

027607

mGBL

mobile Game Based Learning

Specific Targeted Research Project

Information Society Technologies

## D 7.1 Evaluation plan

[revised version]

Due date: March 31<sup>st</sup>, 2007

Actual submission: April 18<sup>th</sup>, 2007

Start date of project: 1. October 2005

Duration: 36 months

[TRIESTE]

Version 1.2

| Project co-funded by the European Union within the Sixth Framework Programme (2002-2006) |   |   |
|--|---|---|
| Dissemination Level  |   |   |
| PU   | Public  | X |
| PP   | Restricted to other programme participants (including the Commission Services)        |   |
| RE   | Restricted to a group specified by the consortium (including the Commission Services) |   |
| CO   | Confidential, only for members of the consortium (including the Commission Services)  |   |

## TABLE OF CONTENTS

|          |   |           |
|----------|---|-----------|
| <b>1</b> | <b>GENERAL DELIVERABLE INFORMATION .....</b>                                      | <b>4</b>  |
| 1.1      | GENERAL DELIVERABLE DESCRIPTION .....   | 4         |
| 1.2      | REVISION HISTORY OF THIS DOCUMENT .....   | 4         |
| 1.3      | EXTERNAL PEER-REVIEW (INTERNAL CHAPTER, NOT PART OF THE PUBLIC DELIVERABLE) ..... | 5         |
| 1.3.2    | General description of the review process.....                                    | 5         |
| 1.3.3    | Comments and Recommendations of the External Reviewers .....                      | 5         |
| 1.3.4    | Justification and actions/corrections taken regarding the external peer-review    | 6         |
| 1.4      | EXECUTIVE SUMMARY OF THE DELIVERABLE.....   | 6         |
| <b>2</b> | <b>INTRODUCTION TO THE REVIEWED VERSION.....</b>                                  | <b>7</b>  |
| <b>3</b> | <b>INTRODUCTION: MGBL PROJECT FRAMEWORK .....</b>                                 | <b>7</b>  |
| <b>4</b> | <b>MGBL EVALUATION FRAMEWORK.....</b>   | <b>8</b>  |
| 4.1      | OVERVIEW.....   | 8         |
| 4.2      | PURPOSES .....  | 11        |
| 4.3      | OVERALL ASSESSMENT CRITERIA .....   | 11        |
| 4.4      | EVALUATING MOBILE USABILITY OF DIGITAL LEARNING MATERIAL .....                    | 13        |
| 4.5      | FACTORS INFLUENCING THE APPROACH .....  | 15        |
| 4.6      | FITTING THE MINISTERIAL DECLARATION OF RIGA OF JUNE 11, 2006.....                 | 16        |
| <b>5</b> | <b>REVIEW OF EVALUATION METHODS AND TECHNIQUES FOR MGBL .....</b>                 | <b>16</b> |
| 5.1      | USER TRIALS .....   | 17        |
| 5.1.1    | Overview.....   | 17        |
| 5.1.2    | Resources .....   | 18        |
| 5.2      | FIELD TRIALS .....  | 19        |
| 5.2.1    | Overview.....   | 19        |
| 5.2.2    | Resources .....   | 21        |
| 5.3      | USERS' QUESTIONNAIRES .....   | 22        |
| 5.4      | STUDENT KNOWLEDGE TESTS .....   | 27        |
| 5.5      | EXPERT REVIEWS .....  | 28        |
| 5.6      | PERSONS INVOLVED IN USER TRIALS .....   | 29        |
| 5.6.1    | USERS.....  | 30        |
| 5.6.2    | USERS HELPERS .....   | 30        |
| 5.6.3    | SELECTION OF USERS.....   | 31        |
| 5.6.4    | SPECIAL CONSIDERATIONS.....   | 31        |
| <b>6</b> | <b>PEDAGOGICAL ISSUES.....</b>  | <b>34</b> |
| 6.1      | EVALUATING THE LEARNING PROCESSES .....   | 34        |
| 6.2      | ACTIVITY THEORY AS A THEORETICAL FRAME OF EDUCATIONAL GAMES .....                 | 35        |
| 6.3      | EDUCATIONAL PROCESS: BLOOM'S TAXONOMY .....                                       | 37        |
| <b>7</b> | <b>MGBL USER TRIALS: IMPLEMENTATION PLAN .....</b>                                | <b>42</b> |
| 7.1      | STEP 1: PLANNING.....   | 43        |
| 7.1.1    | Usability goals.....  | 43        |
| 7.1.2    | Test design.....  | 43        |
| 7.1.3    | Testers and Observers' Roles .....  | 46        |
| 7.1.4    | Number of subjects.....   | 46        |
| 7.1.5    | Procedure description.....  | 47        |
| 7.1.6    | Pilot trials.....   | 47        |
| 7.2      | STEP 2: LOGISTICS .....   | 47        |
| 7.2.1    | Test material .....   | 47        |

|           |   |           |
|-----------|---|-----------|
| 7.2.2     | Tasks.....  | 49        |
| 7.2.3     | Instructions and demonstrations.....                                    | 50        |
| 7.2.4     | Interviews and questionnaires .....                                     | 50        |
| 7.2.5     | Observational measures .....  | 51        |
| 7.2.6     | Time Line .....   | 51        |
| 7.3       | STEP 3: THE TRIAL.....  | 52        |
| 7.4       | STEP 4: DATA ANALYSIS .....   | 52        |
| 7.5       | STEP 5: IMPLICATIONS .....  | 54        |
| <b>8</b>  | <b>USER KNOWLEDGE TEST: IMPLEMENTATION PLAN .....</b>                   | <b>54</b> |
| 8.1       | STEP 1: TESTS DESIGN.....   | 54        |
| 8.2       | STEP 2: TESTING OF THE DRAFT .....                                      | 55        |
| 8.3       | STEP 3: LOGISTICS .....   | 55        |
| 8.4       | STEP 4: DATA ANALYSIS.....  | 55        |
| 8.5       | STEP 5:IMPLICATIONS.....  | 57        |
| <b>9</b>  | <b>DEVELOPING EVALUATION REPORTS.....</b>                               | <b>58</b> |
| <b>10</b> | <b>CONCLUSION: MGBL EVALUATION GUIDELINES.....</b>                      | <b>61</b> |
|           | <b>APPENDIX A: USEFUL TIPS FOR SURVEYS AND INTERVIEWS' DESIGN.....</b>  | <b>67</b> |
|           | <b>APPENDIX B: PROCEDURES FOR THE SELECTION OF THE USER GROUPS.....</b> | <b>71</b> |
|           | <b>APPENDIX C: EVALUATION SCHEDULING .....</b>                          | <b>73</b> |

# 1 General Deliverable information

This section provides general information about the deliverable.

They are:

- General Deliverable Description
- Revision history
- External peer-review (internal, not part of the public deliverable)
- Executive Summary of the Deliverable

## 1.1 General Deliverable Description

|                                  |   |
|----------------------------------|---|
| WP number:                       | <b>7</b>  |
| WP name:                         | <b>Evaluation and Validation</b>  |
| Deliverable number:              | <b>D7.1</b>   |
| Deliverable name:                | <b>Evaluation plan</b>  |
| Responsible work package leader: | <p><b>name: Paolo Inchingolo</b><br/> <b>address: via Valerio 10</b><br/> <b>email: inchingolo@bioing.units.it</b><br/> <b>phone: -</b><br/> <b>mobile: -</b><br/> <b>fax: +39 040 558 3460</b></p> |
| Involved project partners:       | <b>[ARC-sr, Ultralab, PFRI, FFRI, Aster, AZM-LU]</b>  |

Table 1: General Deliverable Description

## 1.2 Revision history of this document

| <b>Date</b> | <b>Version</b> | <b>Description</b> | <b>Author</b> |
|-------------|----------------|--------------------|---------------|
| 31/03/06    | 1.0            | Draft              | TRIESTE       |
| 14/06/06    | 1.1            | Final version      | TRIESTE       |
| 30/03/07    | 1.2            | Reviewed version   | TRIESTE       |

### **1.3 External peer-review (internal chapter, not part of the public deliverable)**

This section contains a description and an overview of the results of the external peer-review of the deliverable. This is an internal chapter (consortium, Project Officer and reviewers) and will be removed within the final public version of the deliverable.

#### *1.3.2 General description of the review process*

Original D7.1 deliverable was reviewed by EU review panel at the end of year 1 activity report and accepted as draft.

This modified D7.1 was reviewed by prof. G. Vercelli.

Reviewer profile: Gianni Vercelli received his Laurea degree in electronic engineering in 1987 and his Ph.D. in computer science in 1992. He was with the University of Trieste, Italy, from 1996 to 1999, and he is currently an Assistant Professor in Computer Science and Multimedia Design at the Education Faculty of the University of Genoa. He is a member of the IEEE Computer Society and of the Italian Association for Artificial Intelligence. His scientific interests are focused on robotics and artificial intelligence, intelligent agents, and multimedia education. He has written more than 70 papers.

#### *1.3.3 Comments and Recommendations of the External Reviewers*

Comments by EU review panel on original D7.1:

“D7.1 is very general, it might have been better if it had provided more concrete guidelines on how to perform the different evaluations and on how to report on the outcomes of the evaluation. The usefulness of the deliverable is questionable.”

“The deliverable is too general, it is recommended to resubmit as a guideline on how to perform evaluation – more focused on the steps to be

performed and on the methods to be used for summarizing the evaluation. In that way the document may have more use for the project.”

Comments by prof. G. Vercelli on modified D7.1:

The document is readable and well structured. No particular objections.

#### *1.3.4 Justification and actions/corrections taken regarding the external peer-review*

The whole deliverable was revised: in this way we hope the document may be more useful for the mGBL project. It provides more concrete guidelines on how to perform the different evaluations: chapters 7 and 8 offer concrete implementation plans, describing step by step how to perform the different evaluations.

Chapter 9 is a general guideline on reporting the outcomes and summarizing the evaluation.

Chapter 10 offers a set of recommendations to perform evaluations described in this report.

After the review by prof. Vercelli we made some minor corrections on some pointed issues (added also some explaining foot-notes).

### **1.4 Executive Summary of the Deliverable**

An evaluation plan (including details of the used methodology and techniques) for the whole project. This evaluation plan contains proceedings to evaluate the following:

- results of the user trials;
- attainment of the aim to provide a platform enabling learning in an effective, efficient and playing way.

## 2 Introduction to the reviewed version

This last version was submitted to meet formal requirements emerged after the first project review.

The whole deliverable was revised: in this way we hope the document may be more useful for the mGBL project. It provides more concrete guidelines on how to perform the different evaluations: chapters 7 and 8 offer concrete implementation plans, describing step by step how to perform the different evaluations.

Chapter 9 is a general guideline on reporting the outcomes and summarizing the evaluation.

Chapter 10 offers a set of recommendations to perform evaluations described in this report.

## 3 Introduction: mGBL project framework

*mGBL addresses a two-fold need in the EU:*

- 1. the need to support decision making in critical situations, not only in a cognitive but also in an emotional way. Examples are career-related decisions, business-related decisions or decisions in the context of health environment (epidemics, disasters, etc.).*
- 2. and as a consequence, the need to build on cutting edge work in the new field of m-learning with research-based development on interactive game based learning using mobile devices.*

*mGBL will prototype a platform for the cost- and time-efficient development and deployment of mobile games. mGBL target audiences are mainly students and younger people, with high interest in mobile technologies and in lifelong learning, and their teachers. mGBL example implementations will be in the fields of e-health, e-commerce and career guidance. A special focus is on the implementation of mechanisms known from marketing and psychology to trigger an emotional learning process.*

*To understand and support this emotional part of decisions by using mobile games is the common umbrella of decisions in the different fields mentioned above.*

*The overall goal of the project is to improve the effectiveness and efficiency of learning in the target group of young people through the development of innovative learning models based on mobile Games.*

*The specific aim of the project is to design, develop and trial a prototype game platform that can be used to efficiently develop games for m-learning, whereby the focus is on the support of decision making in critical situations, not only in a cognitive but also in an emotional way. These games shall firstly directly support learning via opportunities to develop knowledge and cognitive skills in an exciting and inspiring – thus in a highly emotional - way, and secondly indirectly motivate users to refer to other media (e.g. "classic" libraries, scripts, etc.) for learning purposes.*

*In order to support the goal of being able to author suitable mobile games efficiently, a platform will be developed, consisting of an authoring tool, a module for measuring utilization and learning success, and a deployment module. This platform will enable teachers to develop individual mobile games from their existing content (scripts, books etc.) and use predefined game templates fast and easily and to distribute them to their students.*

## **4 mGBL evaluation framework**

### **4.1 Overview**

Extracts from mGBL approved project (WP 7, 2 and 6):

*"User trials at different universities and also at institutions performing educational advice services shall yield qualitative and quantitative data allowing for the measurement of goal achievement (i.e. supporting*

*effective and efficient learning through the utilization of the mGBL platform).*

*Evaluation strategies that may be deployed during user trials include observations, user diaries/log books, pre and post activity/project assessment and focus groups. In addition the consortium will sub-contract the services of an external evaluator, who will closely monitor and evaluate project development and outcomes throughout the project. The evaluator will be recruited from an organisation within the ALADIN network, for example the University of Budapest.*

*Measures of project success to be monitored and analysed in the evaluation tasks will be developed in collaboration with the external evaluator and with user groups involved in the trials. Success criteria will include:*

- at least 50% of vocational education and training organisations in the target fields in the countries of the partner organisations have received disseminated information or attended*
- dissemination events and are therefore aware of the project and the educational potential of state of the art mobile games technologies*
- enjoyment of use of mGBL prototype by at least 75% users involved in trials*
- changes in attitudes to learning and/or learning habits by at least 75% of learners involved in trials, for example enhanced interest in related formal learning opportunities*
- perceived improvements in at least 2 specific critical decisions, by at least 75% of learners involved in trials*

*...*

*The end user-panel will include younger people who are at a decision stage to decide their further education or students in the field of e-health*

*and e-commerce. We plan to set up one-to-one in-depth interviews with minimum  $n=30$  young adults in minimum 3 partner countries (minimum total sample:  $n=90$ ). The sample will be drawn from mGBL target audiences. Each interview will take about 1 hour using a semi-structured interviewer-guide with closed and mainly open-ended questions. evolaris will co-ordinate the study and provide all participants with a questionnaire and a report-template for analyses. Recruitment and conduction of fieldwork has to be organised by each participating partner. The end user panel will also constitute the sample for user trials in WP 6 and will also be part of an online user panel that will be set up. The online panel (with online chats, discussion groups, etc.) will consist of all kinds of potential users of the mGBL platform (students, teachers, IT staff) and will be active throughout the whole project. This way user needs and wants can be constantly monitored and iteratively fed into system specification and development activities. The participants of the online panel thus form a group of co-researchers in a kind of living laboratory. The main task here is to constantly manage the panel, collect input, provide feedback, organise online chats, etc.*

...

*The user trials will also build upon the user panel selected in work package 2 and iteratively inform the specification and development work packages. A first user trial of prototype games will start in month 9 in order to gain fast feedback and 3 successive user trials will provide input for evaluation."*

Under these premises, and from analysis of the full project milestones, in our case there will be a sort of overlapping between the concepts of user trials and field trials, as they were defined in previous paragraphs. From now on we will continue to use the term "User trials" but it must be noted that in this context we mean this includes the field trials.

Also there will be need of strict cooperation between WP2 and WP6, since the focus groups created by WP2 should furnish the users who will participate in the WP6 user trials.

## 4.2 Purposes

User trials at different universities and also at institutions performing educational advice services shall yield qualitative and quantitative data allowing for the measurement of goal achievement (i.e. supporting effective and efficient learning through the utilization of the mGBL platform).

The purpose of this evaluation plan is to provide timely and accurate data for use by the participating organizations with decisions related to the development, research and project management processes.

### The specific sub-purposes:

- To collect **information regarding the acceptability, usability, and effectiveness from the target audience of students** involved in the user trials
- To collect **information regarding the content accuracy, usability, and appropriateness from experts in the areas of education and information technology.**

## 4.3 Overall assessment criteria

The following broad assessment criteria will be used based loosely on the ISO 9241 Part 11 concept of utility and its definition of **usability**:

*Usability is the extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use.*

ISO TS 16071 (Guidance on Accessibility of Human Computer Interfaces) defines **accessibility** in relation to usability as:

*The usability of a product, service, environment or facility by people with the wildest range of capabilities.*

This difference between accessibility of the system and usability of the software is an important distinction to make. Accessing to a service is not enough. In fact, although a Web site is accessible to users, it still may not be sufficiently usable to such users, and therefore, guidelines need to be followed by Web developers to achieve this. The goal of usability is to make users' experience with the mGBL games more efficient and satisfying.

Since more prolonged use will be necessary to assess the following usability aspects adequately, they are expected to assume greater importance during the final evaluation study of the project:

- **Utility** – It is essential to assess whether the software is of value for the target group. It will therefore be critical to assess to what extent the software promotes the user's interest in lifelong learning or the development of decision making skill for various critical situations.
- **Effectiveness** – It will be important to assess how well the software supports users in developing these abilities.
- **Efficiency** – Efficiency is related to the effort required to perform activities, and can be assessed in part by the amount of the time that it takes to perform expected results playing the game and the number of keystrokes needed for system operation. Whilst such performance measures are of value in HCI (Human Computer Interaction) assessment, it is considered that in itself time needed to perform a task is of limited value in this context, and a more

relevant question is whether the game provides advantages over alternative (current) ways of performing the same task.

- **Satisfaction** – This involves an assessment whether the user enjoys using the software and is also related to motivation to continue playing it in the future. As well as subjective opinion, this can also be indirectly measured by monitoring whether the software continues to be used over a period of time.

#### 4.4 Evaluating mobile usability of digital learning material

Whilst the criteria so far discussed are germane to all developing software systems, there has been some work done specifically on the mobile learning area. The University of Tampere has developed some criteria and methods for evaluating mobile learning systems in the Digital Learning project in Finland, financed by TEKES (Finnish National Research Fund, 2002-2004). The mobile learning research group in this project has tried to extend the criteria for pedagogical learning environments to the mobile sets with future research (scenario and Delphi research techniques) and mostly qualitative methods. The derived criteria are divided into two categories: technical and pedagogical mobile usability. Following components of the technical mobile usability were presented:

1. Accessibility,
2. Learnability and memorability,<sup>1</sup>
5. Graphical layout,
6. Reliability,
7. Consistency,
8. Efficiency,
9. Memory load and

---

<sup>1</sup> Learnability and memorability: Learnability concerns novice, and memorability concerns casual expert user (Nielsen, 1993, 31). System that is hard to learn is only valuable for those users who are able to spend time to learn it. System that is impossible to learn has no value for any user.

## 10. Errors.

Components of the pedagogical mobile usability are as follows:

2. Learner activity,
3. Cooperative learning,
4. Goal orientation,
5. Applicability,
6. Effectiveness, and
8. Valuation of previous knowledge.

There are a wide variety of other authors offering complementary and perhaps competing criteria for all the various aspects of evaluating mobile learning; Bates and Poole (2003) have proposed a model for the effective use of technology for teaching in higher education that suggests eight criteria to be used in determining choice of technology. An investigation of whether the right technology has been selected is arguably an important aspect of a comprehensive evaluation of mobile learning. It would therefore have to take account of these criteria, namely:

- the appropriateness of the technology for students
- ease of use and reliability
- costs
- teaching and learning approaches
- interactivity
- organizational issues
- novelty, as a choice not to use existing technology
- speed, i.e. how quickly materials can be developed

We can make some progress on the basis that mobile learning pilots and trials each have their own aims and objectives, and that these have driven evaluation in the sense of defining the outcomes sought by the evaluation

and hence driving the selection and development of the techniques, instruments and protocols used in evaluation.

Perhaps the obvious potential objective for any educational pilot or trial is cognitive change, where students have learnt something new. An evaluation may be looking for this or might be looking for meta-cognitive change, where students have learned something about the process of learning. An evaluation may also be looking for affective changes in students, reflecting changed feelings, values or preferences and it may also look for social changes, perhaps in how students relate to or work with each other, or in how groups of students show increased collective interaction, competences or skills.

Mobile learning takes place in a wider social context and evaluation must also recognise this. The wider social and economic benefits of projects may be evaluated through the eyes of learners and other stakeholders, if the pilot or trial has been funded with a social and economic agenda. There is increasing recognition of such benefits but also the difficulty of evaluating them appropriately (Dewson *et al.* 2002).

#### **4.5 Factors influencing the approach**

The evaluation approach followed in the mGBL project has been shaped by some external factors that one needs to take into account when working in such a dynamic and distributed environment. The factors are:

- The characteristics and culture of the corporate environment in different European countries with focus on learning development
- The distributed working environment within the project is taken place
- The state-of-the-art in gaming / learning technology in Europe today
- A specific orientation on evaluation theory and practise

All these factors are needed to take into consideration, and most likely will have a significant influence on the final results of the study.

#### **4.6 Fitting the Ministerial Declaration of Riga of June 11, 2006**

Considering that the Ministerial Conference “ICT for an inclusive society” of the Austrian Presidency of the European Council and the European Commission held in Riga (Latvia) from June 11<sup>th</sup> to June 13<sup>th</sup>, 2006 has update all the criteria an inclusive application of ICT, both for European and non European citizen; considering that the Ministerial Declaration signed in Riga the 11<sup>th</sup> of June 2006 addresses particular rules to enhance eAccessibility and usability, with a series of actions, including change of legislation, fostering the application of common requirements and standards, European or global, for accessible and usable ICT hardware, software and services, to be supported by appropriate user involvement, and means of demonstrating conformance, encouraging interoperability, open architectures of accessible convergent communications, all the above listed criteria of evaluation will account for the Riga Ministerial Declaration and for the European eAccessibility standards and common approaches in public procurement for ICT product and services, to be explored by the Union by 2007, according the Riga Declaration.

## **5 Review of evaluation methods and techniques for mGBL**

Hereunder, we'll take an exploration of the relevant evaluation method used within mGBL project. The very next steps are to apply selections of these methods to the different phases within mGBL, and ultimately to refine and further develop the methods themselves.

According to the goals of WP7, formative evaluation will focus on the development and implementation of project outputs, in particular on the attainment of the aim to provide a platform enabling learning in an effective, efficient and playful way. **The approach to undertaking formative evaluation will be based on:**

- 1) **Collecting and assessing formative user feedback in a structured and systematic way** (questionnaires, structured interviews, focus group ) in order to test games adapted to the game platform and to collect feedback on game-based learning methods and ideas on possible future actions.
- 2) **The involvement of the experts and teaching professionals** (which will be met during the dissemination events) in order to collect feedback and comments on the game platform and perceptions on the usage of game-based learning in education and life-long learning.

The pedagogical aspects in the mGBL trials are thus referring to effective processes along the learning life cycle. The focus is on the learning enabling environment and processes but also on the contents of the games.

Multiple methods of data collection would be used to answer these questions. These methods include:

- **User knowledge pre- and post- tests**
- **User questionnaires**
- **Expert reviews**

## **5.1 User trials**

### *5.1.1 Overview*

In user trials a product is tested by “real users” trying out the product in a relatively controlled or experimental setting, where they are given a standardized set of tasks to perform. The result can be a “problem list” which contains valuable information for designers regarding the potential for improving the usability of a product. Time spent completing a task or the number and types of errors made in use, is information that can be used to compare two different products or two versions of the same user interface. Subjective statements about acceptance are normally part of the results of such trials.

We will describe a simple and robust “qualitative” approach to such trials, requiring that observers have an understanding of the system to be tested so that they can easily deduct from the user’s behaviors that a problem has been encountered. In doing this, knowledge of the user group is of course also very valuable in interpreting the results of such trials. In this situation, the observer must however, be aware that there is always a possibility of “seeing what you want to see”. Using more than one observer will minimize this problem and is to be encouraged as a general procedure to follow.

User trials are normally applied when a prototype product is running, or a when complete product is to be evaluated. Low-tech mock-ups and prototypes may also be used. They are often used before a final product design has been agreed, and are commonly used on pre-production prototypes. They are often used as a simpler way of evaluating products compared to more extensive field trials, which commonly take place when a more completed product is to be evaluated prior to market release.

### *5.1.2 Resources*

Besides the equipment to be tested and a quiet room where testing can take place, such trials require the investment of a reasonable degree of

effort on the part of the investigator. One should be aware that the whole process may stretch out for several weeks, or even months with a number of testing sessions taking place during that period. The costs of using such an approach will vary with the number of subjects to be tested and number of tasks to be prepared and analyzed. The number of tasks in turn will vary with the complexity of the system to be evaluated. It is common in performing such trials to try and identify a representative set of tasks that users would want to perform using the product, and to use this as a basis for planning the evaluation. These can include commonly occurring tasks, but in addition can also include tasks or events which would be difficult to observe in other settings. For example, in a user trial it is possible to simulate dangerous events which might not be observed in any other forms of investigation i.e. field trials, in order to anticipate how a product would perform in actual use.

A simple way of capturing data is to ask participants to “think aloud” or talk through their interactions with the product, and for an observer to make notes about the problems they experienced. This can be supplemented by a more formal set of questions asked after the trials have been completed. It is also recommended that such studies are conducted by two experimenters, the first being responsible for giving the user appropriate tasks and prompting them for comments, whilst the second records the users’ interactions.

## **5.2 Field trials**

### *5.2.1 Overview*

In field trials a product is tested by users in a ‘real life’ setting (as opposed to testing under artificial laboratory conditions). Both the product and the field trial setting are designed to be as close as possible to actual usage. This often involves installing a particular piece of equipment and then monitoring its performance over a period of time. It is common to

allow users to operate equipment as they would in actual usage, and it is usual to monitor that usage using objective and subjective measures. One common method is to conduct regular interviews with users in order to plot their experiences in using a product. In addition the technique can be used in conjunction with other data capture tools e.g. diary keeping methods. Usage and non usage of equipment can also be recorded in such trials, and in some cases the product itself can keep automatic records of its usage i.e. where the product is computerized and has automatic logging facilities. The result of such an investigation can be a “problem list” which contains valuable information for designers regarding the potential for improving the usability of a product. The use of field trials is very common for the testing of new products prior to their commercial launch. Often this is referred to as Beta testing. A Beta testing site is essentially a test site where new products can be tested before being commercially available, and it is common for developers to use existing customers to assist them in this process. The rationale behind Beta testing is that often it is only possible to identify certain problems by testing equipment in realistic settings approximating actual usage. Consequently by allowing ‘friendly’ organizations or individuals to use new products prior to actual release, potentially expensive mistakes can be identified and rectified. In addition, the use of field trials has good face validity as, unlike laboratory-based user trials, field trials are not conducted in artificial settings. Products are used in the context of the whole environment in which they have to work in and many practical difficulties that would not be apparent in controlled settings can be revealed. Products often interact with their environment in ways which cannot be anticipated in advance, and can therefore be missed if only controlled user trials (See User Trials) are performed. It is difficult to make explicit recommendations about the design of field trials. To a large extent this will depend on the complexity of a particular product and the nature of the anticipated user population. Some products need to be used extensively before users become

accomplished users and the full range of usage is achieved, whilst for other products very short trials are all that are needed. Often this can only be determined by trial and error, and by monitoring a particular trial to see whether there are changes still taking place in user performance over a period of time. Once usage appears to have stabilized for a new product, then the trial is complete.

Field trials are normally applied when a final prototype is available, or a complete product is to be evaluated. Because of the relative time and expense of running field trials it is not common to use them in the early stages of product development, but rather to use them for evaluation purposes.

### *5.2.2 Resources*

Compared with other techniques, field trials are resource intensive as the performance of a product needs to be examined over an extended period. Resources to conduct regular interviews with end users are also usually needed to gain maximal benefit, and the approach is very expensive to apply when large numbers of users are involved. However good results can be obtained from field trials with as few as four participants, and even field trials with single users are also of some value. The duration of a field trial is a factor which needs to be given careful consideration. Trials should be conducted until usage settles down into a regular pattern, and the user is confident that they know how to use the product. Some flexibility is needed in planning field trial duration. For example, if during a field trial it quickly becomes obvious that a product has fundamental flaws then product redesign clearly needs to be addressed. It is reasonable in those circumstances to draw the field trial to a close prematurely and so avoid wasting resources on an extensive study. Following redesign, a further field trial could then be conducted.

Field trials are inherently less controlled than laboratory-based trials, but often some degree of control can be attempted to allow a combination of activities to be observed. One common approach is to allow subjects periods of time when they can use the product freely and as they see fit, this is then interspersed with periods when they are required to perform specific tasks or activities. This allows some control of the tasks to be performed which means that information can also be gained on infrequent activities which might otherwise not be observed. The costs of using the field trial approach will vary with the number of subjects to be tested and the length of time given to the investigation. Field trials also need resources for analysis purposes, and again the resources needed for analysis may also vary considerably with the kind of measures used and with the degree of detail recorded. Automatic data logging can produce a great deal of information which is time consuming to summarize, and in addition it can be time consuming to report the results of regular interviews and usage diaries. As with other techniques, it is important to give some attention to the analysis of results when defining a field trial and a good rule of thumb is to ensure that data captured can be easily summarized. Where possible simple multiple choice or yes/no answers to questions should be considered which allow material to be summarized simply and quickly.

### **5.3 Users' Questionnaires**

Questionnaires are classical methods in empirical research. A prerequisite for the design of questionnaire are proper well-defined objects of interest. Questionnaires are in particular suitable for quantitative analysis and are targeted to a larger group of people. They provide material for statistical analysis. However for an explorative and qualitative analysis of a target group they are less suitable and should be combined with qualitative and explorative methods like interviews, for example.

Of particular relevance to mGBL project, **it is necessary to define:**

- who the different users are
- what kinds of questions they are asking

Different users are interested in different things relevant to the kind of decisions they have to make. In some cases, they have different or competing views about what is important, what constitutes success and how success might be measured.

In consideration with the mGBL goals there are three types (roles) of end users involved in user trials of the mGBL prototype game templates:

1. **Students** – focused group (end users),
2. **Teachers** - as developers (using the platform) of the learning games (trainers)
3. **Experts** – in pedagogy and information technology

After the trials, the user might be asked general questions including usability, acceptability, preferences, difficulties, etc.

The relationships among these three users' types and the evaluation questions are described in the evaluation matrix that appears below.

|                 | Reactions to use the platform | Reactions to game templates | Motivation to learn | Fun      | Content accuracy | Applied knowledge usage | Improvements can be made |
|-----------------|-------------------------------|-----------------------------|---------------------|----------|------------------|-------------------------|--------------------------|
| <b>Students</b> |                               | <b>X</b>                    | <b>X</b>            | <b>X</b> | <b>X</b>         | <b>X</b>                | <b>X</b>                 |
| <b>Teachers</b> | <b>X</b>                      | <b>X</b>                    | <b>X</b>            |          | <b>X</b>         | <b>X</b>                | <b>X</b>                 |
| <b>Experts</b>  | <b>X</b>                      | <b>X</b>                    |                     |          |                  | <b>X</b>                | <b>X</b>                 |

There are several types of questions which can be used. Some of the most common types are described in the following.

- *Multiple Choice Items*

This type of questions provides two or more specific responses from which respondents have to choose. Such scales can be easily analysed, but it is important that the full range of significant alternatives are investigated in pilot work. Small changes in wording may also lead to misunderstandings.

- *Rating Scales*

These are scales that can be used to obtain an indication of both the nature and magnitude of the informants' opinions. Normally a rating scale has between 5 and 7 alternatives with the end points of the scale representing the two extremes of opinion possible. For example a respondent may be asked to rate their agreement with a particular statement ranging from strongly agree to strongly disagree. Such scales are intended to have approximately equal intervals between the different points on the scale, which allows statistical analysis to be used. In practice such scales can be difficult to interpret as often large numbers of subjects are needed in order to demonstrate a significant difference between responses of groups of subjects.

- *Paired Comparisons*

This can be regarded as a special type of rating method, where informants have to decide which of a number of specific design alternatives are most appropriate, by series of comparisons of the items as pairs. This technique can provide highly reliable ratings, but may be a considerable effort for informants when a large number of alternatives are to be compared. In addition paired comparisons by themselves do not give a good indication of absolute levels of feeling, but rather just assist in ranking items on some criteria.

- *Ranking*

Ranking requires that informants order items according to some specific criteria e.g. preference. This can be a simple way of identifying which option is preferred from a number of design options, but is somewhat limited as by itself does not give any indication of the absolute level of feeling. If this type of question is used it is not recommended to use more than ten alternatives.

- *Open Ended questions*

This requires the informants to write their own answer or comment on the question. The approach is particularly likely to be used in an exploratory study, and can be used in order to identify the range of responses that should be used in multiple choice questions. The approach requires more effort both from the informants in order to write the answers, and from the analyst to interpret and systematise them. However such open ended questions can be a rich source of information, when the respondents are motivated enough to fill in the questions fully. It is common for open-ended questions to be left blank, and another problem can result in not being able to read the respondent's writing.

In the design of a questionnaire it is important to decide what content areas are central to the study and to decide what questions should be asked and how they will be presented. Where a large number of respondents is anticipated then questions with a limited number of responses are the most appropriate as they are easier to collate and interpret. However, open ended questions can provide a richer source of information, and should at least be used in pilot studies in order to assist in the process of identifying what alternatives should be provided in any fixed response categories. Before a draft questionnaire is tested out it is good advice to write open ended questions, and then convert them to questions with a fixed set of alternative responses after initial pilot tests

have been conducted. Once completed this draft questionnaire should also be piloted again.

In addition to considering in detail the wording of such questionnaires it is also essential to consider in detail how the results of the study will be analysed. It is a very common problem in behavioural science for investigations to be carried out without sufficient thought given to how the material produced will be analysed. This issue also applies to the use of other techniques e.g. user trials.

A central issue in questionnaire design lies in ensuring that all respondents interpret the questions in the way that an investigator intended. Some guidance on the wording of the questions that would apply to any questionnaire is:

- Use familiar words (for the user) in short simple sentences
- Avoid using negatives where possible, and phrase questions in a positive form. e.g. "Are you unhappy?" is better than "Are you not happy?"
- Avoid the use of technical terms and acronyms, unless the investigator can be certain that they are fully understood by all of the respondents
- Cater as far as possible for all possible responses e.g. in questions with a limited number of options it can be useful to have an "other" category.
- Avoid sensitive issues unless absolutely necessary for the study
- Ensure anonymity, and that the informant understand that the information will be treated confidentially
- Avoid asking leading or biased questions which may imply a correct answer
- Ensure that the purpose behind asking the questions is fully understood, particularly if questions are of a sensitive nature.

Even if a lot of effort is put into the construction of the questionnaire, it is possible that there might be some questionnaire items which could give difficulties. Therefore, it is important to test out the draft on a sample of the respondents before the questionnaire is distributed.

#### 5.4 Student Knowledge Tests

End users (students) of mGBL Game Prototypes will be tested on their learning improvements.

Given that the product under test are games that will probably involve some kind of “**scoring system**” (points, level reached, game finished, rank risen ...) these data could be collected to evaluate user argument knowledge. Otherwise it is well known that students may have varying levels of abilities in playing video-games or different levels of preparation in the topics. Data collection in this sense should not allow having an objective comprehension of changes in user knowledge directly attributable to the content of the games.

The **pre- and post- knowledge tests** will consist of multiple choice questions on the argument of the game. The results of the pre-test will be compared with the results of the test performed after the period of trials. That will give an objective measure of the improvements in the game argument.

The limitations of this evaluation include that a relatively small sample of students is involved. Within this small sample, there are students with a variety of educational and experiential backgrounds. Students may have varying levels of interest in the topics depending on their personal interests and educative histories. These students will differ greatly in terms of their pre-trials knowledge related to e-health or e-commerce. As a result of their experiences, the students may also differ in terms of their expectations and interest in the games.

## 5.5 Expert reviews

One final method that mGBL may use to gather information will be **expert reviews**. Ideally, this evaluation requires three different types of experts in content, usability, and instructional design area. Moreover, experts who can cover three different areas at the same time are preferred.

To the casual observer, usability testing and expert review probably look very similar.

- Both identify and prioritize opportunities to improve the user experience
- Both evaluate applications at both the task level and the detailed-design/presentation level
- There is typically overlap in the overlap in the findings
- When done properly, both yield in concrete recommendations for improving the design

In Usability Testing users should do a specific set of critical and frequent tasks that are central to the goals of the game. The focus is narrower. But critically, the findings provide insight into the user’s conceptual use model for the application/game. Expert Review can’t do that.

|                               | Expert Review                                 | Usability Testing                         |
|-------------------------------|---|---|
| <b>Complimentary benefits</b> | Focuses on what the design brings to the user | Focuses on what users bring to the design |

|                 |   |  |
|-----------------|---|--|
| <b>Benefits</b> | <ul style="list-style-type: none"> <li>• Rapid results</li> <li>• Tactical recommendations</li> <li>• Comprehensive evaluation</li> </ul> | <ul style="list-style-type: none"> <li>• Synthesizes recommendations across the task experience</li> <li>• Contextualizes recommendations to the specific objectives of the site and the limitations of the users</li> </ul> |
|-----------------|---|--|

**Expert reviews are usability reality checks** that results in **immediate and concrete fixes** to improve developers' experience. Fresh eyes review a design against industry standards and current research on how humans think and interact with things. An Expert Review is an important technique because it's often hard for designers to recognize usability trouble spots in their own designs. They created the task flow. They *know* how it works. When followed by Usability Testing, an Expert Review ensures that an organization's usability investment is optimized.

## 5.6 Persons involved in user trials

Two **very** important aspects must be remembered:

- **Privacy statements:** in ANY occasion where data are collected (by interviews, questionnaires etc.) statements must be **signed** by the subject to allow the use of the data, according to the country laws regulating privacy. Participation in these surveys or trials is completely **voluntary** and the user therefore has a choice whether to disclose requested contact information (such as name and mailing address) and demographic information (such as zip code or job title).

- If the user involved is **underage** and local laws prescribe such, **parental signed authorization** could be needed before the tester is involved. This is important as this could be the case in mGBL project, where target users are young adult.

### 5.6.1 Users

End users of the product under consideration are the typical participants in trials like this. The testers should be very clear about what kind of previous experience of the same or similar systems the participants should have. Users may be recruited through users' organizations or by contacting schools or institutions in the area.

In mGBL it was decided that users will be the same panel for user interviews (giving useful input for game specifics), prototype testing (user trials) and successive and final testing (field trials).

Yet **we suggest at least a partial "turnover"** so that in field trials at least 30% of the participants did not test the prototype. In this way a part of the testers will have no previous knowledge of user interface, mechanics, etc. This should allow an evaluation of the product also comparing the results of those who already knew the prototype and beta-testers who see for the first time a more mature product.

### 5.6.2 Users helpers

Many products within rehabilitation technology require that other persons than the end users themselves interact with the system. It is therefore also relevant to test the products with helper of the end user, if they will also operate it.

The mGBL products are not meant to specially address the issues of disabled or impaired persons, yet neither are they precluded to be used by

such persons. In case some of the users involved have need of helpers they should be present. See 3.6.4 for more on this.

### *5.6.3 Selection of users*

Although the ideal is to let a “representative sample” of the user population try out the equipment, this is often difficult and too expensive to apply. After defining the user population, (for example by using the results of the user analysis summation tools) the testers should decide whether to approach the extremes of the distribution (best case – worst case) or the mode (the typical case).

*As stated in project “The end user-panel will include younger people who are at a decision stage to decide their further education or students in the field of e-health and e-commerce.”*

Selection of users will also depend upon the specifics that will be chosen for the game templates and game platform after the expert interviews, which will define with more precision the target audience.

Users should of course be chosen between people with interest in use of mobile technologies. For reasons explained later (see 5.2 Logistics) **at least some of the users should be already owners of mobile devices with good technical characteristics.**

### *5.6.4 Special considerations*

When using such techniques with disabled people, it is also important to remember that such users may require longer periods of time to become comfortable using a new product, and in addition some of the problems users are likely to experience with new products will only manifest themselves after extended periods of use. For these reasons it is useful to extend trials for as long a period as is practical, and to review how

confident one is that the major problems with the product have been identified. One should be aware that it is often valuable to have more than one observer present during the testing, and one should take care that the observers themselves don't interfere with the evaluation process e.g. by guiding the user to agree with a particular opinion. Participating in a user trial requires interest and motivation on behalf the user. One should be aware that many disabled people may not be motivated to participate in these kinds of exercises. On the other hand there are numerous examples of disabled persons finding this kind of activity very stimulating and interesting. One should be aware that many user trials will reveal serious problems with a product, to the extent that particular tasks may not be possible to perform. This may lead to a situation where the user is constantly confronted with his/her disability, and as a result may find the experience demoting. This must be avoided and it is important that the user's self esteem is not affected by his or her eventual lack of ability to operate the device. For this reason it is important to emphasize in any instructions to users that it is the product rather than the user which is being evaluated, and to also always include tasks that all participants can solve.

Problems that can arise are:

- General communication problems: people with communication problems should not be invited to participate in "think aloud" sessions. In such cases one should consider whether it is possible to infer user problems only from participants observed behavior. In many cases it may also be sufficient that the user writes down their experiences after the event.
- Hearing impairment: It is usually possible to present the necessary instruction for user trials as text. Doing a "think aloud" session will normally be difficult if the user is also speech impaired. Some

special arrangements will be needed, for example as text on a screen, if the user needs help or instructions during the testing.

- Blind and visually impaired: unfortunately, given the characteristics of mGBL products (games are most often necessarily vision-based; mobile devices have small displays and generally lack the software support for visually impaired persons that can be implemented on personal computers) persons heavily disabled in this sense should not take part in user trials. Although it would be wise to include at least some persons with limited sight problems to test usability of products (e.g. readable game interface) for this persons, representing a good part of the population even between the young adults target users. Partial disabilities (e.g. daltonism) should be taken in account and recorder as characteristics of the user for the analysis of results.
- Mentally impaired: same issues as for visually impaired. The use of high technology small instruments like mobile devices could be problematic for some categories, and these categories are not the direct target of mGBL project. Yet many mentally impaired can participate in user trials if they are properly motivated, and the trials are carefully designed not to create anxiety or be threatening to participants. In order to ensure that the user feels secure and confident, one should consider that a person knowing the user accompanies him/her during the trials. However, measures should then be taken to assure that the person accompanying the user does not interfere during the trial or attempt to act as a spokesperson on the users' behalf.
- Accessibility considerations. When preparing the user trial it is important to take into consideration whether the physical arrangements, desks, tables etc. need special adaptation to the users mobility aids, for example a wheelchair. If the testing involves the user using weak muscles, it is important to ensure that sessions

are short to prevent excessive fatigue. Interviewing the user in advance of the testing should give appropriate information on this.

Field trials do require a considerable degree of effort and involvement on the behalf of users. There are a number of issues relating to the use of field trials which need careful consideration before a trial is organized. One of the most significant of these issues is a moral concern, in that some consideration needs to be made of how long a product's use will be supported by a developer, and what will happen to the equipment after a trial is concluded. For field trials to be successful it is also important that users can obtain adequate support in the event of problems.

## 6 Pedagogical issues

### 6.1 Evaluating the learning processes

mGBL mobile games should give opportunities to improve the quality and the variety of teaching and learning which would not otherwise be achieved through traditional methods. This is a general claim for the project. We will begin our discussion about evaluative considerations from a pedagogical and psychological point of view. The following elements are of importance:

- students and their **emotional learning processes** are to be at the centre of attention;
- the learning scenario should be enhanced by allowing a rich variety in **communication**;
- focus should be put on younger people, with **high interest in mobile technology**;
- the **individuality of learning styles** should be acknowledged.

For example, Honey and Mumford (2000) suggest that there are likely to be several kinds of student learners. "Activist" students learn best when

they have fresh tasks to confront and constantly seek new experiences. In contrast, the “Reflector” needs time to evaluate material. “Theorists” want to connect what they learn to a wider context and a background of theory. In contrast, a “Pragmatist” is keen to put theory into practice. If a classroom has a mix, then it would be appropriate for the teacher to use some traditional and some innovative techniques.

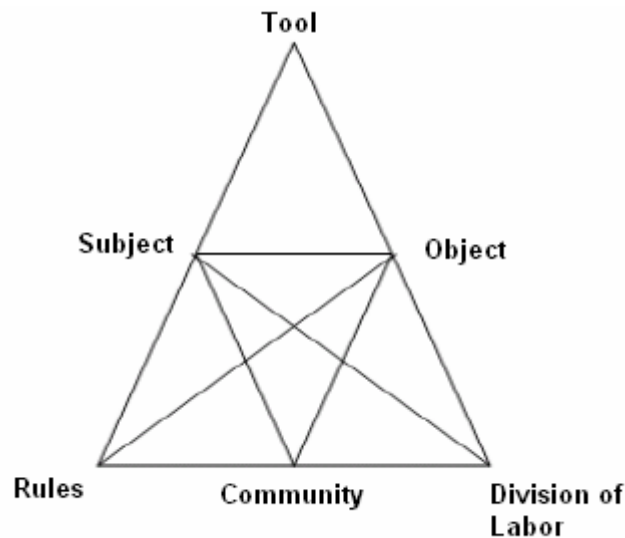
These games offer a chance for different students to bring their strengths to the fore: some will be more attracted to the task involved, whereas others will want to understand the issues in terms of the wider real world. It is visceral, in the sense that students get to have an immediate and sometimes passionate experience of living through problems that mirror the dynamics of everyday life. The situations found in the games should be realistic, even if the setting is contrived, and therefore students can extrapolate from their findings. Depending on the teacher’s objectives in the particular course, there are opportunities for both task completion, observation and reflection on a conceptual level.

## **6.2 Activity Theory as a theoretical frame of educational games**

Kuutti (1995) defined the **activity theory** as ‘philosophical and interdisciplinary framework for studying various practices of human beings as developmental process’ and maintained that human practices are all associated with individual level as well as with social level.

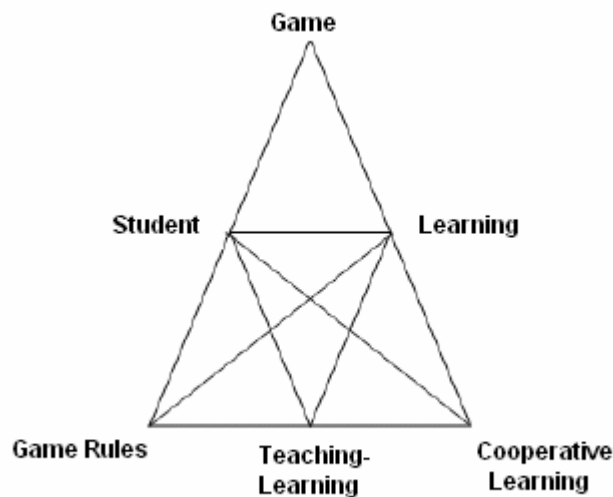
The activity theory is about the relation and meaningful combination between people’s consciousness and their activity. In particular, it emphasizes artefacts, regards computers as important media of human experiences, and believed that all human experiences are visualized by tools and symbolic systems.

Engeström (1987) added communities, as shown in the figure above, in order to solve problems that the relation between individuals and their surroundings are not fully considered.



In an activity, the relation between Subject and Object is mediated by Tool, the relation between Community and Subject by Rules, and the relation between Community and Object by Division of Labor.

The activity theory provides a theoretical frame of **educational games**. The activity theory analyzes not human behaviours but human activities including meaningful contexts. In reality, students do not play a game after learning it, but learn diverse phenomena happening in the game while playing it. In the situation of an educational game, the activity theory provides a theoretical frame in order to develop interface that enables learners to use the educational game intuitively without intentional learning.



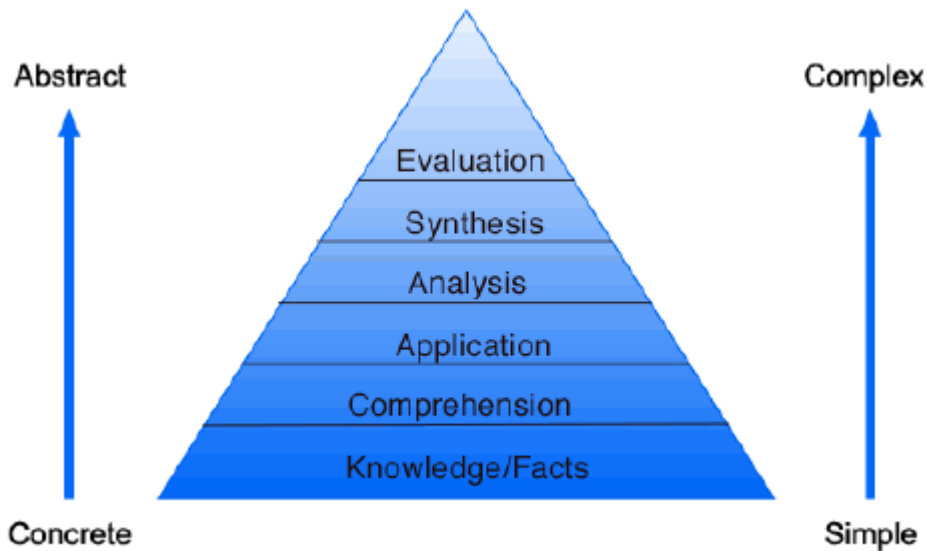
As shown in the figure, we can assume that the relation between Student and Learning is mediated by Game, the relation between Teaching-Learning and Learning by Cooperative Learning and the relation between Student and Teaching-Learning by Game Rules.

### 6.3 Educational process: Bloom's Taxonomy

Following the 1948 Convention of the American Psychological Association, B.S. Bloom took a lead in formulating a classification of "the goals of the educational process". Three "domains" of educational activities were identified. The first of these, named the **Cognitive Domain**, involves knowledge and the development of intellectual attitudes and skills. (The other domains are the Affective Domain and the Psychomotor Domain, and need not concern us here).

Eventually, Bloom and his co-workers established a hierarchy of educational objectives, which is generally referred to as Bloom's Taxonomy, and which attempts to divide cognitive objectives into subdivisions ranging from the simplest behavior to the most complex.

It is important to realize that the divisions outlined above are not absolutes and that other systems or hierarchies have been devised. However, Bloom's taxonomy is easily understood and widely applied.



The classification puts higher priority towards practical aspects of learning where assessment, research and problem solving are put on the top of the pyramid of intellectual complexity. Potential value of cognitive approach in education can be development of intellectuals with practical skills instead of persons possessing mainly factual knowledge. In the age of Internet the factual knowledge is rapidly devaluating. Higher-level knowledge and practical skills are becoming more competitive.

| Competence       | Skills Demonstrated   |
|------------------|---|
| <b>Knowledge</b> | <ul style="list-style-type: none"> <li>• observation and recall of information</li> <li>• knowledge of dates, events, places</li> <li>• knowledge of major ideas</li> <li>• mastery of subject matter</li> <li>• <i>Question Cues:</i></li> </ul> |

|                             |   |
|-----------------------------|---|
|                             | <p>list, define, tell, describe, identify, show, label, collect, examine, tabulate, quote, name, who, when, where, etc.</p>   |
| <p><b>Comprehension</b></p> | <ul style="list-style-type: none"> <li>• understanding information</li> <li>• grasp meaning</li> <li>• translate knowledge into new context</li> <li>• interpret facts, compare, contrast</li> <li>• order, group, infer causes</li> <li>• predict consequences</li> <li>• <i>Question Cues:</i><br/>summarize, describe, interpret, contrast, predict, associate, distinguish, estimate, differentiate, discuss, extend</li> </ul> |
| <p><b>Application</b></p>   | <ul style="list-style-type: none"> <li>• use information</li> <li>• use methods, concepts, theories in new situations</li> <li>• solve problems using required skills or knowledge</li> <li>• <i>Questions Cues:</i><br/>apply, demonstrate, calculate, complete, illustrate, show, solve, examine, modify, relate, change, classify, experiment, discover</li> </ul>   |
| <p><b>Analysis</b></p>      | <ul style="list-style-type: none"> <li>• seeing patterns</li> <li>• organization of parts</li> <li>• recognition of hidden meanings</li> </ul>  |

|                   |   |
|-------------------|---|
|                   | <ul style="list-style-type: none"> <li>• identification of components</li> <li>• <i>Question Cues:</i><br/>analyze, separate, order, explain, connect, classify, arrange, divide, compare, select, explain, infer</li> </ul>  |
| <b>Synthesis</b>  | <ul style="list-style-type: none"> <li>• use old ideas to create new ones</li> <li>• generalize from given facts</li> <li>• relate knowledge from several areas</li> <li>• predict, draw conclusions</li> <li>• <i>Question Cues:</i><br/>combine, integrate, modify, rearrange, substitute, plan, create, design, invent, what if?, compose, formulate, prepare, generalize, rewrite</li> </ul>  |
| <b>Evaluation</b> | <ul style="list-style-type: none"> <li>• compare and discriminate between ideas</li> <li>• assess value of theories, presentations</li> <li>• make choices based on reasoned argument</li> <li>• verify value of evidence</li> <li>• recognize subjectivity</li> <li>• <i>Question Cues</i><br/>assess, decide, rank, grade, test, measure, recommend, convince, select, judge, explain, discriminate, support, conclude, compare, summarize</li> </ul> |

As teachers we tend to ask questions in the "knowledge" category 80% to 90% of the time. These questions are not bad, but using them all the time

is. We must try to utilize higher order level of questions. These questions require much more "brain power" and a more extensive and elaborate answer. Below are the six question categories as defined by Bloom.

- **KNOWLEDGE**

- remembering;
- memorizing;
- recognizing;
- recalling identification and
- recall of information
  - Who, what, when, where, how ...?
  - Describe...

- **COMPREHENSION**

- interpreting;
- translating from one medium to another;
- describing in one's own words;
- organization and selection of facts and ideas
  - Retell...

- **APPLICATION**

- problem solving;
- applying information to produce some result;
- use of facts, rules and principles
  - How is...an example of...?
  - How is...related to...?
  - Why is...significant?

- **ANALYSIS**

- subdividing something to show how it is put together;
- finding the underlying structure of a communication;
- identifying motives;
- separation of a whole into component parts
  - What are the parts or features of...?

- Classify...according to...
- Outline/diagram...
- How does...compare/contrast with...?
- What evidence can you list for...?
- **SYNTHESIS**
  - creating a unique, original product that may be in verbal form or may be a physical object;
  - combination of ideas to form a new whole
    - What would you predict/infer from...?
    - What ideas can you add to...?
    - How would you create/design a new...?
    - What might happen if you combined...?
    - What solutions would you suggest for...?
- **EVALUATION**
  - making value decisions about issues;
  - resolving controversies or differences of opinion;
  - development of opinions, judgments or decisions
    - Do you agree...?
    - What do you think about...?
    - What is the most important...?
    - Place the following in order of priority...
    - How would you decide about...?
    - What criteria would you use to assess...?

## 7 mGBL User Trials: Implementation plan

According to mGBL project plan, some user trials will be carried out as part of the fundamental research into mobile game-based learning design. This chapter offers some recommendations for the evaluations of the mGBL project based on the user trials.

Performing a user trial requires at least five steps.

## 7.1 Step 1: Planning

### 7.1.1 Usability goals

The general usability goals should be determined in advance, for example during a Group discussion and by using the Usability Evaluation Planning tool. The goal of the investigation will normally be to obtain a problem list to provide design feedback, but in addition criteria for acceptable levels of performance and errors might be needed in order to test conformance of the product to such usability goals.

In first mGBL trials the goals should concern **usability, user interface, errors, evaluation of ideas** and other aspects that need to be cleared in prototypal design. These will be “**User trials**” in the sense described in the introduction.

In a later stage, the subsequent tests on fully developed prototypal games with their e-learning contents, aim of tests will be the evaluation of the final goals of mGBL project, that is, as stated before:

- **Enjoyment of use** of mGBL prototype.
- **Changes in attitudes to learning** and/or learning habits, for example enhanced interest in related formal learning opportunities.
- **Perceived improvements in at least 2 specific critical decisions.**

These will be the “**Field trials**” in the sense described in the introduction.

### 7.1.2 Test design

Generally, if the test involves more than one system one must be clear whether the same person should be tested on all systems, or tested on one only. Being tested with the same tasks on two different systems opens the possibility of transfer of experience between the two systems. Generally it is easier to compare two systems tested by different users;

however, if one is interested in subjective preferences for the systems, it is often the best strategy to let the users try both.

This must be taken in consideration in mGBL project, where at least two different kinds of games must be implemented.

**In the first trials, aimed at evaluation of interface, ideas and usability, the typical design of user trials can be used, that is a short session** (about 1 hour or so) for each user.

**This time scale is inadequate for evaluation of aspects like enjoyment, changes in learning attitude, improvement in decisions** that will be evaluated in the final trials. For these the setting of typical field trials should be used, letting the user try the product for an extended period of time.

Generally in field trials decisions need to be made regarding the numbers of users to involve in a trial and the duration of a trial. Often these decisions are based upon the practical constraints of project time scales and resources, but where possible attempts should be made to run trials for a reasonable period. For very simple products field trials of a few days (or even hours) duration can be acceptable, whilst for more complex systems periods of several months might be more appropriate. There are no strict guidelines to follow and judgment is required in order to decide on an appropriate trial period. One approach which can be followed is to arrange trials on a flexible basis, where a provisional trial period is set, but actual experience determines when the trial is drawn to a close.

For the mGBL project products, at first thought, **a time period of one or two weeks should be enough for the user to fully evaluate a mobile e-learning game.**

In general during the planning phase decisions also need to be made regarding **data capture** during the trials and any analysis that may be required. For many field trials it can be sufficient to conduct interviews with the users (and other relevant parties) at regular intervals, noting what users have used the product for and any particular problems they have had. During such interviews it can also be valuable for investigators to ask users to demonstrate their use of the product, and users can be given specific activities to perform, very much in the same way as in user trials. For mGBL we suggest **pre- and post-trial interviews or questionnaires** could be sufficient to our needs, but it could be useful also to collect data regarding use: e.g. if “online” games are tested it could be useful to **collect logs monitoring activity frequency**.

Also, given that the product under test are games that will probably involve some kind of “**scoring system**” (points, level reached, game finished, rank risen ...) these data should also be collected **to evaluate game difficulty** and adjust it consequently. It is well known that a game that is too difficult is often abandoned since it becomes frustrating for the user, while a game too easy becomes quickly boring. Data collection in this sense should also allow to “calibrate” possible adjustable levels of difficulty in the games.

Collecting **log data** monitoring the use of mGBL games will also be very important **to evaluate the enjoyment of use**. In fact in general in field trials decisions also need to be made regarding the extent to which users should be encouraged to use the product in the trial, or whether they should be allowed to decide not to use the product. It is useful to know that end users will not elect to use a product when they have a choice, but it is also important to tease out the reasons for non use, which may have little to do with the actual quality of the product. One solution is to encourage active usage at the beginning of the trial, and to ensure that

users have been adequately trained, but then to allow users to decide for themselves whether or not to continue usage. This requires some sensitivity on the part of the investigator, as it can be easy to give messages to users that they are to blame for not using the product, and in some way “ruining” the trial. To some extent this effect can be minimized by having set periods where users are asked to perform specific activities with the equipment, so that all parties concerned are happy that some empirical data on usage has been obtained. However in a field trial is important to remember that knowing about non use can be as important as usage information, and that emphasis needs to be placed on the fact that it is the product rather than the person that is being evaluated. Under these circumstances non use is a failure of the product rather than of the person.

#### *7.1.3 Testers and Observers’ Roles*

For user trials, normally a usability test will require at least two participants in addition to the user; one to administer the test and one to observe. It is important to stress that the observer should not interfere during the trial. Both tester and observer should know the tasks and their solutions in advance of the testing. If possible, the observer should belong to the design or development team, and the tester should not.

#### *7.1.4 Number of subjects*

Deciding on the number of users may be done in this step; providing that detailed comparisons are not required it should be sufficient to test about 4 - 6 users. However, if the cost of testing every single user is high, one might want to add new users to the group of users that can be called on one by one and to stop testing when a sufficient amount of information is gathered. In mGBL the number of testers anyway has been decided beforehand at the time of project implementation, so this step is unimportant in our case.

### *7.1.5 Procedure description*

The testers guide describes how the test is run from start to end. It prescribes the sequence of tasks to be performed, when and what questions are asked, and how instructions are given. The procedure should also describe how the Observations are made. It is important to include a “help strategy” telling the tester what to do when the subject asks for help or does not know what to do. For example, one approach is to provide the user with the minimum of additional information for the subject to proceed with solving the task, and only to provide assistance if they repeatedly make the same mistakes or are unable to continue with the trial.

### *7.1.6 Pilot trials*

It is a good idea to run pilot trials with easily available subjects; colleagues, students etc. until you are sure that there will be no technical and procedural problems during the testing. It is particularly important to ensure that any prototype equipment will be as reliable as the finished product is anticipated to be, as poor product reliability can seriously influence an evaluation study. In addition to considering in detail the running of such trials it is also essential to consider in detail how the results of the study will be analyzed. It is a very common problem in behavioural science for investigations to be carried out without sufficient thought given to how the material produced will be analyzed.

## **7.2 Step 2: Logistics**

### *7.2.1 Test material*

The result of this step is a set of tasks, scenarios, instructions, interview guides and observation measures.

In our project the needed materials will be at least:

- **High-end mobile devices:** since the games will probably require good characteristics of the device in terms of large display, several digital packed connection types (e.g.: GPRS/EDGE etc.), good user interface etc. **Moreover the final prototypes will be ready at 28 months from here so it can be assumed that the “average” mobile devices at that time will have characteristics equal or superior to current high-end devices** (if we were to design a game for desktop PC, we wouldn't want to create a game apt for a Pentium II CPU and release it when Pentium IV are common, don't we?)
- **Connection services:** will have to be taken in account.
- **Server:** if the game will require the presence of a centralized server, at least a prototypal server must be running and maintained during the field trial period.
- **Paper and pencils.**

In prototype trials one particular important issue regarding mGBL project is that evaluation of the user interface or of possible problems in a game on a mobile device has a very strong bias depending upon confidence of use of the mobile device of the user. A person already accustomed to the commands of his/her cellular phone or PDA will likely have lesser difficulty in using a game than someone accustomed to use of mobile devices but not to that particular model.

It could be useful to have some of the users using their own (high-end) mobile device, if they own one. They should download the game in it and use the mGBL games just like another downloaded game.

Other users should be allowed for the duration of the test to use a mobile device new to them, furnished by the testers. This would simulate a user who buys a new mobile device with mGBL games preloaded in it. Analysis

of results in intuitiveness and efficiency of game interface should keep in account these differences.

### 7.2.2 Tasks

In general for trials it is important that the tasks are representative of the product under consideration. One approach is to **try to select tasks which will be performed frequently** by a typical user, but emphasis should be also placed on those tasks which are also particularly important. Thus any activities which are needed to ensure the safety of users should be looked at closely. In addition user trials can be a good opportunity to investigate critical activities which are likely to occur infrequently, e.g. emergency conditions, which for practical and ethical reasons should not be allowed to occur in a more natural or field setting. The tasks should also have different degrees of difficulty, so that all participants can solve some problems. A test session should normally not exceed one hour and a half, unless the situation demands otherwise. Try to estimate on the basis of pilot trials, the average time it takes to go through the different tasks, and also remember that there is likely to be considerable variability in the ability of different subjects to learn to use the product. For each of the tasks a sequence of "correct" user behaviour must be worked out i.e. those actions which will achieve the task objectives in the most efficient or effective way, and this can then be used a yardstick from which to assess user performance.

It is often a good idea to prepare a **scenario**; that is a description which helps the user to see the tasks to be performed in a context. This can make the testing session more realistic for subjects, and help them imagine the implications of using the technology. "Imagine that you have just got this new device, and then one day..." The scenario should then be integrated in the general instructions given to participants.

### *7.2.3 Instructions and demonstrations*

All instructions to the user should be written in advance even when they are spoken to the user by the experimenter, and instructions should not be improvised during testing. Instructions may be necessary both before and during testing. Remember that lengthy instructions may be difficult for the subject to absorb, so it is important that they are as simple and self-evident as possible. Be prepared to describe the same thing several times in different ways, and avoid using technical terms in the instructions. Written instructions can be used to supplement verbal instructions for some categories of users. Depending on the objectives of the trials, it may also be necessary to demonstrate some aspects of the system in advance of the trial. The demonstration must be seen in connection with the tasks which have been worked out.

### *7.2.4 Interviews and questionnaires*

Before the actual testing the subject will be interviewed for relevant background information, which often includes age, gender, prior experience with similar technology etc. Usually it is necessary to ask for previous experience in the application field and in the usage domain. This is to find out if the subject has used similar systems earlier, if they have experience in using the input device (for instance a mouse, a mobile phone...), etc. It is important to address questions which give information about the users' pre requisites for operating the system, with special emphasis on input/output devices. During testing, the subject might be asked about immediate reactions after having completed each task. After the test, the user will be asked general questions including acceptability, preferences, difficulties, etc. Specific questions about some aspects of the system might also be asked at this stage.

### 7.2.5 Observational measures

To observe and record when the user experiences a problem is likely to be the most useful output from the usability test. It is important that the observer takes notes during testing. Interaction sequences may be so rapid that it can be hard to locate exactly where in an interaction the problem started, and it may therefore be necessary to get users to go through the interaction again, explaining in detail what they were trying to do. This is facilitated by breaking the trials down into a series of smaller tasks limited in duration. Whether or not the user has encountered a problem becomes subject to the observers interpretation. Therefore it is usually very helpful if the user can “think aloud” while operating the system i.e. verbalize what they are thinking and the problems they are experiencing. “Thinking aloud” may however, be difficult for some subjects, and the test should not be dependent on this. The thinking aloud may in itself make the task more difficult to complete. The subject should also be questioned immediately after each task about their experiences of the situation and any problems faced. An ordinary stopwatch is sufficient if one wants to measure the time used to complete each task, as in many cases an accurate estimation of time taken is not needed. Recording the types and numbers of errors may be of particular interest, but counting errors per task requires that the interaction is compared to some model of the correct sequence of user actions. If it is necessary to have detailed recording of particular aspects of the users behaviour one might try to include more than one observer who can concentrate on different user actions, for example visual scanning, mouse usage, keyboard usage, pauses etc. However such approach rapidly becomes resource intensive, and for most purposes simpler method is preferred.

### 7.2.6 Time Line

The **first user trials** will start in month 9<sup>th</sup> of the project, to evaluate the prototype games. **Second user trials** will be conducted from month 15<sup>th</sup>,

to test the prototype platform and game. In month 26<sup>th</sup> will start the **final user trials**, to evaluate the platform and the games, included the contents.

### **7.3 Step 3: The trial**

For user trials, according to the written procedure, subjects are interviewed, given instructions and tasks, they then solve while thinking aloud. One session will normally involve several tasks, and should not take more than one to one and a half hour. It is also important to remember that sufficient time should also be allowed both to introduce the users to the activities required from them, and also to close the testing session, by answering any questions that participants may have. It is also extremely important that the testing and interviewing are performed in a relaxed and friendly atmosphere where the user does not feel threatened or worried about the trials. It is particularly important not to criticize the user in any way for their performance in using the product, and it should be emphasized that it is the product rather than the person that is being evaluated. During the interviews the users' attention should be directed towards the system, and not towards his own shortcomings or failings in operating the system. Observers should record what is happening according to the Observational measures decided on prior to the trials. It is very important that those conducting the trials know the system, the tasks to be performed, and optimal solutions. This is important to support the user in learning to use the product, but is also needed so that any observer can interpret what is taking place during an interaction. Immediately after the trial any notes taken can be transferred to a report form.

### **7.4 Step 4: Data analysis**

Usually a list of problems is the main output from the testing. The problems should be ranked in order of severity. This might be done by

asking users or experts to rate each problem for severity on a 10 point scale. If this is not possible, the analysts should rank the problems according to their own best judgment. If two different systems are compared, the mean time to complete each task is easy to calculate and can be very informative. Frequency of errors per task, if calculated, might be used in the same way. However misleading results can be obtained when trying to compare small samples of users, as often the variability in different users' performance on a given system is high. This means that in order to be reasonably sure that differences in performance are not due to chance, larger numbers of subjects are needed or very large differences in performance need to be observed. One way to reduce this effect is to have the same subjects perform tasks on both systems and then to compare their performances, rather than using different people to perform trials on the two systems to be evaluated. However this is also not without some difficulty, as experience in using one system may have an effect on how the user operates a later one. If detailed comparisons are needed, more rigorous statistical analysis should be applied.

For field trials: they provide a rich source of information about how a product is likely to operate in the real world, but conversely data can be difficult to interpret as a result of this complexity. One way to assist analysis is to use relatively simple reporting forms for problems with usage of equipment, in conjunction with simple interviews and/or questionnaires that can be readily summarized. However it is important not to oversimplify analysis of field trial material, and it can be very useful to periodically conduct extensive interviews with participants in order to summarize the problems experienced in use and any improvements that would be needed. Another technique that can be used in many cases is to bring together participants in a field trial to discuss their experiences collectively. This can be a useful way of summarizing the impressions of a

number of users and can also provide new insights as to possible improvements to the product that may be needed.

## 7.5 Step 5: Implications

In general the recipients of the results from a usability study are system designers/developers, or people who are responsible for making a choice between different solutions. Although the results may be given in a written report, it is very often useful to report the findings during a meeting. In any event, it is important to give a systematic account of the usability problems encountered during testing. If the usability information is to be given to developers or designers who have not been involved in the testing, it is important to focus on the possibilities for improvement, rather than on the problems created by the tested system. The important question is to find the reason for each of the found problems.

## 8 User knowledge test: Implementation Plan

### 8.1 Step 1: Tests design

The pre- and post- knowledge tests will consist of **multiple choice questions** that are designed directly by teachers. The users will answer the questions prior to using the game (the trials) and then again after using the game (the trials).

The purpose of these tests is to provide timely and accurate measures of project success in particular:

- *changes in attitudes to learning and/or learning habits by at least 75% of learners involved in trials, for example enhanced interest in related formal learning opportunities*
- *perceived improvements in at least 2 specific critical decisions, by at least 75% of learners involved in trials*

## **8.2 Step 2: Testing of the draft**

The **draft questionnaires** should be circulated to experts and consultants for comments and suggestions. It is important to allow more than one person to provide feedback on this first draft. Thereafter it is revised and tested again.

## **8.3 Step 3: Logistics**

The knowledge tests will be conducted during the final test bed from month 27 to evaluate the platform, the game and the contents. Prior to the start of the game, all of the students will be given a questions test on e-health/e-commerce problems. After completing the game, a similar test will be administered. Pre- and post-test answers will be compared. A statements survey about the game, based on a Likert scale (1 to 5), will also be completed by the students after playing the game, to assess the students' enjoyment and perception of the educational impact of playing the game.

The survey will be administered to students in paper form, but all survey responses will be anonymous. The students will have a limited time to answer the questions.

## **8.4 Step 4: Data Analysis**

One may conduct appropriate statistical techniques, such as analyses of variance, and examine test score distributions without much concern for the cultural context in which the data were collected, although that may actually be somewhat short-sighted. But the analysis of interview data and the interpretation of descriptions of behaviour related to programs undergoing evaluation cannot be achieved without considerable sensitivity to, and understanding of, the cultural context in which the data are gathered.

Determining an accurate meaning of what has been observed is central in culturally responsive evaluation. Having adequate understanding of cultural context when conducting an evaluation is important, but the involvement of evaluators who share a lived experience may be even more essential. The charge for minority evaluators is to go beyond the obvious.

Knowing the language of a group's culture guides one's attention to the nuances in how language is expressed and the meaning it may hold beyond the mere words. The analyst of data gathered in a culturally diverse context may serve as an interpreter for evaluators who do not share a lived experience with the group being evaluated.

To this end, a good strategy is the creation of review panels principally comprising representatives from stakeholder groups to examine evaluative findings gathered by the principal evaluator and/or an evaluation team. Again, the results of the deliberations of review panels will not lend themselves necessarily to simple, easy answers. Our contention, however, is that they will more accurately reflect the complexity of the cultural context in which the data were gathered.

Disaggregating of collected data is a procedure that warrants increased attention. Disaggregating of data sets is highly recommended because evaluative findings that dwell exclusively on whole -group statistics can blur rather than reveal important information. Worst still, they may even be misleading. For example, studies that examine the correlates of *successful* minority students rather than focusing exclusively on the correlates of those who fail are important. It can be enlightening to scrutinize the context in which data that are regarded as "outliers" occur. The examination of a few successful students, in a setting that commonly produces failure, can be as instructive for program improvement as an examination of the correlates of failure for the majority.

In sum, the data rarely speak for themselves, but rather are given voice by those who interpret them. The voices that are heard are not only those who are participating in the project, but also those of the analysts who are interpreting and presenting the data. Deriving meaning from data in program evaluations that are culturally responsive requires people who understand the context in which the data were gathered.

## **8.5 Step 5: Implications**

Information obtained from the evaluation process will be used as the basis for further discussion within the project. The reporting should follow a common format based on the overall assessment criteria, and be a composite of both expert and user opinion and results of knowledge tests. An additional discussion document will also be generated for use within the consortium, highlighting any mismatches between user requirements together with agreed functional specification and the emerging software.

Distribution and utilization of evaluation outcomes are certainly important components in the overall evaluation process. Moreover, a critical key is to conduct an evaluation in a manner that increases the likelihood that the results will be perceived as useful and, indeed, used. Culturally responsive evaluations can increase that likelihood. Hence, evaluation results should be viewed by audiences as not only useful, but truthful as well (Worthen, Sanders, and Fitzpatrick, 1997).

Information from good and useful evaluations should be widely disseminated. Further, communications pertaining to the evaluation process and results should be presented clearly so that they can be understood by all of the intended audiences. Michael Q. Patton (1991) pointed out that evaluation should strive for accuracy, validity, and believability. Patton (1997) further stated that evaluation should assure that the information from it is received by the “right people.” Building on

his cogent observation we would add that the “right people” are not restricted to the funding agency and project or program administration and staff, but should include a wide range of individuals who have an interest or stake in the program or project.

The distribution and use of evaluation outcomes should be thought through early when preparing an evaluation, that is, during the evaluation-planning phase. Moreover, the use of the evaluation should be firmly consistent with the actual purposes of the evaluation. Further, the purpose of the evaluation should be well defined and clear to those involved in the project itself.

As we talk about diffusion, our discussion comes full circle, and we return to the earliest steps in evaluation design, the evaluation questions. These questions themselves are always keys to a good evaluation—those that would provide information that stakeholders care about and on which sound decisions can be based must always guide the work. The right questions, combined with the right data collection techniques, can make the difference between an evaluation that is only designed to meet limited goals of compliance and one that meets the needs of the project and those who are stakeholders in it. Applying the principles of culturally responsive evaluation can enhance the likelihood that these ends will be met, and that the real benefits of the intervention can be documented.

## **9 Developing Evaluation Reports**

mGBL Evaluation Reports should have the following components:

- An executive summary that provides highlights of the project and evaluation findings.
- The body of the report in which the evaluation and its findings are described in detail.

- Appendices containing additional tables, copies of instruments, or other documentation.

Reports on evaluation findings are most useful for the whole project aims if they contain the following information:

|          |  |
|----------|--|
| <b>1</b> | <b>What the project is about</b>   |
| <b>2</b> | <b>What questions were addressed in the evaluation</b>                                 |
| <b>3</b> | <b>What outcomes were studied and why they were selected</b>                           |
| <b>4</b> | <b>How those outcomes were measured and the quality of the measurement instruments</b> |
| <b>5</b> | <b>Who was included in the study sample</b>  |
| <b>6</b> | <b>How the data were analyzed</b>  |
| <b>7</b> | <b>What it was found</b>   |
| <b>8</b> | <b>What can be concluded</b>   |

**What the project is about:** This section provides a brief description of what was done and why within mGBL project in the period under examination. To the extent possible, the theoretical basis for the activity is described, as is the conceptual model underlying the work.

**What questions were addressed in the evaluation:** This section clearly lays out the central questions that the evaluation addressed. The section indicates why these questions were selected and, if relevant, the project participating organizations that consider them important. If evaluation questions that some might expect to see were eliminated, the reasons for eliminating them should be provided.

**What outcomes were studied and why they were selected:** This section discusses the specific outcomes that were examined and why these outcomes were selected. The relationship between the selected outcome, the evaluation question, and the conceptual model should be

addressed to the extent possible. If several outcomes were considered but rejected, it may be useful to explain the rationale behind their elimination. Factors to consider include relevance to the question being addressed, feasibility of assessing them given time and resources, and local conditions or sensitivities, as relevant.

**How the outcomes were measured and the quality of the measurement instruments:** This section provides a description of how each of the outcomes was examined and the measurement instruments used to gather the data. Where multiple measures of an outcome are used, the section should describe the relationship among them. For example, the measures may look at different aspects of an outcome to provide a fuller picture of that outcome or may look at essentially the same thing and be used for cross-checking and verification. Properties of the instruments should also be provided. That is, it is important to discuss the reliability and validity of the instruments for the purpose for which they are being used.

**Who was included in the study sample:** It is not possible or necessary, to include all mGBL target groups in an evaluation. Some sample will be drawn and data from that sample are used to generalize to the larger population. This section of the report describes the sample and how it was selected and compares the sample to the larger study population. Any limitations in the sample should be explicitly addressed. Limitations could include some deviation in characteristics from the population, such as a preponderance of individuals with more or less experience in a field, or possible sample biases due to lack of cooperation from selected respondents. Information should be provided in enough details so that the reader can judge representativeness.

**How the data were analyzed:** This section describes the data analysis procedures used for each of the data sets. Enough information should be

provided for the reader to judge the appropriateness of the technique for the question and type of data collected.

**What it was found:** This section presents the findings of the evaluation. Most readers find it useful to organize this section by evaluation question. If events transpiring during the evaluation have caused questions to be changed, this approach may not, however, be effective. In such cases, it is important to clarify why questions were dropped and what the implications of these alterations may be.

This section should also explicitly address any issues that were encountered that might pose threats to the validity of the conclusions reached. Possible threats include problems with unmeasured variables, measurement problems, sampling problems, limits in the design, and unanticipated changes in the project's context.

**What can be concluded:** This section presents conclusions with regard to the overall purpose of the project and the questions it was designed to address. To the extent possible, the data from the evaluation are considered in light of the initial theoretical basis for the study and the conceptual model underlying it. Issues that need to be further addressed are identified, and strategies for doing so are suggested, to the extent possible.

## **10 Conclusion: mGBL evaluation guidelines**

In summary, in this report we have seen a wide variety of methods may be applicable to the evaluation of the mGBL project's outputs and how the outcomes of this evaluation should be summarized and reported. In this last chapter we offer some common recommendations to perform the mGBL evaluation based on described techniques.

**Recommendation 1:** For making the results understandable background studies concerning the issues that affect the player experience and form a context for interpretations are needed, e.g.,: the lifestyle, the mobile phone experience, relationship to other mobile games, the places and spaces of gaming. The criteria and methods to collect this background information of the users involved in the user trials should be described in D6.1.

**Recommendation 2:** When choosing the people for doing the evaluations, the advantages and disadvantages of designers themselves being those people should be assessed and other possibilities should be considered. This is simply due to the fact that players are prone to please the designers as common courtesy and due to lack of objectivity of evaluation resulting often problems of validity and reliability. The choice of the evaluators should be documented in the evaluation report (deliverable D6.2, D6.3 and D6.4).

**Recommendation 3:** Evaluations, results, also the negative results or the lack of verification, and their validity and reliability must be documented: the results can be understood and interpreted only if documented in a way or another. For instance simple “todolists” with reasoning and couple of lines about by whom and how the evaluation was carried out, can be very enlightening. The results can be understood and interpreted only if documented.

**Recommendation 4:** The methods should be described and the methodological choices should to be justified or motivated at least. Also criterions of success in evaluations need more definition and clarification. For instance, if an essential evaluation criterion is the degree of players’ engagement in the game, engagement needs to be defined in terms of the game in question. The definition of suitable criteria for usability and

performance testing should be done in strictly cooperation with the technical WP leaders and involving WP3 leader.

**Recommendation 5:** Using evaluation data in redesign cannot be done self-evidently, but it requires careful analysis and making conclusions based on the data, previous experience and knowledge about the contexts of gaming, other similar games and the implications attached to them. Conclusions made – also the negative results or the lack of verification – and their validity and reliability must be documented. Making conclusions may also require breaks between project development iterations that open new views to the game in question. Interpretations require also at least some background information (e.g., about the participants’ demographics gaming habits, previous experiences with mobile phones and mobile games).

In short, methodical considerations must be done in the following areas:

- **Collecting background information** to backup the interpretations and conclusions;
- Considering the **choices of the evaluators**;
- **Making systematic observations and records** in evaluation settings;
- **Defining success criteria** for evaluations;
- **Documenting of evaluations**, describing the methods and justifying them;
- **Interpreting of evaluation results**;
- **Justifying conclusions** properly;
- Compiling **descriptive materials the game and game-play** to players.

## References

R. Cappuccio, F. Di Bono, A. Sillitti, G. Succi, (2004). *Improvement of an e-learning platform through the analysis of usage patterns*. Center for Applied Software Engineering, Free University of Bolzano-Bozen, Italy

Kathleen Straub, Ph.D. (2004). *Cleaning up for the housekeeper or why it makes sense to do both Expert Review and Usability Testing*. Human Factors International

F. Colace, M. De Santo, M. Vento, (2005) *E-Learning Platform: Developing an Evaluation Strategy in a Real Case*. 35th ASEE/IEEE Frontiers in Education Conference.

Joy Frechtling, (2002). *The 2002 User Friendly Handbook for Project Evaluation*, The National Science Foundation.

Worthen, B.R., Sanders, J.R., and Fitzpatrick. (1997). *Educational Evaluation*, Second Ed. White Plains, NY: Longman, Inc.

Patton, M.Q. (1991). Toward Utility in Reviews of Multivocal Literatures. *Review of Educational Research*, 61(3): 287-292.

Patton, M.Q. (1997). *Utilization-Focused Evaluation: The New Century Text*. Thousands Oaks, CA: Sage Publication, Inc.

Bloom B.S.; (1956). *Taxonomy of educational objectives: The cognitive domain*. New York: Longman.

Rosas, Nussbaum, Cumsille, Marianov, Correa, Flores, Grau, Lagos, Lopez, Lopez, Rodriguez, Salinas. *Beyond Nintendo: design and assessment of educational video games for first and second grade students*. Computers & Education 40 (2003) 71–94.

Honey, P. and A. Mumford (2000), *The Learning Styles Helper's Guide* (Peter Honey Publications, Maidenhead).

Kevin Gibson, *Games Students Play: Incorporating the Prisoner's Dilemma in Teaching Business Ethics*. *Journal of Business Ethics* 48: 53–64, 2003.

Kristian Kiili, *Digital game-based learning: Towards an experiential gaming model*. *Internet and Higher Education* 8 (2005) 13–24.

Barry D. Mann, M.D., Benjamin M. Eidelson, Steven G. Fukuchi, M.D., Steven A. Nissman, B.A., Scott Robertson, Ph.D., Lori Jardines, M.D. *The development of an interactive game-based tool for learning surgical management algorithms via computer*. *The American Journal of Surgery* 183 (2002) 305–308.

Ministerial Declaration approved unanimously on 11 June 2006, Riga, Latvia, EU-AT, 1-7.

Syvänen & Nokelainen. (2003) *Criteria For Evaluating Mobile Usability Of Digital Learning Material*.

John Traxler, Agnes Kukulska-Hulme. *Evaluating Mobile Learning: Reflections on Current Practice*. IADIS International Conference: Mobile Learning 2005



## Appendix A: Useful tips for Surveys and Interviews' design

The use of **interviews** as a data collection method begins with the assumption that the participants' perspectives are meaningful, knowable, and can be made explicit, and that their perspectives affect the success of the project. An in-person or telephone interview, rather than a paper-and-pencil survey, is selected when interpersonal contact is important and when opportunities for follow-up of interesting comments are desired.

Two types of interviews are used in evaluation research: **structured interviews**, in which a carefully worded questionnaire is administered, and **in-depth interviews**, in which the interviewer does not follow a rigid form. In the former, the emphasis is on obtaining answers to carefully phrased questions. Interviewers are trained to deviate only minimally from the question wording to ensure uniformity of interview administration. In the latter, however, the interviewers seek to encourage free and open responses, and there may be a tradeoff between comprehensive coverage of topics and in-depth exploration of a more limited set of questions. In-depth interviews also encourage capturing respondents' perceptions in their own words, a very desirable strategy in qualitative data collection. This allows the evaluator to present the meaningfulness of the experience from the respondent's perspective. In-depth interviews are conducted with individuals or a small group of individuals.

**Surveys** are a very popular form of data collection, especially when gathering information from large groups, where standardization is important. Surveys can be constructed in many ways, but they always consist of two components: questions and responses. While sometimes evaluators choose to keep responses "open ended," i.e., allow respondents to answer in a free flowing narrative form, most often the

“close-ended” approach in which respondents are asked to select from a range of predetermined answers is adopted. **Open-ended** responses may be difficult to code and require more time and resources to handle than **close-ended** choices. Responses may take the form of a rating on some scale, may give categories from which to choose, or may require estimates of numbers or percentages of time in which participants might engage in an activity.

Although surveys are popularly referred to as paper-and-pencil instruments, this too is changing. Evaluators are increasingly exploring the utility of survey methods that take advantage of the emerging technologies. Thus, surveys may be administered via computer-assisted calling, as e-mail attachments, and as web-based online data collection systems. Selecting the best method for collecting surveys requires weighing a number of factors. These included the complexity of questions, resources available, the project schedule, etc.

In structured questionnaire, close-ended questions (attitude scales) are used to standardize attitude measurement of single subjects following direct observation. Here are some examples of different types of rating scale.

### Simple checklist

*Can you use the following edit commands?*

|                  | <i>yes</i>               | <i>no</i>                | <i>don't know</i>        |
|------------------|--------------------------|--------------------------|--------------------------|
| <i>duplicate</i> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> |
| <i>paste</i>     | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> |

### Multipoint rating scale

Rate the usefulness of the duplicate command on the following scale?

*Very useful* *Of no use*  

|  |  |  |  |  |
|--|--|--|--|--|
|  |  |  |  |  |
|--|--|--|--|--|

### Likert Scale

Statement of opinion to which the subject expresses their level of agreement.

*"Computers can simplify complex problems."*

|                        |              |                       |                |                          |                 |                          |
|------------------------|--------------|-----------------------|----------------|--------------------------|-----------------|--------------------------|
|                        |              |                       |                |                          |                 |                          |
| <i>Very much agree</i> | <i>Agree</i> | <i>Slightly agree</i> | <i>Neutral</i> | <i>Slightly disagree</i> | <i>Disagree</i> | <i>Strongly disagree</i> |

### Semantic differential Scale

Uses a series of bi-polar adjectives and obtains ratings which respect to each.

Rate the *Beauxarts* drawing package on the following dimensions.

|              |                  |              |                 |                |                 |              |                  |                  |
|--------------|------------------|--------------|-----------------|----------------|-----------------|--------------|------------------|------------------|
|              | <i>Extremely</i> | <i>Quite</i> | <i>Slightly</i> | <i>Neutral</i> | <i>Slightly</i> | <i>Quite</i> | <i>Extremely</i> |                  |
| <i>Easy</i>  |                  |              |                 |                |                 |              |                  | <i>Difficult</i> |
| <i>Clear</i> |                  |              |                 |                |                 |              |                  | <i>Confusing</i> |
| <i>Fun</i>   |                  |              |                 |                |                 |              |                  | <i>Dreary</i>    |

## Rank Order

*Place the following commands in order of usefulness (rank the most useful as 1, the least useful as 4):*

*paste*

*duplicate*

*group*

*clear*

When designing a questionnaire or an interview, considerable attention must be given to some aspects:

- do be clear about the information you want to obtain and ensure there are questions that directly address these issues
- don't risk subjects becoming demotivated, because not interested in the questionnaire or the questionnaire is too long
- don't be lazy, make sure that all questions apply to the product being evaluated
- do provide specific task reference for questions. If questions ask for opinions about particular details of the use of the product, ensure that the task context is clear.
- don't assume that responses will be positive. Although you may think that the product is very good, the questionnaire has to be objective and allow for as many negative comments as positive. Ensure that there is sufficient opportunity for users to justify negative attitudes as positive ones.
- do pilot the questionnaire first

## **Appendix B: Procedures for the selection of the user groups**

The selection of participants whose background and abilities that are representative of the products intended end user is a crucial element of a successful usability evaluation. The evaluation will be valid only if the people evaluated are typical end users of the product, or as close to a selected set of characteristics as possible.

In mGBL it was decided that users will be the same panel for user interviews (giving useful input for game specifics), prototype testing (user trials) and successive and final testing (field trials).

Yet **we suggest at least a partial “turnover”** so that in field trials at least 30% of the participants did not test the prototype. In this way a part of the testers will have no previous knowledge of user interface, mechanics, etc. This should allow an evaluation of the product also comparing the results of those who already knew the prototype and beta-testers who see for the first time a more mature product.

Although the ideal is to let a “representative sample” of the user population try out the equipment, this is often difficult and too expensive to apply. After defining the user population, (for example by using the results of the user analysis summation tools) the testers should decide whether to approach the extremes of the distribution (best case — worst case) or the mode (the typical case).

As stated in project *“The end user-panel will include younger people who are at a decision stage to decide their further education or students in the field of e-health and e-commerce.”*

Selection of users will also depend upon the specifics that will be chosen for the game templates and game platform after the expert interviews, which will define with more precision the target audience.

Users should of course be chosen between people with interest in use of mobile technologies. For reasons explained prior (see 4.2 Logistics) **at least some of the users should be already owners of mobile devices with good technical characteristics.**

### **End user-panel**

We plan to set up one-to-one in-depth interviews with minimum  $n = 30$  young adults in minimum 3 partner countries (minimum total sample:  $n = 90$ ). The sample will be drawn from mGBL target audiences.

#### *E-health game*

The participants will be students of Faculty of Medicine and Surgery, who had enrolled in a degree course in medicine and surgery. Students should be selected from courses of the first 3 years. (In this phase of project the contents are very poor/simple)

#### *E-commerce game*

The participants will be students of Faculty of Economics, who had enrolled in a 3 years degree. Students should be selected from courses of the first 2 years. (In this phase of project the contents are very poor/simple).

## Appendix C: Evaluation Scheduling

| EVALUATION PHASES   | YEAR 1 |     |     |     |     |
|---|--------|-----|-----|-----|-----|
|   | M08    | M09 | M10 | M11 | M12 |
| <i>Evaluation of pedagogical concepts, methodology and instructional design.</i>                        |        |     |     |     |     |
| organization of the evaluation and validation work - e-health   |        |     |     |     |     |
| organization of the evaluation and validation work - e-commerce   |        |     |     |     |     |
| organization of the evaluation and validation work - e-career guidance                                  |        |     |     |     |     |
| experts' interviews on pedagogical concepts, methodology and instructional design for e-health          |        |     |     |     |     |
| experts' interviews on pedagogical concepts, methodology and instructional design for e-commerce        |        |     |     |     |     |
| experts' interviews on pedagogical concepts, methodology and instructional design for e-career guidance |        |     |     |     |     |
| evaluation of generability to other fields  |        |     |     |     |     |
| <b>Report preparation (D 7.2)</b>   |        |     |     |     |     |

| EVALUATION PHASES  | YEAR 1 |     |     | YEAR2 |     |
|--|--------|-----|-----|-------|-----|
|  | M10    | M11 | M12 | M13   | M14 |
| <i>Evaluation of 1st user trials (elaborated games)</i>                |        |     |     |       |     |
| organization of the evaluation and validation work - e-health          |        |     |     |       |     |
| organization of the evaluation and validation work - e-commerce        |        |     |     |       |     |
| organization of the evaluation and validation work - e-career guidance |        |     |     |       |     |
| users' questionnaires from first user trial for e-health               |        |     |     |       |     |
| users' questionnaires from first user trial for e-commerce             |        |     |     |       |     |
| users' questionnaires from first user trial for e-career guidance      |        |     |     |       |     |
| data analysis from first user trial for e-health                       |        |     |     |       |     |
| data analysis from first user trial for e-commerce                     |        |     |     |       |     |
| data analysis from first user trial for e-career guidance              |        |     |     |       |     |
| suggestion for improvement of the games                                |        |     |     |       |     |
| <b>Report preparation (D 7.3)</b>                                      |        |     |     |       |     |

| EVALUATION PHASES  | YEAR2 |     |     |     |     |     |     |     |     |     | YEAR 3 |     |     |     |
|--|-------|-----|-----|-----|-----|-----|-----|-----|-----|-----|--------|-----|-----|-----|
|  | M15   | M16 | M17 | M18 | M19 | M20 | M21 | M22 | M23 | M24 | M25    | M26 | M27 | M28 |
| <i>Evaluation of 2 nd user trials (platform)</i>                           |       |     |     |     |     |     |     |     |     |     |        |     |     |     |
| users' questionnaires from first platform user trial for e-health          |       |     |     |     |     |     |     |     |     |     |        |     |     |     |
| users' questionnaires from first platform user trial for e-commerce        |       |     |     |     |     |     |     |     |     |     |        |     |     |     |
| users' questionnaires from first platform user trial for e-career guidance |       |     |     |     |     |     |     |     |     |     |        |     |     |     |
| data analysis from first platform user trial for e-health                  |       |     |     |     |     |     |     |     |     |     |        |     |     |     |
| data analysis from first platform user trial for e-commerce                |       |     |     |     |     |     |     |     |     |     |        |     |     |     |
| data analysis from first platform user trial for e-career guidance         |       |     |     |     |     |     |     |     |     |     |        |     |     |     |
| suggestion for improvement of the games and of the platform                |       |     |     |     |     |     |     |     |     |     |        |     |     |     |
| <b>Report preparation (D 7.4)</b>  |       |     |     |     |     |     |     |     |     |     |        |     |     |     |

| EVALUATION PHASES   | YEAR 3 |     |     |     |     |     |     |     |     |     |     |     |
|---|--------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
|   | M25    | M26 | M27 | M28 | M29 | M30 | M31 | M32 | M33 | M34 | M35 | M36 |
| <i>Evaluation report on the final test bed incl. content</i>                          |        |     |     |     |     |     |     |     |     |     |     |     |
| users' questionnaires from platform, games & content user trial for e-health          |        |     |     |     |     |     |     |     |     |     |     |     |
| users' questionnaires from platform, games & content user trial for e-commerce        |        |     |     |     |     |     |     |     |     |     |     |     |
| users' questionnaires from platform, games & content user trial for e-career guidance |        |     |     |     |     |     |     |     |     |     |     |     |
| data analysis from platform, games & content user trial for e-health                  |        |     |     |     |     |     |     |     |     |     |     |     |
| data analysis from platform, games & content user trial for e-commerce                |        |     |     |     |     |     |     |     |     |     |     |     |
| data analysis from platform, games & content user trial for e-career guidance         |        |     |     |     |     |     |     |     |     |     |     |     |
| Global validation of the platform   |        |     |     |     |     |     |     |     |     |     |     |     |
| Global validation of the games for e-health   |        |     |     |     |     |     |     |     |     |     |     |     |
| Global validation of the games for e-commerce   |        |     |     |     |     |     |     |     |     |     |     |     |
| Global validation of the games for e-career guidance                                  |        |     |     |     |     |     |     |     |     |     |     |     |
| Cross global validation of m-GBL and its generalization to other fields               |        |     |     |     |     |     |     |     |     |     |     |     |
| <b>Report preparation (D 7.5)</b>   |        |     |     |     |     |     |     |     |     |     |     |     |