A·L·I·C·E
Adaptive Learning via Intuitive/Interactive Collaborative and Emotional systems

Project Number:	**257639**
Project Title:	ALICE: ADAPTIVE LEARNING VIA INTUITIVE/INTERACTIVE, COLLABORATIVE AND EMOTIONAL SYSTEMS
Instrument:	Specific Targeted Research Projects
Thematic Priority:	ICT-2009.4.2:Technology-Enhanced Learning
Project Start Date:	June 1st, 2010
Duration of Project:	24 Months
Deliverable:	**D1.3: Experimentation and Validation Planning**
Revision:	1.0
Workpackage:	WP 3: Experimentation and Validation Planning
Dissemination Level:	Public
Due date:	October 31st, 2010
Submission Date:	October 31st, 2010
Responsible	UOC
Contributors:	CRMPA, MOMA, TUG, UOC

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

SEVENTH FRAMEWORK PROGRAMME

Version History			
Version	**Date**	**Changes**	**Contributors**
0.1	14/9/2010	Initial version	UOC
0.2	30/09/2010	Changes in structure.	UOC
0.3	15/09/2010	Changes in methodology	UOC
0.4	30/09/2010	Feedback from partners	CRMPA, TUG, MOMA, UOC
0.5	15/10/2010	Feedback from internal review	UOC
0.6	25/10/2010	Add Section 2 Theoretical foundations	UOC
0.7	26/10/2010	Add evaluation plans	UOC, CRMPA, TUG, MOMA
0.8	31/10/2010	Add feedback from partners	UOC, CRMPA, TUG, MOMA
1.0	31/10/2010	Move the document to a newer template	UOC

Table of Contents

1 Introduction

This report presents the evaluation plan of the research and technology developed in ALICE as part of the activities of Work Package 1, Task 1.3 of the ALICE project. To this end, a practical method oriented to the testing of the tools developed and its validation in real situations in different educational fields will be followed.

It is worth clarifying at this initial point that the experimentation and validation plan reported here is 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 is instead to be conducted in earlier stages of the whole ALICE evaluation by all participating parties that will develop stand-alone prototypes as a result of their participated research tasks and will test and validate them with the intent of finding software bugs.

Therefore, this document reports the evaluation plan of ALICE considering all individual developments have been tested and integrated into the referenced platform IWT and act atomically as one only e-learning system (i.e., ALICE System).

Two different contexts and two e-learning modalities will be considered to evaluate ALICE System making three evaluation scenarios in all: (i) full e-learning for science teaching at university, (ii) blended learning for science teaching at university, (iii) blended learning for civil defence and emergency at secondary school.

ALICE includes the following 6 work packages, which investigate the main aspects of the project and are candidate to be included in the evaluation plan presented here:

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

These base their research and evaluation goals on (ALICE – Annex, 2010) and (ALICE – D1.1, 2010). 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

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 they have to evaluate ALICE as a whole, tested and validated, with 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

ALICE overall goals and promises are first translated into objectives and research goals for evaluation. These research goals are extracted from a set of scenarios of requirements fully identified and described in the report of the ALICE project (2010b) in order to share in simple terms what is the idea of the final products. For evaluation purposes, the requirements scenarios produced will be extended so as to cover a greater number of cases, taking into account alternative flows not covered in the requirement phase. In turn, each of this extended requirements scenarios will be evaluated by developing specific pilot trials, which will eventually be tested and validated by specific evaluation metrics. Therefore, the evaluation process will be conducted by a set of pilot trials. These pilots will be developed allowing us to perform a complete evaluation of ALICE.

The incorporation of the ALICE technology into secondary and tertiary education entities, and in different contexts of learning, the adaptation to stages of education from face-to-face to full virtual, and mixed, and finally the implantation and use of the developments for real users, will contribute to a very valuable information at the time of evaluating the aims achieved with the project. This information will form the analytical data input for further discussion and interpretation and eventually the promotion and exploitation of the proposed solution by the consortium.

1.3 Document overview

This report is structured as follows. Next Section shows theoretical foundations and background on research methods and evaluation as well as specific techniques found in the literature. This study

already selects those adopted methods and techniques for our purposes of data analysis that best fit the evaluation of the ALICE goals.

Section 3 describes the overall objectives and research question of ALICE that will serve as a reference point to identify what is to be evaluated and how through steps of experimentation and validation processes.

Section 4 shows the full evaluation methodology addressed to evaluate ALICE's research questions and the precise hypotheses formulated. A well-grounded methodology serves as a basis for Section 5 to provide the precise guidelines and procedures to conduct the experimentation and validation activities planned for ALICE.

Finally, Section 6 reproduces and updates the working and contingency plans of Work Package 8 presented in the report of ALICE project (2010c).

2 Theoretical foundations of evaluation methods and techniques

A research oriented learning experience includes a formal and informal process of gaining, utilizing and systematically applying knowledge to an area of interest in order to make sense of the interrelationships between what one knows and what one learns (Denzin, & Lincoln, 1994; Bortz & Döring, 2006). With quantitative reasoning skills, one can integrate deductive logic aspects from multiple knowledge dimensions into program evaluation and research. There is a beginning, middle and an end to this cyclical process which allows for the adjustment of additional information. When approaching evaluation questions, within a particular context, it is important to keep in mind that a scientific, linear model is but one method of organizing information (Guba & Lincoln, 1989).

There is tremendous value in understanding the plural dimensions of both quantitative and qualitative approaches to evaluation methodologies. The context, purpose, and types of research questions asked will define the methodological foundation of a study. Keeping this caveat in mind will eliminate mismatched efforts and results that can only frustrate a beginning student in research (Patton, 1990; Miles & Huberman, 1994).

There is tremendous value in understanding the plural dimensions of both quantitative and qualitative approaches to evaluation methodologies. Wolcott, 1990; Guba & Lincoln, 1989) advocate the necessity of becoming familiar with all other methods in order to appropriately select the method that best fits your area of research and design. The context, purpose, and types of research questions asked will define the methodological foundation of a study. Keeping this caveat in mind will eliminate mismatched efforts and results that can only frustrate a beginning student in research (Trochim & Land, 1982).

Next, an overview and the most important models and methods used in statistics are provided. They will be refined in Section 3.2 with the most appropriate techniques for data analysis in line with ALICE evaluation goals.

2.1 Statistics overview

After considering the context and nature of a research project, the appropriate method of inquiry is selected to help direct the development of specific research questions. The objective of the inquiry is to ask questions in order to retrieve the data or information that is salient to the project (Trochim & Land, 1982). Collecting and analyzing data with quantitative strategies includes understanding the relationships among variables utilizing descriptive and inferential statistics (Mann, 1995; Ostle & Malone, 1988). This process will require a serious research to gain a fuller knowledge base by undertaking statistics or regression analysis (Strauss & Corbin, 1990; Dielman, 1996).

Asking empirical questions in testable forms will involve the traditional use of the Null hypothesis versus the Alternative hypothesis. Test statistics for significance are used to determine if the null or alternative is to be accepted or rejected. The null hypothesis tests for the differences between population means. Inferential logic will establish the standards of your study based on theory and application to reality (Mann, 1995; Ostle & Malone, 1988).

2.1.1 Descriptive Statistics

Descriptive statistics are theoretical postulates used to draw inferences about populations and to estimate the parameters of those populations (Mann, 1995; Dodge, 2003). Measures of central

tendency and dispersion summarize the information contained in a sample and are usually provided in summary form, such as distributions, graphical and or numerical methods. Inferential statistics are based on descriptive statistics and assumptions that generalize to the population from a selected sample. These assumptions focus on the use of continuous data and that the sample is a random representation of the population. Inferences made at large use probabilities and probability distributions. Statistical evidence is especially important to that have a vested interest in evaluation projects.

Probably the most often used descriptive statistic is the mean. The mean is a particularly informative measure of the "central tendency" of the variable if it is reported along with its confidence intervals. (Hill & Lewicki, 2007).

Usually we are interested in statistics (such as the mean) from our sample only to the extent to which they can infer information about the population. The confidence intervals for the mean give us a range of values around the mean where we expect the "true" (population) mean is located (with a given level of certainty, see also Elementary Concepts). If you set the p-level to a smaller value, then the interval would become wider thereby increasing the "certainty" of the estimate, and vice versa; as we all know from the weather forecast, the more "vague" the prediction (i.e., wider the confidence interval), the more likely it will materialize.(Hill & Lewicki, 2007).

2.1.2 Construction of variables

Variables are things that we measure, control, or manipulate in research. They differ in many respects, most notably in the role they are given in our research and in the type of measures that can be applied to them. (Narens, 1981a).

Independent variables are those that are manipulated whereas dependent variables are only measured or registered. The terms dependent and independent variable apply mostly to experimental research where some variables are manipulated, and in this sense they are "independent" from the initial reaction patterns, features, intentions, etc. of the subjects. Some other variables are expected to be "dependent" on the manipulation or experimental conditions. That is to say, they depend on "what the subject will do" in response. Somewhat contrary to the nature of this distinction, these terms are also used in studies where we do not literally manipulate independent variables, but only assign subjects to "experimental groups" based on some pre-existing properties of the subjects. (Hill & Lewicki, 2007).

Variables differ in how well they can be measured, i.e., in how much measurable information their measurement scale can provide. There is obviously some measurement error involved in every measurement, which determines the amount of information that we can obtain. Another factor that determines the amount of information that can be provided by a variable is its type of measurement scale. Specifically, variables are classified as (1) nominal, (2) ordinal, (3) interval, or (4) ratio. (Narens, 1981b).

1. Nominal variables allow for only qualitative classification. That is, they can be measured only in terms of whether the individual items belong to some distinctively different categories, but we cannot quantify or even rank order those categories. Typical examples of nominal variables are gender, race, color, city, etc.
2. Ordinal variables allow us to rank order the items we measure in terms of which has less and which has more of the quality represented by the variable, but still they do not allow us to say "how much more." A typical example of an ordinal variable is the socioeconomic status of families. (Cliff & Keats, 2003).
3. Interval variables allow us not only to rank order the items that are measured, but also to quantify

and compare the sizes of differences between them.

4. Ratio variables are very similar to interval variables; in addition to all the properties of interval variables, they feature an identifiable absolute zero point, thus, they allow for statements such as x is two times more than y. Typical examples of ratio scales are measures of time or space. Most statistical data analysis procedures do not distinguish between the interval and ratio properties of the measurement scales.

2.1.3 Shape of the distribution, Normality

An important aspect of the "description" of a variable is the shape of its distribution, which tells you the frequency of values from different ranges of the variable. Typically, a researcher is interested in how well the distribution can be approximated by the normal distribution. Simple descriptive statistics can provide some information relevant to this issue.

More precise information can be obtained by performing one of the tests of normality to determine the probability that the sample came from a normally distributed population of observations (e.g., the so-called Kolmogorov-Smirnov test (Kolmogorov, 1933), or the Shapiro-Wilks' W test (Shapiro, S. S. & Wilk, 1965). However, none of these tests can entirely substitute for a visual examination of the data using a histogram (i.e., a graph that shows the frequency distribution of a variable).

The graph allows you to evaluate the normality of the empirical distribution because it also shows the normal curve superimposed over the histogram. It also allows you to examine various aspects of the distribution qualitatively (Johnson & Kotz, 1994).

2.1.4 How to measure the magnitude of relations between variables

There are very many measures of the magnitude of relationships between variables that have been developed by statisticians; the choice of a specific measure in given circumstances depends on the number of variables involved, measurement scales used, nature of the relations, etc. Almost all of them, however, follow one general principle: they attempt to somehow evaluate the observed relation by comparing it to the "maximum imaginable relation" between those specific variables. (Hill & Lewicki, 2007).

Technically speaking, a common way to perform such evaluations is to look at how differentiated the values are of the variables, and then calculate what part of this "overall available differentiation" is accounted for by instances when that differentiation is "common" in the two (or more) variables in question. Speaking less technically, we compare "what is common in those variables" to "what potentially could have been common if the variables were perfectly related".

2.2 Methods and techniques for data analysis

The adopted statistical methods and techniques considered for evaluation purposes of this project are studied next in detail from a didactic and pragmatic approach. Further reference to this section is expected during the validation step of the collected data from testing.

2.2.1 Statistical hypothesis testing

A statistical hypothesis test is a method of making decisions using experimental data. In statistics, a result is called statistically significant if it is unlikely to have occurred by chance.

Hypothesis testing is sometimes called confirmatory data analysis, in contrast to exploratory data analysis. In frequency probability, these decisions are almost always made using null-hypothesis tests (i.e., tests that answer the question. Assuming that the null hypothesis is true, what is the probability of observing a value for the test statistic that is at least as extreme as the value that was actually

observed?). One use of hypothesis testing is deciding whether experimental results contain enough information to cast doubt on conventional wisdom (Cramer & Howitt, 2004).

Statistical hypothesis testing plays an important role in the whole of statistics and in statistical inference. For example, Lehmann (1992) in a review of the fundamental paper by Neyman and Pearson (1933) says: "Nevertheless, despite their shortcomings, the new paradigm formulated in the 1933 paper, and the many developments carried out within its framework continue to play a central role in both the theory and practice of statistics and can be expected to do so in the foreseeable future" (Lehmann & Romano, 2005).

2.2.2 t-Test for Independent Samples

The t-test is the most commonly used method to evaluate the differences in means between two groups. For example, the t-test can be used to test for a difference in test scores between a group of patients who were given a drug and a control group who received a placebo. Theoretically, the t-test can be used even if the sample sizes are very small (e.g., as small as 10; some researchers claim that even smaller n's are possible), as long as the variables are normally distributed within each group and the variation of scores in the two groups is not reliably different. As mentioned before, the normality assumption can be evaluated by looking at the distribution of the data (via histograms) or by performing a normality test. The equality of variances assumption can be verified with the F test, or you can use the more robust Levene's test. If these conditions are not met, then you can evaluate the differences in means between two groups using one of the nonparametric alternatives to the t- test.

The p-level reported with a t-test represents the probability of error involved in accepting our research hypothesis about the existence of a difference. Technically speaking, this is the probability of error associated with rejecting the hypothesis of no difference between the two categories of observations (corresponding to the groups) in the population when, in fact, the hypothesis is true. Some researchers suggest that if the difference is in the predicted direction, you can consider only one half (one "tail") of the probability distribution and thus divide the standard p-level reported with a t-test (a "two-tailed" probability) by two. Others, however, suggest that you should always report the standard, two-tailed t-test probability (Hill & Lewicki, 2007).

2.2.3 t-Test for Dependent Samples

The size of a relation between two variables, such as the one measured by a difference in means between two groups, depends to a large extent on the differentiation of values within the group. Depending on how differentiated the values are in each group, a given "raw difference" in group means will indicate either a stronger or weaker relationship between the independent (grouping) and dependent variable.

- **Purpose**

  The t-test for dependent samples helps us to take advantage of one specific type of design in which an important source of within-group variation can be easily identified and excluded from the analysis. Specifically, if two groups of observations that are to be compared are based on the same sample of subjects who were tested twice (e.g., before and after a treatment), then a considerable part of the within-group variation in both groups of scores can be attributed to the initial individual differences between subjects. This fact is not much different than in cases when the two groups are entirely independent, where individual differences also contribute to the error variance; but in the case of independent samples, anything cannot be done since the variation cannot be identified due to individual differences in subjects. However, if the same sample was tested twice, then this variation can be easily identified. Specifically, instead of treating each group separately, and analyzing raw scores, we can look only at the differences between the two measures (e.g., "pre-test" and "post test") in each subject by subtracting the first score from the

second for each subject and then analyzing only those "pure (paired) differences" (Zimmerman, 1997).

– **Assumptions**

  The theoretical assumptions of the t-test for independent samples also apply to the dependent samples test; that is, the paired differences should be normally distributed. If these assumptions are clearly not met, then one of the nonparametric alternative tests should be used.

2.2.4 Analysis of variance

It often happens in research practice that you need to compare more than two groups (e.g., drug 1, drug 2, and placebo), or compare groups created by more than one independent variable while controlling for the separate influence of each of them (e.g., Gender, type of Drug, and size of Dose). In these cases, you need to analyze the data using Analysis of Variance, which can be considered to be a generalization of the t-test. In fact, for two group comparisons, ANOVA will give results identical to a t-test. However, when the design is more complex, ANOVA offers numerous advantages that t-tests cannot provide (even if you run a series of t- tests comparing various cells of the design).

In statistics, analysis of variance (ANOVA) is a collection of statistical models, and their associated procedures, in which the observed variance in a particular variable is partitioned into components attributable to different sources of variation. In its simplest form ANOVA provides a statistical test of whether or not the means of several groups are all equal, and therefore generalizes t-test to more than two groups. ANOVAs are helpful because they possess an advantage over a two-sample t-test. Doing multiple two-sample t-tests would result in an increased chance of committing a type I error. For this reason, ANOVAs are useful in comparing three or more means (Lindman, 1974; Anscombe, 1948).

2.2.5 Factor analysis

The main applications of factor analytic techniques are: (1) to reduce the number of variables and (2) to detect structure in the relationships between variables, that is to classify variables. Therefore, factor analysis is applied as a data reduction or structure detection method (the term factor analysis was first introduced by Thurstone, 1931).

– **Confirmatory factor analysis**

  Structural Equation Modeling (SEPATH) allows you to test specific hypotheses about the factor structure for a set of variables, in one or several samples (e.g., you can compare factor structures across samples).

– **Correspondence analysis**

  Correspondence analysis is a descriptive/exploratory technique designed to analyze two-way and multi-way tables containing some measure of correspondence between the rows and columns. The results provide information which is similar in nature to those produced by factor analysis techniques, and they allow you to explore the structure of categorical variables included in the table (Child, 1973).

– **Combining Two Variables into a Single Factor**

  Correlation between two variables can be summarized in a scatterplot. A regression line can then be fitted that represents the "best" summary of the linear relationship between the variables. If we could define a variable that would approximate the regression line in such a plot, then that variable would capture most of the "essence" of the two items. Subjects' single scores on that new factor, represented by the regression line, could then be used in future data analyses to represent that essence of the two items. In a sense we have reduced the two variables to one factor. Note that the new factor is actually a linear combination of the two variables (Gorsuch, 1983).

– **How many Factors to Extract**

  From considering principal components analysis as a data reduction method, that is, as a method for reducing the number of variables, the question then is, how many factors do we want to extract? Note that as we extract consecutive factors, they account for less and less variability. The decision of when to stop extracting factors basically depends on when there is only very little "random" variability left. The nature of this decision is arbitrary; however, various guidelines have been developed, and they are reviewed in Hill, & Lewicki (2007).

2.2.6 Cluster analysis

Cluster analysis or clustering is the assignment of a set of observations into subsets (called clusters) so that observations in the same cluster are similar in some sense. Clustering is a method of unsupervised learning, and a common technique for statistical data analysis used in many fields (MacQueen, 1967).

Cluster Analysis provides a simple profile of individuals. Given a number of analysis units, for example school size, student ethnicity, region, size of civil jurisdiction and social economic status in this example, each of which is described by a set of characteristics and attributes. Cluster Analysis also suggests how groups of units are determined such that units within groups are similar in some respect and unlike those from other groups.

2.2.7 Questionnaires

A Likert scale (Likert, 1932) is a psychometric scale commonly used in questionnaires, and is the most widely used scale in survey research, such that the term is often used interchangeably with rating scale even though the two are not synonymous. When responding to a Likert questionnaire item, respondents specify their level of agreement to a statement.

A Likert item is simply a statement which the respondent is asked to evaluate according to any kind of subjective or objective criteria; generally the level of agreement or disagreement is measured. Often five ordered response levels are used, although many psychometricians advocate using seven or nine levels; a recent empirical study found that a 5- or 7- point scale may produce slightly higher mean scores relative to the highest possible attainable score, compared to those produced from a 10-point scale, and this difference was statistically significant. In terms of the other data characteristics, there was very little difference among the scale formats in terms of variation about the mean, skewness or kurtosis.

The format of a typical five-level Likert item is:

1. Strongly disagree
2. Disagree
3. Neither agree nor disagree
4. Agree
5. Strongly agree

Likert scaling is a bipolar scaling method, measuring either positive or negative response to a statement. Sometimes a four-point scale is used; this is a forced choice method since the middle option of "Neither agree nor disagree" is not available.

Likert scales may be subject to distortion from several causes. Respondents may avoid using extreme response categories (central tendency bias); agree with statements as presented (acquiescence bias);

or try to portray themselves or their organization in a more favorable light (social desirability bias). Designing a scale with balanced keying (an equal number of positive and negative statements) can obviate the problem of acquiescence bias, since acquiescence on positively keyed items will balance acquiescence on negatively keyed items, but central tendency and social desirability are somewhat more problematic (Wuensch, 2005).

3 Research objectives and promises

The general objective of ALICE is to build an innovative adaptive environment for e-learning combining personalization, collaboration and simulation aspects within an affective/emotional based approach able to contribute to the overcoming of the quoted limitations of current e-learning systems and content. In other words the proposed environment will be interactive, challenging and context aware while enabling learners' demand of empowerment, social identity, and authentic learning experience.

The defined system will be able to effectively involve learners in educational, cultural and informative activities in two specific contexts: university instruction (with particular emphasis on scientific topics) and training about emergency and civil defence (as for example the behavior to take at a personal and collective level when the treat of a big risk shows up e.g. a natural event like earthquake, or a fraudulent one like terrorist attack).

To achieve these aims, ALICE has formulated a series of general open research questions:

- How is it possible to create collaboration conditions and therefore to encourage the learner to choose a collaborative-type education also when collaboration is actually difficult?
- How can the effectiveness of learning actions be supported by interactive simulations and serious games that may be created with low costs thanks to techniques of reusability?
- In what way can the storytelling be integrated with Learning Experiences having contents of different types?
- Eventually, how to create a learning additivity related to the earlier themes, being not the simple sum of various aspects, but a real integration and subsequent super-additivity with respect to single components?

These issues are to be addressed through intensive evaluation effort by piloting all research goals of the project, which in turn means to evaluate each and every ALICE functionalities and potentialities in the form of the requirement scenarios defined in (ALICE project, 2010b). This deductive approach is used to eventually identify the evaluation criteria and metrics that will allow pilots to experiment and validate all requirements of ALICE. To this end, two main evaluation questions will conduct the rest of this document and eventually the whole research methodology of the project:

- What are the ALICE overall objectives and research questions relevant for experimentation?
- What are the ALICE overall objectives and research questions relevant for validation?

Following the mentioned deductive methodology, for each of the research lines of the project, a set of research objectives and promises were formulated in the report of ALICE project (2010a) and are reproduced next. Then, in subsequent Section 5 these general research objectives will be turned into specific evaluation goals that will set the basis of the experimentation and validation procedures.

3.1 WP2. Emotional and affective approaches

- To study models, methods and tools useful in the management of emotional-affective aspects in Intelligent Tutoring Systems.
- To study and define methodologies for the creation of LE (Learning Experience) based on A-Cnt (Advanced-Content) that can complement traditional learning with simulations, collaboration and assessing contents.

- To create two sample Learning Experiences to show the super-additivity between emotional-affective aspects and approaches already adopted in the reference platform, which models the learner and the domain knowledge and uses inductive-experiential teaching model for intuitive and interactive learning.
- To improve the depth of scientific analysis of cognitive science, for studying the laws of physics in a Luna Park, and by analyzing the earthquake.
- To study and analyse the ITS-learner interaction and feedback design tools able to support of the EA-ITS (Emotional-Affective ITS).

3.2 WP3. Live and virtualized collaboration

- To study and analyze models, methods and tools for Collaborative Learning.
- Social Learning and their application in formal and informal/intentional learning contexts (including Emerging Social Web and synchronous communication/collaboration applications).
- To study and define methodologies and techniques Semantic Web-based to provide a machine-readable and machine-understandable representation of knowledge in order to support live collaborative sessions virtualization process.
- To define methodologies to author, manage and execute Collaborative Complex Learning Objects containing Virtualized Collaborative Sessions.
- To design the architecture and developing prototypical tools for supporting the virtualization of Collaborative Sessions in some given sample scenarios and the execution of related CC-LOs.
- To design and develop a sample collaborative learning environment (based on the method of On-Line Meeting) and a sample CC-LO realized by virtualizing the collaboration activities occurred during the live Collaborative Sessions executed in the sample environment.

3.3 WP4. Simulation and serious games

- To develop, from a state of the art, a conceptual framework for the definition of serious games, specifically for Virtual Scientific Experiment (VSE) seen as particular Serious Game.
- To realize new VSE and Learning Experiences of different formative complexity by using some techniques of brick and semiotics reusability, applied to appropriate Complex Learning Objects.
- To analyze the evolution (state transitions) of a material system subject to certain constraints and external stimuli.

3.4 WP5. New forms of e-assessment

- To define an integrated framework for the assessment of Learning Experiences enriched by didactic with high complexity (CLO) such as simulations, collaborative and virtual experiences, story telling as well as by emotional elements.
- To define a set of usable methods to draw information and valuation that can have an impact directly on the adaptability of the learning experience and of its components through an updating of the user's profile and his cognitive model, in terms of knowledge and achieved skills as regards the formative curriculum and a review of didactic methods to be used.
- To experiment the obtained model and the arranged methods on a complex learning experience and on specified components of it such as a simulated experience on an didactic object and to

give a guideline for a correct evaluation.

3.5 WP6. Storytelling

- To investigate the sector of educational research to define and consolidate a Storytelling Design Model (SDM) for the didactic development of a narrative “Story Based Learning Object” (SLO).
- To define a conceptual framework of SDM with particular respect to its specific declination in a SLO in the category of Memorial Stories, referred to a scenario of contextualized learning and linked to big risks in order 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.

3.6 WP7. Adaptive technologies

- To improve and extend existing models, methodologies and components of the reference platform IWT in order to prepare it for a smooth integration of methodological and technological components coming from the preceding work packages.
- To integrate the support for knowledge model contextualisation, for upper level learning goals and for semantic connections between learning Resources.

4 Experimentation and validation methodology

Following the previous view, a comprehensive experimentation methodology is developed in this section for ALICE describing all evaluation activities that will be undertaken throughout the lifetime of the project. The plan includes details on the type of pilot trials, the number and type of participants and the exact duration and aims of each pilot activity. The generic experimentation plan of the project may be further localised to better address the local circumstances pertaining in each experimentation site of user group. Implementation parameters will be 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.

The evaluation methodology takes as inputs the user scenarios from D1.1 of Work Package 1 (ALICE project, 2010b) and performs the definition, integration and testing tasks of the resulting software components.

To pursue these goals, communities of user groups will be organised in selected pilot sites, which will be 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 could enable the exploitation by teachers and students of existing advanced educational technologies. Initially for each scenario of use a devoted user group will be developed drawing from two different contexts, namely Science Teaching at University and Civil Defense and Emergency.

The deployed system and scenarios of its use will be exposed, through demonstration activities, to large 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 iterative tests with numerous users performing authentic technology-enhanced learning tasks will be performed.

Gradually, the size of user groups will be extended by dynamically involving more groups from other subjects and programs. A main part of this plan will be 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 will be determined. The timetable of the proposed activities will be designed in order to be discussed with the users involved.

Next, both a technological and application perspective of the methodology is shown next. Adopted monitoring and reporting tools as well as methods and techniques for analysing the data collected from trials are described in the next sections. Finally, the scenarios of use are located and the roles of the participating pilot sites are presented.

4.1 Technological perspective

This proposed methodology first defines the ALICE system that will be prototyped, then integrates the various prototypes and necessary components of the ALICE system into IWT and finally performs the end-to-end functional testing and evaluation based on pilot trials of the IWT-ALICE integrated prototype. This is further accompanied by the performance of a dry run of certain use case scenarios for the purpose of fine tuning the system.

This step will define the precise hardware and software components required for the experimentation of the ALICE prototypes. It is also responsible for defining the exact methodology that will be followed during the subsequent integration and experimentation stages.

4.1.1 System integration

This step integrates the applications and services that will have been implemented in earlier efforts as well as the necessary hardware in a single prototype implementation of the IWT system as defined in D1.2. For each learning scenario defined in WP1 (ALICE project, 2010b) different configurations will be established, nevertheless most of the work during integration should be common among the different scenarios.

4.1.2 System trial

In this step, the integrated prototypes into IWT will be evaluated by piloting extensively all supported functions against the technical specifications defined earlier in the project (ALICE project, 2010b). If necessary, components interface may be modified in order to provide a more efficient and robust integration among the various ALICE components. This will also provide an insight on the quality of the integrated prototype.

4.1.3 System evaluation

The output of the previous step is fed here to exposing the prototype platform for evaluation. In particular the scenarios of use, defined in (ALICE project, 2010b), will be evaluated and redesigned accordingly, so that they best meet the requirements. More specifically, this procedure can be described as a combination of two basic flows of information that will take place in two phases. During the first phase a linear flow of information will be attempted, extending from the proposition of initial scenarios of use to the consolidation of user input which will facilitate the refinement of these first scenarios.

4.2 Application perspective

Consecutive cycles of intensive implementation of the developed scenarios of use in ALICE will take place in the selected pilot sites. During these trials user input and reactions will be monitored by the research team in various ways, using adopted monitoring and reporting so as to gather data for analysis. In this task, the partnership will initiate the use of the developed system in the participating pilot sites.

The application approach of this methodology includes two experimentation phases. During the first round, introductory workshops will be organised in the participating user organisations involving also experts in the field. The consolidated user input, gathered through thematic discussions and activities during the workshops, will help in refining the scenarios of use and focusing on more detailed user needs. These scenarios will include non-technical descriptions focusing on functionality, interaction and usability considerations.

The second phase will attempt to introduce the users to the refined scenarios of use during a series of design workshops, aiming to develop a detailed user requirements list and the final scenarios that illustrate the context of use, bearing in mind specific user requirements and usability issues defined during the process. Each item of the user requirements list will be associated with specific information gathered during the workshops and interviews as a result of the content analysis of all the transcripts from the workshops and interviews. The consortium will also work to collect additional data through web surveys to verify the value of the users' feedback and to asses the adopted process.

4.2.1 Training workshops

A number of workshops will take place in the participating pilot sites, in accordance with the experimentation plan, in order to prepare the relevant user groups for the upcoming trials.

4.2.2 Piloting with users (first phase)

The first implementation phase will test the approach, the processes and the relevant technological elements to all involved actors and will help them get familiarized with these. In addition, users' reactions to the proposed system and the scenarios of use will be monitored and analysed in detail. Intensive quantitative and qualitative data analysis will take place, using the developed monitoring and reporting tools and the data gathered through them, so that interim conclusions can be drawn and the corresponding improvements can be inserted into the design of the technological solutions and the scenarios of use.

During the validation activities, the tests will mainly involve users to which the consortium has immediate access (e.g. within the partner organisations, in their immediate networks, etc.).

4.2.3 Piloting with users (second phase)

With the experience gained during the first phase of implementation and after the appropriate modifications on both the deployed technologies and the proposed scenarios of use, the partnership will implement the proposed pilot activities within ALICE in the participating pilot sites during one more pilot activities cycle. This time, more user groups and communities will be involved in the trials process, and the user experiences will consist of more complex scenarios of use. This will be an important point on the gradual move of the trials towards less and less guided use by more and more experienced users, with an increasing encouragement of user generated activity on the system, so that the exercise will reveal how the proposed solution can dynamically respond to the emerging user needs.

At the same time, the pilot trials in the second phase will gradually expand, moving from an initial limited number of experiments to a large number of demonstration activities that will follow, thus gradually shifting the focus of work from initial formative evaluation to the more mature phase of validation of the proposed solution through demonstration activities exposing it to a large number of users. Throughout this phase, data will continue being systematically gathered, so that users' activity and reactions to the proposed system and the scenarios of use can be monitored and analysed in detail.

In this phase, the validation effort will also expand to include more users in several other places and contexts, in which demonstration and dissemination activities will be organised by the project. The user basis will become broader to include users from as diverse settings and backgrounds as possible.

4.2.4 Collected data and decisions

Considerations upon the collected data from each phase of trials will be taken, such as the sample and methodology reliability, the study significance with respect to the previous knowledge, the practical application follow up, eventual weaknesses and limitations and eventually the necessity for further development of following research steps.

4.2.5 Integration of results - validation report

Throughout the validation activities performed and peaking towards their end, intensive quantitative and qualitative data analysis of the accumulated data will take place, so that the final version of the technology and of the scenarios of use can be grounded on the final findings of the validation activities with real users.

Finally, the results of all stages of testing and validation will be integrated into the final validation report. This report will also provide significant input to all dissemination and further exploitation efforts of the consortium. It will be one of the key deliverables of the project since it will document the

extensive validation process and its results and will be one of the strongest arguments for the promotion and exploitation of the proposed solution by the consortium.

4.3 Adopted data collection, monitoring and reporting tools

The most important issue while monitoring the pilot trials is the collection and storage of a large amount of event information generated by the high degree of interaction among the participants. Such a large amount of informational data may need a long time to be processed. Therefore, collaborative learning systems have to be designed in a way that classifies and pre-structures the resulting information in order, on the one hand, to correctly collect the teaching and learning activity and, on the other hand, to increase the efficiency during data processing in terms of analysis techniques and interpretations.

To cover these needs for the monitoring and reporting activities during experimentation the following technological tools will be designed and developed so as to facilitate automatic data collection, modelling and exploitation of behavioural and usage patterns during the trials.

4.3.1 Tools for collection and classification of data

An important issue while monitoring on-line learning activity is the collection and storage of large amounts of quantitative information generated by the high degree of interaction among participants during both synchronous and asynchronous instruction. Although the computer has many advantages in terms of storage capacity and data processing, the need to convert the information generated in a workspace into an appropriate computational format represents a major drawback. In addition, qualitative information usually comes in the form of structured and textual questionnaires. The latter must be explicitly provided by the students and is difficult for computers to collect and analyze due to its high degree of informality, so it is manually processed and interpreted.

Collection of large amounts of primitive data from pilot trials will be carried out automatically by regular log files and databases provided by the own ALICE prototypes. Integration of the storage systems of all prototypes will be necessary to unify and normalize the type and data collected. Then, extensions of these storage systems will be developed to efficiently classify and structure this information into spreadsheet-based tools so as to prepare data for later analysis.

4.3.2 Tools for data structure recognition in log files

Online teaching and learning activities produce heterogeneous high-volume data. For the online evaluation and validation this data need to be reduced. Data reduction is a critical problem; there are large collections of documents that must be analyzed and processed, raising issues related to performance and loss-less reduction. Data reduction strategies will be investigated, ranging from data clustering to latent semantic analysis and tensor reduction for related heterogeneous data. Data mining techniques will be used, following the steps of i) pre-processing of data (data selection and preparation to a form useful for data mining methods, applying matrix and tensor decomposition for the extraction and detection of concepts or topics from log files), ii) evaluation of existing data mining methods to apply to classification, clustering, association rules or action (time) sequence patterns.

4.3.3 Tools for user profile modelling

User profiles will be represented by the tools developed using a wide range of techniques, from simple keyword-based files to more intelligent representations involving contextual and semantic attributes. Among others, user profiles could contain representation of long and short term individual user's learning interests. The tools will support automatic user profile creation and maintenance, and analysis of user information preferences.

4.3.4 Tools enabling clustering methods

These tools for the application of clustering methods will help enhance users' teaching and learning experiences, grouping similar actors based on their similarities and helping describe and explore these groups intuitively. For example, by clustering similar student behaviours on the basis of students' interaction with the learning environment, tutors could provide scalable feedback and learning recommendation to learners. Artificial Neural Network model, K-Means algorithm, hierarchical clustering methods will be developed through approaches based on Case-Based Reasoning, as well as 2D and 3D visualization techniques for clustering result.

The variety and complementarily of the above tools will ensure the development of an extensive feedback pool from a variety of users that will interact with the system in the different trial contexts.

4.4 Adopted data analysis methods and techniques

Appropriate methods, techniques and tools for analysing the data collected from experimentation are essential to extract valuable information and knowledge at the time of interpreting the aims achieved with the project. This information will form the analytical data input for the eventual promotion and exploitation of the proposed ALICE solution by the consortium.

Therefore data collected from the extensive piloting in each phase will include a variety of methods and techniques for data analysis, such as questionnaires, hypothesis testing, t-test, analysis of variance, factor analysis and cluster analysis. Other less formal methods, such as interview schedules, observation grids, web tools, procedures for the elicitation of user feedback in workshops and so on will also be considered.

The most relevant and pragmatic methods and techniques for the purpose of extracting knowledge from the collected data of our experiments were already listed and described in Section 2.2, where we will refer to during the stage of data validation and interpretation.

Please note that that proposed list does not intend to be exhaustive. Instead, other evaluation methods and specifically statistical techniques may show up and be used during validation according to the actual type complexity and amount of data coming from the trials.

4.5 Trial scenarios and role of partners

Three different scenarios are considered to exhaustively pilot all features and functionalities of ALICE in different contexts of learning and by participants of each piloting site:

4.5.1 Full e-learning for science teaching at university

The Universitat Oberta de Catalunya (UOC) is an online university born from the knowledge society and whose mission is to provide people with lifelong learning and education and education opportunities. The UOC develops the Virtual Campus as a network community where limitations of space and time are overcome. It also uses a student-centred learning methodology based on the complete personalisation and guidance of the learner. Learners, lecturers and managers interact and cooperate in that network community to create, structure, share and disseminate knowledge. The aim is to help individuals meet their learning needs and provide them with full access to knowledge.

As a leader and innovator in education and technology, the UOC is a benchmark for quality in its academic work and research in e-learning based on Information and Communication Technologies (ICT). The UOC pioneers top quality distance education through the Internet since 1994. About 55,000 students and 2,500 lecturers and tutors are involved in more than 1,200 full on-line official courses from 23 official degrees and other PhD and post-graduate programs. This immense driving force will

serve to exploit ALICE project through its Virtual Campus and eLC by investigating, experimenting and using the resulting innovative adaptive environments for e-learning able to contribute to the overcoming of the limitations of current e-learning and e-training systems and pedagogical models. The resulting environments will be interactive, challenging and context aware while enabling learners' and the public at large's demand of empowerment, social identity, and authentic learning experience.

UOC has the role to evaluate the impacts of the innovative features offered by ALICE inside selected learning and training environments.

4.5.2 Blended learning for science teaching at university

Almost 200 years of research and teaching in society's service make Graz University of Technology (TU Graz) one of Austria's most venerable scientific institutions. The University's success throughout its eventful history has been based upon the achievements of outstanding personalities in science, research and their application. Today engineers have more responsibility than ever for the quality of life of generations to come. This awareness and a modern understanding of technology are the guiding principles of the students, teaching staff and researchers of Graz University of Technology.

The range of courses comprises 16 bachelor's programmes, 29 master's programmes, three teacher-training studies, one diploma study and two doctoral programmes as well as six other postgraduate courses. The Faculty of Computer Science comprises 8 Institutes and covers main subjects such as software engineering, information systems, computer graphics and vision, and security.

As e-assessment is one of the core competences of the institutions at TU Graz, thus there is a main interest in exploiting results on research level as well as on application and service level. Furthermore, TU Graz is involved in the project as application partner and will also implement pilot studies.

4.5.3 Blended learning for civil defence and emergency at secondary school

MOMA S.p.A. born as a no profit spin-off of CRMPA (Centre for Research in Pure and Applied Mathematics Environment). MOMA is involved in research and development, industrialization and marketing of advanced products and services in the ICT sector, with particular focus on the areas of Learning & Knowledge.

In particular, the knowledge acquired through the implementation of projects of industrial research at the highest national and European levels, and of different prototypes, have conveyed into a product named IWT Intelligent Web Teacher, an innovative and user-centric platform for the distance learning, based on the explicit representation of knowledge and extensible in terms of functionality and content. The IWT platform meets the information, training and knowledge needs in several contexts such as the educational, business and professional ones.

MOMA main intention is achieve an enhanced learning environment, built on the IWT platform, developed on adaptive technologies and pedagogical methods, with feature based on affective and emotional approach. The innovative environment, so conceived, enables new forms of learning and of assessment, dynamically adaptive and enhances some basic aspects like collaborative learning, simulation and serious games, storytelling, contributing to the overcome of the quoted limitations of existing e-learning system.

The pilot trials of this scenario will focus on training students of secondary schools about actions and procedures to be performed in case of emergency (e.g. the behaviour to take at a personal and collective level when the treat of a big risk shows up). The pilots will take advantage of the network of secondary schools created by MOMA through the ongoing project InnovaScuola, purposed to foster the use of innovative technologies for learning in Italian schools (about 1000 schools involved). The schools belonging to the network already adopt the IWT platform so, a selection of them, will be involved in the ALICE experimentation.

5 Experimentation and validation procedure

This very crucial step of the project aims to plan the procedures for an extensive experimentation and validation of all aspects of the ALICE system in a wide variety of usage scenarios.

For the purpose to proceed with the evaluation, the project objectives and promises presented in Section 3 were considered to identify and define all ALICE functional requirements reported in (ALICE project, 2010b). Each of these requirements was extracted from learning scenarios proposed by the project's stakeholders in order to meet the real needs for a successful achievement of the objectives of the ALICE System. The requirement scenarios were finally selected, refined, analysed and structured in several points of interests, such as requirement scenario goals and success criteria.

At this point, it is worth reminding here that for evaluation purposes, the requirements scenarios will be extended so as to cover a greater number of cases, taking into account alternative flows not covered in the requirement phase. In turn, each of this extended requirements scenarios will also be evaluated by developing specific pilot trials.

For each piloting phase and (extended) requirement scenario a *goal* or set goals to describe a specific functional outcome is considered. Goals are then formulated in terms of *hypotheses* to be eventually verified by validating a functional outcome.

Then, following the methodology presented in the previous section, the following testing parameters are considered, which will be the same for each requirement scenario, as follows.

- *Apparatus*: technology support to carry out each pilot trial (basically the IWT e-learning system), including other tools to monitor and collect experimental data.
- *Participants*: Specific location and all actors from each piloting site who will be involved in the trials (basically tutors and students). These include the supervisors of the trials.
- *Schedule*: The specific time to run the trials.

These parameters are to conduct the testing *design*, which includes the monitoring and reporting tools to collect and classify data for further analysis.

As for the validation methodology, a similar procedure will occur. After extensive experimentation, all data collected will be analysed and a validation against the same initial hypotheses will result. The propose methodology includes the following parameters:

- *Criteria*: ideal reference situation that should be met in predicting an outcome.
- *Metric*: a variable or set of variables that fits a predefined criteria. All metrics must follow the general quality criteria C1-C5 presented in Section 1.1.

Finally, validation *techniques*, both quantitative and qualitative, will include analysis of variance, t-tests, questionnaires, etc. (see previous section and also Section 2 for further information on adopted statistical methods and techniques). Other procedures will use interview schedules, observation grids, web tools, elicitation of user feedback in workshops collected from experimentation activities.

Nine requirement scenarios and their evaluation procedures are considered. Next, we provide a schema based on the previous items with all the needed data to be carefully followed when running experimentation and validation activities for each requirement scenario.

5.1 R1. Upper Level Learning Goals

- **Evaluation goals**
  - G1.1: to develop a Course Generation System (CGS) able to generate a set of feasible courses starting from a need expressed in natural language by the learner.
  - G1.2: to ensure that generated courses cover the expressed needs and the (optionally) selected skills and contexts (taking into account the available learning material).
  - G1.3: to ensure that the generated courses are personalized on the basis of learner cognitive state and learning preferences.
  - G1.4: to provide an user friendly interface for needs expression, courses 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 CGS.
- **Evaluation hypotheses**
  - H1.1: a set of feasible courses can be effectively and efficiently created starting from a need expressed in natural language and, optionally, a skill and a context.
  - H1.2: the use of the CGS contributes to improve students' motivation.
  - H1.3: the use of the CGS contributes to improve students' understanding of domain concepts.
  - H1.4: the use of CGS contributes to increase students' activity levels.
  - H1.5: the use of the CGS contributes to reduce the time between the emerging of a new learning need and its fulfilment.
  - H1.6: generated courses are considered as a worthy resource by both instructors and students.
- **Experimentation**
  - Apparatus
    - IWT
    - The CGS component
    - An e-learning course on a scientific subtopic
    - A set of additional courses and learning resources on related topics
    - Log files
    - Databases
    - Statistical/Data Analysis software
  - Participants
    - Location: UOC Computer Science Degree
    - Type: Instructors and students
    - Number: 2 instructors, 60 students

- Schedule
  - Start: March 2012
  - End: May 2012
- Design/Methodology

  Students enrolled in a specific on-line course will be asked to study some additional topic not included in the course itself. Three groups will be considered.
  - The first group will be able to use the CGS but will be not able to use standard search facilities to find learning objects and courses in the system resource repository.
  - The second group will be able to use standard search facilities but will be not able to use the CGS.
  - The third group may use both tools.

- **Validation**
  - Criteria
    - C1.1: To evaluate the level of fulfilment of the tool features.
    - C1.2: To evaluate the level of satisfaction of the students that use the CGS.
    - C1.3: To evaluate the increase in students' motivation caused by the use of the CGS.
    - C1.4: To evaluate the increase in students' understanding of key concepts and students' results caused by the use of CGS.
    - C1.5: To evaluate the increase in students' activity levels due to the use of the CGS.
    - C1.6: To evaluate the level of satisfaction of the instructors with the inclusion of the CGS as a learning resource in their courses.
    - C1.7: To evaluate the potential reduction of the time between the emerging of a new learning need and its fulfilment thanks to the CGS.
  - Basic metrics during the pilot
    - M1.1: Number of courses created with the CGS.
    - M1.2: Time employed in creating each course with the CGS.
    - M1.3: Number of students using the CGS.
    - M1.4: Number of visits of learning objects alternative to those included in courses generated by the CGS.
    - M1.5: Students passing the final test and/or with high marks when the CGS is used.
    - M1.6: Students passing the final test and/or with high marks when the CGS is not used.
    - M1.7: Number of students that consider that the CGS is worthy.
    - M1.8: Number of instructors that consider that the CGS is worthy.

- Analysis and techniques
  - Quantitative: Surveys & questionnaires, ANOVA, factor analysis, multiple regression models, etc.
  - Qualitative: Open questionnaires, critical incident method, etc.

5.2 R2. Knowledge Model Contextualization

- **Evaluation goals**
  - G2.1: to develop a Visual Ontology Editor (VOE) for the definition of domain ontologies and contexts with a user friendly interface.
  - G2.2: 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.
  - G2.3: to ensure that generated courses are personalized on the basis of the learner cognitive state and learning preferences.
  - G2.4: to ensure that generated courses allow the effective learning of scientific concepts in selected domains.
  - G2.5: to identify possible ways of improving further the utility of the VOE and related models and algorithms.
- **Evaluation hypotheses**
  - H2.1: 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: the automatically generated courses are compatible with the selected context and are in line with student needs, previous knowledge and learning preferences.
  - H2.3: the use of automatically generated courses contributes to improve students' motivation.
  - H2.4: the use of automatically generated courses contributes to improve students' understanding of domain concepts.
  - H2.5: the use of automatically generated courses contributes to increase students' activity levels.
  - H2.6: automatically generated courses are considered as a worthy educational resource by both instructors and students.
- **Experimentation**
  - Apparatus
    - IWT
    - VOE and Knowledge Model Contextualization components
    - An e-learning course on a scientific topic to be played in two different learning contexts
    - Log files
    - Databases
    - Statistical/Data Analysis software

- Participants
  - Locations: UOC Computer Science Degree, TUG Computer Science Degree
  - Type: Instructors and students
  - Number: 2 instructors, 60 students
- Schedule
  - Start: March 2012
  - End: May 2012
- Design/Methodology

  Students enrolled in two courses about the same topic in two universities will be involved in the experimentation. They will be divided in four groups (two groups for each university). A common subtopic of the course will be selected and will be modelled through a domain ontology. Learning material covering all domain concepts will be selected and indexed with metadata. Two contexts will be defined according to specificities and constraints of the two universities. Four courses will be prepared and assigned to the four groups.
  - The first will be a static course covering the selected subtopic for the first university and will be assigned to the first group.
  - The second will be a static course covering the selected subtopic for the second university and will be assigned to the second group.
  - The third will be a dynamic contextualized course (generated by the system starting from a domain ontology, a context, a target concept and a learner) covering the selected subtopic for the first university and will be assigned to the third group.
  - The fourth will be a dynamic contextualized course covering the selected subtopic for the second university and will be assigned to the fourth group.

- **Validation**
  - Criteria
    - C2.1: To evaluate the level of fulfilment of the tool features.
    - C2.2: To evaluate the level of satisfaction of the instructors that use the VOE.
    - C2.3: To evaluate the level of satisfaction of the instructors with the inclusion of the contextualized courses with their students.
    - C2.4: To evaluate the increase in students' motivation and understanding of domain concepts caused by the use of contextualized courses.
    - C2.5: To evaluate the increase in students' activity levels due to contextualized courses.
    - C2.6: To evaluate the level of satisfaction of the students that use contextualized courses generated by the system.
  - Basic metrics during the pilot
    - M2.1: Number of instructors using the VOE.
    - M2.2: Number of courses created with contextualized ontologies.

        - M2.3: Time employed in creating each course with contextualized ontologies.
        - M2.4: Number of students passing the final test and/or with high marks when they use contextualized courses.
        - M2.5: Number of students passing the final test and/or with high marks when they do not use contextualized courses.
        - M2.6: Instructors that consider that the VOE and contextualized courses are worthy.
        - M2.8: Students that consider that the VOE and contextualized courses are worthy.
    - Analysis and techniques
        - Quantitative: Surveys & questionnaires, ANOVA, factor analysis, multiple regression models, etc.
        - Qualitative: Open questionnaires, critical incident method, etc.

5.3 R3. Semantic Connections Between Learning Resources

- **Evaluation goals**
    - G3.1: to build an editor for Compound Learning Resources (CLRs) that allows efficient building of a CLR even in the case of non-expert instructors (i.e. in a friendly way).
    - G3.2: to playback the generated CLR through a user friendly interface.
    - G3.3: to ensure that a CLR is able to adapt itself on the basis of the context.
    - G3.4: to ensure that a CLR is able to adapt itself basing on teaching and learning preferences.
    - G2.5: to ensure that a CLR allows the effective and efficient learning of scientific concepts in selected domains.
    - G2.6: to identify possible ways of improving further the utility of the CLR and related tools.
- **Evaluation hypotheses**
    - H3.1: a CLR can be effectively created by instructors as well as stored and played by learners through a user friendly interface.
    - H3.2: the use of CLRs contributes to support instructors' task.
    - H3.3: the use of CLRs contributes to improve students' motivation.
    - H3.4: the use of CLRs contribute to improve students' understanding of key concepts.
    - H3.5: the use of CLRs contributes to increase students' activity levels.
    - H3.6: CLRs are considered as a worthy educational resource by both instructors and students.
- **Experimentation**
    - Apparatus
        - IWT
        - Player/Editor for CLRs
        - A CLR on a given scientific subtopic

        - Alternative learning resources on the same subtopic
        - Log files
        - Databases
        - Statistical/Data Analysis software
    - Participants
        - TUG Computer Science Degree
        - Type: instructors, students
        - Number: 2 instructors, 60 students
    - Schedule
        - Start: March 2012
        - End: May 2012
    - Design/Methodology

      Three groups will be considered with equal dimension. The assignment to groups will be random while the participation in the experiment will be on a voluntary basis to build a random sampling. A CLR and alterative learning resources on a given subtopic will be produced.

        - The first group will consist of students who will not use CLR and will only access a standard course made of alternative learning material covering the same subtopic.
        - The first group will consist of students who will not use CLR and will only access a personalized course made of alternative learning material covering the same subtopic.
        - The third group will consist of students who will exclusively use the CLR and will not have access to the alternative learning material.

- **Validation**
    - Criteria
        - C3.1: To evaluate the level of fulfilment of the tool features.
        - C3.2: To evaluate the level of satisfaction of the instructors with the CLR editor.
        - C3.3: To evaluate the level of satisfaction of the instructors with the inclusion of CLRs in their courses.
        - C3.4: To evaluate the increase in students' motivation caused by the use of CLRs.
        - C3.5: To evaluate the increase in students' understanding of key course concepts and students' results caused by the use of CLRs.
        - C3.6: To evaluate the increase in students' activity levels due to the use of CLRs.
        - C3.7: To evaluate the level of satisfaction of students with the inclusion of the SLO in their courses.
    - Basic metrics during the pilot
        - M3.1: Number of instructors using the CLR editor.
        - M3.2: Number of CLRs created.

- M3.3: Time employed in forming new instructors to use the CLR editor.
- M3.4: Time employed in creating each CLR.
- M3.5: Students passing the final test with high marks when CLRs are used.
- M3.6: Students passing the final test with high marks when CLRs are not used.
- M3.7: Instructors that consider that the CLR is worthy.
- M3.8: Instructors that consider that the CLR is worthy.

- Analysis and techniques
  - Quantitative: Surveys & questionnaires, ANOVA, factor analysis, multiple regression models, etc.
  - Qualitative: Open questionnaires, critical incident method, etc.

5.4 R4. Virtualized Collaboration

- **Evaluation goals**
  - G4.1: To build an Authoring System tool for Virtualized Collaborative Session (ASVCS) that is able to build a Virtualized Collaborative Session (VCS) from a threaded discussion (coming from a forum, from a chat or from a set of messages exchanged between different system users).
  - G4.2: To ensure that the aforementioned tool allows efficient building of ASVCS even in the case of non-expert instructors (i.e., in a friendly way and without having to employ too much time).
  - G4.3: To employ the VCS in some of our online and blended courses in order to enhance some aspects of the teaching/learning process.
  - G4.4: To identify possible ways of improving further the utility of the VCS in online and blended courses.
  - G4.5: To create, edit, manage, store and playback the generated storyboard through a user friendly interface.
  - G4.6: To build (automatically) a draft storyboard from a collaborative activity effectively
  - G4.6: To build (automatically) a draft storyboard from a collaborative activity efficiently
- **Evaluation hypotheses**
  - H4.1: A storyboard-based VCS can be efficiently created (in an easy and friendly way for the non-expert instructor) from a threaded discussion.
  - H4.2: Use of VCS contributes to significantly improve students' motivation.
  - H4.3: Use of VCS contributes to support instructors' task.
  - H4.4: Use of VCS contributes to significantly increase students' activity levels, both in individual and collaborative activities.
  - H4.5: Use of VCS contributes to significantly improve students' understanding of key concepts and students' results.
  - H4.6: VCS are considered as a worthy educational resource by both instructors and students.

- **Experimentation**
  - Apparatus
    - IWT
    - Log files
    - Databases
    - Statistical/Data Analysis software
  - Participants
    - Location: UOC-Computer Science degree-GOPI classrooms
    - Type: Instructors and students
    - Number: 3 instructors, 120 students
  - Schedule
    - Start: May 2011
    - End: June 2011
  - Design/Methodology
    - Three groups will be considered. The first group will consist of students who will not use the VCS (but only the traditional forums); the second group will consist of students who will exclusively use the VCS in their course (they will not have access to the original forums); finally, the third group will consist of students who will have access to both the VCS and the classical forums.
- **Validation**
  - Criteria
    - C4.1: To evaluate the level of fulfilment of the tool features.
    - C4.2: To evaluate the level of satisfaction of the instructors with the tool for developing VCS.
    - C4.3: To evaluate the potential increase in students' motivation caused by the use of VCS.
    - C4.4: To evaluate the level of satisfaction of the instructors with the inclusion of VCS in their courses.
    - C4.5: To evaluate the potential increase in students' activity levels due to the incorporation of the VCS.
    - C4.6: To evaluate the potential increase in students' understanding of concepts and students' results.
    - C4.7: To evaluate the level of satisfaction of students with the inclusion of the CVS in their courses.
  - Basic metrics during the pilot
    - M4.1: Number of instructors using the ASVCS tool.
    - M4.2: Number of VCS created with the ASVCS tool.
    - M4.3: Time employed in forming new instructors to use the ASVCS tool.

- M4.4: Time employed in creating each VCS.
- M4.5: Number of students using the VCS.
- M4.6: Number of visits of the VCS.
- M4.7: Number of visits of the traditional forum.
- M4.8: Number of messages submitted by students related to the VCS topics.
- M4.9: Number of messages submitted by students when no VCS is used.
- M4.10: Number of students passing the course and/or with high marks when the VCS is used.
- M4.11: Number of students passing the course and/or with high marks when the VCS is not used.
- M4.12: Number of instructors that consider that the VCS is worthy.
- M4.13: Number of students that consider that the VCS is worthy.

- Analysis and techniques
  - Quantitative: Surveys & questionnaires, ANOVA, factor analysis, multiple regression models, etc.
  - Qualitative: Open questionnaires, critical incident method, etc.

5.5 R5. Storytelling

- **Evaluation goals**
  - G5.1: to build digital storytelling methodologies and tools able to let instructors build a Storytelling Learning Object (SLO) on the basis of the defined storytelling design model.
  - G5.2: to ensure that the aforementioned methodologies and tools allow efficient building of a SLO even in the case of non-expert instructors (i.e. in a friendly way).
  - G5.3: to store and playback the generated SLO through a user friendly interface.
  - G5.4: to ensure that a SLO can be played with different roles and can be adapted basing on the role played by the learner and on his/her user model.
  - G5.5: to ensure that a SLO can perform cognitive and emotional assessment and adapt the story accordingly.
  - G5.6: to ensure that a SLO allows the efficient transmission of lesson learned inside a learning experience on the theme of the risk managements.
  - G5.7: to identify possible ways of improving further the utility of SLOs and related tools in on-line and blended courses.
- **Evaluation hypotheses**
  - H5.1: a SLO can be effectively created by instructors as well as stored and played by learners through a user friendly interface.
  - 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.

- **Experimentation**
  - Apparatus
    - IWT
    - Player/Editor for Narrative Sessions
    - A CLR on a given topic
    - Alternative learning resources on the same topic
    - Log files
    - Databases
    - Statistical/Data Analysis software
  - Participants
    - Location: Istituto Superiore Statale Pitagora/Pozzuoli (NA)
    - Type: instructors, students
    - Number: 2 instructors, 50 students (two classes)
  - Schedule
    - Start: March 2012
    - End: May 2012
  - Design/Methodology

    Three groups will be considered. The experimentation design will be of pre-test/post-test type with an experimental group and two control groups with equal dimension. The assignment to groups will be random while the participation in the experiment will be on a voluntary basis in order to build a random sampling. A SLO and alterative learning objects on a given topic about the behaviour to be taken when a big risk happen will be produced.
    - The first group will consist of students who will not use SLO and will only access alternative learning material covering the same topics.
    - The second group will consist of students who will exclusively use the SLO in their course (they will not have access to the alternative learning material).
    - The third group will consist of students who will have access to both the SLO and the alterative learning material.

- **Validation**
  - Criteria
    - C5.1: To evaluate the level of fulfilment of the tool features.
    - C5.2: To evaluate the satisfaction of the instructors with the system for building

SLOs.

- 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.

- Basic metrics during the pilot
  - M5.1: Number of instructors using the storytelling system.
  - M5.2: Number of SLO created with the storytelling system.
  - M5.3: Time employed in forming new instructors to use the storytelling system.
  - M5.4: Time employed in creating each SLO.
  - 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.
- Analysis and techniques
  - Quantitative: Surveys & questionnaires, ANOVA, factor analysis, multiple regression models, etc.
  - Qualitative: Open questionnaires, critical incident method, etc.

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

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

- G6.5: To ensure that the SG allows the efficient transmission of lesson learned inside a learning experience on the theme of the risk managements.

- **Evaluation 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 contribute to improve students' understanding of key concepts as well as related skills.
  - H6.6: SGs are considered as a worthy educational resource by both instructors and students.

- **Experimentation**
  - Apparatus
    - IWT
    - Player/Editor for Serious Games
    - A Serious Game on a given topic
    - Alternative learning resources on the same topic
    - Log files
    - Databases
    - Statistical/Data Analysis software
  - Participants
    - Location: Istituto Superiore Statale Pitagora/Pozzuoli (NA)
    - Type: Instructors and students of Earth Science
    - Number: 2 instructors, 50 students (two classes)
  - Schedule
    - Start: March 2012
    - End: May 2012
  - Design/Methodology

    A SG and alterative learning objects on a given topic about the behaviour to be taken when a big risk happen will be produced. Three groups will be considered. The experimentation design will be of pre-test/post-test type with an experimental group and two control groups with equal dimension. The assignment to groups will be random while the participation in the experiment will be on a voluntary basis in order to build a random sampling.

    - The first group will consist of students who will not use the Serious Game for acknowledging an evacuation planning; they use only access alternative learning

material covering the same topics.

- The second group will consist of students who will exclusively use the SG in their course (they will not have access to the alternative learning material).
- The third group will consist of students who will have access to both the SG and the alterative learning material.

- **Validation**
  - Criteria
    - C6.1: To evaluate the increase in students' motivation caused by the use of a SG.
    - C6.2: To evaluate the level of satisfaction of the instructors with the inclusion of SG in their courses.
    - C6.3: To evaluate the increase in students' activity levels due to the use of the SG.
    - C6.4: To evaluate the increase in students' understanding of key domain concepts and students' results.
    - C6.5: To evaluate the level of satisfaction of students with the inclusion of the SG in their courses.
  - Basic metrics during the pilot
    - 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.
  - Analysis and techniques
    - Quantitative: Surveys & questionnaires, ANOVA, factor analysis, multiple regression models, etc.
    - Qualitative: Open questionnaires, critical incident method, etc.

5.7 R7. Affective and Emotional Approaches

- **Evaluation goals**
  - G7.1: to build a system that is able to recognize, evaluate and stimulate the emotions and the affective state of a learner in order to support and improve learning.

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

- **Evaluation hypotheses**
  - H7.1: it is possible to create a learning system able to stimulate the affectivity and the emotionality of a learner.
  - H7.2: by recognizing and assisting emotions and affectivity it is possible to improve students' motivation and to create a predisposition to learning.
  - H7.3: by recognizing and assisting emotions and affectivity it is possible to improve students' understanding of domain concepts.
  - H7.4: The visualization and interaction of appropriate learning resources improves the emotional state altered.
  - H7.5: the system for emotional/affective management is considered as a worthy resource by both instructors and students.
  - H7.6: the use of system for emotional/affective management contributes to significantly increase students' activity levels.
- **Experimentation**
  - Apparatus
    - IWT
    - Component for emotional/affective management.
    - A course on a given topic
    - Log files
    - Databases
    - Statistical/Data Analysis software
  - Participants
    - Location: Istituto Superiore Statale Pitagora/Pozzuoli (NA)
    - Type: Instructors, students and experts in cognitive science
    - Number: 2 instructors, 60 students and 1 expert in cognitive sciences.
  - Schedule

- Start: March 2012
- End: May 2012

- Design/Methodology

  Two groups will be considered. The experimentation design will be of pre-test/post-test type. The assignment to groups will be random while the participation in the experiment will be on a voluntary basis in order to build a random sampling.

  - The first group will be involved in a learning experience without the support of the component for emotional/affective management.
  - The second group will be involved in the same learning experience but with the additional component included.

- **Validation**
  - 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.
    - C1.4: 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.
    - C6.6: To evaluate the increase in students' understanding of concepts and students' results due to the affective and emotional support
  - Basic metrics during the pilot
    - 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.
    - M1.7: Number of students passing the final test and/or with high marks when the emotional/affective system is used.
    - M1.8: Number of students passing the final test and/or with high marks when the emotional/affective system is not used.
  - Analysis and techniques
    - Quantitative: Surveys & questionnaires, ANOVA, factor analysis, multiple regression models, etc.
    - Qualitative: Open questionnaires, critical incident method, etc.

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

- **Evaluation goals**
  - G8.1: To provide a WIKI system that can be used collaboratively for writing assignments enriched with self- and peer assessment
  - G8.2: To provide a WIKI system which is enhanced with a visualization tool for individual contributions over time on the level of the assignment
  - G8.3: To provide a tool that allows the learners self-review of their own contribution.
  - G8.4: To provide a tool that allows the learners peer-review others' contributions.
  - G8.5: To ensure that the assessment is recorded and that it appears in the reports managed by the system.
  - G8.6: To receive valuable feedback out of both self- and peer assessment as well as the teachers final review.
  - G8.7: To identify possible improvements for the tool.
- **Evaluation hypotheses**
  - H8.1: Using the tool supports students in their learning process.
  - H8.2: The tool facilitates the work for the instructors.
  - H8.3: The tool allows an efficient and user-friendly management of the assessment.
  - H8.4: The feedback provided by the tool supports students in their learning process.
  - H8.5: Using an enhanced visualization tool as part of the peer review process motivates the students to effectively contribute assignments.
- **Experimentation**
  - Apparatus
    - IWT
    - Statistical/Data Analysis software
  - Participants
    - Location: Classrooms at TUG
    - Type: Instructor and students
    - Number: 1 instructor, approximately 20 to 30 students
  - Schedule
    - Experiment will take place in the summer term 2011 within a regular course at TUG.
  - Design/Methodology
    - Students are asked to build small groups. Each group has to work on a topic assigned by the instructor and to write an essay about this topic. Using the tool, students are asked to self-review their contributions. In addition, individual contributions are peer-reviewed within the small groups (also considering

engagement etc. of the individual student). Furthermore, contribution from each group is going to be peer-reviewed by the other groups to evaluate group performance. Finally, the instructor and the students are asked to fill in a questionnaire in order to evaluate the tool.

- **Validation**
  - 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.
  - Basic metrics during the pilot
    - 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.
  - Analysis and techniques
    - Quantitative: Surveys & questionnaires, ANOVA, t-tests, correlation etc.
    - Qualitative: Open questionnaires etc.

5.9 R9. Assessment in self-regulated learning

- **Evaluation goals**
  - G9.1: To provide a new form of assessment where automatic question generation is used to create assessments for self-regulated learning style.
  - G9.2: 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.
  - G9.3: To ensure that all types of questions provided from the automatic question generator are high in quality.
  - G9.4: To ensure that the answers provided by the tool are relevant and meaningful.
  - G9.5: To ensure that the concepts automatically extracted by the tool from a given text are

relevant.

- G9.6: To ensure that the tool is user-friendly.
- G9.7: To ensure that the tool supports self-regulated learning.
- G9.8: To provide a tool that creates questions using concepts entered by users.
- G9.9: To identify possible improvements for the tool.

- **Evaluation 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 use of the tool is easy even if the user is a non-expert.
  - H9.6: Using the tool supports students' self-regulated learning; i.e., students benefit from the tool during their learning process.
  - H9.7: 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.
  - H9.8: Using the tool has a positive impact on the users' motivation.
- **Experimentation**
  - Apparatus
    - IWT
    - Statistical/Data analysis software (eg., Excel; SPSS)
  - Participants
    - Location: Classrooms at TUG
    - Type: Students
    - Number: approximately 8 to 30 students per study
  - Schedule
    - Studies (1a to 1c) take place in the summer term 2011
  - Design/Methodology
    - Study 1a /Pre-study1: Evaluation of concepts and questions provided by the Automatic Question Generator (AQG))

      At the beginning of the experiment, students are asked to learn an English text. Then they have to extract concepts and questions from this text to become familiar with the topic. A following test also ensures that students have learnt the text well. Students are then asked to evaluate concepts and questions that had been generated from the text in advance, using the AQG. In addition, they also have to evaluate concepts and questions that were extracted by humans. These concepts and questions serve as controls to ensure whether participants worked

on the evaluation task seriously.

- Study1b/Pre-study2: Evaluation of questions provided by the AQG, using concepts that were extracted by students

  The concepts extracted from participants in Study1a are used to generate questions with the AQG. Afterwards, those questions are evaluated by a new sample of students and compared to questions generated by humans.

- Study1c/Pre-study3:

  Students are asked to learn a text using the AQG. At the end of the experiment they have to fill in a questionnaire to evaluate the tool. For instance, they are asked about their experience with the AQG and whether they were supported by the tool with respect to their learning outcomes etc.

- Study 2: Evaluation of the AQG in self-regulated learning

  Half of the students (that had not participated in the studies before) are asked to use the AQG while learning a text (experimental group) whereas the other half has to learn the same text without the support of the tool (control group). However, for a better comparability, the latter group is asked to learn the text by extracting relevant concepts and creating questions from the test. Afterwards, both groups have to attend a test. The experimental group is also asked to fill out a questionnaire in which the usability/functionality of the AQG is evaluated and motivational aspects are inquired.

- **Validation**
  - Criteria
    - C9.1: To evaluate the different question types provided by the AQG.
    - C9.2: To evaluate the quality (i.e., pertinency and terminology) of the questions provided by the AQG.
    - C9.3: To evaluate the level (i.e., difficulty) of the questions provided by the AQG.
    - C9.4: To evaluate the relevance of answers provided by the AQG.
    - C9.5: To evaluate the distractors provided by the AQG for multiple-choice questions.
    - C9.6: To evaluate the concepts provided by the AQG.
    - C9.7: To evaluate questions generated by the AQG, 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.
  - Basic metrics during the pilot
    - 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 AQG.
- 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.

- Analysis and techniques
  - Quantitative: Surveys & questionnaires, ANOVA, t-tests, correlations, etc.
  - Qualitative: Open questionnaires, etc.

6 Working and contingency plans

This section shows the work and contingency plans for the experimentation and validation activities performed in ALICE.

6.1 Working plans

The working plan for all tasks and subtasks of experimentation and validation, including the duration of the tasks and other relevant information, was defined in Section 4.8 of the report ALICE project (2010c).

References to the working plans are necessary to consider some of the risks of the project, as well as the contingency plans described next.

6.2 Contingency plans

In line with Section 5.1 of the report ALICE project (2010c) Risk assessment and contingency planning is carefully taken into account for experimentation and validation activities. The most potential risks found and the corresponding contingency plans are:

Risks	Probability	Impact	Contingency plans
Not respecting the planning and meeting the deadlines of WP 8	low	medium/high	The risk is reduced by the expertise of the partners (technical skills and management experience) that will permit to anticipate planning drifts, and by a development, which starts very soon in the project and will proceed incrementally. The RTM and each WP leader will work to avoid the occurrence of this risk and to mitigate its impact.
Delays in the release of the prototypes of the first cycle, including integration (task 8.1.1. due to end April 2011)	medium	high	Reduce subsequent “refinement” activities or Request a project extension to the EC (to be decided in the next PMB meeting)
Prototype problems during experimentation.	medium	medium/high	.MoMA will take care of bugs during integration of prototypes into IWT and also will technically support experimentation at UOC

References

ALICE project (2010a). Annex I – “Description of Work”. V1.0. May 17th, 2010.

ALICE project (2010b). Deliverable D1.1 - “Requirements”. V1.0. August 31st, 2010.

ALICE project (2010c). Deliverable D11.1 – “Detailed activity plan”. V1.0. October 31st, 2010.

Anscombe, F. J. (1948). "The Validity of Comparative Experiments". Journal of the Royal Statistical Society. Series A (General) 111 (3): 181–211.

Babbie, E. (2004). The Practice of Social Research, 10th edition, Wadsworth, Thomson Learning Inc.

Bortz, J. & Döring, N. (2006). Research Methods and Evaluation. Heidelberg: Springer.

Child, Dennis (1973), The Essentials of Factor Analysis, London: Holt, Rinehart & Winston.

Cliff, N. (1996). Ordinal Methods for Behavioral Data Analysis. Mahwah, NJ: Lawrence Erlbaum.

Cliff, N. & Keats, J. A. (2003). Ordinal Measurement in the Behavioral Sciences. Mahwah, NJ: Erlbaum.

Denzin, N.K., & Lincoln, Y.S. (Eds.) (1994). Handbook of qualitative research. Thousand Oaks, CA: Sage. Chapter 27. Huberman, A.M. & Miles, M.B. "Data Management and Analysis Methods".

Dielman, T.E. (1996). Applied Regression Analysis for Business and Economics. International Thompson Publishing, Inc.: Wadsworth Publishing Company.

Dodge, Y (2003) The Oxford Dictionary of Statistical Terms, OUP.

Cramer, D. & Howitt, D. (2004). The Sage Dictionary of Statistics. p. 76.

Gorsuch, R. L. (1983) Factor Analysis. Hillsdale, NJ: Lawrence Erlbaum.

Guba, E.G., & Lincoln, Y.S. (1989). Fourth generation evaluation. Newbury Park, CA: Sage.

Miles, M.B., & Huberman, A.M. (1994).Qualitative data analysis: An expanded sourcebook. (2nd ed.). Newbury Park, CA: Sage.

Hill, T. & Lewicki, P. (2007). STATISTICS Methods and Applications. StatSoft, Tulsa, OK.

Johnson, N.L. & Kotz, S.; Balakrishnan, N. (1994). Continuous univariate distributions, Volume 1. Wiley.

Johnson, N.L. & Kotz, S.; Balakrishnan, N. (1994). Continuous univariate distributions, Volume 2. Wiley.

Kolmogorov, A. (1933) "Sulla determinazione empirica di una legge di distribuzione" G. Inst. Ital. Attuari, 4, 83

Lehmann, E.L.; Joseph P. Romano (2005). Testing Statistical Hypotheses (3E ed.). New York: Springer.

Likert, R. (1932). A Technique for the Measurement of Attitudes. Archives of Psychology 140: 1–55.

Lindman HR (1974). Analysis of variance in complex experimental designs. San Francisco: W. H. Freeman & Co.. p. 33.

MacQueen, J. B. (1967). Some Methods for classification and Analysis of Multivariate Observations, Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, University of California Press, 1:281-297

Mann PS (1995) Introductory Statistics, 2nd Edition, Wiley.

Narens, L. (1981a). A general theory of ratio scalability with remarks about the measurement-theoretic concept of meaningfulness. Theory and Decision, 13, 1–70.

Narens, L. (1981b). On the scales of measurement. Journal of Mathematical Psychology, 24, 249–275.

Ostle B. & Malone, L. (1988). Statistics in Research: Basic Concepts and. Techniques for Research Workers (4th ed.). Ames: Iowa State University Press.

Patton, M.Q. (1990). Qualitative evaluation and research methods. (2nd ed.). Newbury Park, CA: Sage.

Shapiro, S. S. & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). Biometrika 52 (3-4): 591–611.

Stigler, Stephen M. (1986). The history of statistics: the measurement of uncertainty before 1900. Harvard University Press. ISBN 0-674-40340-1.

Strauss, A., & Corbin, J. (1990). Basics of qualitative research: Grounded theory procedures and techniques. Beverly Hills, CA: Sage.

Trochim, W. and Land, D. (1982). Designing designs for research. The Researcher, 1, 1, pgs. 1-6).

Wolcott, H.F. (1990). "On seeking - and rejecting - validity in qualitative research." In E.W. Eisner and A. Peshkin (Eds.), Qualitative inquiry in education: The continuing debate, (pp. 121-152). NY: Teachers College Press.

Wolfe, C. (1993). Quantitative Reasoning Across a Curriculum. College Teaching, 41, 2-8.

Wuensch, Karl L. (2005). "What is a Likert Scale? and How Do You Pronounce 'Likert?"'. East Carolina University.

Zimmerman, Donald W. (1997). "A Note on Interpretation of the Paired-Samples t Test". Journal of Educational and Behavioral Statistics 22 (3): 349–360.