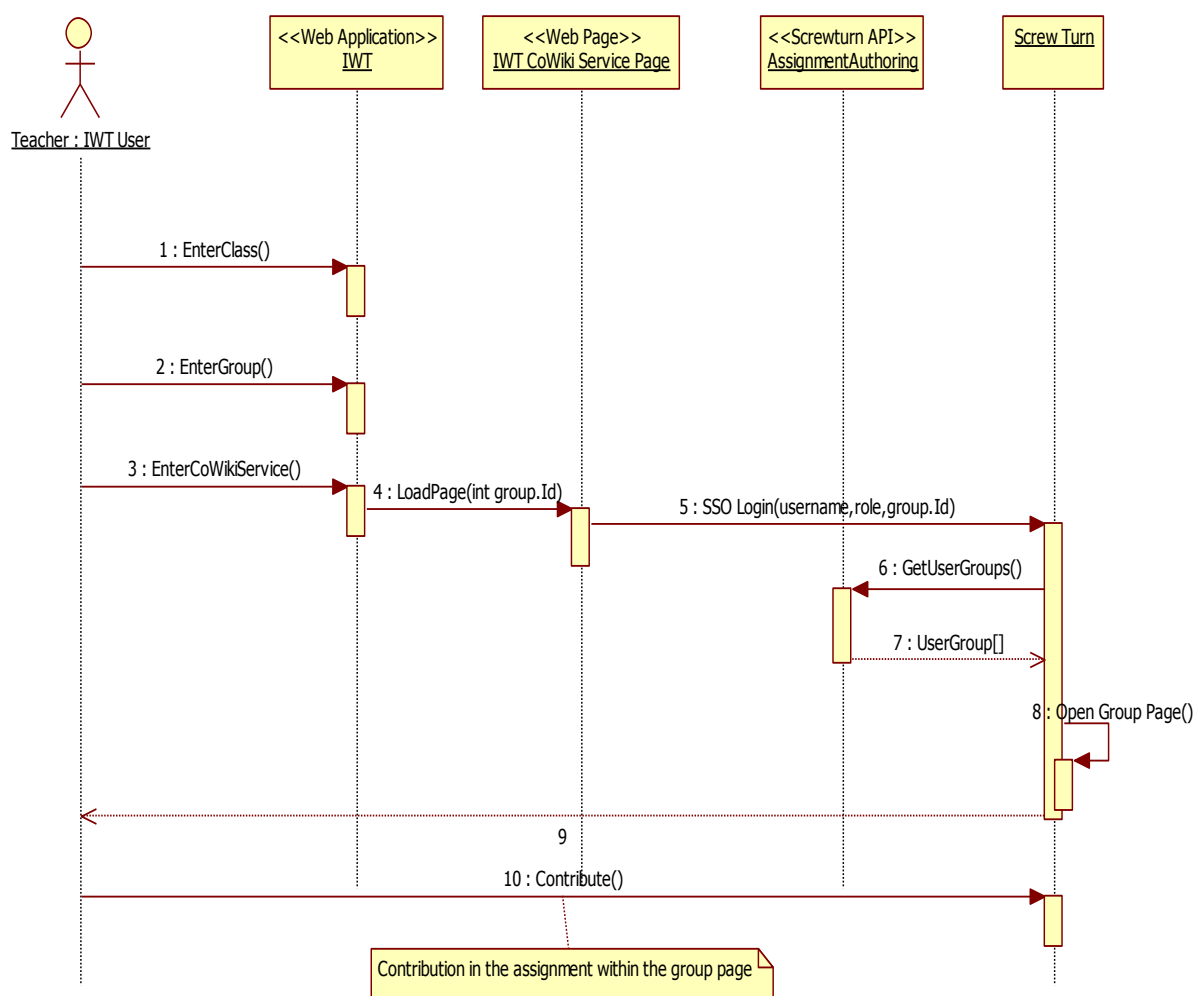


Figure 144. Sequence diagram of student use of Co-writing Wiki

A3.2 Self-regulated Learning

In a self-regulated learning approach, the learner chooses by-itself the concepts he want to acquire, providing to the system some keywords and selecting the proposed material. The system can help the learner in searching the learning content and providing some assessment. On the chosen material in fact the system can automatically assemble an assessment; a question creator tool is used for this purpose.

The *Automatic Question Creator (AQC)* tool can support in the creation of test items or even generates them automatically from the learning content. AQC utilizes an automated process to create different types of test items out of textual learning content, more precisely to create single choice, multiple-choice, completion exercises and open ended questions. AQC is capable to process learning content stored in various file formats, extracts most important content and related concepts, creates different types of test items and reference answers, as well as exports those items in QTI format.

For the sake of the integration of AQC with IWT, a web service (*Question Manager Service*) has been implemented in order to save in IWT the questions created by the AQC (See Figure 145 for the integration sequence diagram). The idea is that IWT starts the creation process invoking the AQC and passing it the learning object to use to extract the questions (step 2 of the sequence).

After having generated the questions the AQC returns them to IWT invoking the *IWT Question Manager Service* methods in the steps 6, 8 and 10 of the sequence.

Finally IWT will generate automatically a new test (Self-regulated test) basing on the received questions (step 13 of the sequence).

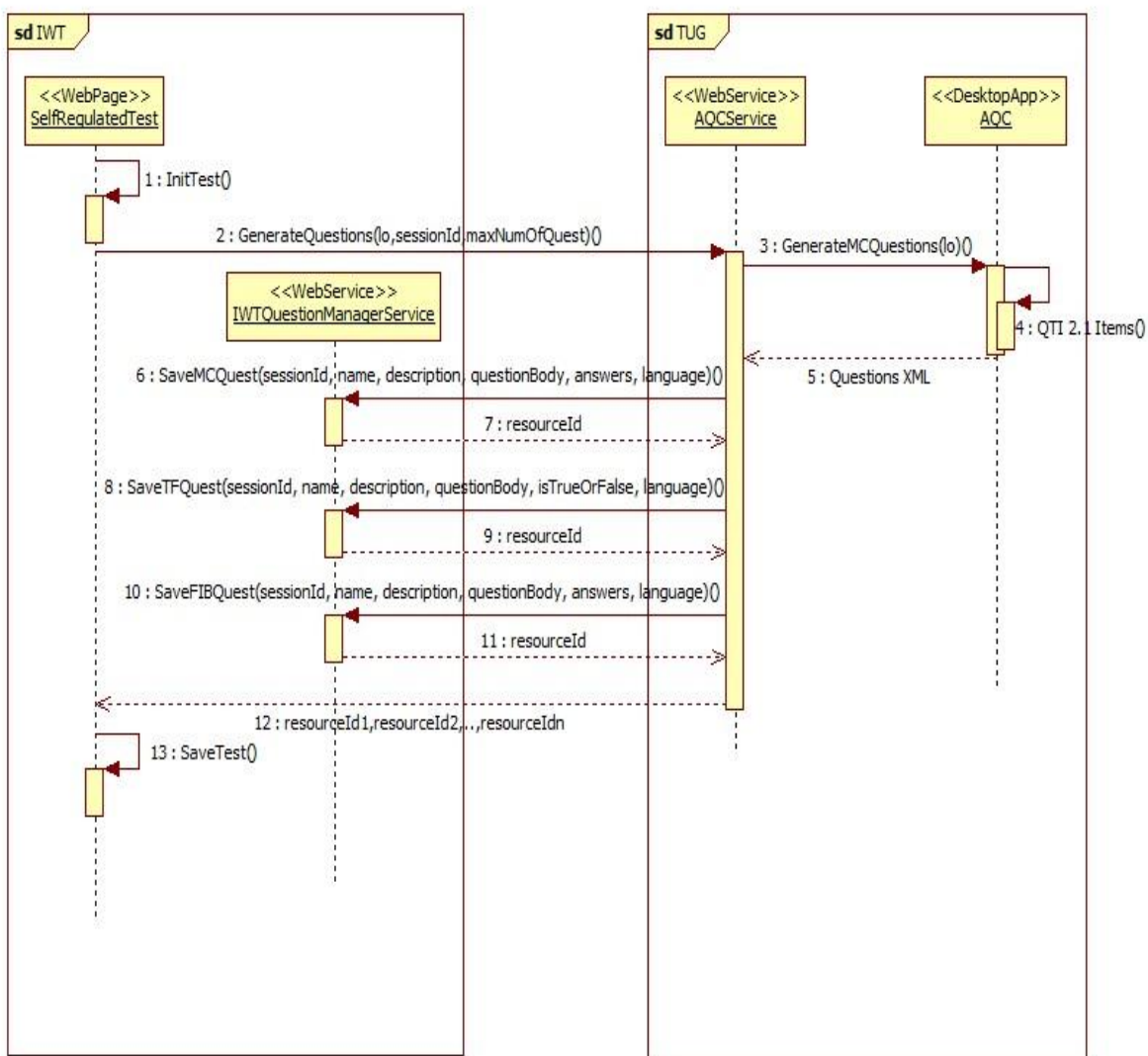


Figure 145 - Cooperation between IWT and AQC

A3.3 Integrated Assessment for Serious Games

Technical Aspects

Since the security model of the used Unity Web Player does not allow to open network communication channels (in contrast to the standalone edition), it was necessary to find an indirect communication method with the evaluation service.

The unity plug-in supports calling JavaScript functions inside the embedding browser window and allows also sending messages back to the plug-in. With that approach it is possible that the containing web page receives the updated log file every time a game event is created.

In order to communicate with the evaluation service, a service proxy library was directly written in JavaScript, which is based on AJAX calls provided through the JQuery library. The script in a second file analysis the log file's XML structure (also with JQuery support) and assembles communication objects that can be used with the web service proxy.

Feedback, if available, is received as response after processing an event and is converted to simple messages that are understood by the Unity Web Player, respectively the internal code that is part of the civil defence game.

Changes to the Example HTML File

In order to make use of the provided JavaScript evaluation client, the following two changes were made to the provided HTML File:

1. Just before the Unity object is created, a new evaluation context ID is requested from the web service and stored in a global variable.
2. When the Unity object sends an event, containing the newest version of the log file, the web service client object is called instead of giving direct feedback. The response call-back then updates the game by sending an appropriate message (depending on the response of the web service). The sample feedback code within the function `OnEventLogUpdate` has been replaced with a call to `processEvent`.

For the whole approach to work, the JavaScript files `EvaluationService.js` and `LogfileConverter.js` have to be included in the Example HTML File (see the following sections for details), as well as the third-party dependencies `jquery-1.7.2.min.js` and `jquery.xdomain.js`.

JavaScript for Calling the Evaluation Web Service

For sending messages to the web service (see section "Implemented Components") it was necessary to create communication objects that match the WSDL specification of the web service.

The following JavaScript object prototypes were implemented to represent the request layout of the web service (source file `EvaluationService.js`):

- EvaluationService (offering the required methods of the web service)
- GameEvent
- EnvironmentChange
- DataUpdate
- PlayerAction
- ActionParameter
- IterationResult
- Feedback
- TextFeedbackContent

To send a request to the web service the composition of these objects is converted to an XML representation that fulfils the WSDL/SOAP specification for the web service. The actual request is done by using appropriate functions of `jquery-1.7.2`. This guarantees a browser independent implementation.

JavaScript for Mapping between Game and Web Service Proxy

Since the game is providing the log file that was discussed in the section “Log File Specification” it is necessary to query the XML data and convert it to the object layout that was discussed in the previous section. This task is also achieved with the library `jquery-1.7.2`. The implementation could be found in the source file `LogfileConverter.js`. By calling the function `processEvent` and providing a string with the log file, the web service will be called implicitly for the latest event in the log file.

Show Cases

Taking the Schoolbag

This showcase demonstrates the feedback, when the player is taking the schoolbag on the table before he or she is leaving the classroom. The virtual character who is accompanying the player will be activated to speak directly with the player if the game receives a “speak” message, activated by a feedback object, received after the action of using the object “bag” was detected:



The following image shows an extract of the underlying assessment model for this show case:

```

5  <behaviour-pattern>
6  <match-action name="UseObject">
7  <compare type="Equal">
8  <get property-name="object_type" />
9  <value type="String">bag</value>
10 </compare>
11 </match-action>
12 <consequences>
13 <feedback frequency="Immediate">
14 <type>Supportive</type>
15 <learning-content-relevance>Essential</learning-content-relevance>
16 <text>
17   You took your bag before leaving the class room. You should not collect your possessions before evacuating.
18 </text>
19 </feedback>
20 </consequences>
21 </behaviour-pattern>

```

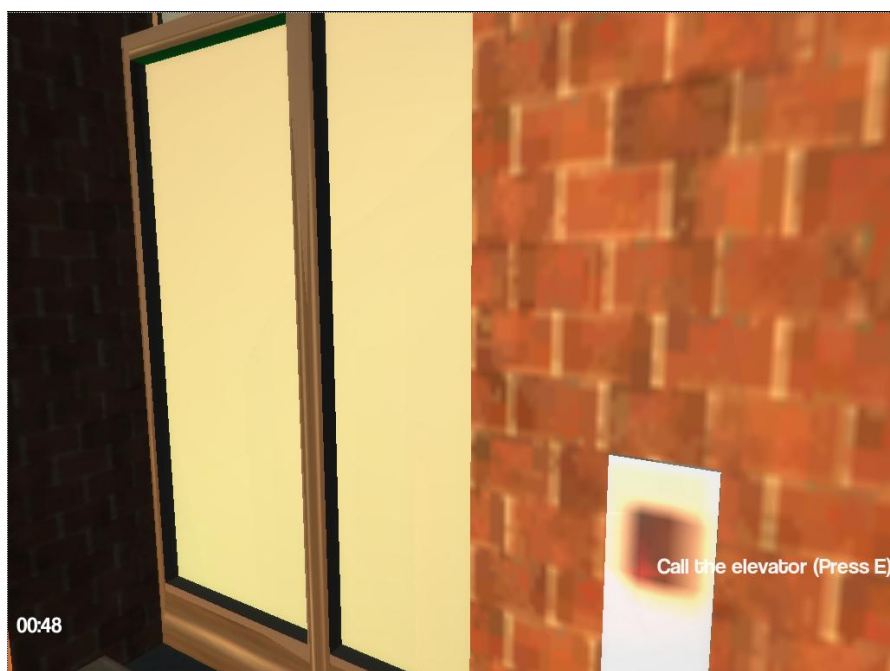
The detection of this action leads to a direct feedback message, expressed through the virtual character:



The message that appears in the green box is directly sent by the web service and not part of the game itself.

Using the Elevator

Another show case demonstrates the direct interference of the virtual character, when the player is trying to call the elevator:



```

152 <behaviour-pattern>
153   <match-action name="UseObject">
154     <compare type="Equal">
155       <get property-name="object_type" />
156       <value type="String">elevator</value>
157     </compare>
158   </match-action>
159   <consequences>
160     <feedback frequency="Immediate">
161       <type>Supportive</type>
162       <learning-content-relevance>Essential</learning-content-relevance>
163       <text>
164         You called the elevator. You should never use an elevator during a fire.
165       </text>
166     </feedback>
167   </consequences>
168 </behaviour-pattern>

```

