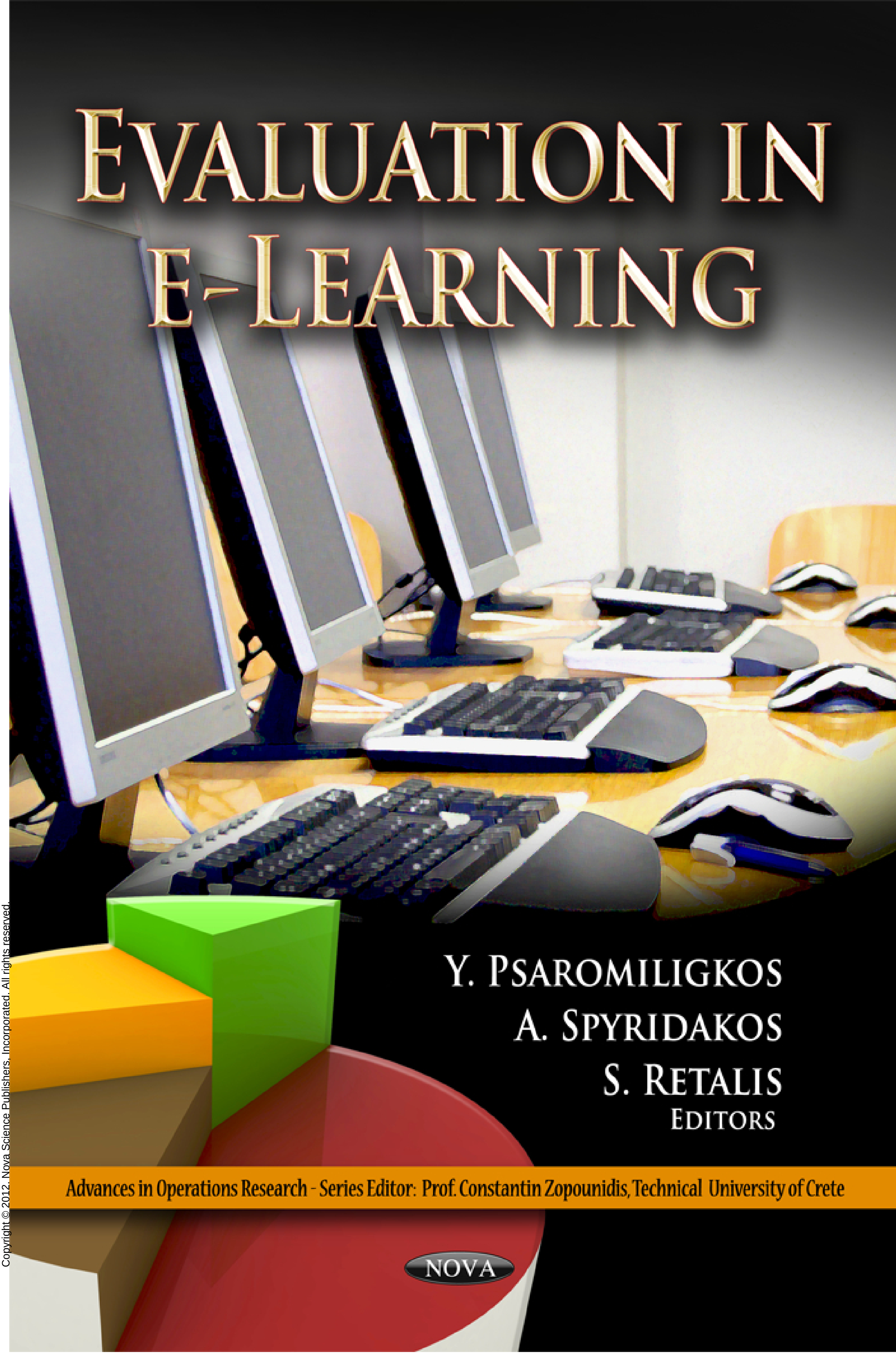

EVALUATION IN
E-LEARNING
Y. PSAROMILIGKOS
A. SPYRIDAKOS
S. RETALIS
EDITORS
Advances in Operations Research - Series Editor: Prof. Constantin Zopounidis, Technical University of Crete
NOVA

ADVANCES IN OPERATIONS RESEARCH

EVALUATION IN E-LEARNING

ADVANCES IN OPERATIONS RESEARCH

***CONSTANTIN ZOPOUNIDIS* – *SERIES EDITOR* –**

TECHNICAL UNIVERSITY OF CRETE

Simulation and Optimization Methods in Risk and Reliability Theory
Pavel S. Knopov and Panos M. Pardalos (Editors)
2009. ISBN: 978-1-60456-658-1

Multiple Criteria Decision Aiding
Constantin Zopounidis, Michael Doumpos, Nikolaos F. Matsatsinis and Evangelos Grigoroudis (Editors)
2010. ISBN: 978-1-61668-231-6

Evaluation in E-learning
Y. Psaromiligkos, A. Spyridakos and S. Retalis (Editors)
2012. ISBN: 978-1-61942-942-0

ADVANCES IN OPERATIONS RESEARCH

EVALUATION IN E-LEARNING

Y. PSAROMILIGKOS
A. SPYRIDAKOS
AND
S. RETALIS
EDITORS

Nova Science Publishers, Inc.
New York

For permission to use material from this book please contact us:
Telephone 631-231-7269; Fax 631-231-8175
Web Site: http://www.novapublishers.com

Library of Congress Cataloging-in-Publication Data

Evaluation in e-learning / editors, Y. Psaromiligkos, A. Spyridakos, S. Retalis.
p. cm.
Includes bibliographical references and index.
ISBN: 978-1-62081-106-1 (eBook)
1. Computer-assisted instruction--Evaluation. 2. Web-based instruction--Evaluation. 3. Distance education--Evaluation. 4. Internet in education--Evaluation. I. Psaromiligkos, Y. II. Spyridakos, A. III. Retalis, Symeon.
LB1028.43.E9426 2011
371.33'44678--dc23
2012004470

Published by Nova Science Publishers, Inc. † New York

CONTENTS

PREFACE

One approach which is becoming very popular nowadays is to measure the effectiveness of e-learning solutions via analysis of data gathered from authentic educational environments. This constitutes the ultimate goal of the underlying book. More specifically, this book discusses and presents several evaluation approaches as they have been applied in real practice. Criticism, best practices, and lessons learned from the application of these approaches are discussed. The book includes innovative contributions and real world evaluation studies in authentic learning environments by experienced researchers in e-learning.

Chapter 1- The purpose of this chapter is to discuss the methodological and theoretical assumptions of qualitative evaluation of e-learning. The authors will begin by discussing the need for evaluation of e-learning and the complexities behind the process. They will then discuss the theoretical foundations of objectivist-quantitative and constructivist-qualitative paradigms. The authors will examine the kinds of situations in which qualitative approaches to evaluation are more appropriate. The key message is that both quantitative and qualitative models have substantial contributions to make to e-learning evaluation. Issues such as types of research questions that are appropriate for qualitative evaluation will be discussed in detail and concrete examples will be presented. The discussion will address the interplay among the evaluator role, data collection procedures, data analysis, and standards of rigor in e-learning evaluation.

Chapter 2- There are various definitions of quality in e-learning. A contextualized definition of quality requires taking into consideration the relevant stakeholders who have different expectations for e-learning. The chapter refers to the factors which influence the quality of training and education on project management online programmes, from two different perspectives: trainers/ teachers and trainees/ students. Two questionnaire-based surveys were conducted and the main results are presented. The first survey aimed at finding out what influences quality of project management education, according to the trainers', professors' and training providers' perspective. 81 % of the responses came from China. The rest were professionals of different EU nationalities: project managers, software developers, financial managers and professors, who are involved in both training on project management, but also as team members or team managers in projects, thus ensuring a balanced overview of both theory and practical issues. The second survey aimed at finding out what quality means for trainees and students. The questionnaire for students was spread to 300 individuals. The response rate was high: The wuthors received 253 responses. The respondents were managers and project leaders from private companies, project managers from the Academy of Economic Studies and National Institute of R&D in Informatics, from Romania and also master students. Although there were small differences of perspective, the authors concluded that both trainers and trainees have the same approach towards a qualitative project management e-education. The authors propose a quality framework based on four pillars, starting from the authors' findings. Furthermore, expectation profiles for both students and teachers are provided, as a method of obtaining maximum satisfaction of all individuals involved in project management e-learning activities, thus improving the education quality.

Chapter 3- Evaluation is a critical and a "must-do" step in every e-learning course since it provides valuable feedback in order to tailor, redesign, improve or accept the course according to the stakeholders' needs. Evaluation by its nature is a complex process and constitutes an ill-structured problem. A series of critical factors are involved such as: a considerable number of different stakeholders (students, tutors, e-learning designers, developers, and so on), the functionality and usability of the available learning technology, the pedagogical approach (learning theories, instructional design models, learning strategies), the delivery mode, constraints (time, budget, people). Moreover, the outcomes of e-learning are difficult to be measured due to time and quality borders. A major goal of an evaluation process is to measure the underlying set of quality factors as perceived by the target stakeholders (e.g. students). The analysis of stakeholders' preferences is an essential factor to the success of the e-learning course because it gives valuable feedback to the decision makers in order to properly redesign/improve the course.

In this chapter, the authors describe in detail the results of a new methodological approach for the analysis of students' preferences in e-learning courses based on the principles of theory of committees and elections and multicriteria disaggregation-aggregation decision aid approaches. The underlying real case study concerns the analysis of the students' preferences of an undergraduate course at TEI Piraeus during the academic year 2009-2010. Multicriteria disaggregation – aggregation approach for discrete alternative actions aims to the assessment of an additive preference value model triggered by students' global preferences on a set of alternative actions, evaluated on a consistent family of criteria. Theory of committees and elections provide the way to structure acceptable global ranking from individual expressed preferences. The case study shows an in depth analysis of students' preference models along with an effective identification of the underlying changes in order to improve the whole course. Also, the authors' new methodological approach seems very promising to be used in a wider and multipurpose evaluation structure (e.g. course, program, curricula).

Chapter 4- This chapter presents a new technique for assessing students' performance in e-Learning environments called "Enriched Assessment Rubric (EAR)". The proposed technique is designed to assist educators to assess/measure students' performance using both quantitatively and qualitatively criteria and taking into consideration the learning products (i.e. tests, individual or group deliverables produced by the students, etc.) as well as the wide range of interaction (between students, between students and educators and between students and resources) performed during all phases of complex interactive scripts, and to report students' final grades in a precise, detailed and accurate way. This technique is called enriched because it is based on the classic and well accepted assessment rubrics, which are enriched with criteria that are related to interaction analysis indicators. In this chapter, the authors highlight the philosophy and the structural components of the "EAR" technique as well as its added value. A pilot study of the application of the proposed technique in an authentic educational environment is fully described that shows how its effectiveness can be evaluated.

Chapter 5- The present chapter tackles the issue of evaluation in CSCL contexts (Computer Supported Collaborative Learning). In particular it proposes a model to evaluate online collaborative learning activities oriented to the development of students' creative skills and attitudes. The model, which encompasses three main spheres, namely the cognitive, the meta-cognitive and the affective spheres, has been tested on two collaborative activities proposed within a real online course, and the results of such test are presented, so that it is possible to make some considerations about the main strengths and weaknesses of the model itself.

Chapter 6- This study focuses on promoting primary school students' understanding of an aspect of the nature of science regarding the distinction between observation and inference. Observations are descriptive elements about natural phenomena that are accessible through senses (or extensions of senses) and about which several observers can reach consensus with relative ease. Inferences are statements about the mechanisms through which phenomena operate and cannot be resolved only through information that is accessible through senses. A scientist can infer models or mechanisms that provide a self-consistent interpretation for observations of complex phenomena. The authors have developed an activity sequence that engages students in studying stories using webcomics as the main instructional medium. The main purpose of the study was to investigate the successes and challenges of digital comics as a medium that scaffolds the learning process and engages students in explicit epistemological discourse. The study reports on results from the first implementation of the activity sequence with a class of sixth graders that draw on various data sources.

Chapter 7- Adaptive learning systems (ALE) are designed towards the main objective of tailoring learning content, system feedback and appearance to individual users – according to their preferences, goals, knowledge, and other characteristics. The implementation of a successful adaptation process is of course demanding. Adaptation is not good per se and poor realisations of adaptation may lead to disappointed users who may reject or disable adaptation mechanisms. This is why the evaluation of ALE needs to be a fundamental and integral part of their development. Evaluation should address the questions whether adaptation works on principle, whether it really improves the system, whether it leads to more effective learning, whether users prefer the adaptive features, etc. The main challenge in evaluating ALE lies in their core characteristic – adaptivity, which results in individual experiences and interactions with the system for each individual user. The attempts of dealing with this challenge are diverse. This chapter provides an overview on existing and suggested methods for evaluating adaptive e-learning. Strengths and weaknesses of the current evaluation approaches are elaborated and relevant topics in the user-centred evaluation of ALE are discussed. The evaluation methodology developed in GRAPPLE, an EC-funded project aiming at developing a generic responsive adaptive personalised learning environment, is outlined as a case study on evaluating adaptive e-learning. In GRAPPLE the concept of 'adaptation quality' is adopted and conceptualized in terms of covering different aspects of adaptive e-learning experiences that are addressed for a holistic evaluation of adaptive e-learning. In sum, the chapter aims at increasing awareness for the importance of careful and properly designed evaluation of ALE. The authros believe that the thorough consideration of the quality of adaptation and the use of evaluation approaches on the basis of mathematical-psychological models and expertise are the ingredients for a sound investigation of the benefits of adaptive e-learning and thus, for contributing to the overall notice, growth, and spread of ALE.

Chapter 8- A hypermedia system gives too much freedom to the user to navigate through the hyperspace. An Adaptive Hypermedia System (AHS) gives personalized content, presentation, and navigation support to help user find the information she needs or complete her tasks. In this chapter the authors focus on AH systems with educational purpose and are interested in evaluating to which extent AHS affect the way students learn. Brusilowsky et al. proposed in 2001 a layered evaluation of AHS, which splits the AHS into layers that are separately evaluated. Ortigosa and Carro proposed in 2003 a continuous empirical evaluation of adaptive courses. Their method is used to identify possible problems by analyzing data obtained from the access log of the students, and also suggest improvements while the course is still happening. The authors' proposal combines the continuous empirical evaluation with the layered evaluation approach. For example, the authors can use the layered evaluation technique to collect data about how much time each student spent in each page of the course and the score each student obtained in a certain test to identify differences between the patterns that lead to good or bad scores. Based on this information, a continuous empirical evaluation can be triggered to suggest some pages to students that have performed poorly in the test. Besides the time spent in each page and test scores, they also suggest other aspects of a distance learning course that can be exploited by the continuous layered evaluation technique.

Chapter 9- In this chapter, the authors introduce *Apex 2.1* (i.e., Assistant for Preparing EXams), a prototype system that provides automated feedback to e-learning students regarding the summaries they wrote about the courses they attended. At first, they present some theoretical underpinnings on how the system has been designed, and then they detail *Apex 2.1* architecture. Eventually, the first results of a validation study involving three groups of stakeholders (students, teacher, administrator) are presented. The utility, acceptability and usability of the system are examined.

Chapter 10- The Delphi technique is researched and presented as a participatory methodology in expert design, development, testing and evaluation of t-learning. It is formed from the evaluation perspective of social relations when defining educational evaluations in e-learning, illustrating it on t-learning as a constitutive part of e-learning.

Television (TV) has been maintaining its predominant broadcasting, centralised transmission and programme production applications. For this reason, its role in education and training has been lagging behind in the past decades. By introducing interactive digital TV, the space became promising for the development of its new applications, and thus, for reinforcing its role in education and training. How t-learning will respond to the challenges and take place within e-learning mainly depends on how potentials of interactive digital TV will be integrated and used in the educational arena. The Delphi technique - proven to be successful in confronting and clarifying the issues of planning and development - has led to a consensus of all the stakeholders to set up technical and pedagogical guidelines for t-learning. The methodology of the Delphi approach to the issue is described and critically assessed in the framework of a real practical case and good practice focusing on t-learning general guidelines and development of a t-learning course. There have been several t-learning courses developed; an in-depth analysis is presented for one of the topics for the area of 'Development of traffic skills'.

When planning t-learning, diverse end-user groups should be considered. TV is the main information medium for groups with no or low access to computer with Internet connection. Some of these groups are having gaps in their digital literacy and therefore are more vulnerable to digital exclusion. On the other hand, young generations which are highly digitally literate are shifting the TV use into the computer use, thus providing a higher level of interactivity. At the end of the authors’ study, the main characteristics of educational evaluation as social practice are highlighted in order to identify the main roles of stakeholders and indicators to be used in e-learning evaluation.

In: Evaluation in E-learning
Editors: Y. Psaromiligkos, A. Spyridakos et al.
ISBN: 978-1- 61942-942-0

Chapter 1

QUALITATIVE EVALUATION IN E-LEARNING

Charalambos Vrasidas[*]
CARDET – University of Nicosia, Cyprus

ABSTRACT

The purpose of this chapter is to discuss the methodological and theoretical assumptions of qualitative evaluation of e-learning. We will begin by discussing the need for evaluation of e-learning and the complexities behind the process. We will then discuss the theoretical foundations of objectivist-quantitative and constructivist-qualitative paradigms. We will examine the kinds of situations in which qualitative approaches to evaluation are more appropriate. The key message is that both quantitative and qualitative models have substantial contributions to make to e-learning evaluation. Issues such as types of research questions that are appropriate for qualitative evaluation will be discussed in detail and concrete examples will be presented. The discussion will address the interplay among the evaluator role, data collection procedures, data analysis, and standards of rigor in e-learning evaluation.

INTRODUCTION

All Instructional design models have evaluation built in the process for designing and implementing education and training activities. Therefore, evaluation should be an integral part of any e-learning project. Evaluation activities can guide all phases of development, including the early stages of idea design, development and prototyping to final development and implementation. However, evaluation is often overlooked and in several cases comes at the end as a "must-do" step to satisfy a client or a sponsor. This book attempts to shed light on the complexities of evaluation and its value for e-learning development and implementation.

The evaluation of e-learning has been faced with a range of approaches including enthusiasm and support to skepticism and cynicism. One of the main challenges faced by

---

[*] www.vrasidas.com.

evaluators is the fact that there is not one model on which e-learning is based. There are many learning theories, numerous instructional design models, and even more delivery options and platforms, which make it difficult to evaluate e-learning. In addition to all this, evaluation has grown into a field that now entails several different models and approaches.

EVALUATION

Evaluation is disciplined inquiry with the aim of evaluating the value of a program, project or system. As a systemic and systematic process it carefully collects data aimed at supporting decision making. There are several kinds of evaluation. Developmental evaluation places the evaluator as part of the development process of the e-learning program. Formative evaluation, places an emphasis on collecting data to improve the program. Whereas, summative evaluation is conducted to make final judgments about the continuation of a program.

There are numerous models and approaches to evaluation (Eisner, 1994, 1998; Hine, 2005; Patton, 2008; Stake, 2004). Several of the ideas in this chapter are in alignment with Stake's (2004) work on responsive evaluation. Stake makes the distinction between standard-based and responsive evaluation; standard-based in the one that is more quantitative in nature, whereas responsive is the one that is more qualitative and interpretive. For the purpose of this chapter, we agree with Reeves and Hedberg (2003) who advocate for a pragmatic rationale for evaluating e-e-learning, with a big focus of evaluation on the utilization of information for decision making. The main driving focus of evaluation will be the information needed by developers, e-learning designers, and policy makers to make the decisions relating to their context. Any instructional designer is often asked to conduct a needs assessment, learner and task analysis, to implement, and evaluate learning designs and program impact. Data collected often come from informal means as well as from formal methods including observations of learners as they accomplish tasks, interviews, document reviews, think aloud protocols and usability testing.

Kirkpatrick's model for the evaluation of training programs is still widely used (Kirkpatrick & Kirkpatrick, 2006). The model focuses on four components: *reaction* of the participants to the training, *learning* changes that occurred, *behavior* changes and improvement in performance, and *overall* impact and results reflected on the organization. This model has been used widely in the evaluation of education and training programs, but its application for e-learning has not been well spread.

Much of the research behind e-learning and the evaluation of effects of technology in teaching and learning have been controversial. One of the key challenges faced by evaluation practitioners is the fact that the methods are used for studying the impact of programs. Another key issue is asking the wrong questions during an evaluation. Given these challenges, it is no accident that numerous studies have shown non-significant difference between traditional and e-learning courses. One of the limitations of traditional quantitative models of evaluation of e-learning is that evaluators focus on the delivery mode and technology, and fail dramatically in capturing and assessing the educational impact of programs and the complexities of e-learning. Several scholars have argued for the need to integrate more qualitative methods in the study of e-learning and particularly for evaluating programs, processes and results (Reeves & Hedberg, 2003; Lagemann, 1999; Vrasidas, 2001).

In the next sections we will provide an overview of the theoretical foundations of education evaluation. We will discuss how the two predominant paradigms of instructional e-learning design (objectivism and constructivism) are aligned with two of the key evaluation paradigms: "objectivist-quantitative" and the "constructivist-qualitative". Depending on the orientation of the e-learning design and evaluators, data collection and analysis will be influenced by these philosophical foundations. We will then focus on the constructivist-qualitative model and discuss the issues of the conceptual framework, data collection and analysis, the role of the evaluator, and standards of rigor. In discussing this model we will intertwine some of the ideas on responsive evaluation (Stake, 2004).

EPISTEMOLOGICAL ASSUMPTIONS: OBJECTIVISM AND CONSTRUCTIVISM

The two major philosophical assumptions of objectivism and constructivism have influenced the instructional design and e-learning field for the last decades. These two systems are placed on a continuum (see Figure 1). For the purpose of this chapter, and while discussing the two ends of the continuum, we will discuss how an objectivist and how a constructivist evaluator approaches the evaluation of e-learning. We will then discuss the two approaches and provide suggestions for practitioners.

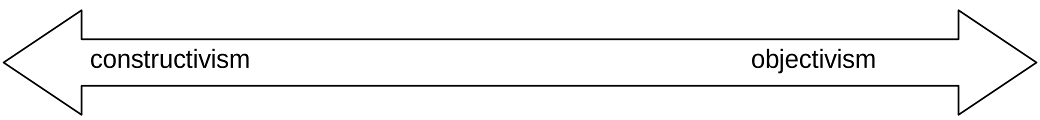


Figure 1. The constructivism-objectivism evaluation continuum.

OBJECTIVISM

Objectivism has dominated the field of instructional design, e-learning and evaluation for almost a century. Most of the traditional approaches to learning and teaching that are based on behavioristic and cognitive theories, share philosophical assumptions that are fundamental in objectivism. The major assumptions of objectivism are: there is a real world fully structured so that it can be modeled; symbols are representations of reality; the human mind processes abstract symbols in a computer-like fashion; human thought is symbol-manipulation and it is independent of the human organism; the meaning of the world exists objectively, independent of the human mind and it is external to the knower (Jonassen, 1992; Lakoff, 1987; Vrasidas, 2000).

An objectivist e-learning evaluator believes that there is one *true* and *correct* reality, which we can find out following the objective methods of science. By studying the world we can identify its structure and entities with their properties and relations, which we can then represent, using theoretical models and abstract symbols. Instructional design models that are based on objectivist philosophy (Dick & Carey, 1996) and behaviorist learning theories are represented by the input-process-output model. All learners are expected to achieve specific objectives. Objective evaluation procedures will be used to determine whether objectives are met and to what degree. Therefore, evaluation in an objectivist approach is goal-driven, places a big focus on measurement and relies heavily on quantitative data collection and analysis. Evaluation in line with this theoretical foundation is standard-based and relies heavily on measures and predefined criteria to assess the quality of a program (Stake, 2004).

CONSTRUCTIVISM

On the other end of the continuum lies constructivism. The fundamental assumption behind constructivism is that knowledge does not exist independently of the learner, knowledge is constructed. The major philosophical and epistemological assumptions of constructivism are the following: there is a real world that sets boundaries to what we can experience, but, reality is local and there are multiple realities; the structure of the world is created in the mind through interaction with the world and is based on interpretation; the mind creates symbols by perceiving and interpreting the world; meaning is a result of an interpretive process and it depends on the person's experiences and understanding (Jonassen, 1992; Vrasidas, 2000).

Constructivists believe that there are multiple truths and realities, and therefore, e-learning programs should be supporting multiple perspectives. Learners interpret their world and educators have to account for the meaning-perspectives of the learners and for their interpretations of the world. Learning is meaning-making. Under a constructivist framework, evaluation is context dependent. That is, the context within which the program is implemented should be taken into consideration during evaluation. Constructivist evaluation attempts to build meaningful constructions of the actors and the context. In Stake's (2004) work, this kind of evaluation is responsive evaluation, and it is supported by interpretive thinking. In his words, "responsive evaluation studies emphasize social issues and cultural values as well as personal and programmatic dilemmas" (p. xvi).

QUALITATIVE METHODS AND SYMBOLIC INTERACTIONISM

The theoretical foundations of constructivist-qualitative evaluation and research approaches are based on the premises that Blumer (1969) posed for *symbolic interactionism.* Firstly, human beings act upon the world on the basis of the meanings that the world has for them. Secondly, the meaning of the world is socially constructed through one's interactions with members of the community. And thirdly, the meaning of the world is processed again through interpretation. The traditional approach to evaluation tends to ignore the importance of meaning, interaction, and interpretation of the actors in shaping behavior. Qualitative inquiry attempts to understand the multiple layers of meaning represented by human actions, how they are interpreted by those involved, and how they influence the quality of programs.

Under a symbolic interactionist conceptual framework, interaction can be defined as the process consisting of the reciprocal actions of two or more actors within a given context

(Vrasidas, 2002). Interaction is not a variable that you take out from a broad context and you examine on its own. Interaction is an ongoing process that resides in a context and also creates the context. Quantitative approaches to research and evaluation use "context stripping" techniques in an attempt to control the influence of context on the findings of the study. There is a reflexive relationship between context and interaction that prevents us from isolating interaction from its context. To examine human action and interaction, it is important to examine carefully the context and the moment-to-moment events that lead to further interaction among people. Traditional approaches to studying interaction in e-learning attempted to examine interaction as a variable isolated from its context. What is important here is that interaction is interaction between actors and not factors. In an e-learning environment, when participants fit their actions to each other's actions, they can respond to each other's messages or they can ignore them.

Discussion of the Objectivist-Quantitative and Constructivist-Qualitative Models

For several decades, the debate between quantitative and qualitative methods has dominated the research and evaluation landscape. Both approaches are based on sets of beliefs, examples, traditions, that guide a certain scientific community on how to conduct its practices. Quantitative and qualitative methods are based on different epistemological assumptions. They coexist and they are used when they are appropriate. Some evaluation questions lend themselves to the use of quantitative methods, whereas some other questions lend themselves to qualitative methods. Both qualitative and quantitative approaches to evaluation have important things to offer towards the understanding of e-learning.

In addition to the epistemological differences discussed earlier, another key difference between the two approaches is their focus and what kind of questions each is more appropriate for. Although quantitative methods use numerical data and advanced statistical techniques for analyzing the data, qualitative approaches do not necessarily exclude the use of similar techniques. The key issue is to carefully decide what makes sense to count and how it can strengthen the plausibility of the evaluator's assertions (Ercikan & Roth, 2006; Stake, 2004; Vrasidas, 2001).

The big difference is in terms of content rather than procedure. Eisner (1994) stated that "Qualities are candidates for experience. Experience is what we achieve as those qualities come to be known. It is through qualitative inquiry, the intelligent apprehension of the qualitative world, that we make sense" (p. 21). Behrens and Smith (1996) have argued, "All quantities are measures of qualities, and the understanding of qualities is not simpler matter" (p. 947). Qualitative evaluation focuses on the meanings of phenomena, whereas quantitative evaluation assumes the meaning and then studies its distribution. To understand the meanings of participants in e-learning programs, the evaluator has to study them in naturally occurring situations and not in highly controlled laboratory settings. Table 1 summarizes the differences between objectivist-quantitative and constructivist-qualitative methods.

Table 1. Comparison of Objectivist-quantitative and constructivist-qualitative methods

Objectivist-Quantitative	Constructivist-Qualitative
Emphasis on knowing	Emphasis on understanding
Emphasis on measurement	Emphasis on interpretation
Objectivity	Subjectivity
Standard-based	Responsive
Context-stripping techniques	Context is important. Study in the natural setting of the phenomenon.
Use objective instruments	The evaluator is the instrument
Large sample	Small sample (case studies)
Evaluation plan is rigid	Evaluation plan is flexible and constantly adapts
Separate, distinct phases of the evaluation	Phases overlap

IMPLICATIONS OF THE CONSTRUCTIVIST-QUALITATIVE PARADIGM FOR THE EVALUATION OF E-LEARNING

One of the fundamental assumptions of qualitative evaluation is that there is not one universal *truth*. There are multiple truths, which derive from the local communities and have particular *local meanings* to the members of that community. This relates to Stake's (2004) responsive evaluation, during which the methods are negotiated with all stakeholders. Meaning and knowledge are socially constructed within a particular context, at a certain time, with specific people. Qualitative evaluation is appropriate when one wants to take into account the views of all stakeholders, and find out more about certain structures of experience, the meaning-perspectives of the actors, and specific interrelationships between actors and environment. The types of questions that are appropriate for qualitative evaluation are those dealing with the choices, actions, and meaning-perspectives of those involved. For example, one might ask:

- What are the specific structures of experience that unfold in this particular e-learning environment?
- What does interaction mean for the learners and the teacher?
- What are the specific patterns of social life and how are they manifested by these particular learners in this e-learning course?
- How are the meanings of these particular learners comparable to the meanings of learners at another e-learning course?

One of the major goals of qualitative evaluation is to discover how all the actions combined constitute an e-learning environment. Important issues that are addressed by qualitative evaluation focus on the nature of teaching with technology, the nature of the classroom as a cultural environment, how it is influenced by technology, and the meanings teachers and learners negotiate throughout in technology based learning and technology-mediated situations. The emphasis of qualitative approaches is not to study the "traits" of the actors or the environment, but to examine the process that takes place and the meanings of the actions of those involved. By examining the local meanings in action, the evaluator asks:

- What is really happening in an e-learning course?
- How does the learner perceive the content of instruction?
- What does learning via e-learning mean for the learners?
- How is interaction formed online and how face-to-face?
- How do the learning platform and technology skills influence the flow of interaction?

The majority of traditional evaluation failed to examine these kinds of questions and mainly attempted to compare various delivery systems without paying attention to the perspectives of teachers and learners. Qualitative evaluation also seeks to shed new light on the comparative understanding of different social settings.

- How do the different levels of a phenomenon interact and influence each other?
- How does this particular e-learning provider's policy on e-learning relate to the design of a particular e-learning course?
- How what happens in a particular setting is compared to what happens in other similar settings?
- How does the interaction that takes place in this e-learning course is compared to the interaction that takes place at another online course in another university?

By comparing the different levels and different settings we can get a better understanding of a given situation and improve e-learning at all levels.

The Conceptual Framework

The conceptual framework guides the process of inquiry, provides a rough guide-map for the study, and communicates the purpose of the study to the evaluator's audience. The categories of the framework are not fixed, but instead they are changing and evolving during the course of the study. From the traditional evaluation perspective, the conceptual framework is rigid and consists of a fixed number of categories with all the relationships among those categories identified in advance. Such an approach relies on the belief that "truth" exists out there and the evaluator can know in advance what it takes to find it by following the step-by-step "scientific method." In qualitative evaluation, the conceptual framework is neither fixed, nor sequential. It is rather fluid and it evolves as the study proceeds. In order for an evaluation study to address the complexities of an e-learning program its design has to be evolving and allowing new relationships to emerge from the data. Evaluation questions are also context sensitive and as such, they are also evolving and might change as the study proceeds.

DATA COLLECTION

The conceptual framework and evaluation questions of the study determine the appropriate data collection procedures. The techniques for data collection include the following

- Observation of face-to-face interaction of users engaging with the e-learning course
- Observation of pairs of students when participating in online synchronous chats
- Semi-structured interviews with the teacher, students, e-learning designers, and policy makers
- Focus groups with participants and stakeholders
- Think aloud protocols with real users
- Usability testing
- Survey questionnaires
- Collection of text transcripts of online discussions
- Collection of messages exchanged online.
- Web and LMS-analytics
- Memo writing and reflections (Anderson & Kanuka, 2003; Hine, 2005)

An important component of the data collection and data analysis procedures is the writing of detailed memos. Memo writing is a common practice in qualitative evaluation. A memo is a theoretical write-up about ideas, concepts, categories and their relationships as they strike the evaluator while in the field and/- or during data analysis. It is the evaluator's note to herself about the data, ideas, method, and the like. It is a narrative representation of the evaluator's understanding of certain aspects of the study. This procedure is used to keep track of emerging ideas and categories, stimulated further analysis and data collection, and serves as a means for the development of assertions and theory integration. As an analytical process, writing memos helps the evaluator in filling out the analytic properties of the descriptive data collected. Subtle connections may also emerge during the process of making memos. In the later stages of analysis, the memo helps in connecting the data, assertions, and the theoretical discussions.

DATA ANALYSIS

Data analysis is a precise and comprehensive process during which the evaluator constructs meaning out of data. During analysis, an account is created from the transaction that takes place between the data and the evaluator. The evaluator as the instrument and interpreter of the world influences the data analysis by making decisions throughout the research process. All analysis is theory-laden since data is interpreted through a theoretical lens, either explicit or implicit (Erickson, 1986; Stake, 1995, 2010).

The two main stages of qualitative data analysis are the inductive and deductive stages (Erickson, 1986; Vrasidas, 2001). Upon entering the inductive stage the evaluator organizes all the transcripts, field notes, summary sheets, and documents and uses data displays, concept maps, and tables to illustrate findings of the study (Miles & Huberman, 1994). The inductive stage of data analysis is very open-ended and it is the stage in which the evaluator generates assertions. After he collects and organizes all the data, he reads through the data three times and tries to gain an overall understanding of what was happening in the e-learning program. As he reads through the data, questions come to mind. He writes notes and memos about those issues and events and begins to generate assertions (Wolcott, 2009). Assertions are propositional statements that indicate relationships and generalizations in the data and which the evaluator believes are true. Examples of assertions are the following: (1) the meanings of interaction can be classified according to the intentions of the participants; (2) the structure of the e-learning program, class size, and prior experience with technology influences interaction. Once he generates assertions from the data as a whole, he enters the deductive stage. In this stage he examines the data and looks for data to confirm or disconfirm his assertions and instantiate or eliminate constructs of the conceptual framework.

Evaluator Role, Choices, and Implications

The evaluator always acts in political, institutional, social, and historical contexts. The evaluator's role is closely defined by a social network of relationships with the client, sponsor, the evaluation team, the design team, and the participants. He has a responsibility to maintain confidentiality and make sure that he does not cause any problems or harm to any of the parties involved. He has to create relationships based on trust and collaboration with the participants. The evaluator is the instrument of the study, and therefore, his or her experience with the subject of study will influence the validity of inferences. Articulation of the background of the evaluator is important in qualitative evaluation. Whereas, in traditional approaches, certain techniques are employed to eliminate possible "evaluator bias".

There is no bias-free point of view in any evaluation approach. We all filter our view of phenomena through our theoretical lens. No matter how hard an evaluator tries, he can never enter a setting as the "fly on the wall." He can never renounce his prior knowledge. However, he can discuss his preconceptions and be aware of their existence and how they influence his view of the world and interpretation of social phenomena. By discussing some of the factors that might have influenced his interpretations, it allows the readers to be co-analysts of the study and reach their own conclusions about the validity of inferences.

Validity and Generalizability

The validity of a statement is often defined as the degree to which the statement is true, credible, trustworthy, or plausible. In quantitative research, the validity of a test has often been defined as the degree to which the test accurately measured what it purported to measure. Such a belief rests on the assumption that there is one truth and the evaluator can know that truth by following the "scientific" method. Qualitative evaluation is more concerned with understanding rather than knowing a phenomenon (Wolcott, 1990), and more concerned with the validation of a claim rather than its validity (Kvale, 1996). Qualitative accounts should be judged on grounds of coherence, plausibility, and whether they help us gain a better understanding of the phenomena under study. The richness of description of data collection and analysis will determine the validity of inferences.

The issue of generalizability often comes up in discussions and criticisms of qualitative evaluation and research. Such criticisms are often used to question the “quality” and “significance” of qualitative studies. How can you generalize from a study in which you had only four students? One cannot reach generalizability without examining the particulars in great detail. Only after the evaluator closely examines the *concrete particulars*, he could reach the *concrete universals* and make claims about the quality of e-learning programs. Generalizability begins within the case at hand. Then, the findings of each study should be compared with other studies that were also examined in great detail in order to see if the results generalize to other similar settings.

The validity of a qualitative account also depends on the comprehensiveness of the description of the procedures followed in the study. The validity of inferences is based on the data as a whole and not upon one single instance. There are multiple perspectives on the nature of credibility of a given account, and multiple sources give the evaluator another way of viewing a phenomenon and insights that can be used to form new questions. Therefore, it is important to use multiple sources for data collection, such as interviews, observations, and collection of LMS analytics.

Qualitative evaluations should be judged for coherence and not for correspondence of the findings with the “objective” world. Qualitative accounts are valid to the degree that they are plausible. The reader should ask:

- How do things tie together and form the whole?
- How do the parts and the whole interact?
- What are the missing pieces?
- Does it make sense?
- Do the categories of the conceptual framework make sense?
- Do they fit the data?
- Is it a comprehensive view of the phenomenon?
- Did the evaluator look at all possible directions?”

As Stake (1995) states, the evaluator should “de-emphasize the idea that validity is based on what every observer sees, on simple replication; emphasize whether or not the reported happenings could have or could not have been seen." (p. 87). The ultimate criterion of the quality of an evaluation is the degree to which it enhances the stakeholders to make informed decisions about the specific e-learning program.

CONCLUSION

Evaluators, e-learning designers, educators, and instructional designers are required to develop e-learning courses and material. E-learning initiatives require interpretive studies to document with detail the day-to-day events that unfold in the online environment and that lead to knowledge construction. The majority of research in e-learning ignored the meanings of participants engaging in e-learning. Qualitative evaluation brings meaning back to the center of inquiry. This chapter presented the theoretical and methodological assumptions of constructivist-qualitative evaluation. Our emphasis has been on the research process with particular focus on issues concerning data analysis, evaluator role, and standards of rigor. The ideas presented are suggestions for e-learning evaluators, but they are not prescriptions on how to conduct e-learning evaluations. Each case is unique and the evaluator will have to make the decisions that best serve his or her purpose.

REFERENCES

Anderson, T., & Kanuka, H. (2003). *e-Research: Methods, strategies, and issues*. Boston: Pearson Education, Inc.

Behrens, J. T., & Smith, M. L. (1996). Data and data analysis. In D. C. Berliner & R. C. Calfee (Eds.), *Handbook of educational psychology* (pp. 945-989). New York: Simon & Schuster Macmillan.

Blumer, H. (1969). *Symbolic interactionism: Perspective and method.* Englewood Cliffs, NJ: Prentice Hall.

Denzin, N. K., & Lincoln, Y. S. (1998). Introduction: Entering the field of qualitative research. In N. K. Denzin & Y. S. Lincoln (Eds.), *Strategies of qualitative inquiry* (pp. 1-34). Thousand Oaks, CA: Sage Publications, Inc.

Dick, W., & Carey, L. (1996). *The systematic design of instruction*. New York: Harper Collins College Publishers.

Eisner, E. W. (1994). *The Educational Imagination: On the Design and Evaluation of School Programs* (3rd ed.). Upper Saddle River, NJ: Prentice Hall.

Eisner, E. W. (1998). *The enlightened eye: Qualitative inquiry and the enhancement of educational practice.* Upper Saddle River, NJ: Prentice-Hall, Inc.

Ercikan, K., & Roth, W. M. (2006). What good is polarizing research into qualitative and quantitative? *Educational Researcher, 35*(5), 14-23.

Erickson, F. (1986). Qualitative methods in research on teaching. In M. C. Wittrock (Ed.), *Handbook of Research on Teaching* (3rd edition ed., pp. 119-161). New York, NY: Macmillan.

Guba, E. G., & Lincoln, Y. S. (1998). Competing paradigms in qualitative research. In N. K. Denzin & Y. S. Lincoln (Eds.), *The landscape of qualitative research* (pp. 195-220). Thousand Oaks, CA: Sage Publications, Inc.

Hine, C. (Ed.). (2005). *Virtual methods: Issues in social research on the internet*. Oxford: Berg.

Jonassen, D. H. (1992). Objectivism versus constructivism: Do we need a new philosophical paradigm? *Educational Technology Research and Development, 39*(3), 5-14.

Kirkpatrick, D. L., & Kirkpatrick, J. D. (2006). *Evaluating training programs* (3rd ed.). San Francisco, CA: Berrett-Koehler Publishers, Inc.

Kuhn, T. S. (1996). *The structure of scientific revolutions* (3rd ed.). Chicago: The University of Chicago Press.

Kvale, S. (1996). *InterViews*. Thousand Oaks, CA: Sage Publications.

Lagemann, E. C. (1999). An auspicious moment for education research. In E. C. Lagemann & L. S. Shulman (Eds.), *Issues in education research: Problems and possibilities* (pp. 3-16). San Francisco, CA: Jossey Bass Inc.

Lakoff, G. (1987). *Women, fire, and dangerous things*. Chicago: University of Chicago Press.

Miles, M. B., & Huberman, A. M. (1994). *Qualitative data analysis* (2nd ed.). Thousand Oaks, CA: Sage Publications.

Patton, M. Q. (2008). *Utilization focused evaluation* (4th ed.). Thousand Oaks, CA: Sage.

Reeves, T. C., & Hedberg, J. (2003). *Interactive learning systems evaluation*. Englewood Cliffs, NJ: Educational Technology Publications.

Stake, R. (1995). *The art of case study research*. Thousand Oaks, CA: SAGE Publications.

Stake, R. (2004). *Standard-based and responsive evaluation*. Thousand Oaks, CA: SAGE Publications.

Stake, R. (2010). *Qualitative research: Studying how things work*. New York: The Guilford Press.

Vrasidas, C. (2000). Constructivism versus objectivism: Implications for interaction, course design, and evaluation in distance education. *International Journal of Educational Telecommunications, 6*(4), 339-362.

Vrasidas, C. (2001). Interpretivism and Symbolic Interactionism: “Making the Familiar Strange and Interesting Again” in Educational Technology Research. In W. Heinecke & J. Willis (Eds.), *Research Methods in Educational Technology* (pp. 81-99). Greenwich, CT: Information Age Publishing, Inc.

Vrasidas, C. (2002). A working typology of intentions driving face-to-face and online interaction in a graduate teacher education course. *Journal of Technology and Teacher Education, 2*.

Wolcott, H. F. (1990). On seeking-and rejecting-validity in qualitative research. In E. W. Eisner & A. Peshkin (Eds.), *Qualitative Inquiry in Education* (pp. 121-152). New York, NY: Teachers College Press.

Wolcott, H. F. (2009). *Writing up qualitative research* (3rd ed.). Thousand Oaks, CA:: Sage Publications.

In: Evaluation in E-learning
Editors: Y. Psaromiligkos, A. Spyridakos et al.
ISBN: 978-1- 61942-942-0

Chapter 2

TOWARDS A CONTEXTUALIZED DEFINITION OF E-LEARNING QUALITY: AN EMPIRICAL STUDY OF ONLINE PROGRAMS IN PROJECT MANAGEMENT

***Constanta-Nicoleta Bodea*[*] *and Maria-Iuliana Dascălu*[†]**
The Academy of Economic Studies, Bucharest, Romania

ABSTRACT

There are various definitions of quality in e-learning. A contextualized definition of quality requires taking into consideration the relevant stakeholders who have different expectations for e-learning. The chapter refers to the factors which influence the quality of training and education on project management online programmes, from two different perspectives: trainers/ teachers and trainees/ students. Two questionnaire-based surveys were conducted and the main results are presented. The first survey aimed at finding out what influences quality of project management education, according to the trainers', professors' and training providers' perspective. 81 % of the responses came from China. The rest were professionals of different EU nationalities: project managers, software developers, financial managers and professors, who are involved in both training on project management, but also as team members or team managers in projects, thus ensuring a balanced overview of both theory and practical issues. The second survey aimed at finding out what quality means for trainees and students. The questionnaire for students was spread to 300 individuals. The response rate was high: we received 253 responses. The respondents were managers and project leaders from private companies, project managers from the Academy of Economic Studies and National Institute of R&D in Informatics, from Romania and also master students. Although there were small differences of perspective, we concluded that both trainers and trainees have the same approach towards a qualitative project management e-education. We propose a quality framework based on four pillars, starting from our findings. Furthermore, expectation profiles for both students and teachers are provided, as a method of obtaining maximum

[*] E-mail: bodea@ase.ro, Web-page: www.ase.ro.
[†] E-mail: cosmicondina@yahoo.com.

satisfaction of all individuals involved in project management e-learning activities, thus improving the education quality.

Keywords: Quality, Project management, training and education, competences

INTRODUCTION

In e-learning, quality is critical, as the study of European Quality Observatory has revealed (Ehlers, Goertz, Hildebrandt, & Pawlowski, 2005). Besides technical requirements, a web-based learning platform has to prove quality. There are different ways to evaluate education programs: formative (programs are evaluated with regard to the consistency between the program plan and implementation) and summative ones (the programs are assessed in terms of their outcome) (Shi & Tsang, 2008). A proactive way to obtain positive results in the evaluation process of an e-learning platform is to establish the features of the targeted users, before delivering learning.

Many definitions of quality in e-learning exist. These definitions are strongly related to what e-learning itself means. A useful set of such definitions is provided by Súilleabháin (2003): "*e-Learning – a term covering a wide set of applications and processes, such as Web-based learning, computer-based learning, virtual classrooms, and digital collaboration. It includes the delivery of content via Internet, intranet/extranet (LAN/WAN), audio- and videotape, satellite broadcast, interactive TV, CD-ROM, and more"(ASTD's Learning Circuits), "e-learning is the systematic use of networked multimedia computer technologies to empower learners, improve learning, connect learners to people and resources supportive of their needs, integrate learning with performance and individual with organisational goal" (Peter Goodyear), "e-Learning is the use of network technology to design, deliver, select, administer, and extend learning."(Eliot Masie).*_ Súilleabháin (2003) also underlines the efforts made by e-learning industry to deal with customers' dissatisfaction with respect to quality issues. We argue that quality in e-learning is not a straightforward problem and has to be dealt consequently, depending on the context of the discussion. A contextualized definition of quality requires taking into consideration the relevant stakeholders who have different expectations for e-learning. Organizations ask for 'adaptable pre-trained workers' and for employees 'to be more involved in self-directed ongoing development' (Garofano & Salas, 2005). This is the main motivation for learning activities in adult life and also the explanation for changes in students' population: more and more students are already employed and have a career path. Huggins identified three motivational factors for adults to participate in a learning activity: possible outcome, learning opportunity and external (Huggins, 2004). The outcome has two directions: the individual and the environment in which she/he works. According to Huggins' study, the adult who starts a learning activity wants to evolve, to improve his confidence, to progress, to be good at his job or he prepares himself for a promotion. The learning process is considered to be a personal challenge, in most cases. Most adults participate at learning activities to support colleagues, 'to be an effective member of the team' or 'to pass the knowledge on to other members' (Huggins, 2004). Other researches underline the monetary and non-monetary benefits of education. In the second category, the following benefits are mentioned: carrying out more interesting and stimulating tasks, increased job stability and autonomy, healthy working conditions, more appropriate tasks (Fabra &

Camisón, 2009). Demirel simply points out the fact that adults have to engage themselves in learning activities, because of the ‘necessity to cope with change’: adults have to permanently renew their life perspective, their conduct and values (Demirel, 2009). All these features of learning population have affected learning processes, which are more competences oriented, more available online, more specific to certain professions. This chapter reveals what quality of project management e-education means, according to the individuals involved in it. The meanings of the quality definition are to be understood and applied to other contexts, if appropriate.

THE RESEARCH CONTEXT

Training is a form of education, more related to the practical aspects of knowledge. Before talking about training in project management, one should think why this kind of education is needed, then how the right kind of courses should be picked up and, finally, about customization issues. Training in project management is needed to provide the individuals with a set of similar terms and language, a set of definitions about processes and models, thus increasing the chance of having projects on time, in budget and with low risks. Picking the right kind of courses depends on the context: whether the individuals who want to participate at training courses need those courses for their general knowledge or they are already in an organization which has specific requirements. If the latter situation is under discussion, then the issue of training customization has to be analyzed: are the individuals who participate at the courses experts or novices? Do they need to learn about a specific process? And so on and so forth. Training in project management contains a considerable amount of case studies, examples, exercises. A training session could aim at developing or improving one of the project manager competences (International Project Management Association, 2006). By developing, we mean that the competence is at a basic form or does not exist at all. By improving, we mean that, from the documentation received from a candidate, we can assume that some competences exist and we can build on those. A training session could refer to one topic or more, thus having an impact on duration of the training (Kanellopoulos et al., 2006), (Lytras et al., 2008). Although it may seem that we are only approaching technical competences, a balanced percentage of theoretical and practical training, backed up by modern teaching methods (Garcia et al., 2003), could lead to the improvement of behavioral competences, as well.

Having established some of the features of training in project management, the paper will further focus on finding out what makes a project management training to be qualitative and whether project management is suitable to be taught online. The chapter presents opinions about what a good e-learning programme in project management should be like, from two different perspectives: teachers (trainers) and students (trainees). Expectation profiles for both categories are also provided. Finally, a quality framework for project management e-education based on a statistics analysis is proposed.

The Research Methodology

The main *objectives* of the research are:

- determining what factors influence quality in project management;
- creating the expectation profiles of both students and trainers, with respect to project management education;
- providing a quality framework for project management e-education;

The instruments used in current research were two *questionnaires.* The main *research question* for both questionnaires is related to what determines the quality of project management training/ programme. The authors argue on determining a good combination of traditional and modern teaching for the project managers-to-be, who do not have the physical time to attend project management traditional courses / programmes. This combination should be practical enough to offer them a ready-to-go information and also will provide them with a certification or a diploma, which is acceptable and valuable by their co-workers and clients.

Other *research questions* are:

- Is training location important for the students and the teachers involved in project management education?
- Does it matter whether the education in project management takes place face-to-face or online?
- Is it important for students to work with certified trainers and for trainers to have certifications and accreditations?
- What qualities should a project management trainer have?
- What is the student's motivation to get involved in project management education?
- What do students prefer regarding teaching methods?
- How should a good trainer be, according to students' expectations?
- The main *hypotheses* for the research are:
- H1: Trainers' or students' needs influence the perception towards the quality of the education programmes in project management.
- H2: Organization environment is affecting the quality of project management education.
- H3: Trainees' motivation influences the quality of project management education.
- H4: Trainers' skills influence the quality of project management education.
- H5: The perspective towards quality of project management education depends on the type of quality assessor: trainer or trainee.
- H6: Project managers and individuals with experiences in project management are the most suitable for being trainers and students in e-learning environments.

Both questionnaires have, as *units of analysis*, the perspective towards quality of project management education, but the *subjects* are different. For the first questionnaire, the trainers are the questioned subjects. For the second questionnaire, our focus is on the students.

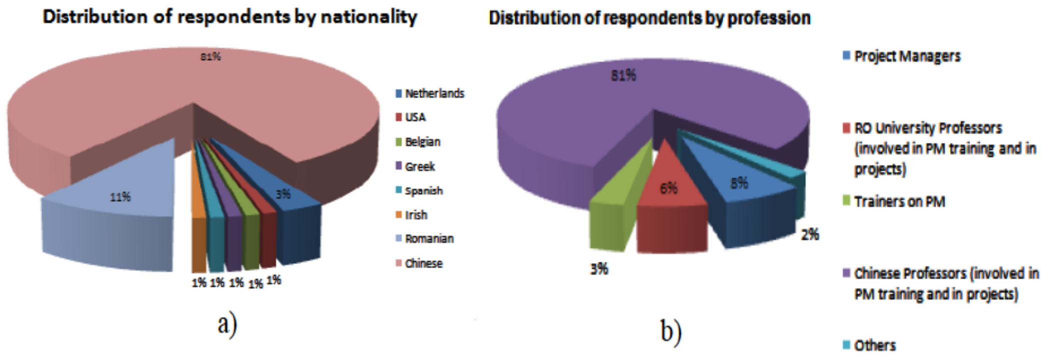


Figure 1. The respondents' distribution for questionnaire addressed to professors, trainers and training providers.

Data collection took place in 2007. The respondents were individuals involved in the activities held by an online master degree programme in project management, organized by the Academy of Economic Studies, the biggest Romanian university in economics and business administration.

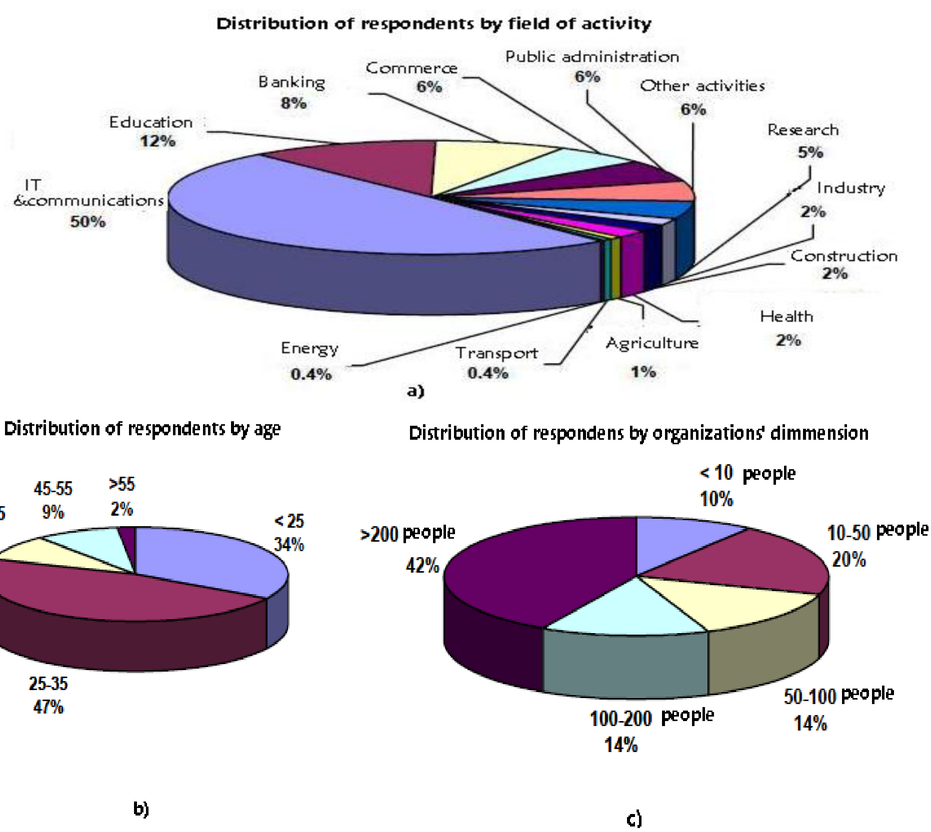


Figure 2. (Continued).

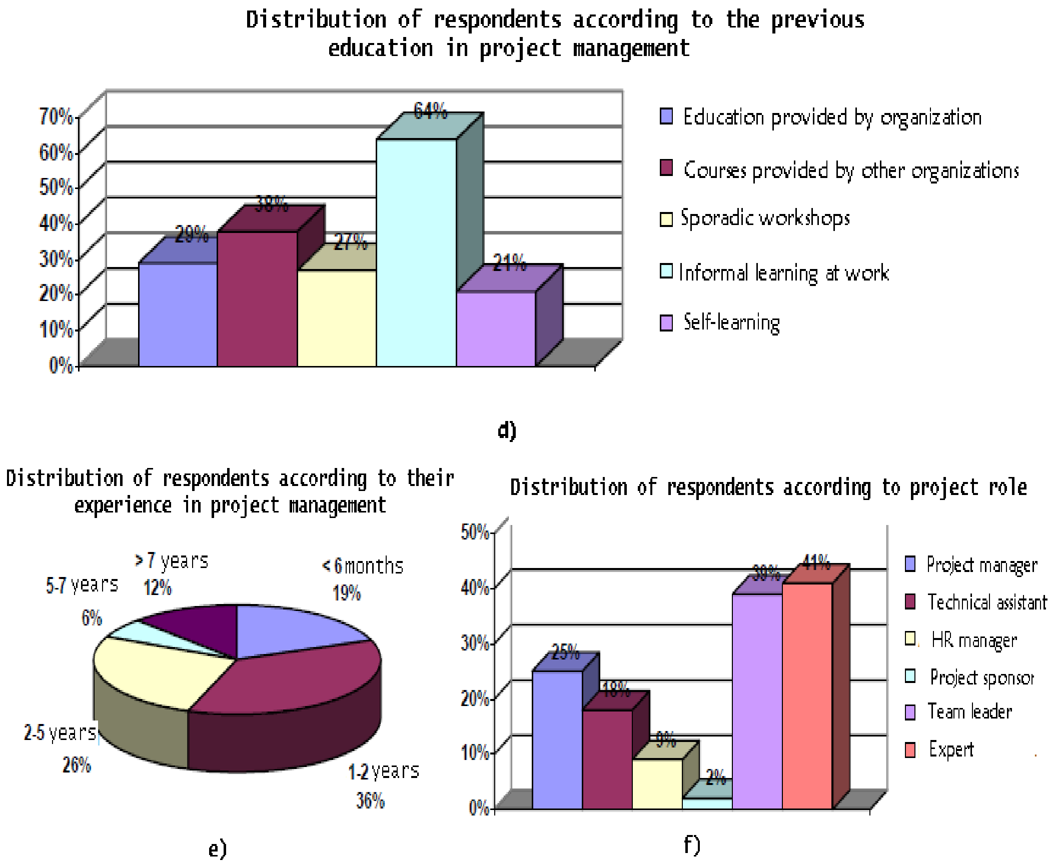


Figure 2. The respondents' distribution for questionnaire addressed to students involved in an e-learning programme.

The questionnaire addressed to professors, trainers and training providers is structured into two main parts, split into questions about general aspects of education and training and particular questions related to project management education and training. The questionnaire was given to 96 professionals engaged in providing project management education.

The respondents' distribution for the first questionnaire (which is available in Annex 1) was analyzed from two different perspectives (see Figure 1):

- *by nationality:* 81 % of the responses came from China; the rest were received from Romania, from a number of professionals of different EU nationalities;
- *by profession:* the percentage of project managers who answered the questions is rather low – 8%; most of the respondents are professors, who are involved in both training on project management, but also as team members or team managers in projects, thus ensuring a balanced overview of both theory and practical issues;

The questionnaire for students was distributed to 300 individuals. The response rate was high: we received 253 responses. The respondents were managers and project leaders from private companies, project managers from the Academy of Economic Studies and National Institute of R&D in Informatics, from Romania and also master students. The different background of our subjects helped us in determining the profile of an individual interested in project management advanced education. The questionnaire comprised 27 questions, structured in 4 sections, as follows:

- first section requested data about organizational environment of the respondents;
- second section asked for previous experience regarding project activities;
- third section tried to identify personal opinions about project related activities;
- forth section revealed users' need with respect to project management education;

The respondents' distribution for the second questionnaire (which is available in Annex 2) was analyzed from various points of view (see Figure 2):

- *by field of activity:* **–** 50% of them were working in IT& communications domain, 12% in education, 8% in banking, 6% in commerce, 6% in public administration, 5% in research, 2% in industry, 2% in construction, 2% in health, 1% in agriculture, 0.4% in transport, 0.4% in energy and 6% in other domains;
- *by age:* **–** almost 50% from the respondents are between 25 and 35 years old, 34% are under 25 and 20% are over 35;
- *by organizations' dimension*: – 42% of respondents work in big organizations, having over 200 employees; 28% of them work in medium organizations, with 50-200 employees and the rest work in small firms, with less than 50 employees;
- *by previous education in project management:* – 64% of respondents said that they learnt while working, 38% said that they were trained by other organizations, 29% learnt thanks to the courses organized by their firms, 27% learnt from workshops and 21% learnt by themselves;
- *by previous experience in project management:* – 55% of respondents have less than 2 years experience in working with project, 26% of them have 2-5 years experience and almost 20% of them have over 5 years;
- *by their role in projects:* – 103 were experts, 98 were team leaders, 63 project managers, 24 human resources managers, 19 technical assistants and 6 sponsors;

RESEARCH RESULTS AND ANALYSIS: TEACHERS' POINT OF VIEW

The survey conducted on professors, trainers and training providers revealed important aspects to the quality of project management education. As shown in Figure 3, the selected location for training is considered important by most respondents – 41 out of the 96. Only 11 state that location is very important, same applies for two who cannot decide. 13 people stated location is not very important, while 27 consider location is neither important, nor un-important. 55 respondents opt for the higher percentage of face-to-face, traditional training. At the extreme, only 14 people consider that open and distance learning (ODL) is more useful than traditional methods (see Figure 4). 24 think an equal split of modern and traditional is useful. Most respondents opt for the network video (52%) and e-platform (28%). 7% recommend case studies, while only 1% opts for e-tutoring, during and after course delivery (see Figure 5). The majority of respondents opt for the outmost importance of accreditation of training providers (national or international). At the other end, 6 respondents do not care at all about this issue.

National accreditation is seen as very important by 43 respondents, when compared to international – only 41 (Figure 6a). The same trend as above applies when talking about trainers' accreditation (Figure 6b). The analysis of the questionnaire shows that the importance of accreditation is mostly stressed for the institution, not for the person, but those who opted for 1 in the above questions, also opted for 1 or 2, in the trainers' answers, as well.

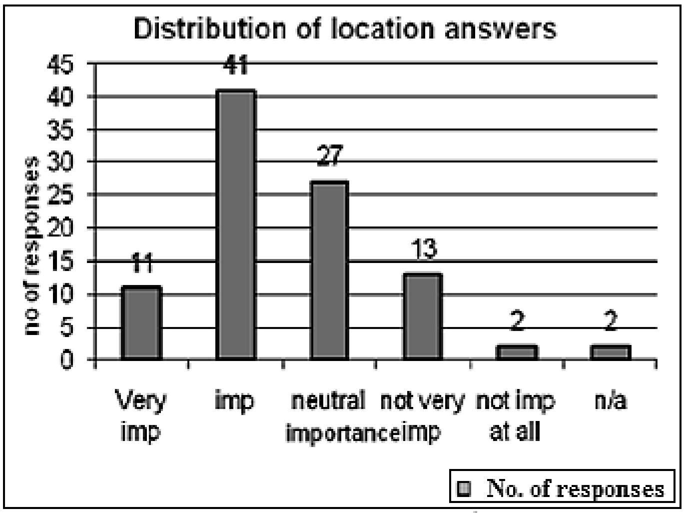


Figure 3. The distribution of answers regarding training location.

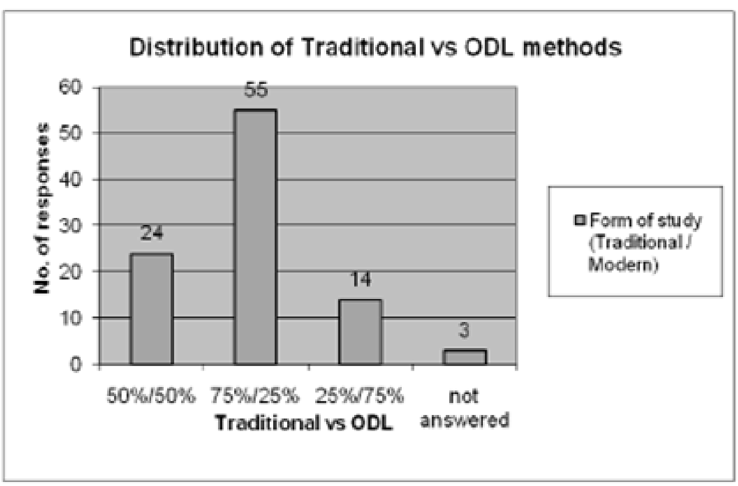


Figure 4. Distribution of answers regarding utility of the ODL methods vs. traditional methods.

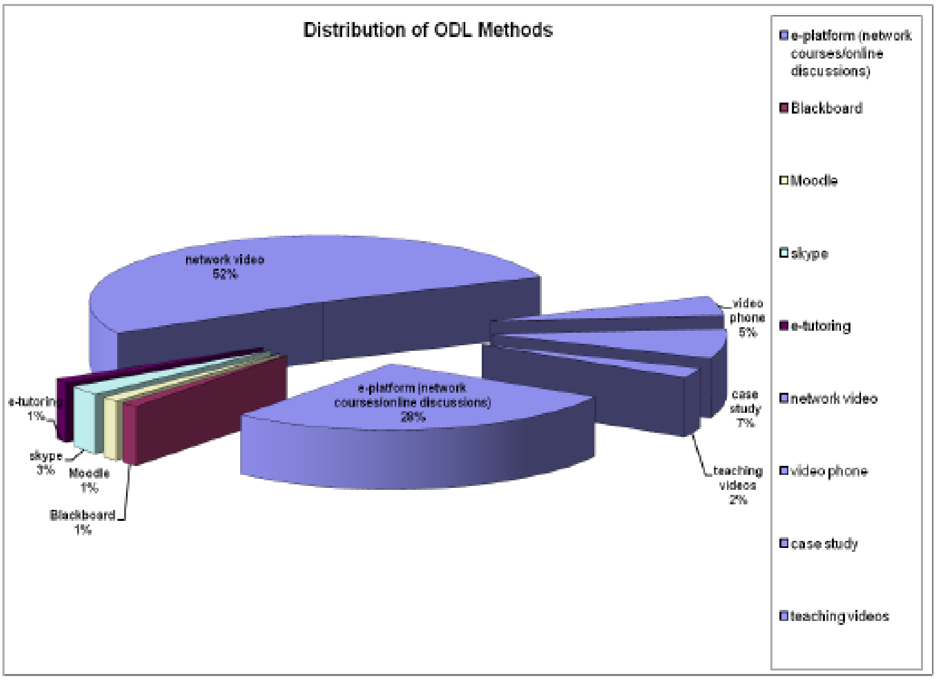
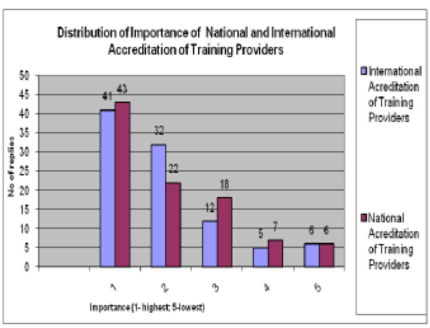
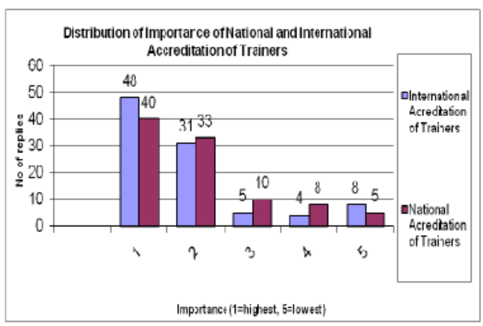


Figure 5. Distribution of answers regarding ODL methods.

With regard to the characteristics of trainers (see Figure 7), 1 person stated language skills, as being important, while 3 others stated pedagogical skills. 81 interviewees state that it is important for trainers to speak from experience, while 70 consider that the technical expertise is also required. 73 say that it would be important that the trainer is certified as Project Manager. PhD is considered important by 28 people.



a)



b)

Figure 6. a) Distribution of answers regarding the provider accreditation; b) Distribution of answers regarding the trainer accreditation.

Candidate selection is very important for 18 people, and important for 45. This leads to a majority of 63 respondents who think that selection of candidates is an aspect of quality assurance. 25 are neutral about this aspect, while 5 feel that candidate selection is not important for training quality. Most of the respondents declare themselves in favor of candidate selection. Most important documents required in the candidate file are: CV (very important for 66 respondents), Diplomas, and Employer’s references (very important for 39 of them, but completely unimportant for 48), as well as an Application Form.

Based on the data obtained from the survey, we could establish in 2008 a small profile for trainers involved in project management activities (see Table 2).

Table 2. Expectation profiles for trainers regarding education in project management

Nationality:

	Accreditation	Teaching method	The possibility of personalization
Romanians	important	face-to-face education distant education (case studies, projects)	useful
Other nationalities	very important	interactive education (e-learning)	very useful

Professions:

	Accreditation	Teaching method	The possibility of personalization
Project managers	important	interactive education (e-learning)	very useful
Professors	very important	face-to-face education	very useful
Trainers	important	face-to-face education	useful

According to our respondents' opinion, the form of addressing the needs of the ones involved in the e-learning activities is the most important factor which determines the quality of the training (see Figure 8). According to the R square value (0.87), the endogenous variables are explained in 87% proportion by the exogenous ones, so the quality indicator is positively influenced by people's needs, organization accomplishments, trainees' motivation and trainers' skills. The research hypotheses H1-H4 are true. The statistics from Figure 8 was made with EViews tool and Ordinary Least Squares method was applied.

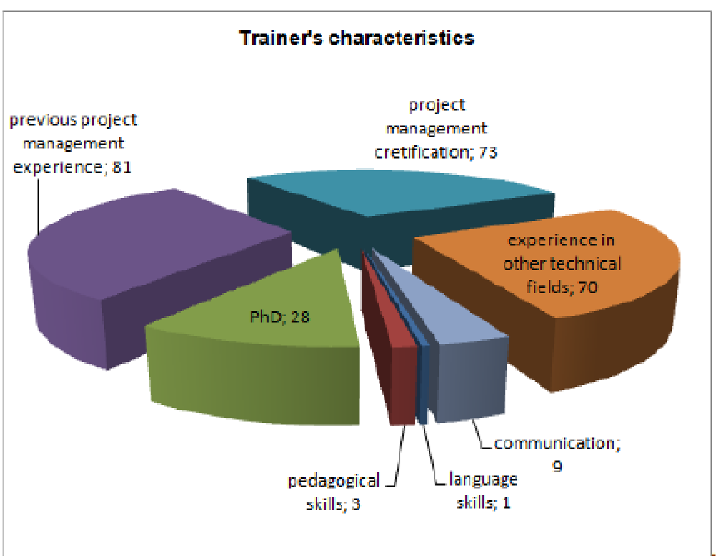


Figure 7. Distribution of answers regarding the trainer characteristics.

Dependent Variable: QUALITY R-squared 0.87	
Variable	**Coefficient**
NEEDS	0.400070
ORGANIZATION	0.135136
TRAINEE_S_MOTIVATION	0.420516
TRAINER_S_SKILLS	0.273471
C	-21.49685

Figure 8. Regression output for the multiple regression model regarding the factors which influence quality of project management trainings.

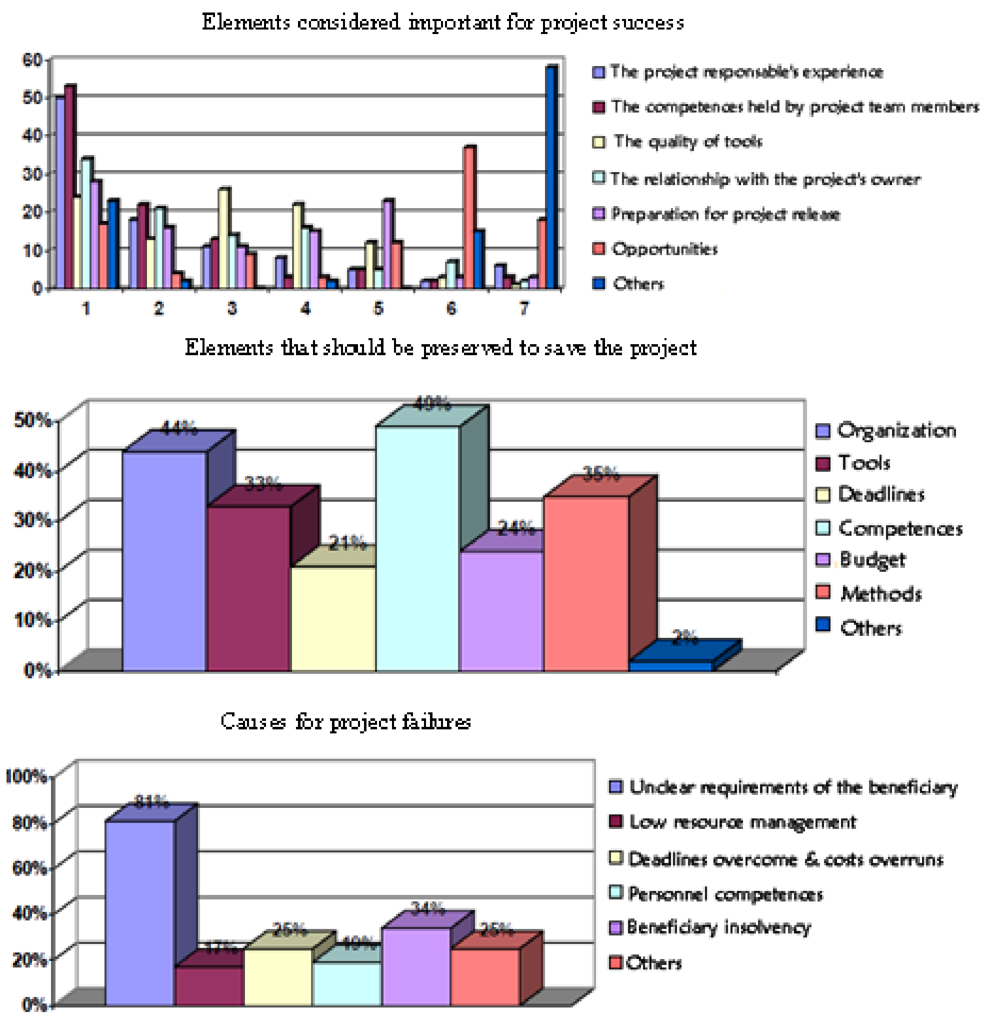


Figure 9. Students' perception about the role of competences within projects.

RESEARCH RESULTS AND ANALYSIS: STUDENTS' POINT OF VIEW

Our second survey revealed different quality attributes of a project management educational programme, depending on business sector, age, experience, role in project management activities of our respondents.

More than 50% students consider that the experience of the project leader and the competences held by the project team are the most important ingredients for the project success (see Figure 9). This aspect is also underlined by other researchers: competence development 'is one of the critical success factors' for project management (Suikki, Tromstedt & Haapasal, 2006). 49% of the respondents said that they would preserve the personnel competences in order to save the project (see Figure 9). 19% of them indicated the lack of competences as the main source of project failure (see Figure 9). The development of key competences to obtain the economic welfare and growth is possible only through systematic and continuous learning. The fact that our respondents understand the impact of competences on project's life proves their awareness of education importance.

The study revealed that more that 50% respondents prefer the combined type of education (on line and class lectures), where as the opinions about a fixed or a variable duration of a course are almost equally divided (see Figure 10).

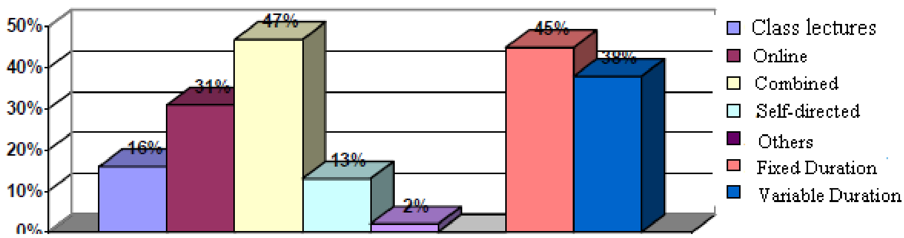


Figure 10. Preferred forms of instruction and training.

The motivation for engaging in a project management education resides in obtaining specialized skills or a professional certification (see Figure 11).

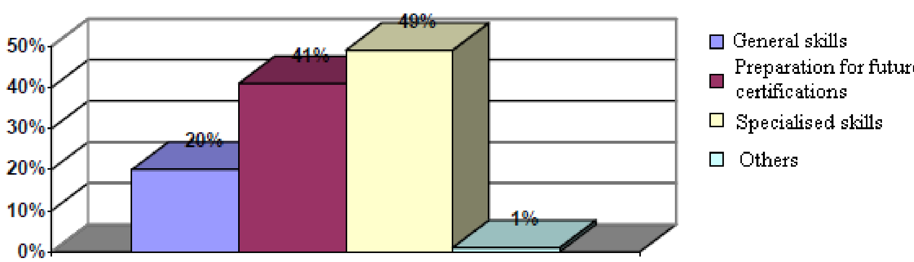


Figure 11. Motivation for engaging in e-learning activities.

The interests towards a special area of project management are diverse enough. Most students prefer interaction methods of education (see Figure 12) and consider that an adaptive e-learning system would be useful or very useful.

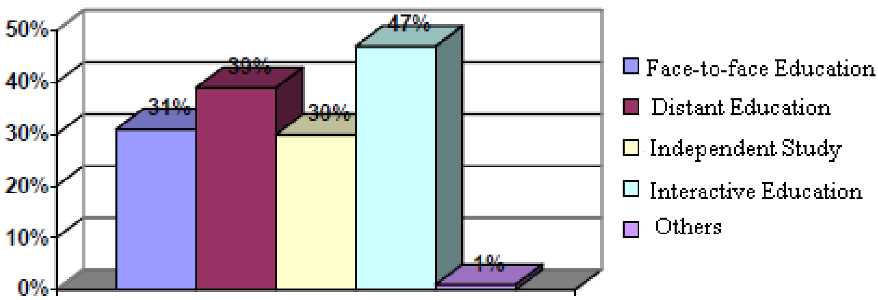


Figure 12. Teaching methods preferred by students.

Table 2. Expectation profiles for students regarding education in project management

Age

	Teaching method	Teaching duration	The possibility of personalization
25 years – 35 years	interactive education (traditional + online)	variable/ fixed	useful (based on learning objectives)
35 years – 45 years	interactive education (traditional + online)	variable	useful (based on previous experience)
>45 years	independent study (just online)	variable	very useful (based on previous experience)

Previous experience in project management

	Learning motivation	Teaching method	The possibility of personalization
6 months – 2 years	specialized skills	interactive education (traditional + online)	useful (based on learning objectives)
2 years – 7 years	certification	interactive education (traditional + online)	useful (based on learning objectives)
>7 years	specialized skills	independent study (just online)	very useful (based on previous experience)

The role in the project:

	Learning motivation	Teaching method	The possibility of personalization
Project manager	specialized skills	interactive education (e-learning)	very useful
Technical assistant	certification	interactive education (e-learning)	very useful
HR manager	certification	face-to-face education	useful
Sponsor	specialized skills	distant education	very useful
Team leader	specialized skills	interactive education (e-learning)	very useful
Expert	specialized skills	interactive education	useful
Field of activity: no influence on the way in which the e-learning quality is perceived.			

The factors for developing a user-centered application are depicted in Figure 13: 68% consider that trainee's objectives are the most important factor which should be taken into consideration when designing an e-learning application, more than 50% consider previous preparation and experience of targeted students to be important when building an e-learning platform, around 30% of the students consider their personal learning style to be important for the success of an e-learning application. All these findings underline the subjective definition of e-learning quality, in students' opinion.

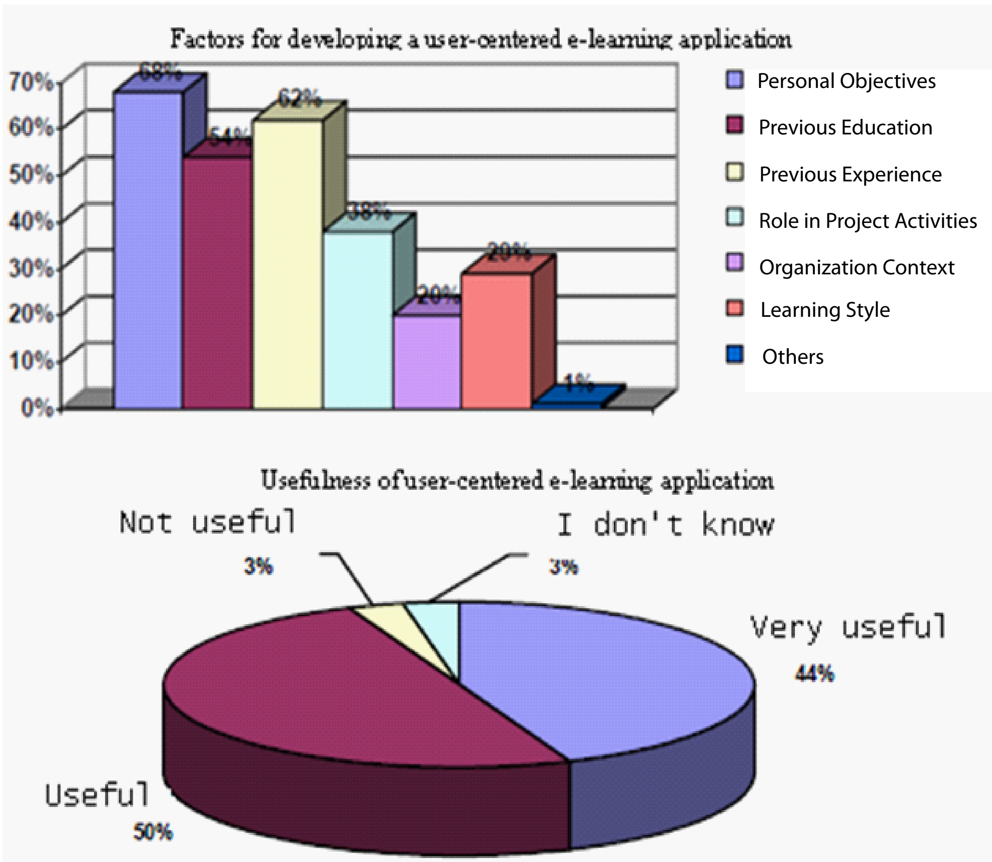


Figure 13. Factors for developing a user-centered e-learning application.

Trainees have a stronger preference for online education: almost all of the interrogated students answered that the online master programmes in project management are a good thing. They justified their option by listing some of the reasons for the emergence of e-learning: a way of overcome time constraints, a good surrogate of traditional learning, an efficient method of learning, flexibility, easy access to information, interactivity (Bodea, 2007). Very few considered the online platforms to be inaccessible, from the technical point of view. They also consider face-to-face meetings occasionally organized to be useful. Other researches revealed that face-to-face approaches and online approaches have equal significance in educational process. The balance might tilt towards one approach depending on the subject to be taught (Kelly et al., 2007). Although most students have a passive attitude towards online discussions, they consider fora an important ingredient of a successful e-learning platform, because they can easily extract information from what others say. Most students think that flexibility in learning environments helps them to obtain better results: they prefer to choose their learning path or, at least, their deadlines for preparing the homework. This finding is also highlighted by Monolescu & Schifter: they explain the request for flexibility and interaction by the need to feel included. (Monolescu & Schifter, 2000)

The instructor’s role is very important to the over-all quality of the education programme. This finding is also supported by previous researches (Guardado & Shi, 2007). The main responsibility of the instructor is to offer support for learning activities (explanations, recommendations). Other responsibilities which were pointed out by the students are: promoting the collaborative learning, monitoring the students’ participation, facilitating/ moderating communication, providing well-structured materials, presenting real case studies, implementing strategies for adult education, being actively involved in discussions with the students. As interactivity is important to 66% of the questioned students, we asked for the most valuable interaction techniques which can be used by a teacher. The following techniques and instruments were identified as being useful: feedback, open and creative questions, team work, debate subjects, short activities that hold a percent in the final grade, online presentations. It is good that students are willing to give feedback to their teachers, as this “can provide a rich source of information to help the instructors evaluate specific elements of course design and structure, make revisions, and assess the effects of those changes” (Brew, 2008).

The responses received from the second survey enlarged our view about quality of project management education and training programmes. The outcome of the study on the students is resumed below:

- age range influences the type, methods and duration of the desired education, as well as the ways of adapting it to the users’ needs;
- previous work experiences influences the type, the methods, the motivation or the area of project management education, as well as the possibility of an adaptive tool of education and the ways of adapting this tool;
- the role in project management activities influences the motivation, the methods, the area of interests and the possibility of an adaptive education platform;
- the business sector has no impact on users’ needs when talking about a project management e-education;

THE SYNTHETIC RESULTS

By comparing the expectation profiles of the trainers (Table 1) and the students (Table 2), we conclude that the project managers are the most suitable teachers for the online environments. They are also good online students. Thus, hypothesis (H6) turned out to be true. The Romanian teachers are still reserved in adopting entirely the e-learning methods, but, because of the great request for such methods, they are willing to adapt themselves. There is a strong compatibility between project managers, as teachers and project managers and team leaders, as students. This finding proves the importance of experience, which is highly realized by individuals involved in project management activities. The concern towards the accreditation of project management programmes is also shared by both trainers and trainees. The criteria for being admitted into such programmes are much fuzzier in trainees' opinion. The students proved to be much more eager to use technology in learning process and accepted very easily the e-learning approaches. This phenomenon is not typical to project management, as other researchers also discovered (Tallent-Runnels et al., 2005), (Young & Norgard, 2006).

In 2009, we added a dummy variable, ISTRAINEE, to the independent variables from previous regression model (see Figure 8). The new variable can take only two values, depending on the type of quality assessor: trainer or trainee. By inserting a new factor, the impact of trainees' needs and trainees' motivation is slightly diminished. The second model has a lower R square (0, 8) (see Figure 14), but the difference is not very significant. Still, the obtained R square is big enough, so we consider hypothesis (H5) to be also true.

Starting from the above regression model, we propose a quality framework for project management online programs, with four pillars and respecting the perspective of all the ones involved in the process (students and teachers):

- respecting trainees' needs;
- providing good organization conditions (logistics and technical platform features);
- respecting trainees' motivation for joining an e-learning program;
- searching highly skilled trainers;

Dependent Variable: QUALITY R-squared 0.8	
Variable	Coefficient
ISTRAINEE	-0.393841
NEEDS	0.379976
ORGANIZATION	0.137221
TRAINEE_S_MOTIVATION	0.355822
TRAINER_S_SKILLS	0.296807
C	-15.67825

Figure 14. Regression output for the multiple regression model regarding more factors which influence quality of project management trainings.

Conclusion

According to our survey, we conclude that quality in project management education and training depends on: organization and logistics – a good organization and combination of teaching methods adds value to training (assessed through Session Feedback Forms), the training need addressed – practical solutions, exercises, case studies (assessed through Session Feedback Forms), trainer's skills – pedagogy, languages, knowledge of the field, experience in the field, trainee's commitment and motivation.

Training providers should promote a coaching service for a limited period after the training sessions. Alumni's databases and fora are desirable solutions. Training providers are not responsible for the performance of the graduate at the work place. Training impact should be assessed through feedback forms, after a reasonable period from session graduation. This chapter can be an important tool in improving quality of project management education, as it identifies the factors which influence it. The proposed quality framework, with four pillars, can be a starting point for improving project management programmes. In the future, the expectation profiles for students and teachers should be better aligned, so that the compatibility analysis offers more relevant information. The chapter highlights the importance of e-learning instruments in project management. Romanians, although slightly reluctant, are willing to accept and use this kind of modern tools, which will ensure the spreading of project management education.

ANNEX 1

Project Management Education and Training Quality Questionnaire 1

Please take the time to fill in the questionnaire below, which is designed for the purpose of defining what quality training and education in project management should look like. After filling in this questionnaire, kindly send it back, by email, to ...

Thank you!

1. General Education and Training Quality Questions

In this section of the PM E&T Questionnaire, we aim at finding out what the conditions for a quality training session are. There are 2 main areas we should cover: training location and organization and e-learning. Each has a number of closed questions and one or more open questions.

1. How important is a physical location in delivery of education and training services? (1 – very important, 5 – not important at all)

1 2 3 4 5

2. If the traditional education and training were combined with modern ODL (open and distance learning) methods, please state the most effective percentages (traditional / ODL):

T/ODL = 50%/50% T/ODL = 75% / 25% T/ODL = 25% / 75%

3. Please state which modern ODL method you find most effective and why

4. What are the preferred forms of instruction and training? Mark the corresponding cells with X.

Class lectures	
Online	
Combined	
Self-directed	
Fixed duration	
Variable duration	
Others	

5. What are the factors for developing a user-centered e-learning application? Mark the corresponding cells with X.

Personal objectives	
Previous education	
Previous experience	
Role in project activities	
Organization context	
Learning style	
Others	

6. Do you consider a user-centered e-learning application in project management to be useful? Mark the corresponding cell with X.

Very useful (based on previous experience)	
Useful (based on learning objectives)	
Not useful	
I don't know	

7. Please state what other types of services should be added to the education and training services. (e.g. catering, accommodation, virtual library, open course-ware, alumni, etc.)

8. When applying e-learning / other ODL methods for training and education, how would you rate the following items for evaluation of results? Please add your own items, if they are not on the list. (1 – very important, 5 – not important at all. Mark the corresponding cell with an X.)

	1	2	3	4	5
Confidence in the candidate's honesty and commitment					
Time estimated for examination					
Difficulty of examination					
Electronic evaluation system characteristics					

9. Please add any other comments and suggestions.

2. Specific Project Management Education and Training Quality Questions

In this section, we aim at finding out what makes a quality and effective Project Management training. There are 5 distinct areas with questions related to key aspects of PM training: training providers, trainers, curriculum, candidate selection (enrolment conditions) and candidate evaluation. Each has a number of closed questions and one or more open questions.

1. Training Providers

1. How important is it for the training provider to be accredited by an international accreditation organization? (1 – very important, 5 – not important at all)

1 2 3 4 5

2. How important is it for the training provider to be accredited by a national body of accreditation (Ministry of Education or of Labour, relevant public bodies)? (1 – very important, 5 – not important at all)

1 2 3 4 5

3. What type of infrastructure should a training provider have to ensure good quality training delivery for 1 project management course? Please tick from the list below:

Own training centre	Yes	No
At least 2 training rooms	Yes	No
Video projectors & screens & laptops	Yes	No
Multiplication & printing equipment	Yes	No
Virtual learning platform	Yes	No
Own library (including e-library)	Yes	No
Other,	please specify	

4. For the development of 1 project management course, how many trainers should the training provider mobilize?

5. With relation to human capital for project management training administration, development and delivery, please tick the necessary resources from the list below:

Administrative Assistant	Yes	No
Trainers	Yes	No
Training Assistants	Yes	No
Certification assistant	Yes	No
Archive manager	Yes	No
PR & Promotion Manager	Yes	No
Other, please specify		

6. What should / can a training provider outsource, in order to make training on project management more qualitative? Please tick from the list below:

Catering	Yes	No
Organization and logistics	Yes	No
Printing and editing of materials	Yes	No
E-learning facilities	Yes	No
Other, please specify		

2. Trainers

1. How important is it for the trainer to be certified by an international certification organization?(1 – very important, 5 – not important at all)

1 2 3 4 5

2. How important is it for the trainer to be certified by a national certification organization?(1 – very important, 5 – not important at all)

1 2 3 4 5

3. Please rate the following characteristics of project management trainers. (1 – very important, 5 – not important at all. Mark the corresponding cell with an X.)

Item	1	2	3	4	5
Proven Project Management experience					
Previous Project Management certification					
PhD					
Technical knowledge / experience in other fields (engineering, constructions, IFI[1], IT, research, soft[2] projects)					

4. In addition to the above, please state below other comments and suggestions.

3. Curriculum

1. How essential do you see the following courses/modules in a Project Management course/program? (1 – extremely essential, 5 – not at all essential. Mark the corresponding cell with an X.)

Item	1	2	3	4	5
Basics of Project Management (Initiation, Planning, Execution, Control & Close-down)					
Project Risk Management					
Human Resources Management					
Financial Management					
Project Quality Assurance					
Information & Documentation Management					
Leadership					
Ethics					
Business					
Project Management Software					

1. In your opinion, *what other topics than the ones above* should be approached in a complete[3] Project Management training course?

2. How should the practical and theoretical knowledge be combined for best results?

[1] IFI – International Financing Institutions, such as Asian Development Bank, World Bank (IBRD, IDA, IFC, MIGA, ICSID), EBRD, IMF, IADB, AFDB and financial institutions of the European Union.

[2] Soft projects means projects in the fields of education, social services, human resources, social inclusion, awareness & PR, capacity building, NGO, environment, etc. By "complete" PM training course, we mean that at the end of the course, a student will be able to develop a project proposal, set out a general strategy for implementation, analyze the crucial aspects regarding project restrictions, limits, risks, deadlines, monitoring indicators, etc.

Practical / Theoretical 50%/50%
75%/25% 25% / 75%

3. What would you consider a quality result of a training course on project management? Please tick from the list below, or add your own:

A complete project proposal Yes No
A project analysis case study Yes No
Other, please specify

4. Please add your comments and suggestions

4. Project Management Training Candidate selection

1. Is candidate selection important for enrolment in a Project Management course? (1 – very important, 5 – not at all important)

1 2 3 4 5

2. Please select the documents you feel would be necessary for candidate selection, from the ones below. Feel free to add any other documents you think may help the selection process. Mark the corresponding cell with an X.

Item	1	2	3	4	5
CV					

Employer’s references					
Standard Application Form of the Training Provider					
Academic Diplomas					

3. Please add any other suggestions or comments.

5. Candidate Evaluation and Follow-up

1. In granting a Project Management diploma / certificate, please rate the importance of the following factors. (1 – very important, 5 – not at all important. Mark the corresponding cell with an X.). Please add any other form of evaluation you consider important.

2. Training providers should assess the impact of project management training. Please find below a number of possibilities – tick the ones you feel are appropriate. Please add your own.

Item	1	2	3	4	5
Written theoretical examination					
Written – development of an application (project proposal, strategy, etc.) of the theoretical aspects studied					
Debate on theoretical aspects					
Debate / interview on a given case study					

Item	1	2	3	4	5
Debate / interview on one of the Candidate's actual projects					
Any number of combinations of the possibilities above					
Only examinations focused on practical aspects of project management – both written and oral					

An information update on the training graduate's activity in project management 6 months after course graduation Yes No

A Company Feedback Form 6 months after training graduation, from the Candidate's employer Yes No

Other, please specify

6. Other Key Aspects

Please add any other suggestions and comments

Personal information:
Age: __________
Graduated faculty: ________________________________

Other completed training programs related to project management (if there are):
____

Experience in project management (number of years): ____

Position in your organisation: ______________________________

Monthly income (below 1500 RON/month, over 1500 RON/month):
____

ANNEX 2

Project Management Education and Training Quality Questionnaire 2

Please take the time to fill in the questionnaire below, which is designed for the purpose of defining what quality training and education in project management should look like. After filling in this questionnaire, kindly send it back, by email, to ………………………….

Thank you!

A. Organizational Environment

1. In which sector your organization belongs to? Mark the corresponding cells with X.

Industry	
Trade	
Energy	
IT	
Transport	
Construction	
Education	
Research	
Banking	
Agriculture & forestry	
Public administration	
Health	
Other (please specify)	

2. Number of employees in your organization. Mark the corresponding cells with X.

<10	
10-50	
50-100	
100-200	
>200	

3. Please specify how the project management (PM) procedures are applied in your organization. Mark the corresponding cells with X.

	Yes	No	I,don't know
Use of PM is limited (just a few departments use PM procedures).			
My knowledge about the use of PM inside the organization is limited to my area of activities.			
The organization doesn't use the term "project".			
Each department has its own PM procedures.			
Project leaders are assigned by a PM office.			
Project leaders are assigned by organization departments.			
There is a planning and control plan for each project.			
There is a planning and control for all the projects.			

B. Previous Experienced regarding Project Activities

4. What is your education in project management? Mark the corresponding cells with X.

Inside organization	
Formal courses at other organizations (ex. master)	
At workplace	
Workshops	
Others	

5. Please specify your PM certifications.

6. Did you lead any project? Mark the corresponding cells with X.

Yes	
No	

7. Since when are you involved in projects' activities? Mark the corresponding cells with X.

Less than 6 months	
1-2 years	
2-5 years	
5-7 years	
More than 7 years	

8. What is your role in projects? Mark the corresponding cells with X.

Director	
Resources administrator	
Team leader	
Technical assistant	
Sponsor	
Expert	
Other (please specify)	

C. Personal Opinions about Project Related Activities

9. What elements are important for project success, in your opinion? Mark the corresponding cells with X.

The project leader's experience	
The competences held by project team members	
The quality of tools used in project management activities	
The relationship with the project's owner	
Preparation for project release	
Opportunities	
Other	

10. What elements should be preserved to save the project? Mark the corresponding cells with X.

Organization	
Tools	
Deadlines	
Competences	
Budget	
Methods	
Others	

11. What are the causes for project failure? Mark the corresponding cells with X.

Unclear requirements of the beneficiary	
Low resource management	
Deadline overcome & Costs overruns	
Personnel competences	
Beneficiary insolvency	
Others	

D. Needs with Respect to Project Management Education

12. What are the preferred forms of instruction and training? Mark the corresponding cells with X.

Class lectures	
Online	
Combined	
Self-directed	
Fixed duration	
Variable duration	
Others	

13. What are the reasons for engaging in e-learning activities? Mark the corresponding cells with X.

General skills	
Preparation for future certifications	
Specific skills	
Others	

14. What are the areas of interest for you? Mark the corresponding cells with X.

Planning & control	
Time management	
Financial management	
Risk management	
Quality management	
Acquisition management	
Post-evaluation activities	
Team selection	

15. What are your favorite teaching methods? Mark the corresponding cells with X.

Face-to-face education	
Distant education	
Independent study	
Interactive education (traditional + online)	
Others	

16. What are the factors for developing a user-centered e-learning application? Mark the corresponding cells with X.

Personal objectives	
Previous education	
Previous experience	
Role in project activities	
Organization context	
Learning style	
Others	

17. Do you consider a user-centered e-learning application in project management to be useful? Mark the corresponding cell with X.

Very useful (based on previous experience)	
Useful (based on learning objectives)	
Not useful	
I don’t know	

18. How important is the instructor’s role in an online educational programme? (1 – very important, 5 – not important at all) Mark the answer.

1 2 3 4 5

19. How is the ideal instructor in an online class? Mark the corresponding cell with X.

Involved (ready to offer answers, advices, explanations)	
Not so involved (I don’t really need an instructor)	
It doesn’t matter.	

20. What is your favourite method of communication with your instructor/ teacher? Mark the corresponding cell with X.

On the plaform, on forums	
On the platform, using online meetings	
Through e-mail	
Face-to-face	

21. If you want to develop any of the answers or to make some comments, please use the spaces below:

Personal information:
Age: __________
Graduated faculty: ________________________________

Other completed training programs related to project management (if there are):

____

Experience in project management (number of years): ____

Position in your organisation: ___________________________

Monthly income (below 1500 RON/month, over 1500 RON/month): ______________________________________

REFERENCES

Arbaugh, J. B. (2004). Learning to learn online: A study of perceptual changes between multiple online course experiences. *Internet and Higher Education*, *7*, 169–182.

Bodea, C. N., & Dascalu, M. (2009). Designing Project Management Tests based on Semantic Nets and Concept Space Graph. *Creating Global Economies through Innovation and Knowledge Management: Theory & Practice, Proceedings of the 12th International Business Information Management Association Conference* (pp. 1232 – 1236). Kuala Lumpur: IBIMA.

Bodea, C. (2007). SinPers-An Innovative Learning System in Project Management. *Proceedings of 21st IPMA World Congress.* Cracow, Poland.

Bodea, C. (2007). An Innovative System for Learning Services in Project Management. *Proceedings of 2007 IEEE/INFORMS International Conference on Service Operations and Logistics. And Informatics* (pp. 417-421). Philadelphia, USA: IEEE.

Bodea, C. N., & Dascalu, M. (2009). *Developing Project Management Competencies through Computer Adaptive Testing*, Proceedings of the International Conference on Knowledge engineering Principles, and techniques, KEPT 2009, Cluj-Napoca.

Bodea, C. N., & Dascalu, M. (2009). *Competency-based E-learning, in the Framework of Knowledge Management*, Proceedings of the 4th International Conference on Business Excellence, 16-17 October 2009, Brasov, Romania, vol. 1, ISBN: 978-973-1747-10-1, 978-973-1747-11-8, Infomarket Publishing House, Brasov, Romania, pg. 44-47.

Brew, L. S. (2008). The role of student feedback in evaluating and revising a blended learning course. *Internet and Higher Education, 11*, 98–105.

Castells, M., & Pekka, H. (2002). *The Information Society and the Welfare State. The Finnish Model.* Oxford: Oxford University Press.

Demirel, M. (2009). Lifelong learning and schools in the twenty-first century. *Procedia Social and Behavioral Sciences*, *1*, 1709–1716.

Ehlers, U.-D., Goertz, L., Hildebrandt, B., & Pawlowski, J. M. (2005). *Quality în e-learning - Use and Dissemination of Quality Approaches in European e-Learning - A study by the European Quality Observatory.* Luxembourg: Office for Official Publications of the European Communities.

Fabra, E., & Camisón, C. (2009). Direct and indirect effects of education on job satisfaction: A structural equation model for the Spanish case. *Economics of Education Review*, 28, 600–610.

Garcia, A. C., Kunz, J., Ekstrom, M., & Kiviniemi, A. (2003). *Building a Project Ontology with Extreme Collaboration and VD&C.* CIFE Technical Report #152, Stanford University.

Gareis, R. (2007). *Happy Projects!* (Romanian version ed.). Bucharest, Romania: ASE Printing House.

Garofano, C., & Salas, E. (2005). What influences continuous employee development decisions? *Human Resource Management Review*, 15, 281–304.

Guardado, M., & Shi, L. (2007). ESL students' experiences of online peer feedback. *Computers and Composition*, *24*, 443–461.

Huggins, K. (2004). Lifelong learning——the key to competence in the intensive care unit? *Intensive and Critical Care Nursing*, 20, 38—44.

International Project Management Association. (2006). *IPMA Competence Baseline.* Nijkerk, Netherlands: IPMA.

Kanellopoulos, D., Kotsiantis, S., & Pintelas, P. (2006). Ontology-based learning applications: a development methodology. *Proceedings of the 24th IASTED International Multi-Conference Software Engineering.* Innsbruck, Austria.

Kelly, H. F., Ponton, M. K., & Rovai, A. P. (2007). A comparison of student evaluations of teaching between online and face-to-face courses. *Internet and Higher Education, 10,* 89–101.

Lytras, M. D., Carroll, J. M., Damiani, E., & Tennyson, R. D. (2008). Emerging Technologies and Information Systems for the Knowledge Society. *First World Summit on the Knowledge Society, WSKS.* Athens, Greece.

Marimuthu, M., Arokiasamy, L., & Ismail, M. (2009). Human Capital Development and Its Impact on Firm Performance: Evidence from Developmental Economics. *The Journal of International Social Research*, 2(8), 265-272.

Markkula, M. (2006). *Creating Favourable Conditions for Knowledge Society through Knowledge Management, eGorvernance and eLearning. (pp. 14-19).* Budapest, Hungary: FIG Workshop on eGovernance, Knowledge Management and eLearning.

Monolescu, D., & Schifter, C. (2000). Online Focus Group:A Tool to Evaluate Online Students' Course Experience. *The Internet and Higher Education*, *2*, 171- 176.

Muijs, D. (2004). *Doing Quantitative Research in Education.* London: Sage Publications.

Shi, Y., & Tsang, M. (2008). Evaluation of adult literacy education in the United States: A review of methodological issues. *Educational Research Review*, 3, 187–217.

Stergiou, N., Georgoulakis, G., Margari, N., Aninos, D., Stamataki, M., Stergiou, E., et al. (2009). Using a web-based system for the continuous distance education in cytopathology. *International journal of medical informatics*, 827–838.

Suikki, R., Tromstedt, R., & Haapasal, H. (2006). Project management competence development framework in turbulent business environment. Technovation, 26, 723–738.

Súilleabháin, G. Ó. (2003). *Principles, Structure and Framework of e-Learning.* Retrieved January 15, 2011, from Learning Services, Ericsson: http://learning.ericsson.net/ socrates/doc/cork.docTallent-Runnels, M.-K., Lan, W. Y., Fryer, W., Thomas, J., Cooper, S., & Wang, K. (2005). The relationship between problems with technology and graduate students' evaluations of online teaching. *Internet and Higher Education*, *8*, 167–174.

Teekaput, P., & Waiwanijchakij, P. (2006). eLearning and Knowledge Management, Symptoms of a Reality. *Third International Conference on eLearning for Knowledge-Based Society*, (pp. 27-33). Bangkok, Thailand.

Turner, R. J., & Simister, S. (2004). *Gower Handbook of Project Management* (Romanian version ed.). Bucharest, Romania: Codecs Printing House .

Young, A., & Norgard, C. (2006). Assessing the quality of online courses from the students' perspective. *Internet and Higher Education*, *9*, 107–115.

In: Evaluation in E-learning
Editors: Y. Psaromiligkos, A. Spyridakos et al.
ISBN: 978-1- 61942-942-0

Chapter 3

ANALYSING STUDENTS' PREFERENCES IN E-LEARNING COURSES USING A MULTI-CRITERIA DECISION MAKING EVALUATION APPROACH

Y. Psaromiligkos[1], A. Spyridakos[1], S. Retalis[2,*], G. Demakos[3,†], E. Zafeiri[3] and C. Kytagias[1]

[1]Technological Educational Institute of Piraeus
Department of Mathematics, Aigaleo, Athens, Greece
[2]University of Piraeus, Department of Digital Systems, Piraeus, Greece
[3]University of Athens Department of Primary Education, Athens, Greece

ABSTRACT

Evaluation is a critical and a "must-do" step in every e-learning course since it provides valuable feedback in order to tailor, redesign, improve or accept the course according to the stakeholders' needs. Evaluation by its nature is a complex process and constitutes an ill-structured problem. A series of critical factors are involved such as: a considerable number of different stakeholders (students, tutors, e-learning designers, developers, and so on), the functionality and usability of the available learning technology, the pedagogical approach (learning theories, instructional design models, learning strategies), the delivery mode, constraints (time, budget, people). Moreover, the outcomes of e-learning are difficult to be measured due to time and quality borders. A major goal of an evaluation process is to measure the underlying set of quality factors as perceived by the target stakeholders (e.g. students). The analysis of stakeholders' preferences is an essential factor to the success of the e-learning course because it gives valuable feedback to the decision makers in order to properly redesign/improve the course.

[*] E-mail: retal@unipi.gr.
[†] E-mail: gdimakos@primedu.uoa.gr.

In this chapter, we describe in detail the results of a new methodological approach for the analysis of students' preferences in e-learning courses based on the principles of theory of committees and elections and multicriteria disaggregation-aggregation decision aid approaches. The underlying real case study concerns the analysis of the students' preferences of an undergraduate course at TEI Piraeus during the academic year 2009-2010. Multicriteria disaggregation – aggregation approach for discrete alternative actions aims to the assessment of an additive preference value model triggered by students' global preferences on a set of alternative actions, evaluated on a consistent family of criteria. Theory of committees and elections provide the way to structure acceptable global ranking from individual expressed preferences. The case study shows an in depth analysis of students' preference models along with an effective identification of the underlying changes in order to improve the whole course. Also, our new methodological approach seems very promising to be used in a wider and multipurpose evaluation structure (e.g. course, program, curricula).

1. INTRODUCTION

Evaluation is a critical and a "must-do" step in every e-learning course since it provides valuable feedback to tailor, redesign, improve or accept the course according to the stakeholders' needs. However, evaluation in e-learning is a multi-dimensional and ill-structured problem because it is an extremely complex process. The complexity is the outcome of some crucial features related to the process and the nature of this kind of evaluation problems. Some of them are described below.

An evaluation process may involve a considerable number of different stakeholders (for example training teachers in a new educational directive or curricula) belonging to target groups with different aspects and cognitive styles (students, tutors, e-learning designers, developers, and so on). It is worthwhile mentioning that these target groups in most of the cases have conflicting interests. This diversity of the target groups leads to different evaluation perspectives as well as to the differentiation of the way of preference on the criteria or characteristics used for the evaluation (Jung 2011; Bodea and Dascălu, 2010; Mitchell and Honore, 2007; Marom, 2005; Katz, 2002; Mumtaz, 2000; Glenn, 1997). Typically, the quality of e-learning is mainly defined from the provider's perspective. According to Jung (2010), the providers' perception in defining quality in e-learning is unequal and ignores the interrelational nature of a quality assurance.

The features offered by the available learning technology along with the underlying usability are critical factors. The e-learning market faces a rapid change over the last years (IDC Report, 2010) and a large list of tools - solutions (both commercial and open source) are available today (http://c4lpt.co.uk/Directory/index.html). Although the existence of a specific set of features offered by a tool is an essential criterion to choose/select the tool, the way these features are implemented or more specifically how the end users perceive the features is also critical. Usability addresses the quality of the functionality of a system in the context of use (Nielsen, 1993; Bevan, 1995). There is a lot of research concerning usability as well as methods and methodologies for conducting usability evaluations (ISO/IEC 9126, 2001; Stefani & Xenos, 2001; Sartzetaki et. al. 2003; Scapin D. & Law E., 2007). It is essential to mention that this is not a trivial task because usability concerns a qualitative characteristic of the underlying product and therefore we can't easily measure it. One of the most well-known usability evaluation methods is the heuristic evaluation introduced by Jacob Nielsen (Nielsen, 2000). A new promising approach is the exploitation of the knowledge captured in design patterns (Georgiakakis et. al. 2009) but till now the development of design patterns in the e-learning field is not matured enough.

The uppermost goal of e-learning is not the functionality or usability of the learning technology but the actual learning. The evaluation of the pedagogical approach is vital and constitutes the linchpin of the underlying context of the evaluation process. However, there are many learning theories as well as instructional design models which complicate the evaluation process. For the last decade objectivism and constructivism are the most predominant paradigms for e-learning, but the evaluation of the underlying pedagogical approach is not an easy task (see chapter "Qualitative Evaluation in e-learning" in this book, by Charalambos Vrasidas). Also, the actual results of an educational programme are difficult to be measured in all of its dimensions. In academic community we see a wide range of studies concerning the term "learning outcome". However, outcomes of e-learning are difficult to be measured due to time and quality borders. Moreover, the definition itself of what constitutes learning outcome is not well established and a range of alternative definitions exist (Prøitz, 2010).

By default e-learning utilizes more or less technology in the educational process. However, there are certain tasks where the delivery option through technology is not the best choice for various reasons (the kind of learning unit, the specific learning strategy, budget limits, etc.). How to design and consequently how to evaluate the delivery mode is not an easy task either. Usually, evaluators focus on the delivery mode and attempt to compare various delivery systems without paying attention to the perspectives of teachers and learners along with the underlying pedagogical approach. These attempts can justify the non-significant difference phenomenon between traditional and e-learning courses shown by numerous studies (Russell, 1999).

A major goal of an evaluation process is to measure the target quality factors as perceived by the selected stakeholders (e.g. students). There are many research efforts and related case studies trying to imprint general preferences and behaviour of stakeholders in on-line environments (Jung, 2011; Paechter and Maier, 2010; Naveh, Tubin and Pliskin, 2010; Williams et al., 2011; Ellis, Ginns and Piggott, 2009; Lee et al., 2009; Lee and Lee, 2008; Sun et al., 2008; Price, Richardson and Jelfs, 2007; Beyth-Marom et al., 2003; Smith, 2005; Liu, Lin and Wang, 2003; Mitchell and Honore, 2007; Westbrook, 2006; Paulus, 2007; Liaw et al., 2008; Tsai, 2008). However, we agree with Jung (2011) in "*...Many claims are made for the advantages of e-learning, but ironically, in a field that emphasizes the importance of learner-centeredness, there have been very few studies looking into the learners' opinions of quality factors.*"

The analysis of stakeholders' preferences is an essential factor to the success of an e-learning environment because it gives valuable feedback to the decision makers in order to properly tailor, redesign, and improve the environment. Despite the above research efforts in defining preferences and behaviour of various stakeholders in on-line environments there is a lack of a theoretical framework for an in depth analysis and assessment of such preferences. Such a framework should respond to decision makers' needs for adjustments at macro, meso and micro-level. The meso-level concerns a typical on-line course, while a macro concerns a series of courses and curricula. A micro-level concerns inner quality attributes of a course such as specific learning activities (e.g. collaborative), learning resources, or services of the on-line environment.

In this chapter we present the initial results of a case study that follows a new methodological approach for the evaluation and analysis of students' preferences in e-learning courses. Our methodology is based on the principles of theory of committees and elections and multicriteria disaggregation-aggregation decision aid approaches. The case study analyses the students' preferences of an undergraduate course at TEI of Piraeus during the academic year 2009-2010. The case study shows an in depth analysis of students' preference models along with an effective identification of the underlying changes in order to improve the whole course. Key points of our methodological approach are: a) the techniques of voting aggregation based on theory of committees and elections. These techniques aim to the assessment of a global ranking as close as possible to the individual rankings on a limited set of alternative prospective actions or decisions; b) the disaggregation – aggregation approaches of Multicriteria Decision Aid Analysis which aim to the construction of an additive value model. This approach leads to the construction of a multicriteria evaluation model based on the global preferences of the evaluators on a selected set of alternative actions and the evaluation of these actions on a consistent family of criteria.

Section 2 describes our new methodological evaluation framework. Section 3 describes in detail the application of our methodology in a real case study at TEI of Piraeus. Finally, conclusions and feature research directions are provided in the last section.

2. A New Methodological Approach

Before presenting our new methodology for the estimation of students' preference models, we give an overview of the underlying theoretical background. Our methodology combines Disaggregation – Aggregation approaches of Multi-criteria Decision Aid Analysis along with techniques of voting aggregation based on theory of committees and elections.

2.1. Multi-criteria Decision Aid Analysis

Disaggregation-Aggregation (D-A) approach of multicriteria decision aid analysis, aim to the assessment of an additive value system. UTA methods (Siskos, 1980; Jacquete-Lagrèze and Siskos, 1982; Siskos and Yannacopoulos 1985) of D-A multi-criteria decision aid family is oriented for ill-structured decision problems for discrete alternative actions under certainty. Actually, D-A approaches are focused on the discovery and picturing of evaluators' behavior and cognitive style into a multi-criteria additive value system. The process is activated by the expression of evaluators' global preferences on a limited set of alternative actions and the evaluation of the alternative actions on the criteria.

The additive value model assessed through interactive feedbacks is described in the following formulae (Keeney and Raifa, 1976):

$$U(g) = \sum_{i=1}^{n} p_i u_i(g_i)$$

$$u(g_{i*}) = 0.\ u(g_i^*) = 1, \text{ for } i=1, 2,..., n$$

$$\sum_{i=1}^{n} p_i = 1$$

$$p_i \geq 0, \text{ for } i=1, 2,..., n$$

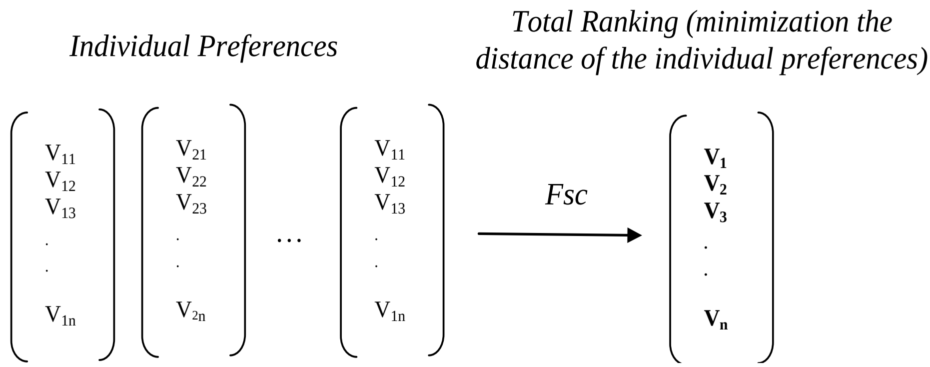
where:

- $g = (g_1, g_2, ..., g_n)$ is the evaluation vector of an alternative action on the n criteria,
- g_{i*} and g_i^* are the least and most preferable levels of the criterion g_i respectively and
- $u_i(g_i)$, p_i are the value function and the relative weight of the i-th criterion.

The assessment of this additive utility model is achieved through the utilization of linear programming techniques. Linear programming techniques are used in order to estimate the ingredients of the value model (weights of the criteria and criteria value functions) where the objective is the minimization of the divergence between evaluators' global preferences and the ranking produced by the additive value models. Also, a post optimality process takes place and the final estimated model is the mean outcome of the maximization of the importance of every criterion. It has to be mentioning that in traditional decision aid problems interactive feedback processes take place, which are employed in cases that the consistency of the assessed additive value model and evaluators' global preferences are considered not satisfactory. The consistency is measured by two indices: the F* function which is the absolute sum of the over(under)estimation errors and the Kendall' τ (varying from -1 to 1) (Kendall, 1970), which measures the degree that the assessed weak order is close to the evaluator ranking. These feedbacks are: i) the modification of the problem formulation; ii) the modification of criteria modelling (insertion or deletion of one or more criteria or modification of the criteria scales); iii) the modification of the selected reference set; iv) the modification of his/her judgment policy; v) the maximization of one or more marginal value functions (taking into consideration the corresponding solutions of the post optimality analysis) and; iv) Trade-off process where local modifications of the marginal value function are allowed in order to eliminate targeted inconsistencies. When the assessed preference model is considered as a satisfactory one, then it can be extrapolated into the whole set of alternatives, where every alternative action is ranked by the assessed value system.

2.2. Theory of Committees and Elections

In order to aggregate individual preferences (expressed in a weak order format) of a limited set of alternative actions to a total ranking as close as possible to the individual ones, we use methodological tools and techniques of theory of committees and elections (Black, 1958; Brams and Fisburn, 2002; Fisburn, 1970; Hwang and Lin, 1988). These techniques are used for collaborative decision making problems where there is a need to make an amalgam of the different individual opinions, attitudes or preferences into a Social Choice Function (SCF).



Let be:

- n, the number of alternative actions and m, the number of decision makers
- a_j, j Є {1,…,n} the set of alternative actions,
- V_{ij} the ranking of i-th DM on alternative action a_j (Ranking can also be expressed in a way a_k P a_l, a_k I a_l and a_l P a_k, k,l Є {1,…,n} and P: preferences, I: Indifference).

a) Borda Function (Fb) (Borda, 1781)

Borda is a simple and easy to use technique, while it is based on the summation of the ranking of every alternative action

$$Fb_i = \sum_{J=1}^{m} V_{ji}, \; i=1,\dots, n$$

The total ranking is estimated by sorting (ascending) the alternative action by Fb_i.

b) Condorcet Function (Condorcet, 1985) is described in the following formulas:

$Fc_{ij} = \#(i, a_j \; P \; a_i \text{ and } j \neq i)$ (Fc_{ij} represent the number of voter who ranked a_i in better position than a_j.) or $Fc_{ij} = \Sigma (p_j, a_j \; P \; a_i \text{ and } j \neq i)$ in case of weighted DMs.

$Fc(a_j) = \min(Fc_{ij}, j \neq i)$, representing the minimum number of DMs preferring a_i from all the other.

The total ranking is assessed by sorting $Fc(a_i)$ (descending).

c) Cook and Seiford (Cook and Seiford, 1978)

This technique aims at the estimation of a ranking that has the minimum distances from all individual rankings. The following process takes place:

- Calculation of the distances d_{ij} (absolute distance of Action a_i from position j).

$$d_{ij} = \sum_{i=1}^{n} |Vij - j|, \; j = 1,\dots, m$$

- Construction of the matrix D = [dij]
- Solve the assignment problem, so as to minimize the summation of distances Σdij

d) EigenVector Function (Hwang and Lin, 1988; Saaty, 1980)

This technique is based on the *estimation* of a vector of priorities and the following steps take place:

- Calculation of the indexes $d_{ij} = \frac{w_{ij}}{w_{ji}}$, where $w_{ij} = \#\left(i, a_i P a_j\right), i = 1,..,n$ and represents the number of DMs which $a_i P a_j$. The matrix D=[d_{ij}], i,j Є {1,…,n}, is reciprocal with positive elements.
- Calculation of the largest eigenvalue of the matrix D.
- Estimation of the eigenvector (W_1, W_2, .., W_n) of the largest eigenvalue.
- The eigenvector represents the priorities (importance) of the relative actions.
- Sorting alternative actions by its priorities values (Descending).

e) Dodgson (Hwang and Lin, 1988)

The final ranking is assessed based on an index representing the minimum preferences changes to be done on an alternative action in order to achieve majority (domination). The following steps take place:

- Calculation of the indexes $d_{ij} = \frac{w_{ij}}{w_{ji}}$, where $w_{ij} = \#\left(i, a_i P a_j\right), i = 1,..,n$, and represents the number of DMs which $a_i P a_j$.
- For every alt. action we calculate

$$FD_i = \sum_{j=1}^{n} \frac{w_{ji} - w_{ij}}{2}, \forall i \in \{1, n\} \wedge w_{ij} < w_{ji}$$

The alternative actions are ranked in an ascending form by the function FDi.

In order to check the consistency of the assessed total ranking the means of individual Kendall's τ (Kendall, 1970) can be used from the estimated total ranking and the corresponding standard deviation. Kendall's τ is an index used for the measurement of the correlation or similarity of two rankings. In this case we can calculate Kendall's τ for every individual ranking compared with the total estimated one. The means and the standard deviation of Kendall's τ show the level of consensus of total ranking with the individual ones.

2.3. A Muti-criteria Evaluation Methodology for E-learning Courses

Each e-learning course is composed of a series of learning modules (units of learning (Koper and Manderveld, 2004; IMS, 2003) that includes a number of learning activities (studying some content, doing a self assessment test, completing a project assignment) according to the underlying learning design. Each learning activity is related to specific learning objectives and may be implemented in a number of ways such as SCORM learning objects, multimedia documents, on-line quizzes, collaboration fora, etc. Students are engaged in the planned activities by following a learning path (or learning workflow) which may be more or less guided or adapted. The IMS-LD specification includes features (properties, conditions) for more sophisticated scenarios for learning activities and workflows to suit the needs and preferences of individual learners (IMS, 2003). How students perceive the underlying learning activities is crucial for their own success as well as the quality of the e-learning course (Jung, 2011; Dondi et al., 2006; Ehlers, 2004; Cashion and Palmieri, 2002). An in depth analyses of students' preferences can help decision makers to identify which alternatives to choose in order to increase the learning effectiveness of the course.

We propose the following steps for the estimation of students' preference models (see Figure 1). Every step shown in Figure 1 requires the completion of the previous one, while there is the capability to work interactively with feedbacks.

The first step concerns the criteria modeling (Roy and Bouyssou, 1993). It is a crucial step for every evaluation method and requires the identification of a consistent family of criteria. In this step, the evaluator should define the criteria that are suitable according to the aims and objectives of the evaluation. The criteria will be the outcome of the decisions that must be taken with respect to what will be evaluated (e.g. all or specific learning modules - units of learning of the course), why (e.g. which learning activities of the target learning modules need improvement in order to increase the learning effectiveness of the course), and for whom (e.g. responsible professor, educational institution policy maker). Multicriteria analysis identifies criteria as real monotone functions capable to be used as a means to measure the evaluators' preferences related to specific points of view. The set of criteria used for the evaluation ought to satisfy (except of the monotonicity) two additional conditions a) mutual independency and b) exhaustiveness in order to be considered consistent (Roy, 1985;1990).

At the second step the data concerning the evaluation of all alternatives on all criteria from all evaluators (students) are collected. This step can be realized using classical data collection techniques such as questionnaires. As in every evaluation study the evaluators should have well understood the target and the scope of the evaluation process in order to express their real opinion and preferences. In our case the students express their opinion on all criteria for all the target learning modules of the course (alternatives).

A crucial step for the effective implementation of our methodological approach is the ranking of alternatives. In this step the students should rank the alternatives (i.e. the target learning modules under evaluation) - according to their overall preference as it has been formulated by their learning experience during the course. However, in cases where there are a large number of alternatives this stage needs an adjustment. The third step serves this purpose and a reference set of alternatives is constructed and subsequently the evaluators rank by ordering the alternatives of the underlying reference set.

As described in the previous section (theoretical background) the ranking of the reference set will be used in order to estimate a representative ranking of the target learning modules as close as to the individual ones. Theory of election provides the tools to acquire a total ranking which is both representative of the students and simultaneously collective enough to the students' attitudes and preferences.

Our approach is supported by a tool that helps not only in the complex calculations but also in the visualization and interpretation of the results which is the scope of the following step. A linear programming problem is solved in order to estimate the marginal value functions and the weights of the criteria. Also, the final matrixes and graphs are constructed and presented which are:

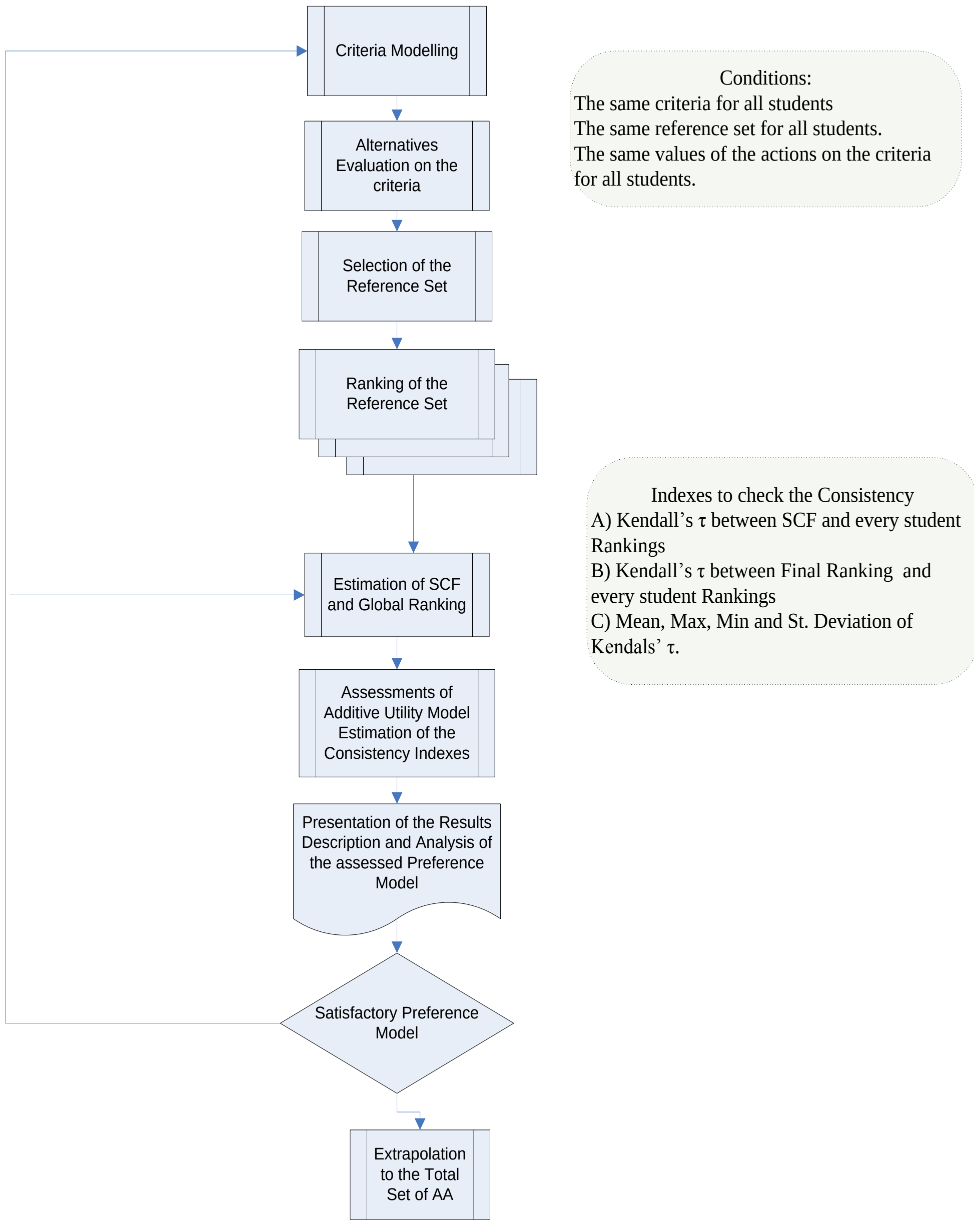


Figure 1. The process for the estimation of students' preference models.

a) The weights of the criteria
b) The marginal value functions
c) A table that includes: i) the target objects of the evaluation, ii) the initial ranking, iii) the global values of the target objects, iv) the assessed ranking and v) the over-under-estimation errors
d) The two indexes that show the degree of consistency of the assessed value model with the initial ranking (f* and Kendall's τ).
e) The ordinal regression curve.

If the consistency of the assessed preference value model is considered satisfactory then we can move to the next step and exploit the results. Otherwise, the evaluation process ought to be re-examined and revised. The revision of the evaluation data can lead to modification of the criteria modelling and subsequent adjustments on the alternatives evaluation on the criteria as well as modification of the ranking data set (e.g. exclusion of inconsistent rankings).

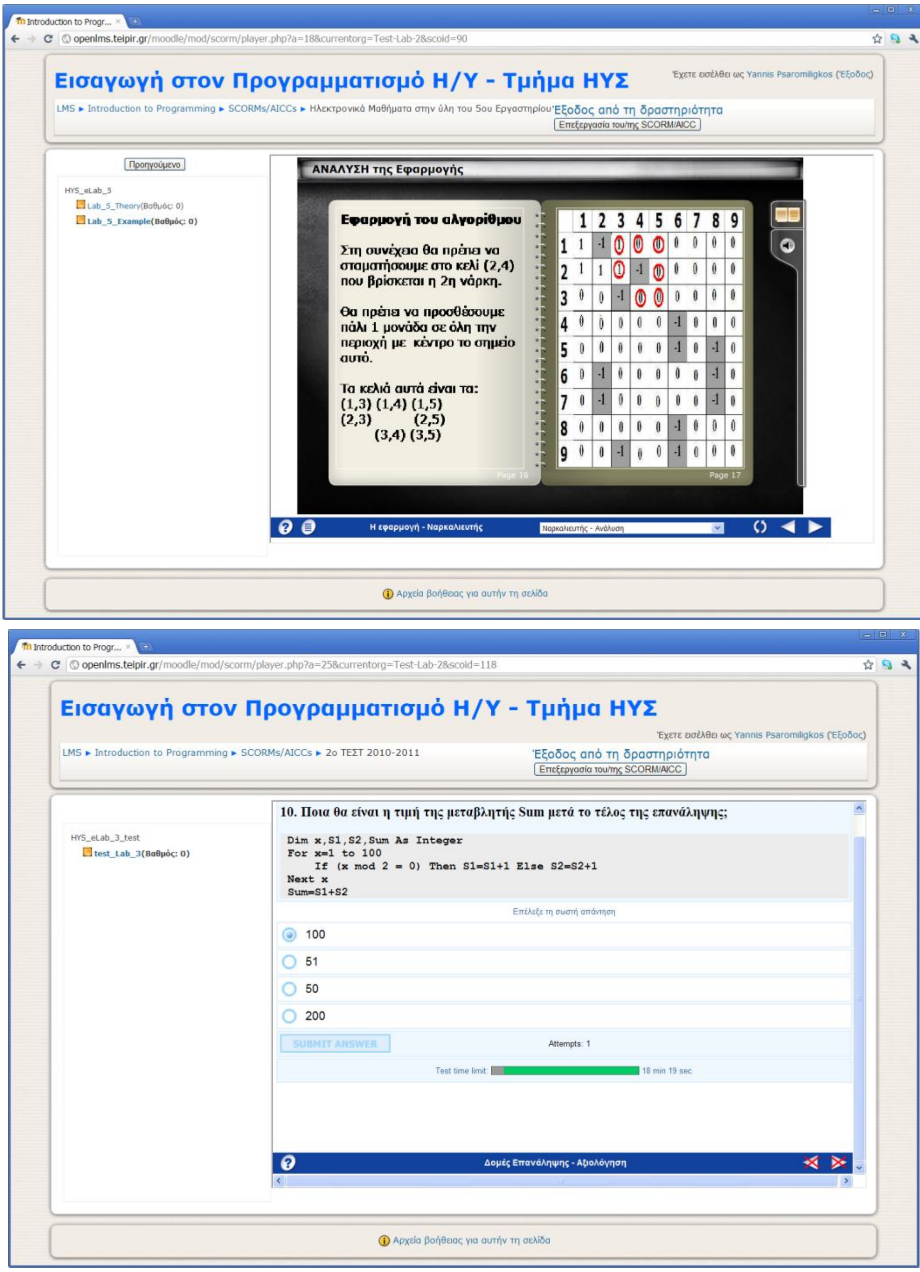
The extrapolation of the results gives a detailed picture of the preference model of the students. For example, the weights of the criteria are a crucial output since the Decision Maker can identify the most important factors of the students' preferences. Moreover, the marginal value functions can guide the Decision Maker to identify where (i.e. the specific learning modules and learning activities) and how (what properties need adjustment) to improve the course. We give a detailed analysis to all the outputs of our methodology in the next section.

3. Case Study

In this section we describe a real case study that follows the proposed methodological approach. The case study concerns an introductory undergraduate programming course at Technological Educational Institute (TEI) of Piraeus during the academic year 2009-2010. The course consists of ten learning modules that cover the whole content. Each learning module concerns a specific topic in programming and includes the following learning activities that should be carried out by the students:

- Study the on-line lessons. The on-line lessons included multimedia content in order to facilitate learning, delivered as SCORM learning objects and they were structured in two units: a theory unit and an example unit (solved exercises) (see Figure 2a).
- Take a self-assessment test. The self-assessment test was also implemented on-line and the first attempt was done at the laboratory in order to be under control. The next attempts could be done by students at their own place (see Figure 2b).
    - Accomplish a group-based project assignment. Each assignment should be delivered at a specific time before the beginning of the next module.
    - Collaborate with other students and tutors. For this purpose the students were supported by appropriate collaboration fora under the delivery platform for both student-student and student-tutor communication.
    - The course was delivered through the Moodle open source learning management system (http://moodle.org).

Our aim was to evaluate the ten learning modules according to the students' preferences in order to investigate where and how to improve the perceived learning effectiveness of the course. So, our primary target was the analysis of students' preferences over the ten learning modules of the course.



(a)
(b)

Figure 2. The structure of the course.

3.1. Subjects and Instruments

We developed an on-line questionnaire which was similar for each learning module of the course (See Appendix – Questionnaire of the Research). After the completion of all activities of each module we asked students to fill in the underlying questionnaire. Moreover, at the end of the course the students had to fill in an extra questionnaire which included a critical point according to our methodology: the ranking of the ten modules of the course according to the students' overall preference as it has been formulated by their learning experience during the course. Moreover, the questionnaire included open-ended questions that attempt to capture an overall judgement about the course. The total number of students which responded to the evaluation study reached 119 (26% women and 74% men) out of the 148 registered in the course which was an extremely high percentage. The profile of the students (previous experience in computers and especially in programming) was similar. All students have come from the technological direction according to the Greek education system for entering tertiary education.

3.2. Criteria Modeling

Following the proposed methodological approach the first critical step is the identification of a consistent set of criteria according to the aims and scope of our evaluation study. The following top six criteria were identified:

- Contribution of the SCORM learning objects to the acquisition of knowledge and skills in programming. This is a qualitative criterion that includes a number of (sub)criteria such as structure, aesthetics, transparency, forgiveness, matching between the metaphors and the learning experiences, informativeness, seamless of content and media, as well as the achievement of the desired learning experiences and learning outcomes (Tessmer, 1995). The answers were measured in a five-point Likert-type scale and they were coded as follows: 5=I absolutely agree, 4=I agree, 3=Neutral opinion, 2=Disagree, 1=I absolutely disagree.
- Contribution of the self-assessment tests to the acquisition of knowledge and skills in programming. A qualitative criterion that explores the suitability of each test according to the learning objectives of the module, the degree of difficulty, and its overall contribution in developing skills in programming. The answers were measured in a five-point Likert-type scale and they were coded like in previous criterion.
- Contribution of the project assignments to the acquisition of knowledge and skills in programming. This qualitative criterion explores similar aspects with the previous criterion, i.e. the suitability of each project assignment according to the learning objectives of the module, the degree of difficulty, if the students enjoyed the projects and they were motivated, and its overall contribution in developing skills in programming. The answers were measured in a five-point Likert-type scale similar to the previous criterion.
- Contribution of the collaboration facilities (forum) to the acquisition of knowledge and skills in programming. A qualitative criterion, too. It explores the contribution of each forum related to the learning objectives of each module in developing skills in programming. The answers were measured in a five-point Likert-type scale similar to the previous criteria.
- Usability of the delivery platform (learning management system). As we mentioned in previous section, the delivery platform was the open source learning management

system Moodle. In this qualitative criterion, we used Nielsen's heuristics in order to measure it (Nielsen, 2000). The answers were measured in a five-point Likert-type scale similar to the previous criteria (See Appendix – Questionnaire of the Research).

- Work load, i.e. the time needed to spend on the learning activities of each module. The answers were measured in a five-point Likert-type scale and they were coded as follows: 5=Extremely High, 4=High, 3=Moderate, 2=Low, 1=Extremely Low.

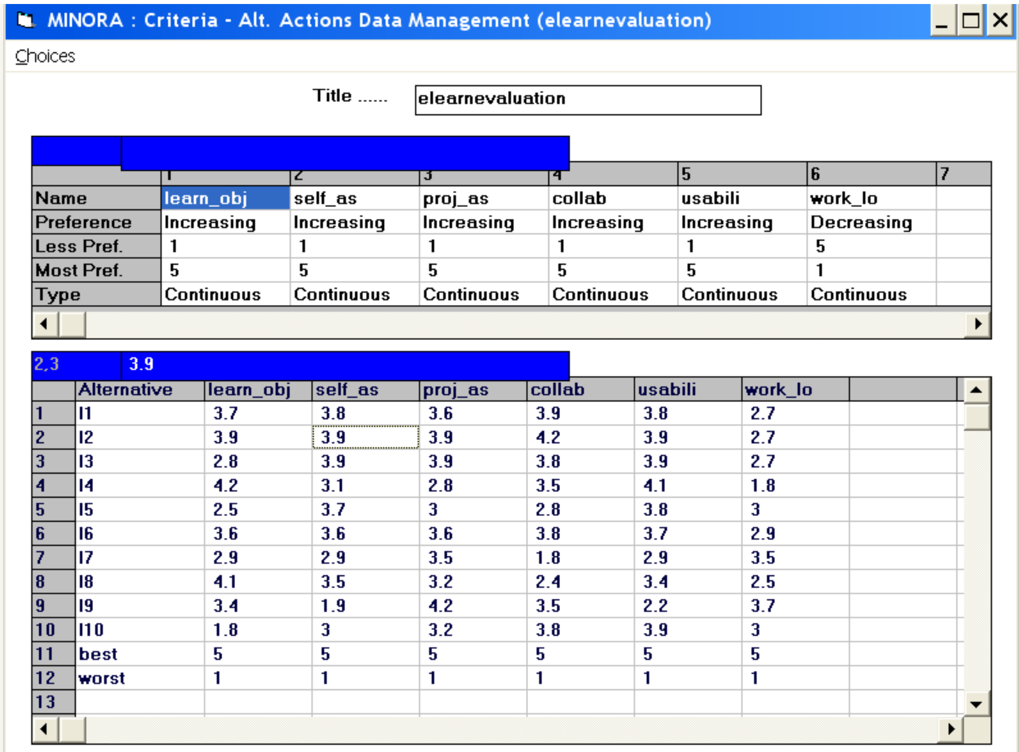
Most of the criteria were operationalized through a multiple set of items (sub-criteria). Each item focuses on a slightly different aspect of the main criterion. In building composite measurement scales, items included were first scrutinized for "face validity." After the data were collected, the validated items in each composite criterion were subjected to a Cronbach's Alpha reliability analysis for internal consistency of the instrument. In arriving at the final composite measurement indexes, every item that substantially lowered the Alpha coefficient was omitted and a new analysis was conducted to arrive at an index having the highest possible reliability measure (>75%). This is not the only test for the consistency check of our criteria. According to our methodology the criteria needs to be mutual independent and exhaustive. Moreover, the output of our multicriteria approach gives a detailed calculation of the over-under-estimation errors where we can identify inconsistencies that need arrangement.

3.3. Preference Model Estimation

After the completion of the ten similar questionnaires (one for each module of the course) on the criteria described in previous section we collected the raw data of the research. We aggregated the sub-criteria into the top six main criteria of our evaluation study and we calculated the final mean values. Figure 3 shows the related tables with the final values as reported by MINORA tool (Siskos et al, 1993).

The students were also asked to rank all the modules (alternatives) of the course. The course consists of ten modules (a small number of alternatives) and therefore there is no need to construct a reference set. The pre-ranking used was estimated utilizing Borda collective voting function (Figure 4). As shown in Figure 4 the pre-ranking of the ten learning modules accentuates the second learning module of the course (lesson 2 – l2 in figure) as the most preferred module by the students. This is also confirmed by the high scores that lesson 2 accomplished in all criteria. This module deals with the basic structures of the underlying programming language. The content analysis of the responses revealed that the students liked very much the examples included in this module as well as the project assignment. The less preferred module was the fifth learning module that deals with "User-defined Data Types". Students found pure the multimedia material and the corresponding examples while the project assignment was difficult.

The students' preference situation is shown at the next figure (Figure 5). The most significant factor of students' preferences constitutes the "Project Assignments" (proj_as) followed by the "SCORM learning Objects" (learn_obj). Students consider the contribution of project assignments to the acquisition of knowledge and skills as the most significant learning activity according to their learning experience. The study of on-line lessons (SCORM learning objects) constitutes the second significant learning activity. These two major findings are also explained by the content analysis of the responses provided by open-ended questions. The students were motivated and learned by doing their projects as well as by studying the on-line lessons. The last activity according to the comments of the students included enough examples (solved exercises) which they mostly liked.



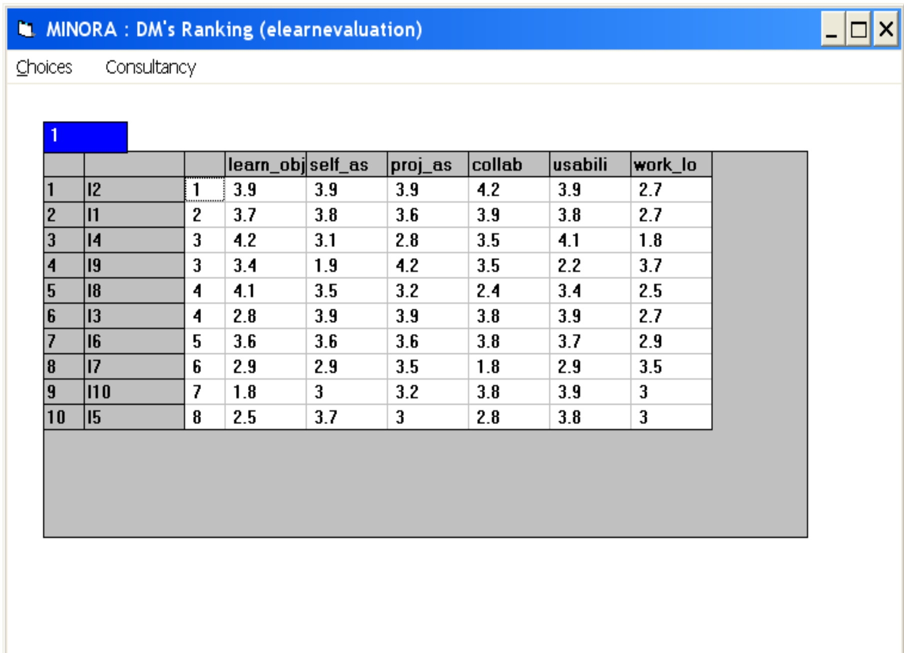
	1	2	3	4	5	6	7
Name	learn_obj	self_as	proj_as	collab	usabili	work_lo	
Preference	Increasing	Increasing	Increasing	Increasing	Increasing	Decreasing	
Less Pref.	1	1	1	1	1	5	
Most Pref.	5	5	5	5	5	1	
Type	Continuous	Continuous	Continuous	Continuous	Continuous	Continuous	

	Alternative	learn_obj	self_as	proj_as	collab	usabili	work_lo
1	I1	3.7	3.8	3.6	3.9	3.8	2.7
2	I2	3.9	3.9	3.9	4.2	3.9	2.7
3	I3	2.8	3.9	3.9	3.8	3.9	2.7
4	I4	4.2	3.1	2.8	3.5	4.1	1.8
5	I5	2.5	3.7	3	2.8	3.8	3
6	I6	3.6	3.6	3.6	3.8	3.7	2.9
7	I7	2.9	2.9	3.5	1.8	2.9	3.5
8	I8	4.1	3.5	3.2	2.4	3.4	2.5
9	I9	3.4	1.9	4.2	3.5	2.2	3.7
10	I10	1.8	3	3.2	3.8	3.9	3
11	best	5	5	5	5	5	5
12	worst	1	1	1	1	1	1
13							

Figure 3. The multicriteria table of the evaluation study.

Similar results have been reported in literature. Especially, the role of content in student satisfaction has been reported in literature as first significant factor (Naveh, Tubin and Pliskin, 2010; Malikowski, 2008; Sun et al., 2008). Also, the majority of students seem to prefer doing well-designed projects both as a means to improve their skills in the subject matter and as a method to evaluate their ability and knowledge (Liu, Lin and Wang, 2003). Moreover, they prefer novel context related to real life (Choi and Song, 1996).

Collaboration criterion (collab) is the third significant learning activity according to the results shown in previous figure. The content analysis of the responses provided by open-ended questions revealed that the students preferred the face to face collaboration despite having enough topics in the forum. Moreover, tutors had to provide more motivation to students in order to use the forum. Finally, self-assessment tests, usability of the delivery platform as well as the work load of the learning activities have all the same weight and do not constitute critical factors for the perceived learning effectiveness of the course. As far as the delivery platform is concerned there were no usability problems because it was user friendly and the students quickly became accustomed to working with it. Similar results have been reported by Tsai (Tsai, 2008). In general, the students considered the self-assessment tests as very easy to accomplish, while the time needed to spend on the learning activities of each module was plenty.



MINORA : DM's Ranking (elearnevaluation)

Choices Consultancy

			learn_obj	self_as	proj_as	collab	usabili	work_lo
1	l2	1	3.9	3.9	3.9	4.2	3.9	2.7
2	l1	2	3.7	3.8	3.6	3.9	3.8	2.7
3	l4	3	4.2	3.1	2.8	3.5	4.1	1.8
4	l9	3	3.4	1.9	4.2	3.5	2.2	3.7
5	l8	4	4.1	3.5	3.2	2.4	3.4	2.5
6	l3	4	2.8	3.9	3.9	3.8	3.9	2.7
7	l6	5	3.6	3.6	3.6	3.8	3.7	2.9
8	l7	6	2.9	2.9	3.5	1.8	2.9	3.5
9	l10	7	1.8	3	3.2	3.8	3.9	3
10	l5	8	2.5	3.7	3	2.8	3.8	3

Figure 4. Pre-Ranking of the ten modules/on-line lessons (l1, l2, …, l10).

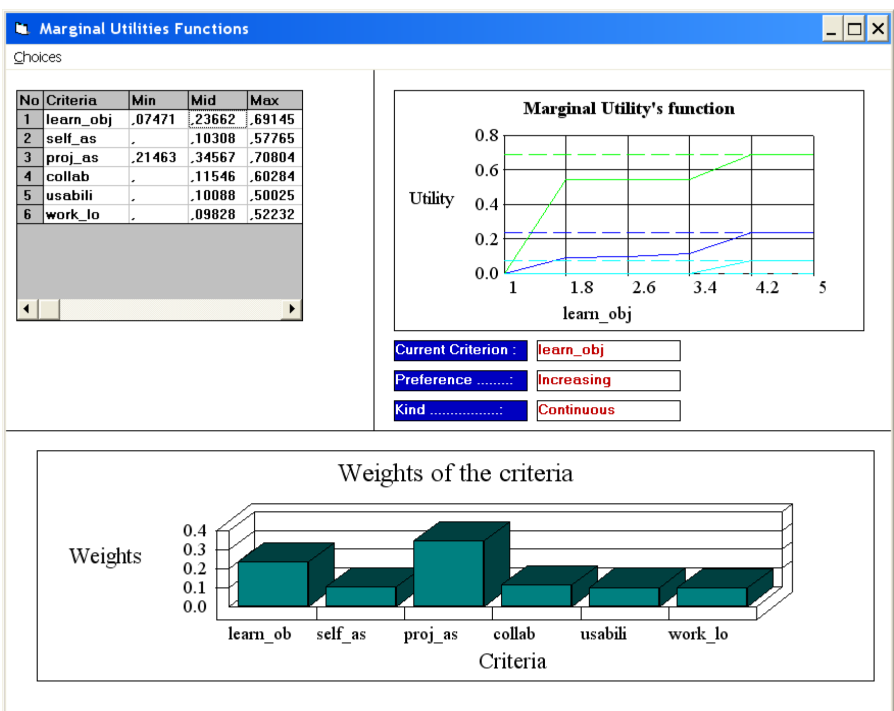


Figure 5. The preference situation of the students.

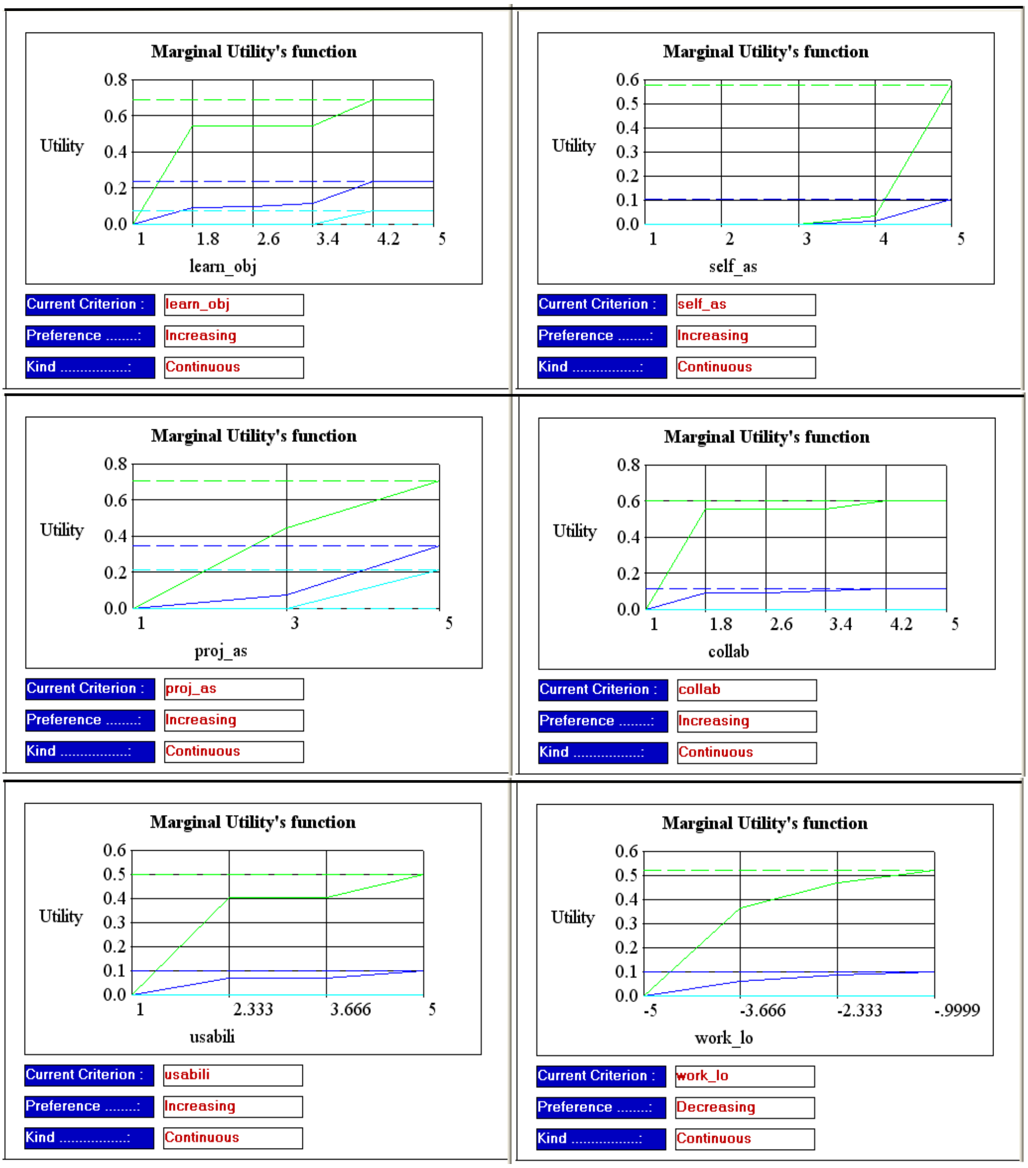


Figure 6. The Marginal Utility Functions of the students' preferences.

The Marginal Utility Functions of each criterion are shown at the following figure (Figure 6). These piece linear value functions capture the variation of students' preference in the climaxes of the criteria. Also, in the graph the minimum and maximum marginal utility functions resulted by the post optimality analysis picturing the stability level of the assessed preference model are presented.

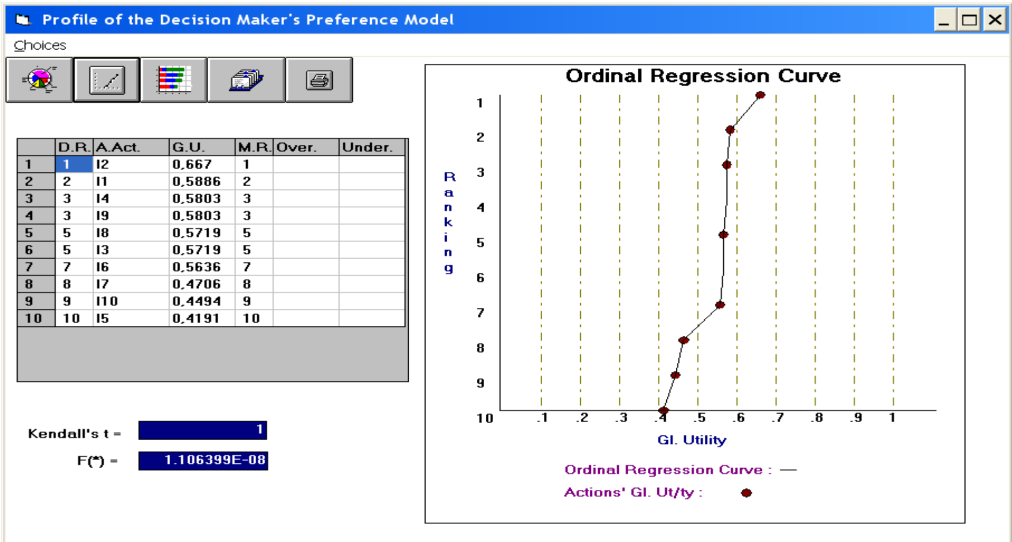


Figure 7. The total ranking of the course modules.

Figure 7 shows the total ranking of the course modules. According to the results the assessed ranking of the ten learning modules is totally consistent with the pre-ranking shown in Figure 4 and it is endured by the values of two indexes F(*) and Kendall's τ used for the consistency measurement. Also, it is obvious from the ordinal regression curve, where x-axis corresponds to the rankings and y-axis to global values, the crooked line (ordinal regression curve) represents the position of the rankings in the plane and the points the position of the alternatives. The corresponding points (alternative actions) are posed on the ordinal regressions curve in the points of the corresponding ranking without any over or under-estimation errors.

By analysing how each general utility function is constructed we can take valuable feedback related to where and how to improve the learning modules. For example, Figure 8 shows the distribution of the utility function concerning on-line lesson 5 (l5). The underlying learning module has come last in the total ranking. According to the professor of the course this is a critical learning module in the programming course and he wants to increase its perceived learning effectiveness. By looking at the pie chart graph, which depicts the marginal utilities, we can easily identify that the best choice for the professor is to improve the learning activity "Project Assignments". This activity-criterion covers a small part of its utility which explains the underlying position in the total ranking. We can have similar feedback by analysing the general utility functions of the other learning modules (see for example lesson 5 in Figure 9).

CONCLUSION

To our knowledge there is no equivalent approach in the area of evaluation of e-learning courses. Multicriteria disaggregation – aggregation approach for discrete alternative actions aims to the assessment of an additive preference value model triggered by students' global preferences on a set of alternative actions, evaluated on a consistent family of criteria. Theory of committees and elections provide the way to structure acceptable global ranking from individual expressed preferences.

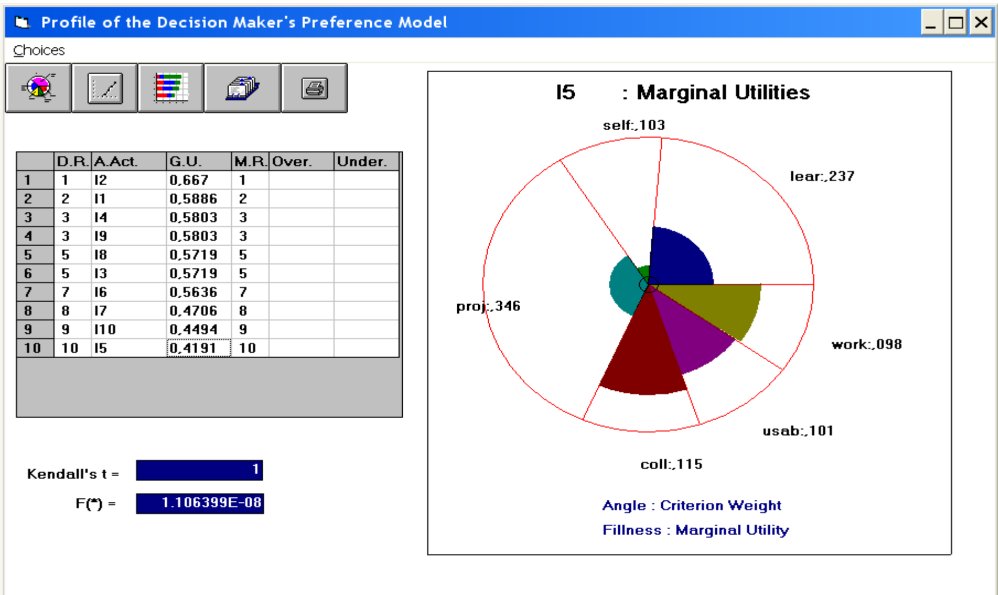


Figure 8. The Marginal Utility Functions of students' preferences for lesson 5.

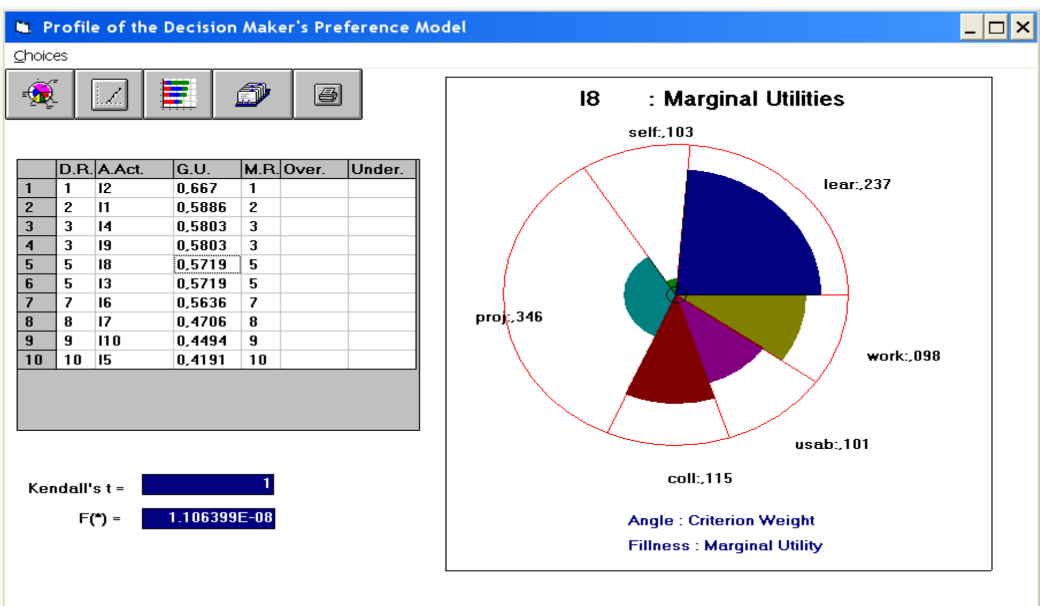


Figure 9. The Marginal Utility Functions of students' preferences for lesson 8.

Evaluation is a process that primarily utilizes information for decision making. In such a context evaluation should focus on information needed by the underlying stakeholders. The final beneficiaries of an e-learning course are the actual students. How the students perceive the course is essential to its success and this also explains why most of the evaluation studies focus on them (Jung, 2011; Naveh, Tubin and Pliskin, 2010; Paechter and Maier, 2010; Harrington and Loffredo, 2010; Lee et al., 2009; Sun et al., 2008; Ehlers, 2004). Our methodology can be used to give a detailed analysis of students' preference models at multiple levels. At the course level (meso-level), as shown in this evaluation study, we can have a detailed analysis of students' preferences concerning all the modules of the course according to the underlying learning activities. At the macro level we could have a detailed analysis of multiple courses or a whole curriculum according to specific services offered by an educational institution. Such a context could be valuable for managers or policy makers. Moreover, our methodology can give valuable feedback at a micro level. For example, a teacher can have a detailed analysis of students' preferences in collaborative learning strategies according to the quality of the achieved individual and group learning outcomes (see chapter "Measuring Students' Performance in e-Learning Environments via Enriched Assessment Rubrics" in this book, by Petropoulou, Retalis and Lazakidou).

The methodological approach proposed in this research work has several advantages:

- it provides a step by step procedure to assess a value model for the evaluation,
- it can handle qualitative data by the exploitation of the marginal value functions of the criteria,
- it can be easily understood and handled by decision or policy makers which are keen on handling and understanding value models, and
- it is based on well structured and time honoured methodological approaches.

Disaggregation–Aggregation approach provides a way to encapsulate and explain stakeholders' preferences. Through a consistent family of criteria the complexity of e-learning can be handled efficiently. The added value is not limited on a ranking assessment. The assessment of beneficiaries' preference model can lead to a wider and multipurpose evaluation structure. Moreover, our approach is supported by a tool to obtain sectors for interventions in order to face weaknesses and to exploit advantages.

Some of our feature research directions are the comparative analysis of preference models of different stakeholders (e.g. preference models of students' versus tutors') and the exploitation of e-learning analytics and educational data mining (Romero et al., 2011; Psaromiligkos et al., 2009) to automatically estimate the preference models of students. This direction could be valuable specifically in adapted learning environments.

APPENDIX - QUESTIONNAIRE OF THE RESEARCH

1. *Contribution of the on-line lesson (SCORM learning object) to the acquisition of knowledge and skills*

The aims and objectives of the on-line lesson were clear.
The on-line lesson explained things clearly.
No effort spent in memorisation of topics of the on-line lesson.
It was easy to find out what was expected in the on-line lesson.
The on-line lesson included examples appropriate to the learning objectives.
The on-line lesson was presented in an attractive way.

The structure of the on-line lesson made the study and assimilation of the subject matter easy.
The on-line lesson could be easily read.
The structure of the on-line lessons was seamless
The way of browsing the on-line lesson was clear and easy to use.
The on-line lesson included enough examples.
The on-line lesson was interactive (not only through browsing)
Invalid selections during browsing of the on-line lesson could be easily corrected.
The navigation schema of the on-line lesson was very good
I developed skills in programming thanks to the on-line lesson.

2. Contribution of the self-assessment test to the acquisition of knowledge and skills

The aims and objectives of the self-assessment test were clear.
The questions were appropriate to the scope of the test.
It was easy to find out what was expected in the questions of the test.
The self-assessment test was difficult to complete.
I developed skills in programming thanks to the self-assessment tests.

3. Contribution of the project assignment to the acquisition of knowledge and skills

The aims and objectives of the project assignment were clear.
The project assignment was appropriate to the objectives of the module.
The project assignment was interesting.
I was motivated by the project assignment to deal with programming.
It was easy to find out what was expected in the project assignment.
The project assignment was difficult to complete.
I developed skills in programming thanks to the project assignment.

4. Contribution of the forum to the acquisition of knowledge and skills

The aims and objectives of the forum were clear.
The forum included enough topics.
I was motivated by the forum to communicate with tutors and students.
I was motivated by the forum to deal with programming.
The forum was difficult to attend.
I developed skills in programming thanks to the forum.

5. Usability of the delivery platform (learning management system)

Are you satisfied with the feedback provided by the system during its usage?
The system always informs me regarding my actions
The system provides information in a natural and logical order
The content of messages and other directives on the screen have explicit meaning.
The sequence of screens is expected.
It is easy to cancel an operation.
There is possibility for transition into other units of the system in an easy and obvious way.
The system always warns me before a critical action.
It is possible to use the system without any other help.
The quantity of information presented on the screen is sufficient.

Consistency/uniformity of user interface exists in all units.
The fault messages always clarify the problem.
It is easy to find help on how to do a specific task.
In general I am satisfied after the system's usage
The system requires often-unjustifiable huge effort on behalf of the user

6. The time needed to spend on the learning activities of the module

REFERENCES

Bevan, N. (1995). Measuring usability as quality of use. *Journal of Software Quality*, 4, 115□130.

Beyth-Marom Ruth, Saporta Kelly, and Caspi Avner. (2005). Synchronous vs. Asynchronous Tutorials: Factors Affecting Students' Preferences and Choices, *Journal of Research on Technology in Education*, Spring 2005: Volume 37 Number 3, pp.241-258.

Black, D. (1958). The theory of Committees and Elections, *Cambridge University Press*, Cambridge.

Borda, J.C. (1781). Memoire sur les elections au scrutin. Histoire de L'Academie Royale des Sciences.

Brams, S.J and P.C. Fisburn. (2002). Voting Procedures, in *Handbook of Social Choice and Welfare*, Volume 1, KJ Arrow, A.K. Sen and K. Suzumura((Rds), Elsevier Science B. V.

Carlson, P.A. (1998). Advanced Educational Technologies – Promise and Puzzlement, *Journal of Universal Computer Science*, 4(3).

Cashion, J., & Palmieri, P. (2002). The secret is the teacher: The learners' view of online learning. Leabrook, Australia: National Center for Vocational Education Research. Retrieved January21,2010,from http://www.ncver.edu.au/popups/limit-Download.php?file=research/proj/nr0F03a.pdf.

Choi Jung-Suk and Song Jinwoong. (1996). Students' Preferences for Different Contexts for Learning Science, *Research in Science Education*, 1996, 26(3), 341-352.

Condorcet, M.J. (1785). Essai sur l'application de l'analyses a la probabilite des decision rendue a la pluralite des voix. Paris.

Cook W.D. and L.M Seiford. (1978). Priorities ranking and consensus formation, *Management Science*, vol. 20, No. 16, pp. 59-73.

Crawford, C. (2004). Non-linear instructional design model: Eternal, synergistic design and development. *British Journal of Educational Technology*, 35, 413–420.

Dondi, C., Moretti, M., & Nascimbeni, F. (2006). Quality of e-learning: Negotiating a strategy, implementing a policy. In U. D. Ehlers & J. M. Pawlowski (Eds.), *Handbook on quality and standardization in e-learning* (pp. 31–50). Berlin: Springer.

Ehlers, U. (2004). Quality in e-learning from a learner's perspective. *European Journal of Open and Distance Learning*, I. Retrieved January 14,2010,from http://www.eurodl.org/materials/contrib/2004/Online_Master_COPs.html.

Ellis, R. A., Ginns, P., & Piggott, L. (2009). E-learning in higher education: Some key aspects and their relationship to approaches to study. *Higher Education Research & Development*, 28, 303−318.

Fisburn P.C. (1972). A comparative analysis of group decision methods, *Behavioral Science*, vol 16, no 6, pp538-236.

Georgiakakis P., Retalis S. and Psaromiligkos Y., (2009). “Design patterns for inspection-based usability evaluation of e-learning systems”, Technology-enhanced learning: design patterns and pattern languages, *Sense Publishers*, ISBN 978-94-6091-061-9.

Glenn A. Brand. (1997). Training Teachers for Using Technology, *Journal of Staff Development*, Winter 1997, Vol. 19, No. 1.

Harrington, R., Loffredo A. D., (2010). MBTI personality type and other factors that relate to preference for online versus face-to-face instruction, *Internet and Higher Education*, 13 (2010) 292–297, Elsevier Ltd.

Hwang, C-l. and M-J. Lin (1988). Group Decision Making under Multiple Criteria, in *Lectures Notes in Economics and Mathematical Systems*, Springer-Verlag,, Berlin, Heidelberg.

IDC Report 2010. (2010). Worldwide and U.S. Corporate eLearning 2010-2014 Forecast: Changing Patterns of Consumption, August 2010.

IMS Global Learning Consortium (2003). IMS Learning Design Information Model. Version 1.0 Final Specification, retrieved January 14, 2010, from http://www.imsglobal. org/learningdesign/ldv1p0/imsld_infov1p0.html.

ISO/IEC 9126 (2001). Quality Characteristics and Guidelines for the User, International Organization for Standardization, Geneva.

Jacquet-Lagrèze, E. and Y. Siskos. (1982). Assessing a set of additive utility functions for multicriteria decision making, *European Journal of Operational Research*, Vol. 10, No.2, pp. 151-164.

Jung Insung. (2011). The dimensions of e-learning quality: from the learner’s perspective, *Educational Technology Research & Development*, 59:445–464.

Katz J. Y. (2002). Attitudes affecting college students' preferences for distance learning, *Journal of Computer Assisted Learning*, Volume 18, Issue 1, pp.2-9.

Keeney, R.L. and H. Raifa. (1976). Decision with multiple objectives: Preferences and value tradeoffs, *J. Wiley*, NY.

Kendall, M.G. (1970). Rank correlation methods, Charles Griffin and Company Limited, London.

Koper R. and Manderveld J. (2004). Educational modelling language: modelling reusable, interoperable, rich and personalised units of learning. *British Journal of Educational Technology*, Vol. 35, No 5, 537-551.

Lee, B. C., Yoon, J. O., & Lee, I. (2009). Learners' acceptance of e-learning in South Korea: Theories and results. *Computers & Education*, 53, 1320−1329.

Lee, J. K., & Lee, W. K. (2008). The relationship of e-learner's self-regulatory efficacy and perception of e-learning environmental quality. *Computers in Human Behavior*, 24, 32−47.

Liaw, S. S., Chen, G. D., & Huang, H. M. (2008). Users' attitudes toward web-based collaborative learning systems for knowledge management. *Computers & Education*, 50, 950−961.

Liu Y., Lin F., Wang X. (2003). Education practice and analysing behaviour of students in a Web-based learning environment: an exploratory study from China, *Online Information Review*, Vol. 27 Iss: 2, pp.110 – 119.

Malikowski, S. R. (2008). Factors related to breadth of use in course management systems *The Internet and Higher Education*, 11(2), 81−86.

Mitchell A., Honore S. (2007). Criteria for successful blended learning, *Industrial and Commercial Training*, Vol. 39 Iss: 3, pp.143 – 149.

Molenda, M. (2003). In search of the elusive ADDIE model. *Performance Improvement*, 42(5), 34.

Moore, M., & Kearsley, G. (2005). Distance education: A systems view (2nd edition) *Belmont, CA: Thomson Wadsworth*.

Mumtaz S. (2000). Factors affecting teachers' use of information and communications technology: a review of the literature, *Journal of Information Techology for Teacher Education*, 9:3, 319-342

Nielsen, J. (1993). Usability Engineering. *Academic Press*, London.

Nielsen, J. (2000). Designing Web Usability: The Practice of Simplicity, *New Riders Publishing*, Indianapolis.

Paechter Manuela, Maier Brigitte. (2010). Online or face-to-face? Students' experiences and preferences in e-learning, *The Internet and Higher Education*, Volume: 13, Issue: 4, Publisher: Elsevier Inc., Pages: 292-297.

Paulus, T. M. (2007). CMC modes for learning tasks at a distance. *Journal of Computer-Mediated Communication*, 12, 1322−1345.

Price, L., Richardson, J. T. E., & Jelfs, A. (2007). Face-to-face versus online tutoring support in distance education. *Studies in Higher Education*, 32, 1−20.

Prøitz S. Tine, (2010). Learning outcomes: What are they? Who defines them? When and where are they defined?. *Educational Assessment, Evaluation and Accountability*, Volume 22, Number 2, 119-137.

Roy, B. and D. Bouyssou. (1993). Aide multicritère à la décision: M⊖thodes et cas, Economica, Paris.

Roy, B. (1985). M⊖thodologie multicritere d'Aide a la Decision, Economica, Paris.

Roy, B. (1990). The outranking approach and the foundations of ELECTRE methods, in Reading in Multiple Criteria Decision Aid, Bana e Costa (Ed.), Springer, Berlin, pp. 155-183.

Russell L. Thomas. (1999). No Significant Difference Phenomenon, North Carolina State University, Raleigh, NC, USA, ISBN 0-9668936-0-3.

Saaty, T. L. (1980). The Analytic Hierarchy Process, New York, McGraw-Hill, 1980.

Sartzetaki, M., Psaromiligkos, Y., Retalis, S. & Avgeriou, P. (2003). Usability evaluation of e-commerce sites based on design patterns and heuristic criteria, 10th International Conference on Human–Computer Interaction, Hraklion Crete, June 22–27.

Scapin D. & Law E. (Eds.) (2007). Review, Report and Refine Usability Evaluation Methods (R3 UEMs), Proceedings of the 3rd COST294-MAUSE International Workshop, Athens, March 5, 2007.

Siskos Y., and Yannacopoulos D. (1985). UTASTAR, an ordinal regression method for building additive value functions, *Investigacao Operacional* 5, no 1, pp 39-53.

Siskos, Y. (1980). Comment modéliser les préférences au moyen de fonctions d'utilité additives, *RAIRO Recherche Opérationnelle*, 14, pp. 53-82.

Siskos, Y., A. Spyridakos and D. Yannacopoulos. (1993). "MINORA: A multicriteria decision aiding system for discrete alternatives", *Journal of Information Science and*

Technology, in: Y. Siskos and C. Zopounidis (eds), *Special Issue on Multicriteria Decision Support Systems*, vol. 2, no 2, pp. 136-149.

Stefani, A., & Xenos, M. (2001). A model for assessing the quality of e-commerce systems, PC-HCI 2001 Conference.

Sun, P., Tsai, J. R., Finger, G., Chen, Y., & Yeh, D. (2008). What drives a successful e-Learning? An empirical investigation of the critical factors influencing learner satisfaction. *Computers & Education*, 50(4), 1183−1202.

Tessmer, S. (1995). Formative multimedia evaluation. *Training Research Journal*, 1, 127-149.

Tsai Chin-Chung, (2008), The preferences toward constructivist Internet-based learning environments among university students in Taiwan, *Computers in Human Behavior* 24 (2008) 16–31.

Westbrook, V. (2006). The virtual learning future. *Teaching in Higher Education*, 11, 471−482.

Williams D. D., South B. J., Yanchar C. S., Wilson G. B., Allen S. (2011). How do instructional designers evaluate? A qualitative study of evaluation in practice, *Educational Technology Research & Development*, Springer, DOI 10.1007/s11423-011-9211-8.

In: Evaluation in E-learning
Editors: Y. Psaromiligkos, A. Spyridakos et al.
ISBN: 978-1- 61942-942-0

Chapter 4

MEASURING STUDENTS' PERFORMANCE IN E-LEARNING ENVIRONMENTS VIA ENRICHED ASSESSMENT RUBRICS

Ourania Petropoulou, Symeon Retalis and Georgia Lazakidou*
University of Piraeus, Department of Digital Systems, Piraeus, Greece

ABSTRACT

This chapter presents a new technique for assessing students' performance in e-Learning environments called "Enriched Assessment Rubric (EAR)". The proposed technique is designed to assist educators to assess/measure students' performance using both quantitatively and qualitatively criteria and taking into consideration the learning products (i.e. tests, individual or group deliverables produced by the students, etc.) as well as the wide range of interaction (between students, between students and educators and between students and resources) performed during all phases of complex interactive scripts, and to report students' final grades in a precise, detailed and accurate way. This technique is called enriched because it is based on the classic and well accepted assessment rubrics, which are enriched with criteria that are related to interaction analysis indicators. In this chapter, we highlight the philosophy and the structural components of the "EAR" technique as well as its added value. A pilot study of the application of the proposed technique in an authentic educational environment is fully described that shows how its effectiveness can be evaluated.

Keywords: E-learning Environments, Students' Assessment Performance, Enriched Assessment Rubric (EAR), Interaction Analysis Indicators

* E-mail: rpetro@biomed.ntua.gr, retal@unipi.gr, glazakidou@gmail.com.

1. INTRODUCTION

The main goal of education in today's Knowledge Society is not only the transferring of useful and necessary knowledge in each subject, but rather the development and cultivation of strong cognitive, metacognitive, social and communication skills (Campbell & Gibson 2008). These skills will enable each student to become an independent thinker and active citizen of the 21st century. Educators are trying to enhance the traditional learning paradigm with the use of advanced networked technologies. Thus, e-learning environments are being created to offer students stimulating learning experiences. Students are engaged in learning activities which require individual and group tasks, co-construction of knowledge through collaboration and social networking, communication (synchronous or asynchronous), self-assessment and use of various learning resources, etc. (Lazakidou & Retalis 2010).

Students' assessment in such interactive e-learning environments is a quite difficult and complex issue (Shepard 2000; Quellmaz & Kozma 2003; Petropoulou et al., 2007; Johnson et al., 2009). As Roberts (2006) stated "from the instructor's point of view, in an e-learning environment, without the advantages that are naturally provided by face-to-face contact and body language, it can be more difficult to gain an accurate assessment of the learning achieved". Educators face many difficulties to capture, track and assess the multiple and various individual and collaborative learning activities performed by all students (Daradoumis et. al., 2006). Current research studies in the field of students' assessment in eLearning environments show that there is a need for new techniques to assess learners' performance, capable of effectively supporting the educators to holistically assess both products of learning and the very complex process of learning (grid interactions between students, between students and educators and between students and learning resources) that differs from the traditional way (Chan & Van Aalst 2004; Morgan & O'Reilly 2006; Zinn and Scheuer 2006; Joosten-ten Brinke et al., 2007; Roberts & McInnerney 2007).

The aim of the chapter is to present a new usable technique of assessing the students' performance in e-learning environments, called Enriched Assessment Rubric (EAR) with Interaction Analysis Indicators, which meets the following criteria: a) it assesses (grades) in a holistic and structured way, both the products of the learning process (tests, individual, group deliverables produced by the students) and the complex collaborative learning process, b) it is based on the classic assessment rubrics techniques enriched by criteria that are related to the interaction analysis indicators, c) it is supported by an innovative interaction analysis tool, d) it can be applied by educators in authentic e-learning environments with effectiveness, while guaranteeing the educators' subjective satisfaction (Petropoulou et al., 2009).

The structure of this chapter is the following. Firstly we present the philosophy, the structural components and the added value of the "EAR" technique. Then we describe, through a pilot research study, the degree of acceptance and usability of the "EAR" by educators in primary and secondary education. Finally, we draw conclusions about the proposed technique and we discuss the plans for future research.

2. THE EDUCATIONAL CONCEPT OF ENRICHED ASSESSMENT RUBRICS

According to modern pedagogical theories, students' assessment is widely recognized as a fundamental part of the instructional process. Inevitably, the dynamic penetration and utilization of e-learning environments in the educational practice marked the beginning of a new era in the field of the assessment of students' performance (Rovai 2000; Stiggins 2004; Underhill 2006; Mödritscher et al., 2006; Darling-Hammond & Adamson 2010). Leaving behind the psychometric assessment models of the previous century, i.e. the "sterile" measurement of cognitive objectives achievement through tests and quizzes, the students' assessment is now defined as a systemic process aimed at measuring the individual and group activities, the level and the nature of collaborative knowledge construction and the broad spectrum of cognitive, metacognitive and social skills, developed during the learning process with as much objectivity, accuracy and completeness as possible (Spada et al., 2005; Pozzi et al., 2009; Partnership for 21st Century Skills 2009; Arvaja et al., 2011).

In daily practice in e-learning environments, educators use various techniques in order to assess students' individual and group performance, such as test/quiz self-assessment, peer-assessment, e-portfolio, observation, assessment rubrics, etc. The last decades rubrics assessment have become very popular, a recognized trend in all levels of education (Shepherd & Mullane 2008; Andrade & Valtcheva 2008). A rubric is an easily applicable qualitative scoring guide that seeks to assess students' performance at learning tasks based on the sum of a full range of criteria rather than a single numerical score. A rubric has the form of the table containing the learning criteria in the vertical axon and the gradations of the qualitative levels of performance (scale), not the quantitative levels, in the horizontal one (see Figure 1).

Educators use rubrics as an effective dynamic assessment tool to clarify learning goals in an explicit and comprehensible way, to communicate the learning goals to their students, to provide students with instructive feedback, to judge final products (i.e. individual, group deliverables produced by the students, etc.) in terms of the degree to which the goals have been achieved, etc. (Hafner & Hafner 2003; Lantz 2004; Buzetto-More & Alade 2006; Blommel & Abate 2007; Shepherd & Mullane 2008; Allen & Knight 2009; Andrade & Valtcheva 2008).

Although the potential benefits arising from the widespread use of the rubrics, are several for both the educators and students, the major disadvantage is that they focus on the unilateral assessment of the learning products (individual and team deliverables produced by students). Thus, in their current form they cannot help educator assess the learning process in detail, e.g. the level and nature of collaboration, the correlations and interactions that occur during the teaching process, etc.

More recently, the international scientific community in the field of e-learning has proposed the interaction analysis as a modern assessment technique of complex computer supported collaborative learning activities (Daradoumis et al., 2006; Bravo et al., 2008; Persico et al., 2009). The interaction analysis process consists of: a) collecting, extracting and filtering all the data (about students' learning interactions) retrieved by the logfiles of the e-learning platforms system, b) analyzing these data according to a set of indicators (Interaction Analysis Indicators) and c) producing reports in the form of simple statistical tables, or visualizations such as bar charts, pie charts, etc. The interaction analysis indicators are measurable metrics reflecting in quantitative terms the “volume” and "quality" of interactivity that occur during the e-learning process (Dimitracopoulou et al. 2004; Pena-Shaff & Nicholls 2004; Soller et al., 2005; Collazos et al. 2007; Bravo et al. 2008).

Although various interaction analysis indicators have been proposed and evaluated by educational researchers in laboratory environments/experimental conditions, frameworks and tools need to be created in order to systemize and to classify them into categories in order that educators (real practitioners) in authentic e-learning environments can use them as efficient and effective assessment media (Vogiatzaki et al., 2006; Reffay & Betbeder 2009). This need was the motivation for creating a new technique for assessing students’ performance in eLearning environments called "Enriched Assessment Rubric (EAR)". The new technique is based on the idea of enriching the traditional assessment rubrics that value the learning products qualitatively with interaction analysis indicators that measure students’ learning behavior in an e-learning environment.

The EAR is designed to assist educators to assess/evaluate both quantitatively and qualitatively the learning products (i.e. individual, group deliverables produced by the students, etc.) as well as the wide range of interaction (between students, between students and educators and between students and resources) performed during all phases of complex interactive scripts, and to report students’ final grades in a precise, detailed and as accurate as possible way.

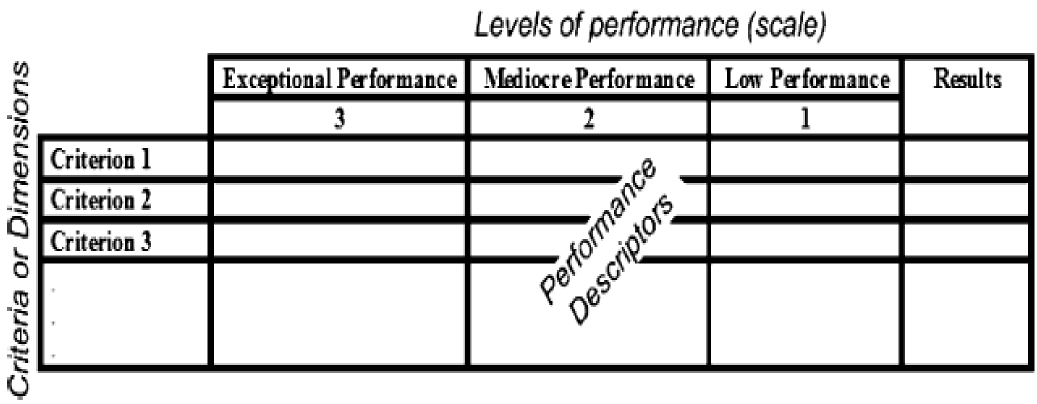


Figure 1**.** An Example of an Assessment Rubric.

2.1. The Structural Components of Enriched Assessment Rubrics

Structural components of the proposed technique are the following: a) the proposed set of interaction analysis indicators, b) the method of design and development of EAR and the EAR developed exemplary, and c) the interaction analysis tool that developed to support the EAR, which are described below.

Table 1. Set of Interaction Analysis Indicators used in the EAR

A. Learning Products	
A1.	The correctness of the learning product (e.g. final solution of the problem)
A2.	The completeness of the learning product
A3.	Final degree of the student per activity (e.g. final test/final report),
A4.	Total number of correct answers in relation to the insufficient or erroneous ones per student (in connection with his/her role)
A5.	Final degree of team per activity
A6.	Final degree of team in the total of activities
A7.	Observation of timetable of activities per student
A8.	Observation of timetable of activities per team
A9.	Duration of completion of an activity (e.g. in the resolution of problem)
B. Learning Process B.1- Interaction: learner-learner & learner-teacher	
B1. 1	Total number of students' messages (written) per activity
B1.2	Total number of team's messages (written) per activity
B1.3	Total number of files shared per student per activity
B1.4	Number and type of students' contributions per activity (e.g. messages argumentation analysis)
B1.5	Comparison of the average of messages (reading) in an assigned activity per student
B1.6	Comparison of the average of messages (reading) in an assigned activity per team
B1.7	Flow of communication (e.g. to one or to many)
B1.8	Total number of discussion threads per student
B1.9	Average of discussion threads per team
B1.10	Degree of central position per student per activity and per phase of instructive scenario
B1.11	Degree of density of social networks developed per team
B1.12	Total number of demands (help –support- encouragement) per student/team towards teacher/schoolmates
B.2 – Interaction: learner-learning resources	
B2. 1	Number of visited resources per student
B2. 2	Number of visited resources per team
B2.3	Popular resources per student
B2.4	Resource's history per student
B2.5	Popular resources per team
B2.6	Total number of visited resources according to his role
B2.7	Total number of visited resources of other roles
B2.8	Number and type of resources that students propose
B2.9	Successive courses of learning resources access per team per activity (theory, example, exercise)
B2.10	Clustering of students (view learning resources)
B2.11	Clustering of teams (view learning resources)

Classification of Interaction Analysis Indicators for the EAR

Adopting the classic Moore's model of learning interactions Moore (1989), we propose a set of indicators which is categorized on two distinct axes: a) the products (quantitative and qualitative analysis) and b) the learning process (interaction between students, between students and educators and between students and learning resources) (Tables 1). The choice and the classification of indicators were based on the following criteria: a) IA that have been proposed by educational researchers (e.g. Spada et al. 2005; Villasclaras-Fernández 2010) b) IA that have been proposed by educators themselves (e.g. Mazza 2004; Zinn & Scheuer 2006; Bratitsis and Dimitracopoulou 2010) and c) IA that is included in specialized IA tools (e.g. Bratitsis et al. 2008; Romero and Ventura 2010).

Design and Development the EAR

The proposed method leads/supports educators through a series of separate sequential steps to design and develop EAR for learning scripts (or per phase) that apply in everyday educational practice (Figure 2).

Towards the greatest possible support of educators we suggest that there should be ready-made forms of combined indicators for the various learning strategies such as the Think-Pair-Share (TPS) and the e-ARMA. Figure 4 depicts the proposed combination of interaction analysis indicators for the well known Think-Pair-Share (TPS) strategy.

Figure 5 shows an extract of an exemplar enriched assessment rubric which was developed for the 'Think" phase of TPS strategy, using the combination of indicators which are shown in Figure 4.

The CoSyLMSAnalytics Tool

For the purpose of implementing the EAR we have developed an interaction analysis tool called CoSyLMSAnalytics (Figure 5) which supports the quantitative and qualitative analysis, such as:

- Traffic analysis and extract of statistical information by student activity, course, etc.
- Production and categorization of clusters (clustering and classification) of students' behavior.

The tool extracts interaction data among students as they have been recorded in the Moodle Learning Management System (LMS). The data are processed and presented by visual indicator. Also, the tool interoperates with other external tools of statistical analysis and data processing (e.g. Excel, SPSS). The tool CoSyLMSAnalytics makes a first-level analysis and then exports the data into appropriate codified form that can be processed further using more sophisticated analytical tools of social networks-SNA (e.g. NetDraw) and produce sociograms (Petropoulou et al., 2010).

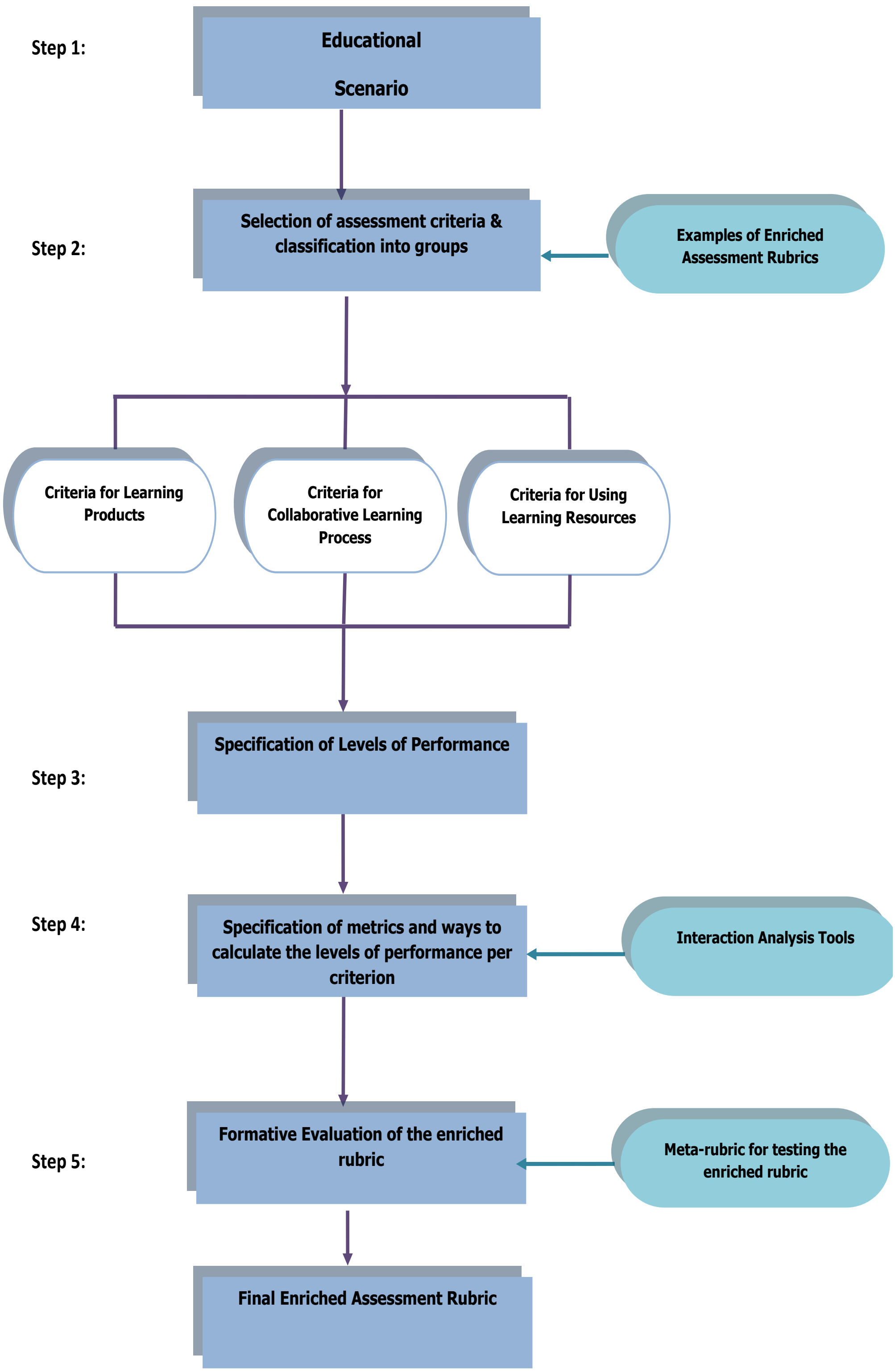


Figure 3. Description of Design Steps of the EAR.

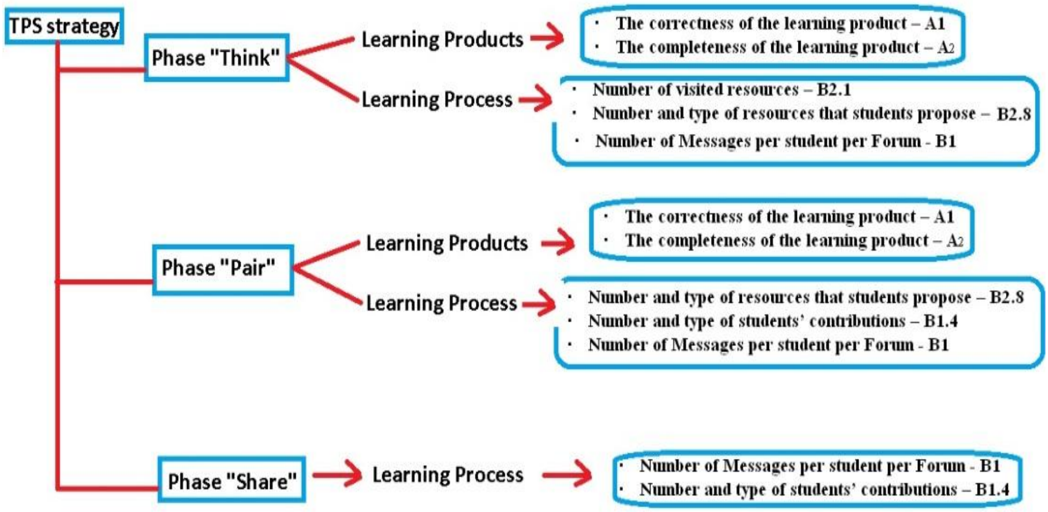


Figure 4. Combined indicators interaction analysis for the TPS strategy.

Criterion	Excellent Performance	Medium Performance	Low Performance
Accuracy of Facts per question (Content)	For almost all the posed questions (more than 6) the supportive facts are reported accurately.	For most of the given questions (4-6) the supportive facts are reported accurately.	For only few questions the supportive facts are reported accurately OR NO facts are reported OR most are inaccurately reported.
Making use of the learning resources (IA indicator: Clustering of students (view learning resources per question)	Learner visited almost all the given learning resources (e.g. more than 8)	Learner visited some of the given learning resources (e.g. 5-8)	Learner did not visit the given learning resources (e.g. less than 5)

Figure 5. An extract of an exemplar enriched assessment rubric for Think phase.

3. Pilot Research Study

In order to assess the degree of acceptability and usability by educators of the EAR technique and its structural elements (design method and development of EAR, proposed framework of interaction analysis indicators and exemplary EAR), we designed one pilot study in the form of specialized seminar. This seminar was attended by 12 teachers (three from primary and nine from secondary Education). More specifically, we organized and implemented a specialized seminar lasting three (3) weeks covering the period from May to June 2009 entitled "Evaluation of student performance in e-learning environments using the EAR technique". The seminar was targeted at educators of primary and secondary education with: a) several years of teaching experience b) substantial experience in e-learning and educational design c) experience in attending hybrid forms training (face to face lecture and access to an electronic platform).

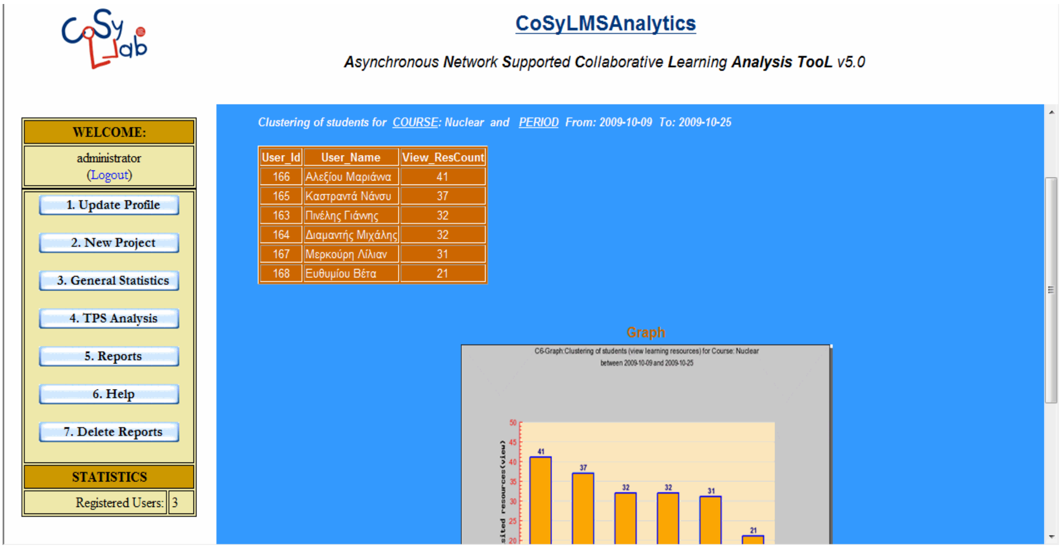


Figure 6. Screenshot of the CoSyLMSAnalytics Interaction Analysis Tool.

3.1. Phases of the Evaluation Study

The implementation process of the seminar consisted of two distinct phases. During the first phase there was a three-hour "face to face" meeting with educators at the Department of Digital Systems, University of Piraeus. The purposes of the meeting were to: a) create a climate of mutual cooperation and trust between educators and researchers, which is essential for ensuring the necessary conditions for efficient completion of the application, b) specify with clarity and detail the research objectives of this study c) give the opportunity to the educators to understand and familiarize teachers with the EAR technique.

During the second phase educators having access to all the various supporting material (examples of the EAR, rubrics templates, interaction analysis indicators, the method for designing rubrics, and the CoSyLMSAnalytics tool), were asked within two (2) weeks to: a) create technology-supported collaborative learning scenarios for the subject of their choice, b) design, develop and electronically submit their scenarios and accompanied enriched rubrics, c) complete and deliver a questionnaire for evaluating the proposed EAR technique. The questionnaire included questions except those that were relevant to the profile of education (individual and service information, etc.), mainly to investigate: a) the degree of clarity, practicality of both the method of design and development of EARs and the exemplary EAR and b) the degree of accuracy and completeness of the same EARs technique.

3.2. Evaluation Findings

Educators who participated in the pilot study successfully developed well-designed enriched assessment rubrics which could be used to holistically assess the learning products and the students' learning behavior during all the phases of the proposed complex collaborative learning scripts which they had scripted. From the analysis of responses in the questionnaires the following key findings were revealed:

The majority of educators (percentage 81.82%) recognize that the existing interaction analysis indicators are sufficient to holistically assess the students' individual and group performance. Most of the educators agreed that both the method of design and development of EARs as well as the exemplary EARs given to them were understandable and easy-to-use (percentage 63.64%). Additionally, the same positive attitude (percentage 63.64%) was stated towards the usefulness of the classification of the interaction analysis indicators. Finally, in an open-ended question, educators were asked to mention impressions, comments, or suggestions about the EAR technique, positive comments were reported such as: "Extremely original technique, the idea to rate the students by what they produce and how they interact impressed me!" and "An impressive combination of the exemplary method EAR, leads to effectiveness". It is positive that the educators claimed that the EAR technique could be applied to e-learning courses of various subjects. Another important issue emphasized was that EAR requires more time and effort from the teachers in order to be applied in real practice than the classic rubrics or the typical grading mechanisms of tests and oral presentations. The existence of ready-made examples of enriched rubrics as well as usable rubrics authoring tools and usable interaction analysis tools will help them apply more easily the EAR technique.

CONCLUSION

This paper presented a new technique for assessing students' performance in eLearning environments which appear to be effective. The idea of using the interaction analysis indicators for measuring learners' performance has been recorded in very few studies in recent years. Bravo et al.'s, (2008) study is of particular interest because it suggests a significant but highly complicated for teachers approach of use of indicators for assessing the learners' performance involved in complex scripts of collaborative problem solving through computer.

The new technique is technically supported through the tool CoSyLMSAnalytics and can be easily implemented by educators. The development of new tools for the use of EAR capable of providing, through a unified (integrated) environment, the educators and the student's ability to use the EAR smoothly and quickly to the educational process. This means that a clear (transparent) integration or interface of interaction analysis tools with tools for creating and viewing rubric. These new tools proposed either to derive data from learning management systems (Learning Management Systems-LMS) and serve as independent tools or be embedded in learning management systems. The immediate/upcoming plans are to repeat the experiment to evaluate the effectiveness and usability of the EAR technique by experienced educators who participate in the training program second level funded by the Act

"Training teachers to use and application of ICT in teaching practice ' and coordinated by the Educational Training Agency (O.EP.EK.).

ACKNOWLEDGMENTS

This research project-related is related with the PAKE in Attica and Central Greece owned in the project "Implementation of Trainers' Education" which is implemented under the Act "Training teachers to use and application of ICT in teaching practice" co-financed by Greece and the European Union (ESF) and coordinated by OEPEK (Teacher Training Agency).

REFERENCES

Allen S., & Knight J. (2009). A Method for Collaboratively Developing and Validating a Rubric. *International Journal for the Scholarship of Teaching and Learning*, 3(2), 1-17.

Andrade, H. & Valtcheva, A. (2008). Promoting Learning and Achievement through Self-Assessment. Theory into Practice, http://www.informaworld.com/smpp/title~db=all~content=t775653706~tab=issueslist~branches=48, v48 48 (1), 12 -19.

Arvaja, M. (2011). Analyzing the Contextual Nature of Collaborative Activity. In S. Puntambekar et al., (Eds.), Analyzing Interactions in CSCL: Methods, Approaches and Issues, (pp. 25-46) Springer Science.

Blommel M., & Abate A. (2007). A rubric to assess critical literature evaluation skills. *American Journal of Pharmaceutical Education*, 71(4), 1–8.

Bratitsis, T., & Dimitracopoulou, A. (2010). Interpretation of Computer Based Interaction Analysis Indicators: a significant issue for enhancing collaboration in Technology Based Learning. In Kock, N. (Eds.), *Interdisciplinary Perspectives on E-Collaboration: Emerging Trends and Applications, Advances in E-Collaboration*, Book series, (pp. 31-59). IGI-Global, USA.

Bratitsis, T., Dimitracopoulou, A., Martínez-Monés, A., Marcos, A., & Dimitriadis, Y. (2008). Supporting members of a learning community using Interaction Analysis tools: The example of the Kaleidoscope NoE scientific network. In P. Diaz, Kinshuk, I. Aedo, E. Mora (Eds.), *The 8th IEEE International Conference on Advanced Learning Technologies*, (pp. 809-813). IEEE Computer Society.

Bravo, C., Redondo, A., Verdejo, F., & Ortega, M. (2008). A framework for process-solution analysis in collaborative learning environments. *International Journal of Human-Computer Studies*, 66 (11), 812-832.

Buzzetto-More N., & Alade A. (2006). Best Practices in e-Assessment. *Journal of Information Technology Education*, 5, 251-269.

Campbell, K. & Gibson, S. (2008). The evolution of assessment in distance education. In T. Evans, M. Haughey, & D. Murphy (Eds.), *The International Handbook of Distance Education* (pp. 341-365). Bingley, UK: Emerald Group Publishing Ltd./Elsevier.

Chan C., & Van Aalst J. (2004). Learning, assessment and collaboration in computer-supported environments. In. P. Dillenbourg (Series Eds.) & J. W. Strijbos, P. A.

Kirschner R. L. Martens (Vol. Eds.): Computer-supported collaborative learning: Vol 3. *What we know about CSCL: An implementing it in higher education* (pp. 87-112). Boston, MA: Kluwer Academic.

Collazos, C., Guerrero, L., Pino, J., Renzi, S., Klobas, J., Ortega, M., Redondo, M., & Bravo, C. (2007). Evaluating collaborative learning processes using system-based measurement. *Educational Technology & Society*, 10 (3), 257-274.

Daradoumis T., Martínez-Monés A., & Xhafa F. (2006). A layered framework for evaluatingon-line collaborative learning interactions. *International Journal of Human-Computer Studies*, 64(7), 622-635.

Darling-Hammond, L. & Adamson, F. (2010). Beyond basic skills: The role of performance assessment in achieving 21st century standards of learning. Stanford Center for Opportunity Policy in Education, Stanford University.

Dimitracopoulou, A., Martinez, A., Dimitriadis, Y., Morch, A., Ludvigsen, S., Harrer, A., Hoppe, U., Barros, B., Verdejo, F., Hulsof, C., Fessakis, G., Petrou, A., Lund, K., Baker, M., Jermann, P., Dillenbourg, P., Kollias, V., & Vosniadou, S. (2004). State of the art on interaction analysis: Interaction analysis indicators. ICALTS Project Deliverable: D.26.1. Retrieved 27 October 2006, from http://www.rhodes.aegean.gr/ltee/kaleidoscope-icalts/Publications/D1 State of the Art Version_1_3 ICALTS_Kal NoE.pdf

Hafner C., & Hafner M. (2003). Quantitative analysis of the rubric as an assessment tool: An empirical study of student peer-group rating. International Journal of science education, 25, (1509-1528).

Johnson R., Penny J., & Gordon B. (2009). Assessing performance: designing, scoring, and validating performance tasks. Guilford Press.

Joosten-ten Brinke D., Van Bruggen J., Hermans H., Burgers J., Giesbers B., Koper R., & Latour I. (2007). Modeling assessment for re-use of traditional and new types of assessment. *Computers in Human Behavior* 23, pp. 2721–2741.

Lazakidou, G., & Retalis, S. (2010). Using computer supported collaborative learning strategies for helping learners acquire selfregulated problem solving skills in Mathematics. *Computers and Education*, 54(1), pp. 3-13.

Mazza, R. (2004). Using information visualisation to facilitate instructors in web-based distance learning. Unpublished Dissertation Ph.D., University of Lugano, Switzerland.

Mödritscher F., Spiel S., & Garcia-Barrios V. (2006). Assessment in E-Learning Environments: A Comparison of three Methods. In C. Crawford et al. (Eds.): *Proceedings of Society for Information Technology and Teacher Education International Conference* 2006, (108-113). Chesapeake, VA: AACE.

Moore, G. (1989). Three types of interaction. *The American Journal of Distance Education*, 3(2), 1-6.

Morgan, C., & O'Reilly, M. (2006). Ten key qualities of assessment online. In Hricko, M. & Howell, S.L. (Ed.), *Online assessment and measurement: Foundations and challenges*. (pp. 86-101). Hershey, PA & London: Information Science Publishing

Partnerships for 21st Century Skills (2009). P21 Framework Definitions explained: White paper. Retrived from http://www.21stcenturyskills.org/documents/p21_framework_definitions_052909.pdf

Pena-Shaff, J., & Nicholls, C. (2004). Analyzing student interactions and meaning construction in Computer Bulletin Board (BBS) discussions. *Computers and Education*, 42, 243-265.

Persico D., Pozzi F., & Sarti L. (2009). A model for monitoring and evaluating CSCL. In Juan, A., Daradoumis, T., Xhafa, F., Caballe, S., Faulin, J. (Eds.): *Monitoring and Assessment in Online Collaborative Environments: Emergent Computational Technologies for E-learning Support* (pp. 149-170). IGI Global.

Petropoulou, O., Altanis, I., Retalis, S., Nicolaou, A., Kannas, C., Vasiliadou, M., & Pattis, I. (2010). Building a tool to help teachers analyse learners' interactions in a networked learning environment. *Educational Media International*, 47(3), 231- 246.

Petropoulou, O., Vasilikopoulou, M., & Retalis, S. (2009). Enriched Assessment Rubrics: A new medium for enabling teachers easily assess students' performance when participating to complex interactive learning scenarios. *Operational Research International Journal*, 11(2), 171-186.

Petropoulou, O., Lazakidou, G., Retalis, S., Vrasidas, C. (2007). Analysing interaction behaviour in network supported collaborative learning environments: a holistic approach. *International Journal of Knowledge and Learning*, 3(4/5), 450 – 464.

Pozzi, E. Manca, S. Persico, D. & Sarti, L. (2007). A general framework for tracking and analysing learning processes in computer-supported collaborative learning environments. *Innovation in Education and Teaching International*, 44(2), 169-179.

Quellmalz E., & Kozma R. (2003). Designing assessments of learning with technology. *Assessment in Education*, 10(3), 389-406.

Reffay C., & Betbeder M. (2009). Extending validation of tools and analyses in CSCL situations: How to collaborate on interaction analysis? In Pre-conference Workshop *on Interaction analysis and visualization: analysis` methods, tools and research questions*, Computer Supported Collaborative Learning (CSCL 2009), Rhodos, Greece.

Roberts T. (2006). Self, Peer and Group Assessment in E-Learning. Idea Group Inc

Romero, C., & Ventura, S. (2010). Educational Data Mining: A Review of the State of the Art. *IEEE Transactions on Systems, Man and Cybernetics, Part C: Applications and Reviews*, 40(6), 601-618.

Rovai, A. (2000). Online and traditional assessments: what is the difference? *The Internet and Higher Education*, 3, 141-151.

Shepard, L. (2000). The role of assessment in a learning culture. *Educational Researcher*, 29(7), 4-14.

Shepherd, C. & Mullane, A. (2008). Rubrics: The Key To Fairness In Performance Based Assessments. *Journal of College Teaching & Learning*, 5(9), 27-32.

Soller, A., Martinez, A., Jermann, P., & Muehlenbrock, M. (2005). From mirroring to guiding: a review of state of the art technology for supporting collaborative learning. *International Journal on Artificial Intelligence in Education*, 15(4), 261-290.

Spada, H., Meier, A., Rummel, N., & Hauser, S. (2005). A new method to assess the quality of collaborative process in CSCL. In T. Koschmann, D. Suthers, & Chan, T. W. (Eds.), *Proceedings of the CSCL 2005*, (pp. 622-631). Mahwah, N.J: Lawrence Erlbaum Associates.

Stiggins, R. (2004). New Assessment Beliefs for a New School Mission. *Phi Delta Kappan*, 86(1), 22-27.

Underhill F. (2006). Theories of Learning and Their Implications for On-Line Assessment. *The Turkish Online Journal of Distance Education*, 7(1), 165-174.

Villasclaras-Fernández, E. (2010). A design process supported by software authoring tools for the integration of assessment within CSCL scripts. Unpublished Dissertation Ph.D., University of Valladolid, Spain.

Voyiatzaki, E., Margaritis, M., & Avouris, N. (2006). Collaborative Interaction Analysis: the teacher's perspective. Proceedings of the *6th IEEE International Conference on Advanced Learning Technologies*, (pp. 345-349), 5-7 July, Kerkrade, Netherlands.

Zinn, C., & Scheuer, O. (2006). Getting to Know your Student in Distance-Learning Contexts. In Proceedings of the *European Conference on Technology Enhanced Learning (EC-TEL 2006)*, v. 4227, (437-451), Lecture Notes in Computer Science, Springer-Verlag Berlin Heidelberg.

In: Evaluation in E-learning
Editors: Y. Psaromiligkos, A. Spyridakos et al.
ISBN: 978-1- 61942-942-0

Chapter 5

EVALUATING ONLINE COLLABORATIVE LEARNING ACTIVITIES ORIENTED TO CREATIVITY DEVELOPMENT

Francesca Pozzi and Michela Ott*
Istituto Tecnologie Didattiche – CNR, Genoa, Italy

ABSTRACT

The present chapter tackles the issue of evaluation in CSCL contexts (Computer Supported Collaborative Learning). In particular it proposes a model to evaluate online collaborative learning activities oriented to the development of students' creative skills and attitudes. The model, which encompasses three main spheres, namely the cognitive, the meta-cognitive and the affective spheres, has been tested on two collaborative activities proposed within a real online course, and the results of such test are presented, so that it is possible to make some considerations about the main strengths and weaknesses of the model itself.

Keywords: CSCL, evaluation, creativity, cognitive, affective, meta-cognitive

1. INTRODUCTION

The value of online collaborative learning has been increasingly recognised and investigated by researchers and designers of the Computer Supported Collaborative Learning field (CSCL) (Dillenbourg, 1999; Kanuka & Anderson, 1999; The Cognition and Technology Group at Vanderbilt, 1991; Scardamalia & Bereiter, 1994). The alleged benefits of CSCL range from promotion of critical thinking and conceptual understanding to enhancement of motivation and development of group problem solving abilities.

* {pozzi, ott}@itd.cnr.it

In this research field, the development of effective methods and tools to monitor and evaluate online collaborative learning is of paramount importance. To address this issue, a number of approaches have been developed and tested in the past years, most of which are based on the analysis of interactions occurring among students during the learning process. In most studies quantitative information extracted directly from the CMC (Computer Mediated Communication) system are juxtaposed to more qualitative approaches, usually relying on the content analysis of the messages exchanged (Henri, 1992; Hara et al., 2000; Rourke et al., 2001; Lally, 2002; Aviv et al., 2003; Lipponen et al., 2003; Martinez et al., 2003; Daradoumis et al., 2004; Schrire, 2006; Weinberger & Fischer, 2006; ICALTS Kaleidoscope JEIRP[1]).

In 2009 Persico et al. proposed a model mainly inspired by Garrison and Anderson's (2003), which encompasses four dimensions: the participative, the social, the cognitive and the teaching dimensions. In the model, in order to bridge the gap between the theoretical framework and its practical applications, suitable indicators have been identified for each dimension, consisting of quantitative or qualitative variables that allow the analysis of each dimension according to specific objectives. These indicators express the actual manifestations of the four dimensions in a learning community (Persico et al., 2009).

Within the CSCL research field there is also another issue which has recently drawn the researchers' attention, which concerns whether and to what extent *creativity* can be stimulated through online collaborative learning activities.

As a matter of fact, at a general level it has been already recognized that creativity can be regarded as something which can be *potentially* stimulated through adequate learning tools and methods (UNESCO, 1972; Nickerson, 1999; Csikszentmihalyi, 1997; Torrance et al., 1989). More recently, many researchers have also stressed that creativity may spring out of a single mind, but – even more frequently – may stems out from interactions among people while working together, sharing paradigms and know-how (Fischer, 2005; Fischer et al., 2005). Many researchers have claimed that each idea or solution generated by a person may stimulate further ideas and solutions in others (Fink et al., 2010) and that creative performances increase as a result of idea sharing and/or idea exchange (Fink et al., 2010; Paulus and Nijstad, 2003; Dugosh and Paulus, 2005).

Thus, nowadays there is a growing tendency in educational contexts to propose not only individual, but also collaborative learning activities oriented to foster students' creative attitudes and skills, by recognizing that the creative process may well take place due to interactions of a student with the learning environment and with her/his peers as well.

In addition to this, it has also been recognized that collaborative techniques or patterns (such as the Brainstorming, the Jigsaw, and others) which are often used in educational contexts to scaffold collaboration, not only are highly compatible with creative techniques (Sie et al., 2010), but often the two categories of techniques share a common structure, made of an alternation between divergent and convergent phases (Forster & Brocco, 2010). As a consequence of this, we are assisting to the convergence of two different research threads: the one focused on collaboration and on how to scaffold it, and the one investigating creativity and the ways to foster it.

[1] ICALTS (Interaction and Collaboration AnaLysis supporting Teachers and Students Self-regulation) is a Jointly Executed Integrated Research Project of the Kaleidoscope Network of Excellence, website at http://www.rhodes.aegean.gr/ltee/kaleidoscope-icalts/

Within this new scenario, a need emerges, i.e. to define new methods and tools able to evaluate those particular collaborative learning activities oriented to the development of students' creative skills and attitudes (Burleson, 2005; Edmonds & Candy, 2002).

This chapter is an attempt to fill in this gap: drawing on the one hand on the experience gained in the field of CSCL evaluation, and, on the other, in the field of creativity development, the authors of this chapter refer to an evaluation model and test its applicability to two online collaborative learning activities, with the aim to investigate whether and to what extent these activities fostered students' creative attitudes and skills. The chapter illustrates the results obtained during the testing phase and discusses the main shortcomings and strong points of the model.

2 Cognitive, Meta-cognitive and Affective Spheres in Learning Activities Oriented to Creativity Development

There seems to be a certain agreement among researchers (Amabile, 1996; Sternberg, 1999; Torrance et al., 1989) that creativity is grounded on:

- *cognitive* capacities. i.e. understanding data and facts and building knowledge,
- *meta-cognitive* abilities, i.e. the capacity of perceiving and elaborating weaknesses and strengths of own reasoning and/or actions, and
- an *affective* involvement in the tasks to be performed, which implies positively accepting the task and actively work to reach the intended goal.

Drawing on such agreement, the three above mentioned spheres are considered crucial for the evaluation of all the learning experiences aimed at developing skills and attitudes oriented to creativity. Thus cognitive capacities, meta-cognitive abilities and affective involvement, are the main components of the proposed model (Ott & Pozzi, 2010). For each sphere, the model identifies also a number of indicators, which are described in the following.

In particular, for the *cognitive sphere*, three fundamental indicators are identified by referring to the New Taxonomy of the Educational Objectives proposed by Krathwohl (2002). Krathwohl defines creativity as the ability of "putting elements together to form a novel coherent whole or make an original product" and considers it the top educational objective to be met. Following the arguments put forward by this author, in the proposed model the three cognitive indicators of creativity are: *Generating*, a process which involves the mental representation of the problem at hand (whatever it could be), in all its aspects and details, possibly making comparison with other problems/situations (instantiated by actions such as: combine, estimate, compare, state, and so on); *Planning*, namely the process of figuring out and mentally designing problem solutions or even defining methods and plans to achieve a goal (instantiated by actions such as: predict, infer, hypothesize, design, define, and so on); *Producing*, which is the process dealing with the actual enactment of what was generated and then planned and which may give rise to the creative act or product (instantiated by actions such as: build, enact, apply, test, verify, and so on) (Ott & Pozzi, 2010).

As to the *affective sphere*, by referring to the existing research in the affective domain field (Bloom et al., 1956; Rovai et al. 2009), two indicators are adopted in the proposed model, able to account for students' attitudes towards: *Receiving* (or paying attention to stimuli); this indicator is denoted by involvement and immersion in learning activities and includes being curious, motivated, trying over and over, and so on; *Responding* (or reacting to stimuli); this indicator refers to the actual expression of positive/negative feelings: satisfaction, joy, disappointment, excitement, depression, fear, and so on (Ott & Pozzi, 2010).

Lastly, for the *meta-cognitive sphere*, by referring to the work of both Kim et al. (2009) and Murphy (2008), the model identifies three main indicators, namely those related to students' capabilities of: *Monitoring* the enacted learning process, which implies the attitude and the ability of recalling and evaluating one's own cognitive process, by also evidencing strengths and weaknesses: *Regulating* one's own behavior on the basis of the perception/understanding of previously performed actions (which also means reviewing, controlling and tuning the activities by carrying out possible improvements, etc.); *Evaluating* one's own activities/performance from the viewpoint of the final outcome; this implies acquiring the awareness of what has been done by criticizing single actions in the light of a comprehensive estimation / judgment of the results obtained (Ott & Pozzi, 2010).

The model illustrated above has been already tested by the authors for the evaluation of individual educational activities carried out in the field of game-based learning (Ott & Pozzi, 2010a). After that experience, the authors wanted to test it also in a completely different context, namely to evaluate an online collaborative learning experience.

3. Context

In the past years the Institute Tecnologie Didattiche (ITD) – CNR has designed and run several editions of a blended course on the topic "Educational Technology" for the "SSIS", which is the Italian institution providing initial training to secondary teachers.

Main educational goal of these courses is the promotion of instructional design competence development, with special focus on the evaluation and selection of learning strategies, techniques and tools and on the implementation of educational technology in the school context.

The target population, made of adults who often teach at schools, makes the CSCL approaches particularly fitting to the context. During online activities student teachers are usually subdivided in groups (typically 20-25 persons per group) and are engaged in tasks (discussing a topic, solving a problem, studying a case, etc.) with concrete outputs to produce, which act as catalysts of interaction and collaboration among them.

The idea to adopt this context to test the evaluation model came from the fact that during one edition of the course, a group of students claimed they judged the CSCL approach they were implementing, able to foster their creative attitudes and skills, as opposed to more traditional, individual approaches.

The authors of the evaluation model, then, decided to select one edition of the SSIS course, and try to apply the model itself, to investigate whatever and to what extent this was able, even in such a context, to capture aspects of the learning process ascribable to creative attitudes and skills.

Within the selected edition, the class was composed of 21 students coordinated by a tutor.

Interactions among students and with the tutor occurred within the CMC system Moodle[2].

During the course students were proposed, among others, two online activities, lasting 3 weeks each; the two activities were each based on a collaborative technique, namely a Role Play and a Discussion. The Role Play activity concerned the use of webquests in the classroom, while the Discussion activity concerned the use of blogs in educational contexts. In this chapter the authors won't provide further details about the two proposed activities, as this would go beyond the aims of the chapter: the two activities were selected for their being "exemplar" and the model was applied to them so to test its ability to capture any signals of creative attitudes/skills emerging from students during the collaborative learning process..

Table 1. An excerpt of the grid operationalizing the evaluation model

Sphere	Indicator	Actions
COGNITIVE SPHERE	Generating	combines
		estimates
		compares
		…
	Planning	predicts
		infers
		hypothesizes
		…
	Producing	builds
		enacts
		applies
		….
AFFECTIVE SPHERE	Receiving	is curious
		is motivated
		is frightened…
		….
	Responding	expresses joy
		expresses disappointment
		…
META-COGNITIVE SPHERE	Monitoring	is aware of the process
		reflects on the process
		….
	Regulating	controls the process
		adjusts the process
		…
	Evaluating	judges the process
		evaluates the outcome
		…

[2] http://moodle.org/

4. METHOD

In order to gather data within the test, content analysis techniques have been used to analyze the messages exchanged among students during the online collaborative activities. In particular, two coders were involved, who were in charge of reading all the messages, segmenting them into units, and then assigning to each unit one of the indicator of the model. The unit of analysis chosen was the "unit of meaning" (Henri, 1992).

As to the process of associating indicators to units, it was necessary to produce and provide the coders with a grid, able to operationalize the indicators of the model (see Table 1).

The grid contained a list of actions referred to each indicator, so that the coders could count on a common guideline on how to associate indicators to units.

The total number of messages exchanged by the students during the examined activities and thus analyzed by the coders, run into 439 (209 messages exchanged during the Role Play + 230 messages exchanged during the Discussion). Within the 439 messages, 1517 units were detected by the coders.

The inter-rater reliability between the two coders (i.e. the agreement between the two) was calculated on a sample of 140 messages (30% of the total messages), and resulted 0,87 (Holsti coefficient) and 0,82 (percent agreement).

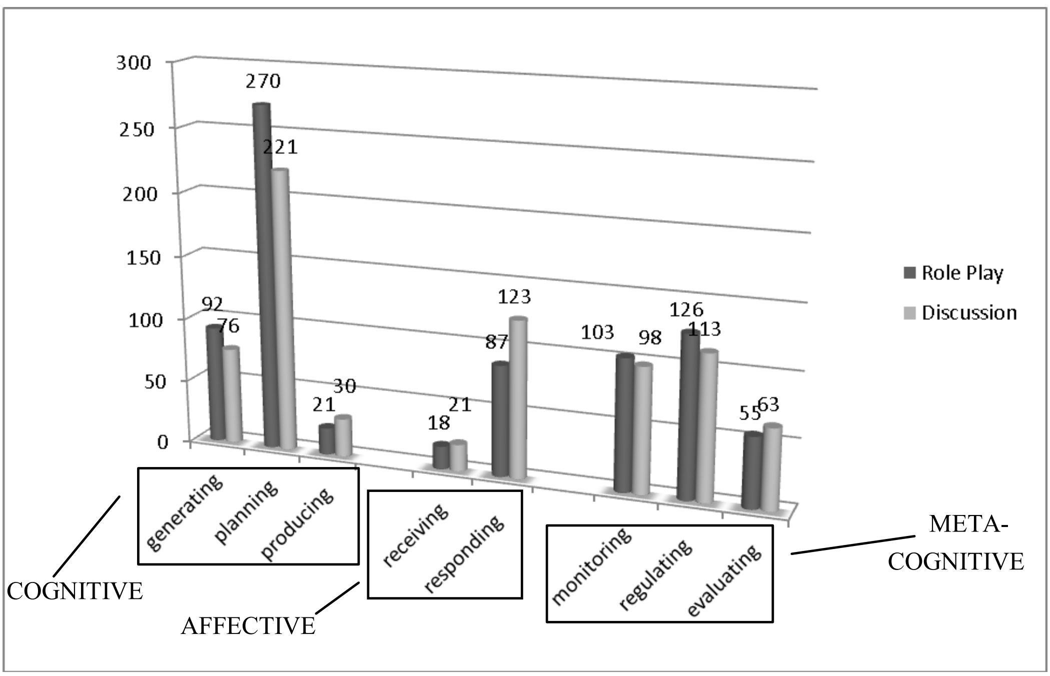


Figure 1. Indicators of the model emerged during the Role Play and the Discussion.

5. Results

The present section illustrates the main results obtained by the content analysis of the messages exchanged by the students during the two online collaborative learning activities.

In particular Figure 1 shows the number of units detected by the coders for each indicator of the model during the Role Play and the Discussion.

As one may note in Figure 1, the highest indicator is “Planning” (cognitive sphere) and this applies to both the proposed activities. “Generating” and “Producing” indicators are both there, in the Role Play and in the Discussion activities, but they are far less developed than the “Planning” indicator.

The high levels of the “Planning” indicator gives a very good signal, as the abilities to predict, infer, synthesize, and so on, are at the heart of the creative process and the fact that in both the proposed activities they strongly emerged, seems to confirm that collaboration and exchange among peers – per se – support students in developing and using these abilities.

On the other hand, the fact that the “Producing” indicator is not very high in any of the activities, can be easily explained by considering that none of them explicitly envisaged an application phase, so it was difficult for the students to put into practice what they have planned during the former phases.

As to the affective sphere, the “Responding” indicator is always higher than the “Receiving” indicator. To be noted that the actions under the “Receiving” umbrella (see grid in Table 1) were particularly difficult to be detected by the coders using only the content analysis, as this particular indicator is more easily detectable by a direct observation of students while carrying out the activity. On the contrary, “Responding” indicator can be detected by looking at what students have expressed in their messages.

Lastly, within the meta-cognitive sphere, it seems that “Monitoring”, “Regulating” and “Evaluating” indicators all have been - to some extent - detected during the two proposed activities.

The indicators in the Figure 1 follow a similar trend during the two activities and the differences in values are not so evident. Still, some differences exist when looking at the various indicators of the model singularly. For example, the “Role Play” activity shows a better capacity to develop the cognitive aspects as for the “Generating” and “Planning” indicators, while the “Producing” indicator, even if rather low in both the activities, is a little more developed during the “Discussion” activity.

The “Discussion” activity reports higher values in the affective dimension (both for “Receiving” and “Responding” indicators).

Finally, as to the meta-cognitive aspects, they are more triggered during the “Role Play” activity. The “Monitoring” and “Regulating” indicators report higher values in the “Role Play” activity rather than the “Discussion” activity. The “Evaluating” indicator only gave higher values during the “Discussion” activity.

All in all, these data indicate the “Role Play” activity as more able to foster the cognitive and meta-cognitive skills, while the “Discussion” activity seems to be more effective as far as the affective sphere is concerned.

6. Discussion

In this chapter the authors have presented a model, aimed to evaluate learning activities oriented to creativity development. The model has been tested to evaluate two online collaborative activities, to see whether and to what extent it was able to capture possible creative attitudes and skills enacted by students while engaged in the collaborative activities themselves.

The model is composed of three main components (spheres) and a set of indicators in each sphere; these have been identified on the basis of the literature in the field and on some previous experiences of the authors.

During the experiment, the model turned out to be sound: all the spheres and indicators identified in the original version of the model, have found a counterpart and a confirmation in the interactions occurred among the students during the two collaborative activities, so that we can claim that, thanks to the evaluation model, it is possible to shed light, even in collaborative learning contexts, on students' creative attitudes and skills, and to give a measure to the extent to which, during a collaborative activity, students show one creative attitude/skill or the other.

Besides, both the coders, interviewed after their coding experience, confirmed that applying the model was rather easy and that the grid containing the list of actions associated to each indicator helped a lot to operationalize the model. Nonetheless, during the coding process some indications came from the coders on how to improve and enrich the grid and these will be implemented for a new version of the grid itself.

Overall the model can be judged positively, as it turned out to be able to capture those skills and attitudes which are considered potential agents of creativity. In particular, it was evident that looking at the creativity-oriented skills and attitudes (namely the cognitive and meta- cognitive skills and the affective attitudes) as a whole and in their specificities, is a key point: looking at the eight indicators as a whole allows one to have a general overview on the students' behaviours, attitudes, skills in respect to the enacted process; nevertheless, at the same time, the possibility to investigate separately the three aspects may better serve the purpose of changing the tack and fine-tuning subsequent educational interventions.

Moreover, the model helps in stressing that collaborative techniques adopted (such as the Discussion, the Role Play, or even others) may differently stimulate one or the other spheres and/ or indicators.

On the other hand, a shortcoming of the model is that its indicators do not give enough account of the various contributions provided by the group members and – in other terms – don't reflect the interactions, by flattening the "creative discourse". This certainly derives from the context where the model was originally conceived, where the work was mainly carried out individually (Ott and Pozzi, 2010a). This aspect is something that need further work in the future.

Also, the model – as it is now – is very much focused on the process and does not take into consideration the final product of the learning activity, which should in same way be considered.

The testing phase is still ongoing and further evaluation tools are now being implemented to integrate the grid based on the model, the idea being that further data are needed to demonstrate and /or strenghten the suitability of the proposed model.

CONCLUSION

This chapter tackles the issue of how to evaluate online collaborative actvities oriented to the development of students' creative attitudes and skills.

In particular a model is illustrated, which was previously developed and experimented by the authors themselves in a different context. In this study, the model is used to evaluate the capacity of an online "Discussion" and a "Role Play" activity, to foster students' creative attitudes and skills. All the spheres and indicators of the model have been "enacted" during the two activities and this confirms that the model can be applied to collaborative learning contexts as well. Although the model is at a preliminary stage of testing, there are proofs to be rather effective. However, further testing and improvements is needed.

Another important aspect emerged from the experiment presented in this chapter, is that "Planning" abilities have strongly emerged from both the "Role Play" and the "Discussion" activities. This highlights once more that collaboration and idea sharing may help students to accommodate data and facts into new scenarios and this is often a preliminary step to the creative act. Although these results are not comparable with the results presented by Ott & Pozzi (2010a), still during the individual game-based activities proposed there, "Planning" was far less developed, as to confirm that working in isolation does not necessarily foster such component, while working in groups does.

This final consideration has led the authors to consider the possibility in the future to further test the model also within a double scenario: by proposing a collaborative activity to a group of students, while asking to another control group to carry out a comparable activity at individual level. This could hopefully lead to further insights on the model effectiveness, and on the process of creativity development.

REFERENCES

Amabile, T.M. (1996). *Creativity in context*. Westview Press.

Aviv, R., Erlich, Z., Ravid, G., & Geva, A. (2003). Network analysis of knowledge construction in asynchronous learning networks. *Journal of Asynchronous Learning Networks*, 7(3), 1-23.

Bloom, B., Englehart, M. Furst, E., Hill, W., & Krathwohl, D. (1956). *Taxonomy of educational objectives: The classification of educational goals. Handbook I: Cognitive domain*. New York, Toronto: Longmans, Green.

Burleson, W. (2005). Developing creativity, motivation, and self-actualization with learning systems. *International Journal of Human-Computer Studies*, 63(4-5), 436-451.

Csikszentmihalyi, M. (1997). *Creativity: Flow and the Psychology of Discovery and Invention*. Harper Perennial.

Daradoumis, T., Martinez-Monés, A. & Xhafa, F. (2004). An integrated approach for analysing and assessing the performance of virtual learning groups. *Lecture notes in Computer Science*, 3198, 289-304.

Dillenbourg, P. (Ed.) (1999). *Collaborative Learning: Cognitive and Computational Approaches*. Pergamon Press.

Dugosh, K.L., & Paulus, P.B. (2005). Cognitive and social comparison processes in brain storming, *J. Exp. Soc. Psychol*. 41, 313–320.

Edmonds, E., & Candy, L. (2002). Creativity, Art Practice, and Knowledge. *Communications of the ACM*, 45 (10), 91-95.

Fink, A., Grabner, R. H., Gebauer, D., Reishofer, G., Koschutnig, K., Ebner, F. (2010). Enhancing creativity by means of cognitive stimulation: Evidence from an fMRI study. *NeuroImage*, 52, 1687-1695.

Fischer, G. (2005). Social creativity: Making all voices heard. In *Proceedings of the HCI International Conference (HCII)*. Accessed December 2010: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.85.6605&rep=rep1&type=pdf

Fischer, G., Giaccardi, E., Eden, H., Sugimoto, M., & Ye, Y. (2005). Beyond binary choices: Integrating individual and social creativity. *International Journal of Human–Computer Studies*, 63, 482–512.

Forster, F., Brocco, M. (2010). A unified process model for creativity-technique based problem solving processes. In S. Retalis & P. B. Sloep (Eds.), *Proceedings of idSpace Workshop on Methods & Tools for Computer Supported Collaborative Creativity Process: Linking creativity & informal learning* (held within EC-TEL 2009). September, 30th, 2009, Nice, France. Accessed December 2010: http://CEUR-WS.org/Vol-536/

Garrison, R. & Anderson, T. (2003). *E-learning in the 21st century. A framework for research and practice*. London and New York: Routledge Falmer.

Hara, N., Bonk, C.J., & Angeli, C. (2000). Content analysis of online discussion in an applied educational psychology course. *Instructional Science*, 28, 115-152.

Henri, F. (1992) Computer conferencing and content analysis. In *Collaborative Learning Through Computer Conferencing: The Najaden Papers* (ed. A. R. Kaye), 115-136. New York, Springer.

Kanuka, H., Anderson, T. (1999). Using Constructivism in Technology-Mediated Learning: Constructing Order out of the Chaos in the Literature. *Radical Pedagogy*, 1(2).

Kim, B., Park, H., & Baek, Y.(2009). Not just fun, but serious strategies: Using meta-cognitive strategies in game-based learning. *Computers & Education*, 52, (4) 800-810.

Krathwohl, D.R., (2002). A Revision of Bloom’s Taxonomy: An Overview. *Theory into Practice*, 41(4), 212-218.

Lally, V. (Ed) (2002). Elaborating collaborative interactions in networked learning: a multi-method approach. In *Proceedings of the Networked Learning Conference 2002*, 26th-28th March 2002, University of Sheffield.

Lipponen, L., Rahikainen, M., Lallilmo, J. & Hakkarainen, K. (2003). Patterns of participation and discourse in elementary students’ computer-supported collaborative learning. *Learning and Instruction*. **13**, 487–509.

Martinez, A., Dimitriadis, Y., Rubia, B., Gomez, E. & De La Fuente, P. (2003). Combining qualitative evaluation and social network analysis for the study of classroom social interactions. *Computers and Education*, 41(4), 353-368.

Murphy, E. (2008). A framework for identifying and promoting metacognitive knowledge and control in online discussants. *Canadian Journal of Learning and Technology*, 34(2). Accessed December 2010: http://www.cjlt.ca/index.php/cjlt/article/view/491/222

Nickerson, R.S. (1999). Enhancing creativity. In: Sternberg, R.J. (Ed.) *Handbook of Creativity*. Cambridge University Press, Cambridge.

Ott, M. and Pozzi, F. (2010). Towards a model to evaluate creativity-oriented TEL activities. *Procedia Social and Behavioural Sciences Journal*, 2, 3532-3536, Elsevier.

Ott, M., Pozzi, F. (2010a). Digital games as creativity enablers for children. *Behaviour & Information Technology*, 1.

Paulus, P. B., & Nijstad, B. A. (2003). *Group creativity: innovation through collaboration*. Oxford University Press, Oxford.

Persico, D., Pozzi, F. & Sarti, L. (2009). A model for monitoring and evaluating CSCL. In: Juan, A.A., Daradoumis, T., Xhafa, F., Caballe, S., Faulin, J. (Eds.) *Monitoring and Assessment in Online Collaborative Environments: Emergent Computational Technologies for E-learning Suppor*t. IGI Global.

Rourke, L., Anderson, T., Garrison, R. & Archer, W. (2001). Methodological Issues in the Content Analysis of Computer Conference Transcripts. *International Journal of Artificial Intelligence in Education*, 12, 8-22.

Rovai, A. P, Wighting, M. J., Baker, J.D., & Grooms, L. D. (2009). Development of an instrument to measure perceived cognitive, affective, and psychomotor learning in traditional and virtual classroom higher education settings. *The Internet and Higher Education*, 12(1), 7-13.

Scardamalia, M., Bereiter, C. (1994). Computer support for knowledge-building communities. *The Journal of the Learning Sciences*, 3(3) 265-283.

Schrire, S. (2006). Knowledge building in asynchronous discussion groups: Going beyond quantitative analysis. *Computers & Education*, 46, 49-70.

Sie, R. L. L., Bitter-Rijpkema, M., Sloep, P. B., Retalis, S. (2010). Knowledge sharing strategies for collaborative creativity. In S. Retalis & P. B. Sloep (Eds.), *Proceedings of idSpace Workshop on Methods & Tools for Computer Supported Collaborative Creativity Process: Linking creativity & informal learning* (held within EC-TEL 2009). September, 30th, 2009, Nice, France. Accessed December 2010: http://CEUR-WS.org/Vol-536/

Sternberg, R.J. (Ed.) (1999). *Handbook of Creativity, Cambridge University Press*, Cambridge UK.

The Cognition and Technology Group at Vanderbilt (1991). Some thoughts about constructivism and instructional design. *Educational Technology*. 31(10), 16-18.

Torrance, E. P., Glover, J. A., Ronning, R. R., & Reynolds, C. R. (Eds.) (1989). *Handbook of creativity: Perspectives on individual differences.* NY: Plenum Press.

UNESCO (1972). Learning To Be: The World of Education Today and Tomorrow, UNESCO, Paris. Accessed December 2010: http://unesdoc.unesco.org/ images/0000/000018/001801E.pdf

Weinberger, A. & Fischer, F. (2006). A framework to analyze argumentative knowledge construction in computer-supported collaborative learning. *Computers & Education*, 46, 71-95.

In: Evaluation in E-learning
Editors: Y. Psaromiligkos, A. Spyridakos et al.
ISBN: 978-1- 61942-942-0

Chapter 6

A DESIGN BASED RESEARCH PERSPECTIVE ON THE DEVELOPMENT OF A TEACHING LEARNING APPROACH FOR PROMOTING EPISTEMOLOGICAL AWARENESS: AN EXAMPLE FROM WEBCOMICS

Rodothea Hadjilouca, Maria Costoullou and Constantinos P. Constantinou [*]
Learning in Science Group, University of Cyprus

ABSTRACT

This study focuses on promoting primary school students' understanding of an aspect of the nature of science regarding the distinction between observation and inference. Observations are descriptive elements about natural phenomena that are accessible through senses (or extensions of senses) and about which several observers can reach consensus with relative ease. Inferences are statements about the mechanisms through which phenomena operate and cannot be resolved only through information that is accessible through senses. A scientist can infer models or mechanisms that provide a self-consistent interpretation for observations of complex phenomena. We have developed an activity sequence that engages students in studying stories using webcomics as the main instructional medium. The main purpose of the study was to investigate the successes and challenges of digital comics as a medium that scaffolds the learning process and engages students in explicit epistemological discourse. The study reports on results from the first implementation of the activity sequence with a class of sixth graders that draw on various data sources.

[*] E-mail: c.p.constantinou@ucy.ac.cy.

1. Introduction–Theoretical Framework

1.1. Epistemological Awareness as a Component of Science Learning

The term *epistemological awareness* refers to understanding the nature of science (NOS), i.e., how scientific knowledge is generated, organized and justified (AAAS, 1993; Driver, Leach, Millar, & Scot, 1996; Flick & Lederman, 2004; Kang, Scharmann, & Noh, 2005; Lederman, 1992; McComas, 2002, NRC, 1996). Appreciating explicit aspects of how scientific knowledge is constructed and therefore developing the ability to effectively engage with socio-scientific issues that affect human life in an increasingly scientific and technological world constitutes a core learning objective of science education (AAAS 1993; Driver et al, 1996, Lederman, 2007; McComas, 2002; NRC 2007; Osborne, Collins, Ratcliffe, Millar, & Duschl, 2003; Sandoval, 2005). At the same time, it constitutes a challenging area for the science education research community since limited attention in conventional teaching practice has been carried out, especially in lower educational levels, and there is a need for developing learning environments that promote epistemological awareness (Akerson & Volrich, 2006, Kang, Scharmann & Noh 2005; Lederman 2007, Sandoval & Morrison 2003). Existing research findings suggest that epistemological awareness is not an emergent outcome of students' mere interaction with conventional science education learning environments (Kang et al., 2005; Lederman, 2007) but it needs direct and explicit teaching (Akerson, Abd-El-Khalick, & Lederman, 2000; Driver et al., 1996; Lederman, 2007) through purposely designed activity sequences (Khishfe & Lederman, 2006; Lederman, 2007; McComas, 2002; Sandoval & Morrison, 2003).

Despite the lack of consensus in the philosophy of science on a single viewpoint of how science operates (Alters, 1997), it should still be possible and useful to help students develop a set of fundamental ideas while excluding the intricacies of the underlying discourse which is not necessarily relevant to school science nor opposes to the aims of helping students formulate an initial version (simplified but prerequisite) account of how science works. This perspective is consistent with the consensus that seems to exist within science education on such a set of ideas (Abd-El-Khalick, Bell & Lederman, 1998; Lederman, 2007; McComas, 2008; Osborne et al., 2003). For example, these ideas include, amongst others, the durable, albeit tentative, nature of scientific theories, the central role of empirical evidence in science and the distinction between observations and inferences. The latter aspect is the focus of the present study.

1.2. Distinction between Observation and Inference

Scientific knowledge is a product of both observation and inference and neither is theory-neutral. Observation and inference are science process skills and both lead to products of the scientific enterprise. *Observations* are descriptive elements about natural phenomena that are accessible through our senses or extensions of them and about which several observers can reach consensus with relative ease (Lederman, 2007). *Inferences* are statements about phenomena that are not accessible through our senses, but a scientist can infer models or mechanisms that provide a self-consistent interpretation for observations of complex phenomena (Lederman, 2007). The process of inferring is an attempt to explain or speculate about a number of observations (interpretation).

Learners cannot observe all phenomena thus they cannot always draw on observations to explain phenomena. Therefore, learners need to distinguish between observation and inference and eventually develop the skills of observing and inferring in order to construct explanations for phenomena (Hanuscin & Rogers, 2008). Research evidence suggests that students tend to hold naive understandings about the distinction between observation and inference (Khishfe & Abd-El-Khalick, 2002; Khishfe, 2008; Khishfe & Lederman, 2007; Papadouris, 2009). Specifically, they fail to appreciate the role of human creativity in the development of science, by mostly viewing science as a rigidly structured field that follows a predefined process of deducing scientific theories (directly) from careful observation (Kang et al., 2005; Lederman, 2007; Papadouris, 2009).

1.3. Teaching and Learning about the NOS

A main finding that has emerged from the existing research on teaching and learning about the NOS refers to the significance of students' systematic engagement in explicit epistemological discourse, as an important factor facilitating their understanding of the NOS (Clough, 2006; Lederman, 2007; McComas, 2002). Additionally, the literature provides useful insights into possible ways of integrating this in science teaching. Specifically, it suggests three possible approaches, as follows: (a) *nonintegrated (de-contextualized) activities* in which the elaboration of NOS aspects is disconnected from the science content (e.g., Lederman & Abd-El-Khalick, 2000; Bell, 2008; among others), (b) *integrated (contextualized) activities* that seek to embed the epistemological discourse in the context of the elaboration of the science content (e.g., Khishfe & Lederman, 2006; Walker & Zeidler, 2003; Zeidler, Walker, Ackett, & Simmons, 2002), (e) engaging students in epistemological discourse while *studying episodes drawn from the history of science* (e.g., Abd-El-Khalick & Lederman, 2000; Kim & Irving, 2010; Olson, Clough, Bruxvoort, & Vanderlinden, 2005; Rudge & Howe, 2009; Solomon, Duveen, Scot, & McCarthy, 1992; Straits & Nichols, 2007; Yip, 2006).

1.4. Webcomics

Comics have appeared as a new, promising instructional medium which provides a context for Clark and Paivio's (1991) dual coding theory, which supports the importance of imagery and narration in cognitive operations. Pairing written and visual elements is a key reason for using comics in education (Mayer & Moreno, 1998). The interaction of pictures and text that tell a story humanize a subject and in consequence create an intimate connection between students and characters of a comic story (Morrison, Bryan & Chilcoat 2002; Versaci, 2001). Therefore, their use in education has the potential to re-engage reluctant readers and unmotivated students and also to provide scaffolding towards understanding difficult concepts and ideas (Keogh & Naylor 1999, Morrison, Bryan & Chilcoat 2002, Yang 2003).

Webcomics is a plurimedia environment that combines text and imagery as well as hypermedia and streaming elements. Webcomics can be an interactive means for studying and for self-assessment. Children having read some pages of the webcomic book could be asked to respond to an online assessment test, e.g., a test that has the form of unfinished dialogues among the characters. Also, creating a webcomic book has the advantage of infinite canvas, easy access, wide distribution and the potentials of hypermedia (McCloud, 2006). For all the above reasons, the webcomic medium appears as a promising instructional medium for science education in favor of engaging students in explicit epistemological discourse by utilizing authentic stories from the history of science in order to develop students' epistemological awareness. The above issues need research and in this context the ability for producing innovative teaching and learning materials is provided.

1.5. Design Based Research as a Framework for Developing Research-informed Learning Environments

Design-based research (DBR) is an umbrella of methodologies that encompass theoretically framed, empirical research of learning and teaching based on particular designs for instruction. DBR serves the dual goal of developing effective learning environments and using such environments as natural laboratories for studying learning and teaching and it therefore, facilitates the binary interaction of research and innovation and the study of theoretical claims in situ. DBR is viewed as engineering field (Brown, 1992) that includes the following components: iteration, emergent phenomena, variation in developmental trajectories of intellectual growth, customization of teaching, outcome, climate and system variables (Collins, Joseph, & Bielaczyc, 2004). The DBR paradigm is scientifically and technologically oriented since it is framed in terms of empirical investigations of teaching and learning processes with respect to particular designs for instruction and simultaneously pursues in developing effective learning environments and using them to study learning and teaching (Sandoval & Bell, 2004).

DBR provides a methodical framework for modifying the intervention as it unfolds (in order to respond to the dynamic, contingent nature of decision making during teaching, and thus make the intervention "work" in a complex setting) without placing at risk the need for empirical control that leads to theoretical connections and provides answers to the question "what makes a specific intervention successful in a specific setting?". The extent to which interventions can be refined to meet their theoretical expectations, reveals valuable insight on the validity of theoretical claims. Hence, DBR can inform theoretical understandings of teaching and learning without having to exclude or ignore the complexity of learning environments (a common problem with controlled experiments).

DBR appears to have important advantages. Firstly, it contains explicit mappings of theoretical conjectures and assumptions; as a result the constraints of ecological validity are made transparent (generalizability). Secondly, it provides rich descriptions of "the context" in a way that intangible aspects of teaching interventions become problematized and thus accessible (e.g., how to organize a classroom discussion, how to lead a group dialogue, etc). Thirdly, it bridges laboratory studies of learning/teaching with studies of complex instructional interventions. Fourthly, it allows the isolation of variables at one level without ignoring the complexities introduced by the remaining levels of a complex system such as schooling (e.g., individual, group interactions, classroom environments, school settings, learning communities, systemic constraints, social priorities and specifications, etc).

The present study was implemented in the DBR framework and specifically applied a first cycle of a curriculum development process proposed by Papadouris and Constantinou (2009) that encompasses a series of steps for developing science curricula as outlined in the following figure.

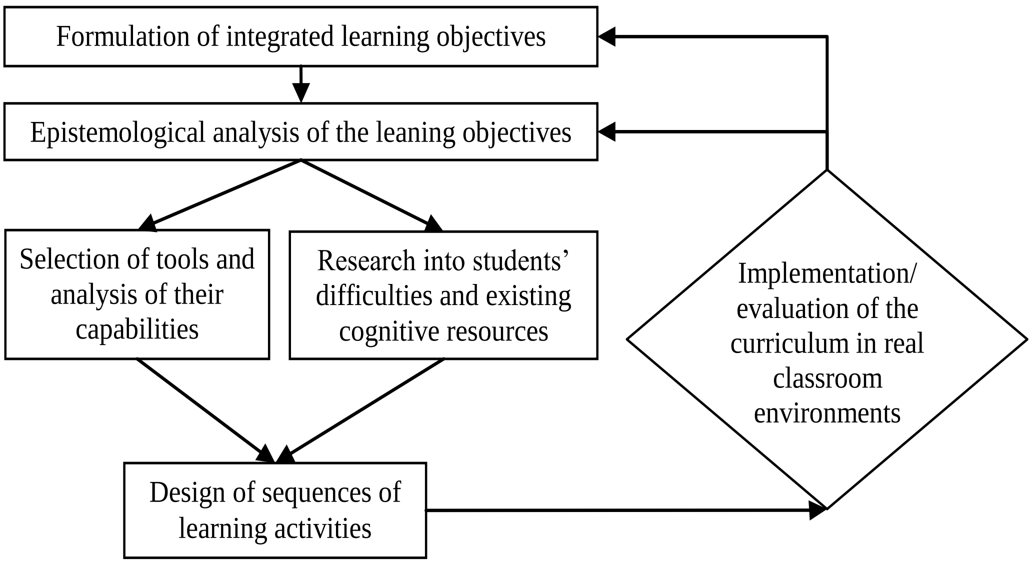


Figure 1. The methodology for the development of science curricula proposed by Papadouris and Constantinou (2009).

The above methodology is aligned with and largely relies on the DBR paradigm (Barab, 2006; Brown, 1992; Collins, et al., 2004; Confrey, 2006) and was adapted in order to fit to the present study's expectations.

Research Question

The present study was interested in developing computer-enhanced teaching and learning materials that utilize webcomic narratives, as a context for initiating epistemological discourse about NOS. The study aimed to answer the following research question: How can webcomics be used in the science classroom in order to engage students in epistemological discourse regarding the distinction between observation and inference?

Research Design

The research was carried out in four phases where we followed a DBR approach, according to the following figure.

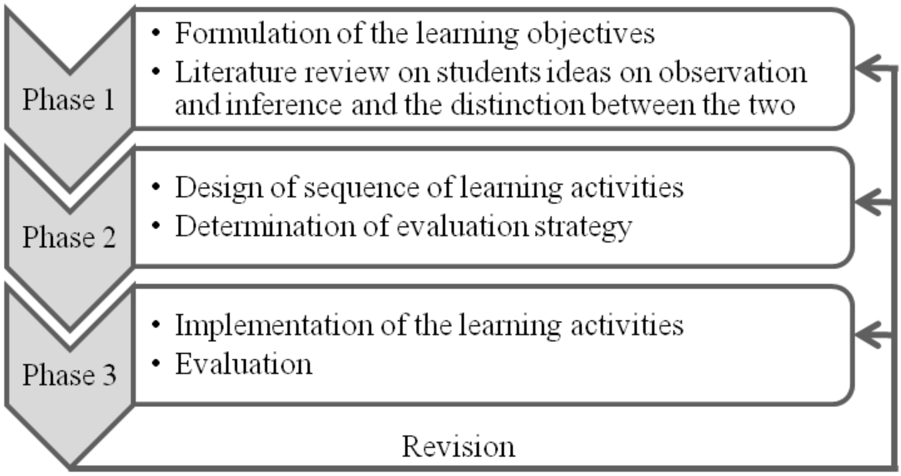


Figure 2. Research design.

SYNOPSIS OF THE LEARNING MATERIALS

Learning Objectives

Students were intended to appreciate specific facets of the distinction between observation and inference: (a) observations as descriptive statements about natural phenomena that are accessible through our senses, (b) inferences as invented statements that provide explanatory frameworks for observations, (c) that it is possible to formulate a number of alternative inferences for a specific observation; disagreement on inferences is a productive scientific activity, (d) that disagreement on observations needs to be resolved through repetition of the data collection process and associate procedures, (e) that the validity of inferences can be assessed based on the extent to which they can account for observations, (f) the distinction between discovery and invention and associate the former with observations and the latter with inferences, (g) the idea that while scientists draw on observations they also need to use their creativity to formulate theoretical interpretations.

Learning Process

Students work in groups of 2-3 members and are guided by a constructionist activity sequence that guides them to reflect on and further develop comics within a content of scientific stories. The whole teaching and learning approach falls under the Physics by Inquiry pedagogy (McDermott et al., 1996). Within this pedagogical framework teaching does not include any lecture, but is conducted through small, autonomous, group investigations. The teacher acts as a coordinator of group and class semi-socratic dialogues by posing questions and identifying difficulties encountered by students and helping them to overcome these difficulties by initiating appropriate (epistemological, in this case) discourse.

ComicLab

The learning activities were designed so as to be carried out using the ComicLab (http://webcomicbookcreator.com/). ComicLab is a windows software application appropriate for young learners to easily create digital comics offering a number of capabilities. Using ComicLab, students can construct their own multimedia comic stories by inserting backgrounds, characters, dialogue boxes and adding hyperlinks, sounds and videos into the text balloons. ComicLab provides infinite canvas while panels as well as characters and borders can be maximized, minimised, flipped horizontally or vertically and rotated. Text can be formatted and comic stories can be saved and further edited. Users can easily export their comic stories as interactive flipping books and/or as PDF files. In addition, comic books can be published on the World Wide Web.

Synopsis of the Activity Sequence

The activity sequence included three parts. In the *first part* students use a learning guide that helps them learn how to use the tool and become familiar with its capabilities.

The *second part* consists of a nonintegrated set of activities that were adapted from literature suggestions (Bell, 2008; Papadouris, 2009), which address differences between observation and inference. In this activity students observe patterns of fossil tracks, make observations and inferences and in sequence revise their observations and inferences as more of the fossil is gradually exposed. An incomplete set of comic strips is provided presenting the story of four children that are trying to interpret patterns of fossil footprints. Each group is supplementary given written paper instructions in a stepwise form for running and completing the webcomic story.

During this exploration characters are engaged in a discourse and try to define observation and inference and distinguish between the two. Students navigate through the comic book story and participate in the characters' dialogues by expressing their agreement or disagreement with them. Students' opinion is explicitly asked at various points where they are asked to complete dialogue balloons of a character or add a comic strip. In this way learners do not hesitate to express their opinion since if it might later be concluded to be wrong, then the critique will be referred to the comic character (Stephenson & Warwick, 2002). Additionally, stop signs are placed at specific points of the comic story where each group should discuss their opinion with the teacher before moving on.

In the *third part* of the learning activities, students are given the opportunity to create from scratch and share their own comic stories (Morisson et al., 2002). Groups are given specially structured stories concerning scientists' efforts in inventing interpretations for their observations of natural phenomena. Each group studies one of the following stories that are provided in paper form: (a) Why did dinosaurs disappear?, (b) Why do rosehips grow only in the Hionistra peak?, (c) Why did eels become extinct in Cyprus? Next, students are instructed to create a webcomic that narrates the story and also specific criteria are given to them concerning specific points someone reading the comic should be able to learn about. At the end of this activity each group presents the comic to their classmates. The whole process provides further opportunities for elaboration on the targeted objectives through explicit epistemological discourse, which is intended to help students connect the processes of observing and inferring with the scientific work.

RESEARCH PROCEDURE

Sample

The activity sequence was implemented with a class of sixth grade students (N=17) of an urban elementary school in the school's computer lab and lasted four 80minute sessions.

Figures 3 and 4. Pieces of the webcomic used in the second part of the activities.

Data Sources

The research is evaluated with respect to students' beliefs and attitudes concerning webcomics as a learning medium and students' awareness of the distinction between observation and inference. Multiple sources of data were collected prior to, during and after the implementation through: groups' webcomic stories, semi-structured interviews with some of the participants (approx. 30%), videotaped lessons, teacher's notes and non-participant observer's notes on students' interaction with webcomics.

Data Analysis

Data were organized and gradually processed until the formation of specific aspects of successes and challenges of webcomics in science learning. These aspects were used so as to obtain empirical grounding from the various data sources. The reliability of the results was promoted through crystallization of data from multiple sources for each aspect.

Results and Discussion

Aspect 1. Development of Epistemologically Informed Views Concerning the Distinction between Observation and Inference

Analysis of data from the students' webcomic book stories (third part of the activity sequence), led to the conclusion that there was some improvement in the learners' understanding. Students could identify observations and interpretations for a natural phenomenon in every story. Specifically, they were able to rephrase information extracted regarding the addressed aspects using suitable words to express observations (e.g. see, notice, observe) and interpretations (e.g. believe, think, maybe). Additionally, pictures inserted were related to the dialogue in each story. Thus it could be inferred that students understood that interpretations (text) should be based on observations (pictures inserted). The potential of webcomics in pairing written and visual elements was of vital importance for the promotion of the specific learning objectives regarding observation and inference.

Aspect 2. Engagement in Epistemological Discourse

During students' exploration of the incomplete comic book story where comic characters were trying to distinguish between observation and inference (second part of the activity sequence), it was observed that all groups were actively engaged in scientific discourse about the addressed aspects. Also, in the activity where they were asked to compose their own stories (third part of the activity sequence), students were also engaged in epistemological discourse concerning the characters' discussion and which information should each character communicate regarding the two aspects. A representative example from a group dialogue follows:

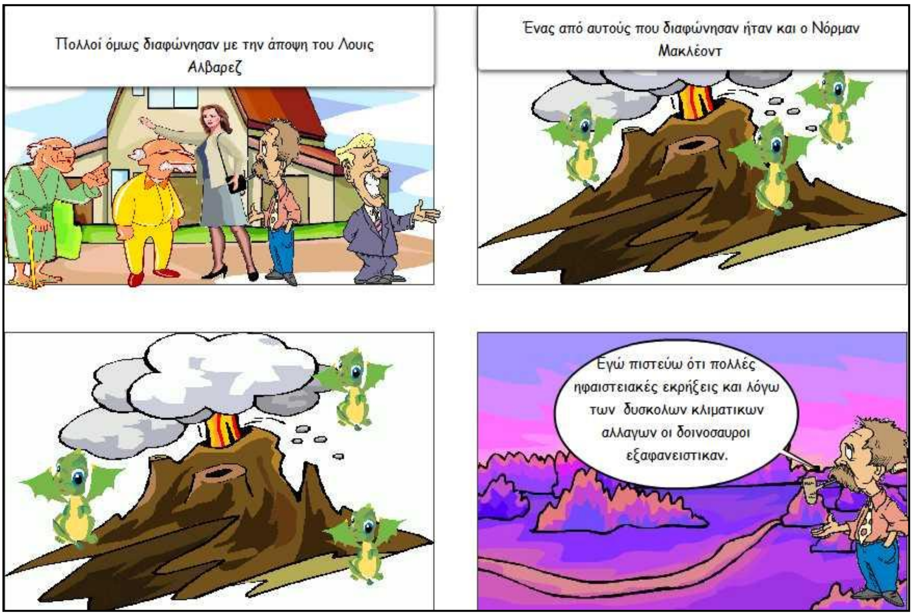
Student 1: (while typing text in a comic character's speech balloon) I can see ....can you help me?

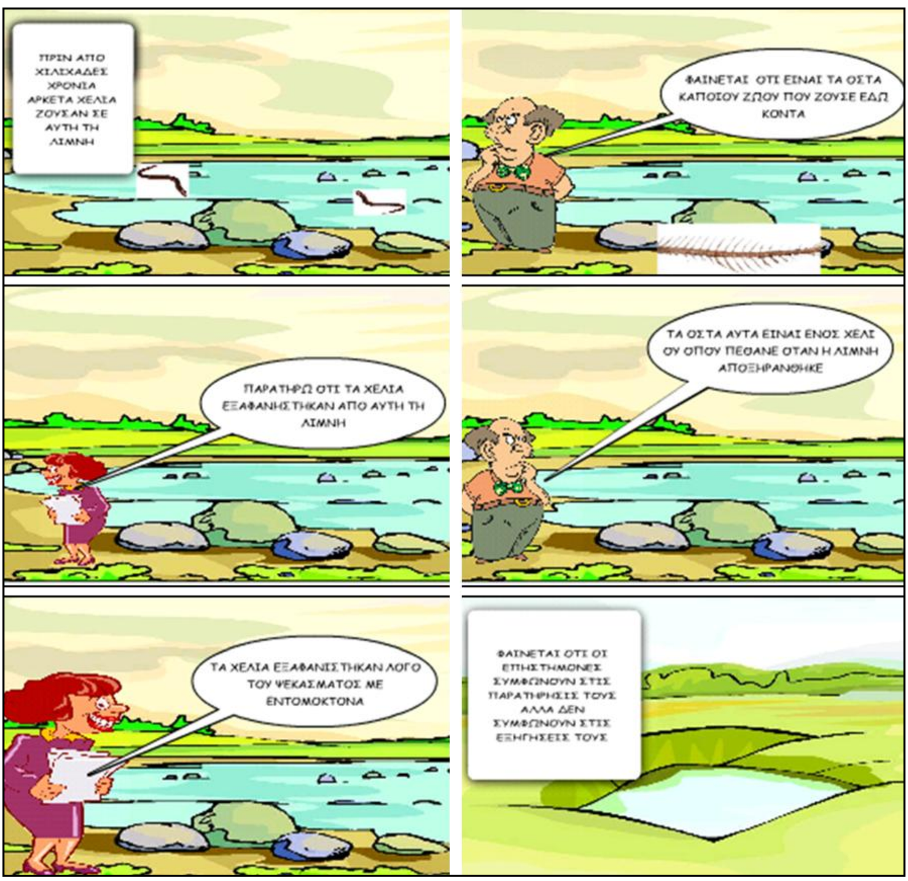
Student 2: Write the following: In the picture I can see the feet of a frog....

Student 1: But, they are not feet! We are not certain that what we observe are the feet of a frog.

Student 3: These are tracks.

Student 1: If we say that these are the feet of a frog it would mean that we are interpreting what we observe. In order to make an observation we should state only what we can see and in this case these are just prints or tracks.





Figures 5 and 6. Pieces from students' stories constructed in the third part of the activities.

Aspect 3. Development of Analytical and Critical Thinking Skills

During the activity where students constructed a story, each member of the group had to explain to his/her peers the reasoning for choosing a particular character to be added in the webcomic or for the selection of a specific background, so as to support the information obtained from the story in order to sustain the scientists' observations or interpretations. Students' debates while constructing a story provoked their thinking and helped in sustaining a productive critical discourse within the group. A representative excerpt from the independent observer's notes follows:

> "The pictures they want to use are related to the story they have read. The fact that the girls in the group are justifying the use of a particular picture and not another indicates that they feel the need to persuade their classmates about their choice." (non participant observer's notes).

Aspect 4. Stimulate Metacognition

Retelling a scientific story had the advantage of determining what is important from their readings in relation to the learning objectives. The construction of a comic story required students to identify key ideas, summarize and reorganize information in a webcomic book format where understandings of the addressed NOS aspect would have been reflected. Therefore, it can be concluded that comic book activities in the science classroom could prompt self-explanation and in consequence develop metacognitive skills. Video data document group discussions where students tried to recall information from the lesson, concerning the differences between observation and inference, in order to decide what information from the story they had read can be classified as observation and what can be classified as interpretation. One such example follows where students appreciate that scientists should interpret based on their observations of the phenomenon under study and they base the structure of their comic story based on this element.

Student 1: We should erase this sentence that we wrote and replace it with an observation and next with an interpretation of that observation

Student 2: But here we have already written an observation.

Student 3: I also agree that we should erase it.

Student 2: Ok, we should write "I observe that this plant grows only in high peaks since there it is colder"

Teacher: Have you written about the scientists' observations concerning the flora of rosehips?

Student 1: Yes, we wrote their observations and no we will write their interpretations for these observations.

Aspect 5. Motivating

Evidence from students' responses during interviews and also from video data, support that utilizing webcomics in science teaching motivates students during the lesson and promotes students' active engagement within the learning process. Data derived from non participant observer notes state that *"…all students in this group want to actively participate in the construction of the webcomic story"*. Additionally, a representative quotation of a student interview transcript indicating webcomics' motivating potential is the following:

"During the lesson the atmosphere of the classroom was positive and you felt that you wanted to learn more..." Furthermore, some students proposed that webcomics should be used in more subject lessons and justified their opinion by expressing that utilization of webcomics in the learning process makes the lesson more interesting, entertaining and fascinating than other forms of learning.

Researcher: Would you like to propose how webcomics could be used in other lessons?
Student: During the Greek Language subject.
Researcher: How could it be used in the Language subject?
Student: Instead of reading texts in books we could read webcomics.
Researcher: Is there any other way in which webcomics could be used?
Student: In Maths.
Researcher: How?
Student: Webcomics' characters instead of just texts could give problems and we could solve them.

(Student interview)

Aspect 6. Facilitation of Cooperative Learning

Analysis of data derived from all sources supports the argument that webcomics promote learning through cooperation. The purpose of constructing a webcomic story using information drawn from a story of the history of science required the development of positive interdependence among the learners of each group so as to successfully achieve the completion of the assignment.

"After the initiative of the boy on the left, team members have a responsibility according to the requirements of the assignment." (non participant observer's notes)

However, it might be concluded that students' success in group cooperation was a result of each member's character and personality since the teacher reported that in some cases students were not able to work cooperatively within their group. On the other hand, interview evidence supports that webcomics enhance cooperative learning since for example, a student who had expressed difficulties in cooperation with a group, in the particular activities where webcomics were utilised, showed a more positive attitude in the group.

Researcher: Do you think that the use of webcomics in the learning process enhanced group cooperation?
Student: Yes, because I personally find difficulties in working cooperatively with my classmates.... In the activities using webcomics, I felt that it was like playing a game where we all had to work together in trying to win. So, in the webcomics activities we had to work cooperatively in order to understand certain things.

(Student interview)

Aspect 7. Capabilities of the Webcomic Tool

The Webcomic tool provides capabilities that enhanced students' learning in the science classroom. The capability of selecting background, characters and pictures that are related to the context of their story provided an opportunity for learners to be creative and precise in what they wanted to communicate in relation to observations and interpretations mentioned in the stories they had read. In the webcomic format this is better facilitated by integrating text with pictures. Also, the characters' size could be changed according to the requirements of the student-creators. The following two excerpts from video transcripts provide evidence for this conclusion.

Excerpt 1

"The student on the left is reading the story to the group: 'According to this scientist this finding supports the opinion that dinosaurs disappeared when a meteorite crashed on earth'. He continues to say: 'We should insert this picture showing a meteorite'"

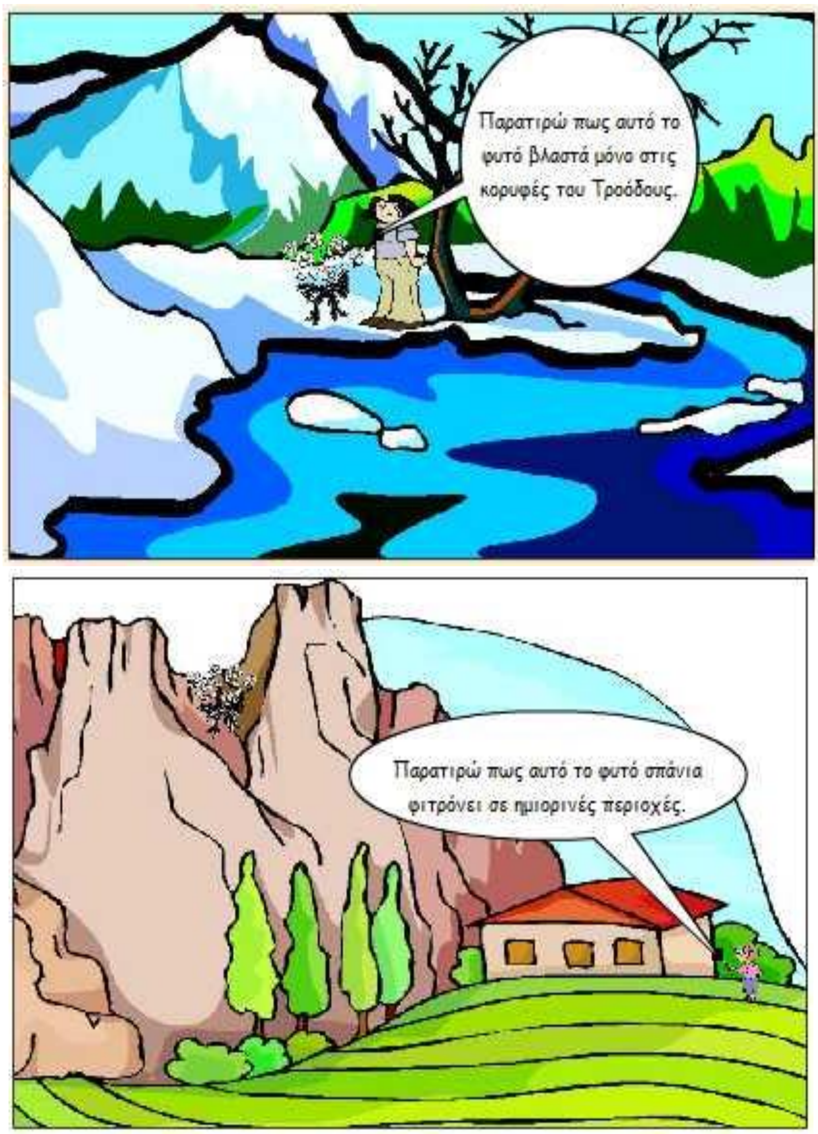


Figure 7. A representative example of a group's story illustrating how students tried to keep the essence.

Excerpt 2

Student 1: We should insert a picture with a dark background.
Student 2: Why?
Student 1: Because when a meteorite crashes there is darkness everywhere.

However, this capability had also led to students' attention being distracted since some groups wasted time on selecting backgrounds even though a particular folder including pictures related to the story was at their disposal. On the contrary, some groups that had expressed their disappointment on the limited number of pictures each folder included and were allowed to search for pictures in a folder containing a greater variety of pictures were not destructed but managed to complete their assignment. Nevertheless, some students encountered a technical difficulty with the software application while they were trying to insert backgrounds or characters to their stories which resulted in disappointment.

An additional capability of the webcomic tool is that its format helped students to purposely simplify their narratives so as to retain the essence. Students discussed in their groups what is important to be included in the dialogue of the webcomic and consequently were refining their understanding concerning observation and inference. Non participant observer's notes state:

> 'The girl is saying to the boy showing a paragraph in her handout: 'We should not include all of these!'
>
> The boy on the left answers: 'Of course we should not. We will just write what is important'.

Conclusion

Utilizing webcomics in the design and implementation of the learning activities developed in this study has facilitated learning in various ways. Consistent with previous research, results have shown that using comics in science education promotes collaborative learning (Keogh & Naylor, 2004), motivates (Yang, 2003) and actively engages students in the learning process (Keogh & Naylor, 1999). In addition, the students' involvement with webcomic activities has engaged them in explicit epistemological discourse about the targeted NOS aspect, i.e., the distinction between observation and inference. However, some of the software's capabilities proved to be a restraining factor during the learning process. Improvement of the webcomic tool's functions might prove useful in resolving these difficulties. The learning activities developed are the initial version of a teaching-learning sequence that needs to be more thoroughly investigated with respect to its effectiveness in accomplishing its learning objectives and be refined and re-implemented in a future research study.

ACKNOWLEDGMENTS

Work presented in this paper has been partially supported by the European Union through the Life Long Learning Programme in the project EduComic (ref num 142424-2008-GR-COMENIUS-CMP).

REFERENCES

AAAS. (1993). *Benchmarks for science literacy*. New York: Oxford University Press.

Abd-El-Khalick, F., Bell, R. L., & Lederman, N. G. (1998). The nature of science and instructional practice: Making the unnatural natural. *Science Education, 82*(4), 417-436.

Abd-El-Khalick, F., & Lederman, N. (2000). The influence of history of science on students' views of nature of science. *Journal of Research in Science Teaching, 37*(10), 1057-1095.

Akerson, V. L., Abd-El-Khalick, F., & Lederman, N. (2000). Influence of a reflective explicit activity-based approach on elementary teachers' conceptions of nature of science. *Journal of Research in Science Teaching, 37*(4), 295-317.

Akerson, V. L., & Volrich, M. L. (2006). Teaching nature of science explicitly in a first-grade internship setting. *Journal of Research in Science Teaching, 43*(4), 377-394.

Alters, B. J. (1997). Whose nature of science? *Journal of Research in Science Teaching, 34*(1), 39-55.

Barab, S. (2006). Design-based research: a methodological toolkit. In *The Cambridge Handbook of the Learning Sciences* (Keith R. Sawyer ed., pp. 153-169). New York: Cambridge University Press.

Bell, R. L. (2008). *Teaching the nature of science through process skills: activities for grades 3-8*. Boston: Pearson Education Inc.

Brown, A. L. (1992). Design experiments: Theoretical and methodological challenges in creating complex interventions. *Journal of the Learning Sciences, 2*, 141-178.

Clark, J. M. & Paivio, A. (1991). Dual coding theory and education. *Educational Psychology Review, 3(3), 149-170.*

Clough, M. P. (2006). Learners' responses to the demands of conceptual change: considerations for effective nature of science instruction. *Science Education, 15*(5), 463-494.

Collins, A., Joseph, D., & Bielaczyc, K. (2004). Design research: theoretical and methodological Issues. *Journal of the Learning Sciences, 13*(1), 15-42.

Confrey, J. (2006). The evolution of design studies as methodology. In *The Cambridge Handbook of the Learning Sciences* (Keith R. Sawyer ed., pp. 135-151). New York: Cambridge University Press.

Driver, R., Leach, J., Millar, R., & Scot, P. (1996). *Young people's images of science*. London: Open University Press.

Flick, L. B., & Lederman, N. G. (Eds.). (2004). *Scientific inquiry and the nature of science*. Dordrecht: Kluwer Academic Publishers.

Hanuscin L. D. & Rogers Park A. M. (2008). Learning to observe and infer. *Science and Children, 45(6), 56 – 57.*

Kang, S., Scharmann, L. C., & Noh, T. (2005). Examining students' views on the nature of science: results from Korean 6th, 8th, and 10th graders. *Science Education, 89*(2), 314–334.

Keogh, B., & Naylor, S. (1999). Concept Cartoons, teaching and learning in science: an evaluation. *International Journal of Science Education, 21*, 431–446.

Keogh, B., & Naylor, S. (2004). Children's ideas children's feelings. *Primary Science Review, 82*, 18-20.

Khishfe, R., & Abd-El-Khalick, F. (2002). Influence of explicit and reflective versus implicit inquiry-oriented instruction on sixth graders' views of nature of science. *Journal of Research in Science Teaching, 39*(7), 551-578.

Khishfe, R., & Lederman, N. (2006). Teaching nature of science within a controversial topic: integrated versus nonintegrated. *Journal of Research in Science Teaching, 43*(4), 395-418.

Khishfe, R. & Lederman, N. (2007). Relationship between instructional context and views of nature of science. *International journal of Science Education, 29(9), 939-961.*

Khishfe, R. (2008). The development of seventh graders' views of nature of science. *Journal of Research in Science Teaching, Online first.*

Lederman, N., & Abd-El-Khalick, F. (2000). Avoiding de-natured science: activities that promote understandings of the nature of science. In W. F. McComas (Ed.), *The Nature of Science in Science Education* (pp. 83-126): Kluwer.

Lederman, N. G. (1992). Students' and teachers' conceptions of the nature of science: a review of the research. *Journal of Research in Science Teaching*, *29*(4), 331-359.

Lederman, N. G. (2007). Nature of science: past, present, and future. In S. K. Abell & N. G. Lederman (Eds.), Handbook of Research on Science Education (pp. 831-879). New Jersey: Lawrence Erlbaum Associates.

McDermott, L. C., & the Physics Education Group (1996). Physics by Inquiry, New York: Wiley.

Kim, S. Y., & Irving, K. E. (2010). History of science as an instructional context: student learning in genetics and nature of science. *Science & Education, 19*(2), 187-215.

Mayer, R. E. & Moreno, R. (1998). A split-attention effect in multimedia learning: Evidence for dual processing systems in working memory. *Journal of Educational Psychology, 90, 312-320.*

McCloud, S. (2006). *Making Comics*, Northampton MA Kitchen Sink Press Inc.

McComas, W. (2008). Seeking historical examples to illustrate key aspects of the nature of science. *Science & Education, 17*(2), 249–263.

McComas, W. F. (Ed.). (2002). *The Nature of Science in Science Education. Rationales and Strategies* (Vol. 5). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Morrison, T. Bryan, G. & Chilcoat, G. (2002). Using student-generated comic books in the classroom. *Journal of Adolescent & Adult Literacy, 45*(8), *758 - 767.*

NRC. (1996). *National Science Education Standards*. Washington, DC: National Academy Press.

NRC. (2007). *Taking science to school: learning and teaching science in Grades K-8* (Duschl, R. A., Schweingruber, H. A., & House, A. W. ed.). Washington D.C.: The National Academies Press.

Olson, J. K., Clough, M. P., Bruxvoort, C. N., & Vanderlinden, D. W. (2005). *Improving students' nature of science understanding through historical short stories in an* *introductory geology course*. Paper presented at the Eighth International History, Philosophy, Sociology & Science Teaching Conference (IHPST), Leeds, UK.

Osborne, J., Collins, S., Ratcliffe, M., Millar, R., & Duschl, R. (2003). What "ideas-about-science" should be taught in school science? A Delphi study of the expert community. *Journal of Research in Science Teaching, 40*(7), 692-720.

Papadouris, N., & Constantinou, C. P. (2009). A Methodology for the Integration of Computer-Based Learning Tools in Science Curricula. *Journal of Curriculum Studies, 41(4),* 521-538.

Papadouris, N. (2009). *Combined development of understanding energy mechanisms and decision-making skills and their connection to the nature of science*. Unpublished PhD Thesis (ISBN 978-9963-689-26-2), Department of Educational Sciences, University of Cyprus (in greek).

Rudge, D., & Howe, E. (2009). An explicit and reflective approach to the use of history to promote understanding of the nature of science. *Science & Education, 18*(5), 561-580.

Sandoval, W. A. (2005). Understanding students' practical epistemologies and their influence on learning through inquiry. *Science Education, 89*(4), 634– 656.

Sandoval, W. A., & Bell, P. (2004). Design-based research methods for studying learning in context: introduction. *Educational Psychologist, 39*(4), 199-201.

Sandoval, W. A., & Morrison, K. (2003). High school students' ideas about theories and theory change after a biological inquiry unit. *Journal of Research in Science Teaching, 40*(4), 369-392.

Solomon, J., Duveen, J., Scot, L., & McCarthy, S. (1992). Teaching about the nature of science through history: Action research in the classroom. *Journal of Research in Science Education, 29*(4), 409-421.

Stephenson, Ph. & Warwick, P. (2002). Using concept cartoons to support progression in students' understanding of light. *Physics Education, 37, 135-141*.

Straits, W. J., & Nichols, S. E. (2007). Using historical nonfiction and literature Circles to develop elementary teachers' nature of science understandings. *Journal of Science Teacher Education*, *18*(6), 901-912.

Versaci, R. (2001). How Comic Books Can Change the Way Our Students See Literature. *One Teacher's Perspective English Journal, 9(2), 61 - 67.*

Walker, K. A., & Zeidler, D. L. (2003). Students' understanding of the nature of science and their reasoning on socioscientific issues: a web-based learning inquiry. *Paper presented at the Annual Meeting of the National Association for Research in Science Teaching*, Philadelphia, PA.

Yang G. (2003). *Comics in Education*. Online version of the final projectproposal for Masters of Education degree of the author at California State University at Hayward. Retrieved from www.humblecomics.com/Comicsedu/index.html.

Yip, D. Y. (2006). Using history to promote understanding of nature of science in science teachers. *Teaching Education, 17*(2), 157-166.

Zeidler, D. L., Walker, K. A., Ackett, W. A., & Simmons, M. L. (2002). Tangled up in views: beliefs in the nature of science and responses to socioscientific dilemmas. *Science Education, 86*(3), 343-367.

In: Evaluation in E-learning
Editors: Y. Psaromiligkos, A. Spyridakos et al.
ISBN: 978-1- 61942-942-0

Chapter 7

TAILOR-MADE OR UNFLEDGED? EVALUATING THE QUALITY OF ADAPTIVE E-LEARNING

Christina M. Steiner[1*] and Dietrich Albert[1,2]
[1]Graz University of Technology, Knowledge Management Institute, Graz, Austria
[2]University of Graz, Department of Psychology, Graz, Austria

ABSTRACT

Adaptive learning systems (ALE) are designed towards the main objective of tailoring learning content, system feedback and appearance to individual users – according to their preferences, goals, knowledge, and other characteristics. The implementation of a successful adaptation process is of course demanding. Adaptation is not good per se and poor realisations of adaptation may lead to disappointed users who may reject or disable adaptation mechanisms. This is why the evaluation of ALE needs to be a fundamental and integral part of their development. Evaluation should address the questions whether adaptation works on principle, whether it really improves the system, whether it leads to more effective learning, whether users prefer the adaptive features, etc. The main challenge in evaluating ALE lies in their core characteristic – adaptivity, which results in individual experiences and interactions with the system for each individual user. The attempts of dealing with this challenge are diverse. This chapter provides an overview on existing and suggested methods for evaluating adaptive e-learning. Strengths and weaknesses of the current evaluation approaches are elaborated and relevant topics in the user-centred evaluation of ALE are discussed. The evaluation methodology developed in GRAPPLE, an EC-funded project aiming at developing a generic responsive adaptive personalised learning environment, is outlined as a case study on evaluating adaptive e-learning. In GRAPPLE the concept of 'adaptation quality' is adopted and conceptualized in terms of covering different aspects of adaptive e-learning experiences that are addressed for a holistic evaluation of adaptive e-learning. In sum, the chapter aims at increasing awareness for the importance of careful and properly designed evaluation of ALE. We believe that the thorough consideration of the quality of adaptation and the use of evaluation approaches on the basis of mathematical-psychological models and expertise are the ingredients for a sound investigation of the

[*] E-mail: {christina.steiner, dietrich.albert}@tugraz.at

benefits of adaptive e-learning and thus, for contributing to the overall notice, growth, and spread of ALE.

1. INTRODUCTION

Adaptive learning systems (ALE) acknowledge the fact that learners who get in touch with a learning system or a course are very different – featuring different goals, starting conditions, knowledge, experience, or interests. ALE try to respond to this diversity by overcoming the traditional 'one size fits all' approach and are designed and developed towards the main objective of tailoring information provided by the system, the systems' feedback and appearance to the needs of the individual user (e.g. Brusilovsky, 1999; 2001; De Bra, Aroyo, & Cristea, 2004). An ALE, like any other adaptive system, automatically alters aspects of the interface, functionality, or presented information. For doing that, the system gathers and processes information about the user in order to establish a user model, a representation of the user and his/her characteristics in the system. The user model is built based on implicit or explicit user input. User characteristics represented in the user model may be knowledge and competence, goals, background and experience, preferences, interest etc. (e.g. Brusilovsky, 1996, 2001). This information serves adaptation decision making for tailoring to the diverse needs of different individuals or groups of learners and their changing needs over time. The adaptation may, for instance, consist in adaptive selection of learning content or activities, adaptive presentation of information, or adaptive navigation support etc. (for the latest taxonomy on adaptation techniques see Knutov, De Bra, Pechenizkiy, 2009). Figure 1 presents a visualisation of the adaptation process consisting of user modelling and adaptation decision making.

ALE has a great potential to increase learning performance and satisfaction (e.g. Leutner, 1992). The realisation and implementation of a successful and effective adaptation process is, however, a challenge. If the underlying theoretical structures or the inferences made by a system are invalid, user modelling and/or adaptation decision making might not be appropriate and might end up in inappropriate adaptation. As a result, the adaptation or guidance provided by the system might be actually worse than no guidance (De Bra, 2000). This is why the evaluation of adaptive systems needs to be a fundamental and integral part of their development. Defining and realising a sound, reliable and useful evaluation methodology may even be considered as one of the challenges of designing an ALE as such (Van Velsen, Van der Geest, Klaasen, & Steehouder, 2008).

This chapter aims at creating awareness for the important role of evaluating the quality of adaptive e-learning and is structured as follows: First, the challenges of evaluating ALE are pinpointed. Then, evaluation approaches for adaptive systems are discussed and an overview of the current situation of evaluating adaptive systems is given. The most common evaluation methods and evaluation topics are presented. The concept of adaptation quality is presented as an evaluation topic covering different aspects of adaptation and thus, building a suitable basis for a comprehensive evaluation of the benefits of adaptivity in learning systems. The evaluation methodology for investigating different perspectives and aspects of adaptation developed in GRAPPLE, a European project aiming at the development of a generic, responsive, adaptive personalised learning environment, is sketched as a methodological case study. Finally, the chapter concludes with a wrap up and an outlook on future research.

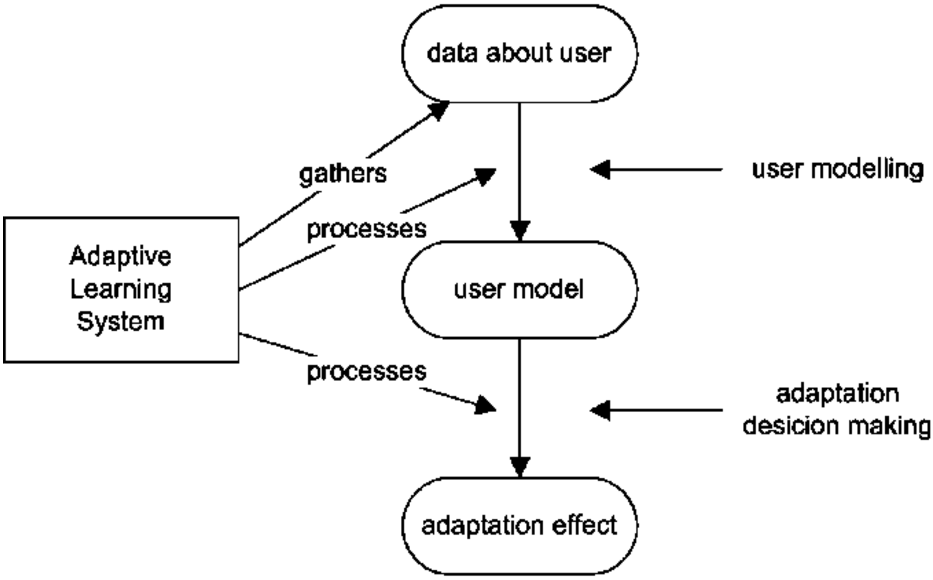


Figure 1. Adaptation process in adaptive learning systems (adapted from Brusilovsky, 1996).

2. THE CHALLENGES IN EVALUATING ADAPTIVE LEARNING SYSTEMS

Evaluation of ALE needs to address the question whether adaptivity improves the system, whether users prefer the adaptive version and features, whether the exploitation of adaptation and adaptation choices are able to improve interaction, and whether adaptation leads to improved learning (e.g. Gena, 2005). Thus, the evaluation of user-adaptive systems in general and of ALE in particular, needs to follow a user-centred approach with users as main addressees of the application and thus as main sources of information.

In the evaluation of e-learning, in general, a distinction needs to be made between content and interface of a system. These two aspects need to be differentiated in evaluation in order to draw meaningful conclusions (Gena, 2005). This is an issue relevant for the evaluation of learning systems in general, whether they are adaptive or not; and it is a shortcoming identified in evaluation research. In studies on the educational effectiveness of distance learning the effects of the technology or media and the learning task are mixed up (e.g. Clark, 1994). As a result of confounding these two aspects, it can actually hardly be determined which effects a particular technology has and which effects are due to a certain learning task or content. In the context of an adaptive system it needs moreover to be taken into account that the learning content provided to different users will be different (Gena, 2005). It might be the case that one and the same learning system provides different courses with different learning contents and probably also applying different adaptation strategies. As a result, the evaluation outcomes when investigating adaptive learning will always be made up by the combined effect of the learning system itself and the concrete learning content and adaptation strategy applied.

The main challenge of evaluating an adaptive system lies in its core characteristic, i.e. adaptivity. Adaptation results in individual experiences and interactions with the system for each individual user. And exactly this inclusion of personalised output complicates the evaluation of ALE (Van Velsen et al., 2008). The personalised learning delivery through an ALE makes the evaluation of the system's effectiveness difficult. Gena and Weibelzahl (2007) characterise evaluation approaches and frameworks of adaptive systems as being still in an exploratory phase, with a strong need for new approaches by combining different techniques and identifying metrics for appropriately assessing adaptivity. Due to the insufficient consideration of adaptation in existing evaluation frameworks, current evaluation approaches still and oftentimes apply basic evaluation tools without taking explicitly into account adaptive systems' features (Paramythis, Totter, & Stephanidis, 2001). It is, however, this very issue of adaptivity that needs to be thoroughly examined in order to give evidence of the added value of adaptive learning and to justify the effort of creating adaptive functionality.

3. Evaluation Approaches for Adaptive Systems

In the evaluation of adaptive systems so-called 'with and without adaptivity evaluation designs' are best known and widely applied, where an adaptive instance of the system is compared to a non-adaptive counterpart (e.g. Höök, 1998). This type of comparison is of particular interest in order to investigate the effectiveness of adaptation, e.g. the benefit of adaptive learning in case of an ALE. The problem with this approach is that adaptation is naturally designed into the system and constitutes its core part, such that the non-adaptive version most probably cannot be optimal in any way. Furthermore, no insights are provided into why the adaptive system is better (or not) than its non-adaptive version. Another option is to compare the adaptive system with a different, non-adaptive (or adaptive) system. This approach, however, confounds the specific effects of (a certain) adaptation with the differences in design or look and feel of the different learning systems. As a result, benefits in the evaluation results for the adaptive system cannot be traced back solely to the adaptation technology of the system. A comparison to traditional teaching is also possible; this way of comparison, though, naturally features very different contextual conditions. A fair means of comparison can be reached if an adaptive system is built in terms of adding adaptive functionality to a conventional system, e.g. a learning management system (LMS). This is for example the case in the GRAPPLE project (www.grapple-project.org), where the ALE is integrated with different LMS but also exists as stand-alone learning platform. In this way, a 'with and without' approach can be realised, where the LMS with integrated adaptive functionality is contrasted with the LMS only. In this way, both conditions feature the same general design, such that differences in evaluation outcomes can better be attributed to adaptation.

Another, quite sophisticated and elaborated version of a with and without approach for evaluating the effectiveness of adaptive system behaviour has been taken in the context of adaptive digital educational games by the adoption of a so-called yoked-control design (Harmatz & Lapuc, 1968; Kickmeier-Rust, Marte, Linek, Lalonde, & Albert, 2008). This design was applied for investigating the effectiveness of adaptive problem solving support

(Augustin, Hockemeyer, Kickmeier-Rust, & Albert, 2011; Kickmeier-Rust, & Albert, 2010). The main idea of the yoked-control is to have comparable pairs of participants, whereby one individual of each tuple receives an adaptively tailored treatment, while the second individual of the tuple receives exactly the same treatment as the first one, i.e. independent from his/her behaviour. In this way, identical and comparable conditions are created, but only one group of learners is provided with actual adaptive functionality. Consequently, an experimental design can be created that realises a methodologically fair comparison for evaluating the effects of adaptive functionality.

Another common evaluation approach is to consider adaptivity as an integral part of the system, an approach that does not depend on the presence of and comparison with a non-adaptive counterpart (Oppermann, 1994). Such a non-comparative approach of evaluation will be criteria-based, i.e. identifying basic criteria for ascertaining whether the adaptive system works or provides an added value. It is then determined through evaluation how well the ALE actually meets each criterion. The problem in this case is that it is not possible to assess the degree to which the different factors contribute to success or failure.

More recent attempts for evaluating adaptivity are so-called approaches of modular or layered evaluation. The main idea behind this approach is that adaptation should not be treated as a monolithic or singular process behind the scenes. Following a layered evaluation approach, the aim is to break down adaptation into its constituents or components for separate evaluation, rather than treating it as one singular process. Two main and distinct phases of adaptation can be distinguished on a high level (compare also Figure 1) and have been modelled for layered evaluation (Karagiannidis & Sampson, 2000; Brusilovsky, Karagiannidis, & Sampson, 2001, 2004; Totterdell & Rautenbach, 1990):

a) Interaction assessment – referring to the system’s conclusions and inferences concerning user-computer interaction and thus, to the user characteristics represented in the user model.
b) Adaptation decision making – referring to the specific adaptations that are selected and realised based on the results of the assessment phase.

In current evaluation practice, these distinct aspects of adaptation are often not taken into account but rather the adaptation is evaluated as a whole. Usually, this is done by comparing the whole adaptive application with its non-adaptive counterpart. If the evaluation results argue for a successful adaptation, a reasonable conclusion is that both phases have been successful. In case of an unsuccessful adaptation, however, no information is available whether both or only one (and which) of the phases have been unsuccessful (Brusilovsky, 2001). In an approach of layered evaluation the success of adaptation is decomposed and investigated on different layers reflecting these two main phases of adaptation – i.e. by evaluating separately the assessment phase (“Are the conclusions drawn by the system concerning user characteristics and stored in the user model valid?”) and the adaptation decision making (“Are the adaptation decisions based on the assessment results valid and meaningful?”). On the first layer, the user modelling process is evaluated. This can be done for instance by user tests, or by experts monitoring users during their work with the system and comparing experts’ estimations of user characteristics with the conclusions drawn by the system and stored in the user model (Brusilovsky, 2004). Users may also be asked to evaluate whether the system’s conclusions reflect their actual characteristics and needs. At the second layer the adaptation decision making is evaluated, which can be done through user testing applying specific scenarios, e.g. evaluating knowledge-based adaptation by direct knowledge assessment, or goal-based adaptation by giving the user a particular goal. The aim of evaluation is then to identify whether the adaptation is helpful given the goal or knowledge level. Moreover, users and experts may be asked to evaluate whether the adaptation contributes to the quality of interaction (Brusilovsky et al., 2004).

Brusilovsky et al. (2004) demonstrated the application of layered evaluation by revisiting an earlier empirical study on an adaptive learning system (Brusilovsky & Eklund, 1998). While the original evaluation had shown that the adaptation process as a whole had failed to achieve the desired benefit, authors now tried to find answers to the question for the reasons. Instead of evaluating the effect of adaptation as a whole, the layered evaluation provided the opportunity to make a step-by-step evaluation of the different layers of adaptation. Thereby, the two main stages of interaction assessment and adaptation decision making were distinguished. Thus, the goal of the post-evaluation was to check how well the system could predict the user knowledge level, on the one hand, and to investigate whether the adaptation in terms of adaptive link annotation was meaningful. It could be determined that the interaction assessment process worked reasonably well, but the decision to use adaptive link annotation turned out to be an inappropriate method of adaptation for the respective group of learners and their navigational strategy. In conclusion, while users' learning progress could be validly assessed by the system, another method of adaptation for the given students group should have been applied and/or a group of students who can benefit from the existing adaptation method should have been sought to properly investigate the benefit of the respective way of adaptation.

Other modular approaches to adaptive system evaluation propose and define more granular evaluation layers or modules. Paramythis, Totter, and Stephanidis (2001), for example, identify seven adaptation modules, which are comprised of one or more adaptation stages. Each module can be addressed in evaluation separately or in combination:

Module A1 - interaction monitoring, interpretation/inferences, and modelling. The evaluation goal is to ensure that the system's models inferred from interaction assessment are optimal.

Module A2 - explicitly provided knowledge and modelling. This is similar to A1, but without automatic assessment of interactions.

Module B - adaptation decision making. The goal of evaluating this module is to ensure that adaptation decisions are appropriate and accepted.

Module C - applying adaptations. This is complementary to module B and may refer to timeliness, obtrusiveness, user control of/over adaptation – factors that can be considered as directly contributing towards acceptance of module B.

Module D1 - modelling and transparent models. Here the goal of evaluation is to ensure that users' perception of the models match their actual state.

Module D2 - adaptation decision making and transparent adaptation. The goal is to ensure that the recommendation/adaptation of the system matches the users' needs.

Module E - automatic adaptation assessment. This is to evaluate whether the system shares the same view/perception on the success or failure of adaptations as the users.

Weibelzahl (2001; Weibelzahl & Lauer, 2001) proposed an evaluation framework consisting of six steps: correctness of input data acquisition, correctness of inference, appropriateness of adaptation decision, change in system behaviour when system adapts, change of user behaviour when system adapts, and change and quality of total interaction. In an attempt to integrate previous approaches suggested for layered evaluation Paramythis and Weibelzahl (2005) devised a merged framework distinguishing also six layers of evaluation:

- Collection of input data: gathering and assembling user interaction data
- Interpretation of the collected data: processing of the previously collected input data to acquire meaning
- Modelling of the current state of the world: deriving new knowledge about the user and interaction context
- Deciding upon adaptation: decision on the necessity and type of adaptation given the current user model
- Applying adaptation decision: actual realisation of adaptation on the basis of the previous decision
- Evaluating adaptation as a whole: evaluation of the primary adaptation theory, in addition to the individual adaptation stages

In sum, all the suggested approaches for layered evaluation have in common that they “recognise that adaptive systems are similar to each other at some level of consideration” (Brusilovsky et al., 2004, p. 405). The separate evaluation steps or layers distinguished in these approaches refer to information processing steps within the adaptation process or to different components involved in the adaptation process. Such evaluation approaches are in general more demanding and laborious, but entail the benefit of providing more detailed evaluation results, such that possible causes of adaptation failure or success can better be identified. The various frameworks for layered evaluation that have been proposed differ with respect to their granularity level, i.e. the number of steps/components that are distinguished and taken into account for evaluation. The models that distinguish more fine-granular evaluation layers can be related and summed up by the two main phases of adaptation distinguishing interaction assessment and adaptation decision making. The consideration of more detailed evaluation layers corresponds to the identification and modelling of particular evaluation goals and not necessarily to distinguished steps/stages of adaptation. Distinguishing the individual evaluation layers, however, especially on a rather fine grained level, turns out to be quite challenging. This might be the reason for the fact that in evaluation practice this differentiation is often not taken into account (Chin, 2001) and often rather serves as a reference point instead of actually guiding the evaluation process (Brusilovsky et al., 2004). The selection of measurement methods and evaluation techniques appropriate for addressing the different layers appears demanding and complicated. In recent work, Paramythis, Weibelzahl, and Masthoff (2010) present a framework and methods for layered evaluation unifying previous approaches and discuss practical issues in the employment of these methods. Overall, layered evaluation is a reasonable approach of tackling the problem of evaluating adaptivity and can help to identify problems in the adaptation process. The basic level of decomposition into the two main adaptation phases seems a generally practicable and suitable approach. It distinguishes the two main steps of the adaptation process and therefore provides a manageable granularity level, but nevertheless stands out from traditional approaches of evaluation that completely neglect the decomposition of adaptation.

4. The Current Situation of Evaluating Adaptive Systems

When consulting the literature on adaptive systems, there is still a lack of reporting on the empirical evaluation of those systems. Chin (2001) reviewed articles in the field of user modelling and user-adapted interaction and came to the conclusion that there is an insufficient extent of empirical evaluations, and in existing evaluations oftentimes shortcomings with respect to proper design and execution can be recognised. Nevertheless, he identified an encouraging upward trend demonstrated by an increasing amount of reports on evaluation attempts in more recent years. Weibelzahl (2003) also tried to get an overview of publications on evaluating adaptive systems and compiled a synopsis of evaluation studies, with the result that only few of them comply with methodological standards. In a later and even broader review on the current status of evaluating user-adaptive systems, Gena (2005) also denotes the empirical evaluation research on adaptive systems as “very fragmented” (p. 13). Her review yielded that only a portion of papers and presentations included some kind of evaluation. The most recent literature review on the evaluation of adaptive systems was carried out by Van Velsen et al. (2008) and aimed at reflecting on the quality and suitability of the current evaluation practice and at providing suggestions for improvement. Similar to the previous reviews, the authors conclude that evaluation reports often lack empirical value. They suggest orienting evaluation on the iterative design process of an adaptive system – accordingly the goals of evaluation will change from supporting decisions on design, to identifying problems and issues for improvement, to verifying the quality as an adaptive system becomes more and more mature (see Figure 2). The phases suggested by this model can actually be referred to as requirements analysis, formative and summative evaluation (e.g. Gena & Weibelzahl 2007; Harvey, Higgison, & Gunn, 2000). In each phase different evaluation topics or variables will be of particular interest and different methods will be most appropriate.

5. Evaluation Methods

Van Velsen et al. (2008) systematically analysed the evaluation methods used in adaptive system evaluations and identified the most commonly used methods. These are questionnaires, interviews, log data analysis, focus groups and group discussions, thinking-aloud, and expert reviews. In the sequel, an overview of these and further evaluation methods is provided.

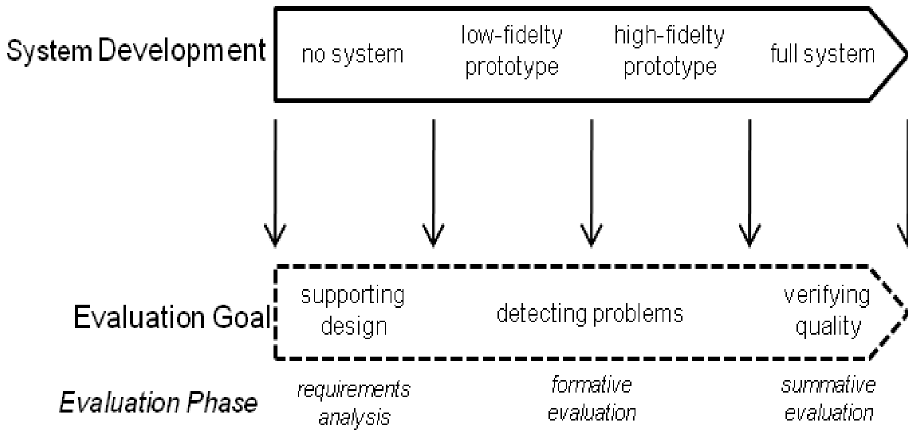


Figure 2. Design process and associated evaluation goals and phases (adapted from Van Velsen et al., 2008).

5.1. Questionnaires

Questionnaires are the most frequently used type of evaluation instrument. They are easy to administer and allow collecting and obtaining feedback from a large number of people within a short time. However, questionnaires are limited in the kind of data they can gather; they are best used for collecting unambiguous, factual data – rather than opinions or descriptions that require explanation or justification. Besides, with the use of questionnaires there is commonly the risk of low response rates. Questionnaires are more objective than interviews because the questions to persons are standardised. But this standardisation prevents explaining any points in questions that respondents might misinterpret (Harvey, 1998).

There are many different forms of questionnaires, from factual to opinion based, from tick boxes to free text responses. To clarify which form is the best and the most useful for a given purpose, it is important to be clear about the aim of the questionnaire. Closed questions generally only allow a limited range of answers. This requires the questionnaire designer to anticipate possible and likely questions. More open questions require more thinking and time on part of the respondent. These questions allow gathering responses that would not have been anticipated. However, they are less likely to be answered and also require additional effort in data analysis.

In the context of software and e-learning evaluation a broad range of questionnaires exist, such as on the user interface and usability (e.g. IsoMetrics – Gediga, Hamborg, & Düntsch, 1999) or on the affective, motivational dimension of user experience (e.g. Instructional Material Motivation Survey – Keller, 1993). Thus, when planning an evaluation one can draw from these existing questionnaires in accordance with the evaluation questions defined. In case of very specific evaluation interests, however, it will oftentimes also be necessary to adapt existing questionnaires or to create new data collection instruments.

5.2. Interviews

An interview is a directed conversation with an individual using a list of different questions. Interviews are suited for the collection of self-reported experiences, opinions, satisfaction, preferences and behavioural motivations (Gena, 2005). Kvale (1996) defines research interviews as “attempts to understand the world from subjects’ point of view, to unfold the meaning of peoples’ experiences, to uncover their lived world prior to scientific explanations” (p. 1).

Different types of interviews can be distinguished according to how structured they are (Harvey, 1998; Wilkinson & Birmingham, 2003):

- A structured (or standardised) interview means that all respondents are asked the same questions, with the same wording and in the same sequence. Often a fixed range of answers is provided. This sort of interview is very restrictive; the data gathered is easy to analyse, code and compare. However, this approach bears the risk of losing important and unanticipated information.
- In a semi-structured interview questions are predetermined, but there is sufficient flexibility that allows shaping the flow of information. An interview guide is used that provides a list of key questions to be covered; the wording and possibly also the order of questions are left to the interviewer’s discretion. The aim of this interview type is that the interaction is kept focused, covering the same topics, but allowing individual experience to emerge. There is room to ask for clarification if questions or answers are unclear, to probe for further information and to include or elicit issues that were initially not considered.
- An unstructured interview is purely informal and conversational and allows full freedom in carrying out the interview and in responding without the need to follow an interview guide. However, this technique requires a lot of time to get systematic information and it can be extremely difficult to analyse the collected data.

Interviews are usually conducted face-to-face; other options are e.g. phone interviews, online interviews, tape interviews or self-recorded audio taped interviews. In sum, interviews are an effective way for giving insight into people’s perception, thinking, and opinion. Ambiguous replies that might have been collected by questionnaire can be clarified in the course of an interview. Interviews allow gathering in-depth data from stakeholders, but are quite effortful, which usually restricts the number of people that can be involved and considered with this evaluation instrument.

5.3. Log Data Analysis

System log data is a step-by-step recording of user interaction with a learning environment (e.g. mouse clicks, menu calls, operations on objects, time spent, etc.). It can be regarded as an indirect form of observation and produces reliable statistical data related to a range of issues such as usage patterns, user preferences, navigational choices, usability, integration strategies and perceived usefulness of courseware. However, data collected through system logs are in general not able to answer 'why' questions and further evaluation is necessary to find explanations for occurring phenomena (Cook, 2002).

5.4. Focus Group

The focus group method is a research technique that gathers data through group interaction on a topic determined by the researcher (Morgan, 1996). A focus group can be regarded as a structured group discussion that addresses a particular topic and is moderated by a group leader. In the context of e-learning evaluations, focus groups are usually moderated meetings with end-users discussing their experience with a learning technology. The moderator can interact directly with the respondents and can also gain nonverbal information to complement verbal answers.

With this method, data can be collected more quickly and cost-effectively and considering a wider range of views than with individual interviews. The main idea behind the focus group method is that group processes can help people to explore and clarify their views in ways that would be less easily accessible in an interview. Well-worked group dynamics can take the conversation in new and often unexpected directions by bringing out users' spontaneous reactions and ideas.

Focus groups are generally suited for obtaining a large amount of interactive information. This method is able to give information about how groups or people think about a specific topic. In interactive system development and evaluation focus groups are a powerful tool to discover what users want from the system and how they operate the system. Therefore, focus groups can in sum be considered as a useful instrument for improving the planning and design of new learning technologies, or for evaluating existing ones.

Aside from face-to-face focus groups, online focus groups have proven to be a viable alternative (Reid & Reid, 2005). These are basically focus groups conducted in online settings (Gaiser, 2008). They allow a relatively easy and inexpensive conduction of group discussions, with the possibility of having participants taking part from different places and even at different points in time. An online focus group may be realised in different ways and with different technologies; synchronous or asynchronous means of communication are applicable.

5.5. Thinking-Aloud

The think-aloud method is characterised by asking people to think aloud, i.e. to express their thoughts, while working on a task or solving a problem (Someren, Barnard, & Sandberg, 1994). Users are encouraged to articulate whatever they are thinking, doing, and feeling while performing a task in question. By doing this, the user just has to concentrate on the task: he/she is asked to give a concurrent account of his/her thoughts without any interpretation or explanation of actions. This is performed nearly automatically and does not interfere with the task performance (Ericsson & Simon, 1993). Observers take notes of everything that the user says, without interrupting, interpreting, or justifying actions. Alternatively, or in addition, the session can be audio- and/or video-recorded. Two different experimental procedures of the think-aloud method can be distinguished (Hannu & Pallab, 2000): concurrent thinking aloud (i.e. during the execution of the task) and retrospective thinking aloud (i.e. after having accomplished a task).

The method of thinking-aloud allows finding out about the specific problems a person has by interacting with an interface by delivering a high-quality, qualitative user feedback. The data gathered in this way allows understanding the cognitive processes of the user by making explicit what is implicitly present in users who perform a specific task. Although there are some drawbacks, like the guidance given to participants, observer influence, and the complexity of data analysis, the richness of the data collected is able to outweigh these constraints and therefore the think-aloud method has a high potential and value in the field of e-learning evaluation (Cotton & Gresty, 2005). It has proven to provide reliable results on explicit intentional behaviour and expectations of people towards technical system designs (Boren & Ramey, 2000).

5.6. Expert Reviews

In the field of usability evaluation, there is a range of evaluation methods that try to predict a system's usability based on models, specifications, or early prototypes (e.g. Ling & Salvendy, 2005; Nielsen & Molich, 1990). These approaches involve rather low costs and are usually carried out by domain or usability experts. Such evaluations provide an efficient way for detecting problems with limited resources; however, they bear the risk of biases if experts are evaluating their own system (Gena & Weibelzahl, 2007).

So-called 'heuristic evaluation' involves a small set of evaluators that investigate a system's user interface and look in order to identify possible problems or violations of design principles. This method furthermore appears suitable for considerations from a pedagogical point of view. The more general approach of 'expert reviews' refers to evaluations carried out by experts who mimic the role of inexperienced users and try to identify in particular usability problems in this way. This is done by simulating system usage and is actually similar to so-called cognitive walkthroughs, where an evaluator constructs task scenarios and role-plays of system usage in order to identify shortcomings. The 'parallel design' approach refers to the procedure of exploring and comparing different design alternatives of a system before implementing and developing one further.

5.7. Observations

Observations involve the observation of users while interacting with an e-learning technology (directly or via video, synchronously or asynchronously). It is suitable for collecting information on navigation issues, anomalies in the content, and impact on behaviour. Observations can be used to ensure efficiency and descriptive data during software initiation and to document program activities and processes. They are suitable in the following cases:

- When wanting direct information.

- When trying to understand an ongoing behaviour, process, unfolding situation or event.
- When written or other data collection procedures seem inappropriate.

For doing observations checklists and structured observation sheets, as well as scaled ratings, or narrative comments for recording users' responses to particular features are commonly applied. Through an observation utilising a systematic methodology and recording, the likelihood of receiving specific and valuable feedback can be considerably increased. A systematic observation can therefore be seen as an approach to quantifying behaviour observed in real context (Bakeman & Gottman, 1986).

5.8. Goal Attainment Scaling

Goal Attainment Scaling (GAS) aims at assessing the attainment of personal unique goals in a quantitative way and originated in clinical contexts (Kiresuk & Sherman, 1968; Kiresuk, Smith, & Cardillo, 1994). In the beginning each participant/person is asked for his/her subjectively important goals - in the present case regarding a learning system. During or after the interaction with the system each person rates on a rating scale to which extent each of his/her subjective goals were reached. The aggregation of these subjective ratings of goal-achievement can be considered as an indicator of the overall usefulness of the learning system.

Even though GAS originated and is currently mainly applied in the therapeutic and clinical context, this technique is also acknowledged as an instrument for educational applications, like evaluating the effectiveness of training programs and monitoring student progress (e.g. Cusick & Ottenbacher, 1994; Roach & Elliott, 2005). This approach can be used for evaluating a broad spectrum of educational interventions or methods/applications. When building a highly personalised environment that is adaptive to individual needs and goals of the users, the main aims are optimal adaptive and individualised support of the users in order to maximise motivation and learning. Thereby, the users of the system will normally have multiple individual goals or expectations. The GAS method can be applied in this case as follows: Users are asked to define their goals in terms of the functionalities of the learning environment. Another variant would be to have goals pre-defined by the evaluator. During or after the application the users rate to what extent they feel that the specific goals were reached and how important they consider the individual goals. Both indices together can be used for a weighted index of the quality of adaptation.

5.9. Analysis of Textual Data

The text evolving from learning activities such as email, computer mediated conferencing, or chat can be analysed for evaluation purposes. This evaluation method is suitable for assessing student participation and interaction, and the effect of teachers' interventions on learner's thinking (Cook, 2002). Besides, the text-based technologies are also suitable for examining the quality of contributions and the extent of interaction that take place.

5.10. Cost Analysis

Cost analyses assess the costs of creating, implementing or using a new program (Sewell & Marczuk, 2004). They also may improve understanding of program operation and tell what levels of intervention are most cost-effective. Factors such as authoring and administration time, technical support, hardware costs, programming, etc. are included in cost analyses. Any educational intervention does not only create costs, but also an outcome. Thus, in general a cost analysis is done through assessment of the cost effectiveness by measuring costs (e.g. personnel, initial investment, technical support, operating costs, etc.) against outcomes (motivation, learning, transfer). This kind of cost-benefit analysis is usually extremely difficult, as benefits can be intangible and hard to quantify in cost terms.

5.11. Simulations, Scenarios and Prototypes

In the first stages of system design there are a couple of methods available that are suitable for first formative evaluations by gathering initial feedback on design from users – to identify possible mistakes and aspects for potential improvement (for an overview see e.g. Gena, 2005). These approaches are intended for getting initial hints that can be used to inspire further steps towards implementation but are not sufficient for a complete system evaluation. One variant of such evaluation approach is a so-called ‘Wizard of Oz simulation’, where the user interacts with an emulated system that acts through control of the experimenter. Prototypes provide part of the features of the intended system and can in principle be interactive and software-based, but also static and paper-based. Scenario-based designs only rely on textual or visual descriptions explaining and illustrating usage scenarios in a step-by-step manner.

5.12. Information Retrieval and Filtering Metrics

In the context of information retrieval and information filtering systems, evaluation metrics have been developed in order to investigate performance quality of such systems (e.g. Herlocker, Konstan, Terveen, & Riedl, 2004). The objective of these systems is to extract user-relevant content from the overall set of contents, i.e. content adaptation. Evaluation therefore addresses the selection process and measures developed for this purpose are in principle also applicable to ALE. The measures used are commonly based on a dichotomic classification of content according to whether it is relevant or not relevant to the user or learner.

For example, the precision of a content selection process refers to the accuracy of content selection and is assessed as the ratio between user-relevant content and all content presented to the user. It provides information on the accuracy of content selection or recommendation. A slightly different measure is the so-called recall, which is given by the ratio between user-relevant contents and the whole content available in the system. This measure indicates how good the system is in finding valuable information. Utility metrics and error rate uncover and address cases of content that was selected/recommended but was not relevant or likable for the user and cases of content that was not suggested but would have been relevant.

5.13. Application of Evaluation Methods in the Field of Adaptive Systems

The above described methods have been applied in the context of evaluating adaptive systems to a diverse extent. We want to quote here some prototypical examples of studies, where specific methods have been used or are further elaborated in the context of adaptive systems. This existing work provides an insight to the possible challenges and benefits to be faced with the application of a specific evaluation method and to the implications for the use in the domain of adaptive learning. Further references to a whole range of empirical studies on the evaluation of adaptive systems can be found in the review of Van Velsen et al. (2008), as well as in Weibelzahl (2003) and Chin (2001).

Van Velsen, van der Geest and Klaasen (2007) tried to identify specific strengths and weaknesses of interviews, thinking-aloud, and questionnaires in a study on the usability of an adaptive system. They examined what kind of (unique) issues interviews, questionnaires, and thinking-aloud are able to elicit. Questionnaires and interviews proved good methods to identify usability issues. Compared to these methods, thinking-aloud could be shown to elicit more comments from participant on perceived relevance of the information presented by the system, which may be considered as an important condition for user acceptance. The authors come to the conclusion that the use of a combination of thinking-aloud with interviews or questionnaires would be the most recommendable approach for evaluating an adaptive system.

Log data analysis was used by Kickmeier-Rust et al. (2008) for investigating the effects of individualised feedback and interventions provided during problem solving tasks of an adaptive educational game (Kickmeier-Rust & Albert, 2010). Through the analysis of users' interactions in the game logged by the system, the level of appropriateness of each intervention, learners' response times after system feedback, and the actions performed following different types of interventions could be determined. Overall, the method proved suitable for a detailed analysis of adaptive micro-level feedback and interventions on problem solving behaviour and learning performance.

In the field of adaptive systems, methods that realise an early involvement of users in evaluation and that may inform design and development have also been researched. Georgiakakis, Psaromiligkos, and Retalis (2006) have proposed a method for carrying out scenario-driven usability evaluation of adaptive and personalisable web-based systems. Design knowledge of experts in terms of design patterns is thereby used as a basis for the evaluation. In this way, a scenario-based heuristic usability evaluation can be performed in an effective way and by minimizing the preparatory phase for evaluation. The authors also provide a concrete example demonstrating the different steps of this approach. Masthoff (2006) proposes an approach that was inspired by the Wizard of Oz method. Applying the respective approach, users are taking over the role of the adaptive system or a component of it in a first step, and judge the performance of others in a second step. Through this procedure of early involving users in evaluation, their actions and rationale can be exploited for the design of adaptive algorithms as well as for the definition of criteria for further evaluation. This method, which is called 'the user as wizard', has been illustrated by a number of case studies showing its concrete application for different types of adaptive systems (Masthoff, 2006).

Current research tries to integrate and combine evaluation approaches used in the context of personalised information retrieval and of adaptive hypermedia system in order to develop a comprehensive evaluation methodology for adaptive systems (Lawless, Connor, & Mulwa, 2010). While evaluation of information retrieval focuses on accuracy, numerical performance, and effectiveness of retrieving information or content, evaluation of adaptive hypermedia systems largely involves users as integral part of evaluation. A combination of both evaluation traditions is assumed to provide a broad spectrum of assessment of an entire adaptive system. The practical application and validation of this unified approach to evaluation in future research shall prove its potential and significance.

6. Evaluation Topics

Chin (2001) identified in his review four broad topics that are commonly addressed when evaluating adaptive systems: adaptive-hypermedia/information filtering, student modelling, plan recognition/mixed-initiative interaction, and user interfaces/help systems. These actually refer to different types of adaptive systems and their according adaptation mechanisms. Van Velsen et al. (2008) identified evaluation topics in the field of adaptive system evaluation on a more specific level of variables assessed in evaluations. The most commonly addressed evaluation variable turned out to be usability, an evaluation topic referring to the actual use of an adaptive system. This was followed by appropriateness of adaptation, which concerns the output of an adaptive system. Furthermore, perceived usefulness and intention to use resulted as frequently addressed evaluation topics. These variables concern the adoption of a system and actually constitute aspects of user acceptance. In the sequel, an overview of these popular evaluation topics for adaptive systems is given. In addition, learning effectiveness is discussed as an evaluation topic that is especially relevant for the evaluation of ALE.

6.1. Usability

Usability is a multidimensional construct that can be examined from different perspectives. A multitude of definitions and attributes of usability can be found in the literature (e.g. Jeng, 2005; Nielsen, 2003). Many definitions and studies on usability relate usability to ease-of-use or user-friendliness and consider it from an interface effectiveness point-of-view. The most popular and widely cited definition is the one of the International Organisation for Standardization (ISO). The ISO 9241 standard defines usability as "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" (1998, p. 3). Effectiveness is the accuracy and completeness with which users are able to achieve certain goals in the environment. Indicators for effectiveness are quality of solution or error rates. Efficiency is the relation between accuracy and completeness with which the users achieve certain goals and the resources used for it. Indicators for efficiency include task completion time and learning time. Satisfaction is the user's comfort and acceptance with the use of the system.

In the context of adaptive systems, discussions have emerged about the usability of such adaptive interfaces. In fact, adaptive systems and their interfaces are at risk of violating central usability principles (e.g. through automatically adapting the interface), such as providing the user with control over the system, making the system and its behaviour predictable, and making its inner workings and principles transparent and understandable (Höök, 2000; Gena & Weibelzahl, 2007). It has also been pointed out that usability of adaptive systems may require a new way of defining and addressing usability, which differs from general usability principles. An aggravating factor is that often the design and development of adaptive systems seem to focus on inference strategies rather than on interface design (Gena, 2005). On the other hand, it has been argued that adaptivity – rather than causing problems concerning usability – can be the solution to usability problems, as one single interface cannot be designed to meet the usability requirements of all user groups of a system (Benyon, 1993). Usability and adaptivity goals need to be taken into account concurrently in the evaluation of an adaptive system; they partly seem to conflict somehow, although under certain conditions it is also possible to match usability and adaptivity objectives (Gena & Weibelzahl, 2007).

6.2. Appropriateness of Adaptation

Appropriateness of adaptation refers, as the wording is speaking for itself, to the question whether the adaptation realised by the system is appropriate and corresponds to the users' needs. Broadly speaking, this involves the question whether user modelling is done in a meaningful and valid way, i.e. whether the inferences drawn by the system to a user's characteristics are correct, as well as the question whether the adaptation decisions taken are appropriate, i.e. whether the actual adaptation is realised suitably and serves and responds to the given user model.

For a detailed consideration of the appropriateness of adaptation and the different steps of the adaptation process a layered evaluation approach should be taken. The processes of user modelling and adaptation decision making are independent, but nevertheless closely connected as the latter one grounds on the interaction assessment of the first phase. For a successful adaptation it is critical that both phases work appropriately. If, for example, the user modelling is poor and leads to incorrect assumptions with respect to user characteristics, the best adaptation will not work. Similarly, a good user modelling will be of no use if the adaptation decisions taken based on this information are wrong or inappropriate. The separate consideration of these two processes as layers of evaluation allows getting a deeper insight into the success of each phase of adaptation, and thus facilitates the identification of problems in case of an unsuccessful adaptation (Brusilovsky et al., 2004).

6.3. User Acceptance

Perceived usefulness and intention to use have been identified by Van Velsen et al. (2008) as variables that are very commonly addressed in adaptive systems' evaluation. These topics constitute important aspects of user acceptance. According to Dillon (2001) user acceptance can be defined as the willingness demonstrated within a user group to employ an information technology for the tasks it is designed to support.

The introduction of a new technological system, such as a learning system, does not inherently translate into individuals actually using the system – independent of how technically sound it is (Hirschheim, 2007). This is, why in the last decades many theoretical approaches have been devised aiming at a better understanding of this issue – and at deriving indications on how to prompt users to successfully adopt a new system.

A variety of theories and research on acceptance can be found in the literature (for an overview see e.g. Dillon & Morris, 1996; Venkatesh, Morris, Davis, & Davis, 2003), ranging from case studies of accepted technologies to the individual psychological characteristics of acceptors and resistors. One well-known theoretical model in this context is the Technology Acceptance Model, which was devised by Davis (1986; Davis, Bagozzi, & Warshaw, 1989). The model is grounded on the assumption that the perceived usefulness and perceived ease of use of a technology strongly correlate with user acceptance. Both are factors influencing users' decision about how and when to use a new technology. As a result, user acceptance is considered as a strong predictor for the intention of using a system and thus, its actual use.

6.4. Learning Effectiveness

In addition to the evaluation topics outlined above, in case of ALE naturally the benefit for learning is of great interest. The added value of adaptive learning experiences for objective learning performance and subjective learning experience is therefore an important topic when evaluating ALE.

Learning effectiveness refers to the objective or subjective knowledge benefit in the subject domain when dealing with an e-learning environment or course. (Note: Learning effectiveness in this sense needs to be clearly differentiated from the effectiveness of learning how to use the software, which constitutes an aspect of usability.) In the context of an adaptive system it needs to be taken into account that the assessed learning performance and effectiveness relates to the specific underlying adaptation strategy.

Objective learning effectiveness refers to objectively assessed learning performance in order to derive a statement about the quality of learning and teaching that the system intends to promote. There are different approaches for investigating objective learning effectiveness. One possible approach uses a specified learning goal as a starting point, based on which it can be investigated whether this learning goal is reached by the use of the system (in terms of a criterion-based evaluation without comparison) or, rather, how effectively it can be reached e.g. in terms of time spent for achieving the learning goal in comparison to other systems/teaching approaches (i.e. comparison-based evaluation). The time spent as a measure of learning effectiveness may also be investigated with respect to mastering problems or tasks during or after the learning phase. It is, though, debatable, whether time spent can and should be used as a valid measure of learning effectiveness. The time spent on learning or a task likely does not give an appropriate impression of the quality of learning as such – a user who spends less time has not necessarily learned more effectively but maybe more superficially, and a user who needs much time does not necessarily learn ineffectively, but possibly very thoroughly. Furthermore, correct and incorrect user actions while learning and interacting with the system may be analysed through log file analysis in order to get information on learning performance. External parameters of success, like grades, etc. may also be used as indicators for learning effectiveness. Another valuable and commonly applied approach is to operationalise learning effectiveness in terms of knowledge gained by using the system. In this way, learning effectiveness is most directly assessed through the actual learning outcome. This requires taking into account and measuring prior knowledge before the learning in addition to assessing learning results after the learning experience with the system. In this way the increase of knowledge can be determined as a relevant variable of learning effectiveness by comparing prior knowledge with performance after the learning phase.

Objective learning effectiveness may be addressed at different levels – targeting either retention or transfer. Retention tests are in general more common and easier to administer. Transfer relates to the learning effectiveness in terms of ability to translate or apply the knowledge acquired to different contexts or tasks (e.g. problem solving) and thus, in general, involves higher order cognitive processes. Another important issue is the consideration and assessment of the sustainability of the achieved learning outcome. In addition to diagnosing learning outcome immediately after the interaction with the learning environment (short-term learning), it is also desirable to investigate and prove learning effects on a longer term, i.e. after a reasonable time interval. This is even more urgent for learning technologies that explicitly aim at supporting lifelong learning (e.g. Knapper & Cropley, 2000). In this case, evaluation should take into account also the aspect of lifelong learning, which may be done through longitudinal studies on learning performance, but also by querying users on the perceived potential of the system to support lifelong learning.

Subjective learning effectiveness refers to the perceived learning effect or benefit due to the interaction with a learning technology. This aspect of learning effectiveness can be investigated using subjective estimations of learning success, which are generally applicable whatever specific knowledge domain or subject matter is concerned. The goal attainment scaling approach constitutes a suitable evaluation method for this purpose. This would mean, that learners are asked to make explicit their subjective goals and expectations regarding learning before the actual learning process and use of the system. After the learning experience learners rate the subjectively perceived extent to which each goal has been reached. Another way to assess subjective learning effectiveness is by using specific scales or subscales of questionnaires or also via interviews.

7. Adaptation Quality as a Comprehensive Evaluation Concept

To emphasize the importance of adaptation in the context of evaluating ALE, in the GRAPPLE project adaptation quality has been conceptualised as a central and comprehensive evaluation topic. This topic is, of course, partially intertwined with the other evaluation topics presented in the previous section.

A fundamental issue that is to be kept in mind in the context of an ALE is that adaptation is not good per se – and poor realisations of adaptations may lead to disappointed users rejecting and disabling adaptation mechanisms (López-Jaquero, Montero, & González, 2008). Adaptation quality needs to be investigated by applying and elaborating state of the art procedures for the evaluation of adaptive systems. We have adapted the concept of adaptation quality in order to represent a comprehensive construct covering different aspects of the quality of adaptation in an ALE. Adaptation quality actually links to the evaluation topic ‘appropriateness of adaptation’ as presented above and commonly addressed in evaluations of adaptive systems. It constitutes, however, a more systematic approach by clearly distinguishing and investigating different aspects playing a role for the appropriateness and quality of adaptation.

7.1. Aspects of Adaptation Quality

When thinking about an ALE, one important aspect of the quality of such a system’s adaptation is, of course, whether it brings an added value for learning. Learning effectiveness (as described above) can therefore be regarded as one indicator for adaptation quality. If an ALE is able to demonstrate benefits for learning this indicates the success of adaptation.

Furthermore, adaptation quality can be considered with respect to subjective aspects and users’ perceptions and feedback relating to adaptation. Relevant questions in this context are, for example, whether the adaptation provided is perceived as helpful for the user or whether it corresponds to users’ preferences. These constitute positive facets of adaptation. Asking users to estimate adaptivity effects, however, might be inadequate and difficult in some situations – users possibly have no anchor of what good or bad interaction or adaptation means, they might not have any experience with the usual and non-adaptive way, and they might not even be aware of the adaptation at all (Gena & Weibelzahl, 2007). So, it is important to care for appropriate conditions when asking for subjective assessments and feedback on adaptivity effects. Users should be made aware of the concept of adaptation in general, and the adaptation process should be transparent and visible to the user. Given these preconditions qualitative user feedback on adaptation quality can provide valuable information.

Possible negative facets of adaptation should also be considered when evaluating the quality of adaptation. The adaptive features provided by a system, e.g. adaptation of presentation or interface, may involve negative aspects for the user – because of the changes that are caused by the adaptation. The cognitive effort that is due to adaptation needs to be taken into account. Cognitive load means that through increasing or decreasing the amount of information presented to the user due to adaptation the load of cognitive processing may change. Adaptation necessarily leads to changes in the user interface, which might also cause additional load for cognitive processing, involving additional cognitive load or discontinuity (López-Jaquero et al., 2008). Discontinuity can be seen as another relevant aspect of cognitive effort, which refers to incidents, where the user is somehow disrupted in his current task and/or forced to divide attention (e.g. between different entities, windows etc.) because of the adaptation. In general, it is of course desirable that an adaptive system does not put additional cognitive load on the learner. On the contrary, with suitable adaptation cognitive load should be leveraged by making the learner aware of relevant information – through the recommendation and selection of appropriate contents and the use of appropriate presentation modes.

To sum it up, relevant aspects of adaptation quality of ALE are:

a) Learning effectiveness
b) Cognitive effort
c) User feedback on adaptation quality (i.e. whether the adaptation realised is helpful, whether adaptation improves the quality of the system, whether adaptation corresponds to users' preferences etc.)

As can be seen, the concept of adaptation quality completely focuses on the benefits adaptation brings to a system. Besides that, also users' general perception of and experience with the user interface and the general acceptance of the new learning technology are critical. Consequently, in addition to adaptation quality, usability and user acceptance are important evaluation topics that should be addressed in a proper evaluation of an adaptive learning system (compare Figure 3).

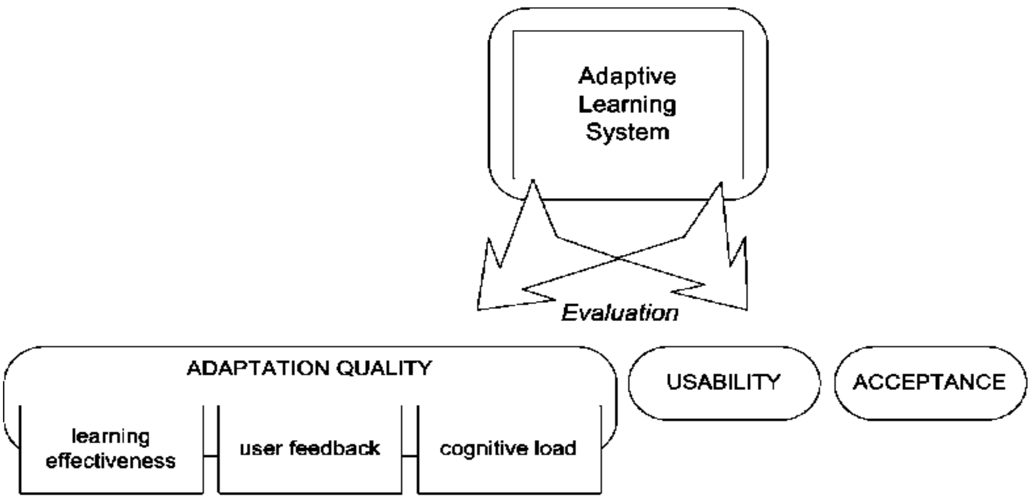


Figure 3. Relevant topics for the evaluation of an adaptive learning environment.

7.2. Assessing Adaptation Quality

For the evaluation of learning effectiveness as an indicator of adaptation quality, objective as well as subjective learning effectiveness may be addressed. Possible approaches on how to measure this evaluation topic have been outlined in the subsection on learning effectiveness above.

For gathering user feedback on the adaptation quality of a system, goal attainment scaling (Kiresuk et al., 1994) is a possible approach. In this way, the extent to which a system is capable of adapting in a way that is perceived helpful and beneficial by the user can be evaluated. This is done through the definition of goals (by the users themselves or, alternatively, by the evaluator) before actually using the system. After system use, the users rate the extent to which their respective goals were reached and how important the specific goals were for them. Alternatively, users can explicitly be queried (through questionnaires, interviews, focus groups) whether the adaptation was helpful for them in accomplishing their task, whether adaptation improved the system, etc. Another option is to analyse thinking-aloud protocols of users collected while working with the system.

Cognitive effort or load "is generally considered a construct representing the load that performing a particular task imposes on the cognitive system" (Sweller, Van Merrienboer, & Paas, 1998, p. 266). Basically, there are three main categories for measuring this mental effort, subjective, physiological, and task- or performance-based indices. The technique of using subjective rating-scales has proven to be useful in instructional research, while physiological approaches did not (Paas, Van Merrienboer, & Adam, 1994). Consequently, for the purpose of evaluating a learning technology, subjective techniques are most suitable. These are grounded on introspections of individuals concerning their cognitive processing and the amount of mental effort expended. This can be done by asking users to report their invested mental effort on a rating scale "by translating the perceived amount of mental effort into a numerical value" (Sweller et al., 1998, p. 267). Such scales contain labels and assigned numerical values with categories corresponding to and ranging from 'very low mental effort' to 'very high mental effort' (e.g. Hart & Staveland, 1988; Marcus, Cooper, & Sweller, 1996; Paas et al., 1994).

8. Evaluation in GRAPPLE: A Methodological Case Study

In this section we want to sketch the evaluation methodology developed and used in the GRAPPLE project for comprehensively evaluating ALE. GRAPPLE aims at bringing together the world of LMS and ALE in order to offer a life-long learning solution (De Bra, Smits, van der Sluijs, Cristea, & Hendrix, 2010). The evaluation methodology foresees that all evaluation activities are carefully aligned with training measures on the GRAPPLE tools and facilities, as an according training on the functionality of the software provided and how to use it is considered as a crucial prerequisite for meaningful evaluation. Through the realisation of sufficient training phases it can, for example, be ensured that users' subjective assessments of usability are not confounded with learnability.

The phases of evaluation are strongly interconnected with the development process (compare Figure 4) and thus, correspond to the model suggested by Van Velsen et al. (2008). Each evaluation phase is characterised by a particular evaluation focus. In the requirements analysis phase the needs and expectations of users towards ALE from an authoring and learning perspective have been elicited through semi-structured interviews (Harrigan, Kravcik, Steiner, & Wade, 2009; Höver & Steiner, 2009).

GRAPPLE offers a generic and open approach to adaptive e-learning and does not only aim at supplying an environment for consuming adaptive courses, but also provides a set of rich authoring tools (GRAPPLE Authoring Toolset – GAT) for modelling knowledge domains and setting up adaptive course structures. As a consequence, not only the aspect of adaptive learning is addressed in evaluation, but also an in-depth evaluation of the GRAPPLE authoring tools is carried out in order to ensure a successful authoring process, which is actually the main prerequisite for the realisation of adaptive learning experiences. Users addressed in this line of evaluation are authors or instructors from university as well as business contexts. They are the target audience of the authoring tools, which shall serve the translation of educational intentions into adaptive story lines. Usability and user acceptance are the main evaluation topics in the evaluation of GAT. Especially in the formative evaluation phase the collection of extensive qualitative feedback allows to identify issues for further improvement that can be fed back into the development cycle. GAT is also addressed in summative evaluation, to give evidence of the impact of software improvements on users' subjective assessments. Topics addressed in the mixed-method evaluation approach for GAT are usability and functionality of GAT and user acceptance. As the main topic of this chapter is not on authoring, but on the adaptive e-learning experiences that are based on it, the evaluation methodology for GAT is not presented here in detail.

The evaluation of the GRAPPLE Adaptive Learning Environment (GALE) also starts in the formative evaluation phase but is focused more intensely in summative evaluation. GALE is in general applicable as a stand-alone learning environment as well as integrated with different open-source and commercial LMS. The evaluation topics addressed are adaptation quality, usability, and user acceptance. A multi-method approach is adopted for evaluations, combining questionnaires and objective learning assessments as well as qualitative methods such as focus groups or thinking aloud. The main aim is not only to gather evaluation outcomes in terms of scores providing a numerical assessment of the quality of adaptation and the perception of the system, but also to gather an understanding of the interaction and thinking processes when working with the learning environment.

A two-tier approach of evaluating GALE has been proposed in order to realise a systematic investigation of adaptation effects, on the one hand, as well as the analysis of the practical usefulness of GALE, on the other hand. Evaluation workshops shall provide the opportunity to realise systematic comparison studies with GALE. They are carried out in laboratory settings under controlled conditions. Deployments in real-life settings shall serve the investigation and demonstration of the significance and applicability of GALE in educational practice. The following subsections present these two evaluation perspectives in more detail.

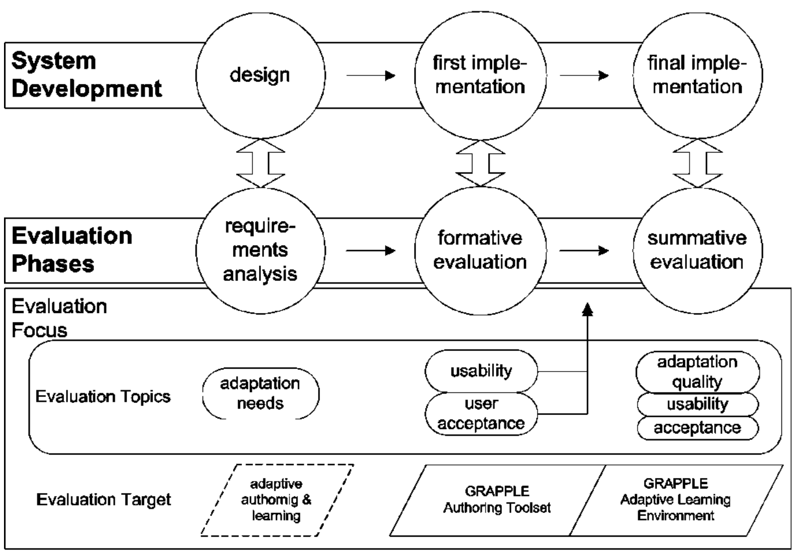


Figure 4. Evaluation phases and focuses in GRAPPLE.

8.1. Comparative Evaluation

Workshops for comparative evaluation purposes consist in a foregoing training introducing the concept of adaptivity and the functionality and use of GALE and a subsequent learning session in which participants are asked to work through a prototypical adaptive course. This learning phase is followed by the evaluation phase in a narrower sense, where data on usability, acceptance, and the different aspects of adaptation quality is collected.

In this scope, comparison studies are conducted by systematically varying certain factors for investigating the effects. One important comparison is of course the investigation of the added value of adaptation. To this end, a comparison between adaptive (experimental condition) and non-adaptive (control condition) learning experiences is conducted through a between-subjects study design. In this case, however, learners in the control group receive no training on adaptivity to avoid frustration. Comparisons between adaptive and non-adaptive learning are realised in terms of presenting the experimental group with an adaptive course in GALE and the control group with a non-adaptive version of the same course with the same learning contents and layout, but without adaptation. This is considered an appropriate and fair way of realising a ‘with and without’ evaluation approach, as except for the presence or absence of adaptation the conditions for learning are the same for both groups. Another approach of comparative evaluation is the comparison between different adaptation strategies or techniques. Moreover, the investigation of probable differences in user opinions and feedback depending on the presentation of GALE as a stand-alone environment and GALE integrated in an LMS is of interest in case of the GRAPPLE project. The different versions might lead to different subjective assessments on the users’ side. In this case, of course prior experience with an LMS has to be taken into account. These comparative evaluation studies would in principle also be doable in a real-world educational context. The realisation of different groups as foreseen in these comparative designs, however, might raise ethical issues, as not all learners are provided with exactly the same conditions for learning.

Furthermore, in GRAPPLE an approach of layered evaluation has been developed, which allows a sound investigation of the quality of adaptation by considering individual components of the adaptation process. To this end, the layered evaluation model of Brusilovsky et al. (2004) has been adopted, which distinguishes the two high-level distinct processes of adaptation: User modelling and adaptation decision making. A possible design for a comparative evaluation study separately investigating these two adaptation layers has been elaborated. This design assumes an adaptive course using knowledge-based adaptation, e.g. adaptive annotation. For evaluating the different adaptation layers three conditions are realised, two experimental and one control condition:

- Experimental condition 1 (Ex1): Tailored adaptation
- Experimental condition 2 (Ex2): Arbitrary adaptation
- Control condition (Contr): No adaptation

For the two experimental groups matched pairs are realised, i.e. creating pairs of participants who have similar characteristics (e.g. gender, background). The experimental condition ‘tailored adaptation’ realises full adaptive functionality, i.e. learners receive adaptation according to their knowledge as inferred by the system. The second group, 'arbitrary adaptation', receives adaptation that is not tailored to the individual learners' knowledge but is rather more or less arbitrary. This is realised in terms of a yoked-control design (e.g. Kickmeier-Rust et al., 2008) by providing each member of this group the exactly same adaptation behaviour as the other member of the matched pair from Ex1 has received. Consequently, participants of Ex2 receive adaptation, but not tailored to their personal knowledge. To realise this experimental condition the implementation of a record-and-play functionality is required. The third group is presented with the learning content without adaptation of the system, while still realising user modelling in the background. A visualisation of the design and the evaluation aspects and hypotheses addressed in this kind of evaluation study is represented in Figure 5.

For investigating the adaptation layer of user modelling the results on a learning test assessing objective learning effectiveness are compared with results of user modelling. If user modelling is successful, then the knowledge inferred by the system for a learner should correspond to his objective performance in the learning test. This can be investigated for the overall group of learners. For investigating adaptation decision making, the results on adaptation quality of the different conditions are contrasted. Learning effectiveness is assumed to be significantly higher for learners receiving tailored adaptation than for the other two groups of learners. Furthermore, cognitive load is assumed to be significantly lower for tailored adaptation than for the other two groups. When contrasting user feedback given in the two experimental groups on adaptation quality, it is expected that results should be better for tailored than for arbitrary adaptation.

In addition to the consideration of the different aspects of adaptation quality, also results for usability and acceptance collected in the different conditions can be compared. This does not serve the evaluation of the individual adaptation layers, but rather only the comparison of the different versions of the system (i.e. conditions). Optimally, the results on these evaluation topics should also be best for the version realising tailored adaptation.

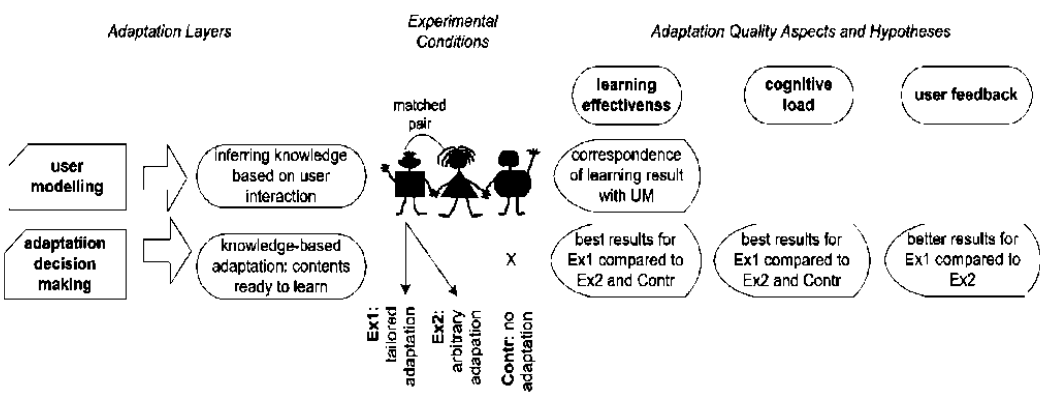


Figure 5. Evaluation study design for a layered evaluation of adaptation processes.

The suggested design allows an appropriate comparison between different groups. The two experimental groups feature equivalent environmental conditions (i.e. both members of one matched pair receive the same adaptation) and therefore allow evaluating whether the provision of adaptation is more effective than receiving no adaptation. Effects of adaptation are isolated from other possible factors that may influence the effectiveness of adaptation, such as the number of changes due to adaptation.

Evaluation outcomes in the comparative evaluations generally need to be considered with reference to the respective course(s) used and the specific adaptation strategies and methods applied therein.

8.2. Real-life Evaluation

Deployments of GALE in real-life settings constitute a continuation of deployments of GAT. Adaptive courses created by instructors with GAT in real-life settings are used as a basis for real-life evaluation of GALE. Hereby, the main objective is to investigate the applicability and added value of GALE in educational practice. In particular learning effectiveness and learners' opinion on the use of GALE in educational practice are considered. The general idea is to convey the learning contents of part of an actual course or lecture, e.g. at university, through an adaptive course. Also in this case foregoing training measures are realised, in order to provide learners the necessary knowledge about adaptivity, and GALE functionality and use. Subsequent to this training, the adaptive course is presented to the learners. During the learning phase, which is realised at distance in a self-regulated manner, a continuous and immediate available support system is provided to learners in order to overcome any technical issues or problems in the use of GALE and thus, to avoid confounding learnability with user feedback in evaluation.

A main difference of these real-life evaluations to comparative evaluations is that firstly, there are no different experimental groups realised and compared, and secondly, the learning contents and general conditions of training and evaluation are not constant over different deployments. As adaptive courses created by different instructors from different backgrounds and institutions and giving different types of lectures are used, there is a large variety in the knowledge domains covered, as well as in the adaptation strategies realised in the courses. If benefits (or drawbacks) of adaptive learning can be identified across this diversity, these outcomes argue for a generalisable effect of adaptive learning experiences with GALE.

The adaptive learning experiences in the real-life settings are generally longer and more intensive than those carried out in the comparative evaluation studies and thus, allow the examination of the different evaluation topics after a longer-term use of and interaction with GALE in educational practice. This refers to usability and acceptance, but also to adaptation quality, in particular learning effectiveness.

In sum, the GRAPPLE evaluation methodology constitutes a comprehensive framework for evaluating the creation and consumption of adaptive courses. Evaluations following this methodology are aligned with the iterative design and development process of a generic ALE, they consider different aspects of the benefits of adaptation and its underlying technology, and they realise systematic comparisons as well as the examination of the practical relevance and value of an ALE.

CONCLUSION

The evaluation of the quality of adaptive learning and its underlying technology is an important part of a sound design and implementation process of ALE. Today, the research in adaptive systems is still in a rather early stage and correspondingly also evaluation approaches taking into account the requisites and challenges of evaluating adaptive systems are not yet fully elaborated (e.g. Van Velsen et al., 2008; Weibelzahl, 2005). When examining existing evaluation approaches, methods, and topics and when planning an evaluation it has to be considered that there is no universally applicable sound evaluation methodology. Each ALE is a unique environment and evaluation plans need to be tailored to the specific purpose of evaluation in each case. The contextual conditions of future educational settings, in which an adaptive learning technology is to be applied, need to be taken into account and reflected throughout evaluation. When reviewing the current evaluation practice one gets the impression that evaluation is often carried out by applying traditional evaluation approaches without explicitly addressing and investigating the system's adaptation and its added-value. Often it seems that evaluation addresses the general benefit and learning effectiveness of a system without explicit and systematic investigation of adaptation or any means of comparison. In other cases, traditional methods of learning programme evaluation are adapted to the level of ALE or are even applied in an unreflected manner. In sum, reports on evaluations of adaptive systems have shown to be quite often of limited empirical value and/or lacking specification of the applied evaluation design and procedure. Nevertheless, a growing recognition and interest for evaluating adaptive systems can be observed. This increasing awareness of the need for evaluation is an important step towards sound evaluation methodologies.

This chapter provided an overview of today's most common evaluation methods and topics in the field of ALE. The evaluation methodology developed in the GRAPPLE project was presented as a methodological case study that realises a comprehensive and multi-perspective investigation of ALE. This methodology considers the investigation and quality assurance of the authoring tools underlying the realisation of adaptive courses as a highly important issue that needs to be taken into account even before addressing the quality of adaptive learning itself. The evaluation approaches addressing adaptive learning in GRAPPLE accommodate for the need of comparative designs for a systematic investigation of the benefits of adaptation, as well as for real-life evaluation in order to demonstrate applicability and significance in educational practice.

Evaluation studies should benefit from the application and triangulation of quantitative and qualitative methods and should be an integral part of every development phase of an ALE. Evaluation needs to accompany the iterative cycle of system design and development from requirements elicitation and support of design decisions to testing the benefit and final achievement of a finished system. As Gena and Weibelzahl (2007) point it out, there is a need for bringing evaluation of adaptive systems to the level of more rigorous methodological approaches of subject sampling, statistical analysis, and experimental design and procedures. Even though big steps forward and towards more and empirically sound evaluations of adaptive systems have been made in recent years, there is still a lot to do. New approaches that allow a sound and comprehensive evaluation of ALE are needed. This can be reached by combining different techniques, by further adapting existing techniques to the purpose of adaptive systems, by developing appropriate experimental designs, by exploring new ways and metrics of assessing adaptation and its quality. In addition to the consideration of evaluation aspects that are relevant for adaptive systems, in general, in case of ALE the learning perspective needs to be emphasised and addressed systematically.

Different evaluation perspectives need to be followed for an ample evaluation of adaptive learning. User-centred evaluation is of course critical in order to gain an understanding about users' experiences with and opinion about an ALE, to elicit ideas for further improvement of the system, and to gather information about the benefit of adaptation. In addition to and even before that, the scientific soundness and value of adaptivity and its underlying methodological approaches need to be evaluated. This means, that the theoretical structures, assumptions, and processes underlying adaptation need to be tested and verified. On the one hand, this concerns the validity of user modelling processes – it needs to be investigated whether the used indicators and rationale for interpreting interaction behaviour in a certain way (e.g. reading time as an indicator for knowledge, reaction time as an indicator for motivation etc.) are valid. On the other hand, this refers to the structures and considerations underlying adaptation decision making (e.g. whether the knowledge domain structures underlying a course adequately reflect the world). We see a great potential for advancing evaluation of adaptive systems through the application of formal psychological frameworks (e.g. Albert & Lukas, 1999; Doignon & Falmagne, 1999). By the use of such theoretical approaches possible ways for continuous and semi-automatic evaluation procedures can be researched. This may involve for instance non-invasive, continuous assessment techniques and the interpretation of user interaction and log-data (e.g. Augustin et al., 2010). Future research in this direction will support evaluation in gaining an in-depth view on adaptivity and a longer-term picture about the value of adaptation.

Finally, sound evaluations of the quality of adaptive e-learning are important for demonstrating the potential and benefit of ALE. Thoroughly planned and designed evaluations can lead to significant evaluation results, which in turn are able to lead to more appropriate and successful adaptive learning systems. Moreover, the provision of sound evidence of the benefit of ALE is highly important in order to give good reason for the costly and effortful implementation of adaptation. As a result, evaluation can strengthen the status of ALE and stimulate interest in and the acceptance and use of adaptive e-learning environments.

ACKNOWLEDGMENT

The work presented in this paper is partially supported by the European Community (EC) under the Information and Communication Technologies (ICT) programme of the 7th FP for the GRAPPLE project (contract no. 215434).

REFERENCES

Albert, D., & Lukas, J. (Eds.). (1999). *Knowledge spaces: Theories, empirical research, and applications*. Mahwah, NJ: Lawrence Erlbaum Associates.

Augustin, T., Hockemeyer, C., Kickmeier-Rust, M. D., & Albert, D. (2011). Individualized skill assessment in digital learning games: Basic definitions and mathematical formalism. *IEEE Transactions on Learning Technologies, 4, 138-148.*

Bakeman, R. & Gottman, J.M. (1986). *Observing behavior: An introduction to sequential analysis.* Cambridge: Cambridge University.

Benyon, D. (1993). Adaptive systems: a solution to usability problems. *User Modeling and User-adaptive Interaction, 3*, 65-87.

Boren, M. T., & Ramey, J. (2000). Thinking aloud: Reconciling theory and practice. *IEEE Transactions on Professional Communication, 43(3)*, 261-278.

Brusilovsky, P. (1996). Methods and techniques of adaptive hypermedia. *User Modeling and User Adapted Interaction, 6*, 87-129.

Brusilovsky, P. (1999). Adaptive and intelligent technologies for web-based education. In C. Rollinger and C. Peylo (Eds.), *Special Issue on Intelligent Systems and Teleteaching, Künstliche Intelligenz*, 4, 19-25.

Brusilovsky, P. (2001). Adaptive hypermedia. *User Modeling and User-Adapted Interaction, 11*, 87-110.

Brusilovsky, P. (2004). Adaptive educational hypermedia: From generation to generation. In *Proceedings of 4th Hellenic Conference on Information and Communication Technologies in Education* (pp.19-33).

Brusilovsky, P. & Eklund, J. (1998). A study of user-model based link annotation in educational hypermedia. *Journal of Universal Computer Science, 4*, 428-448.

Brusilovsky, P., Karagiannidis, C., & Sampson, D. (2001). The benefits of layered evaluation of adaptive applications and services. In S. Weibelzahl, D. Chin & G. Weber (Eds.), *Proceedings of the Workshop on Empirical Evaluations of Adaptive Systems at the International Conference on User Modeling* (pp. 1-8). Freiburg: Pedagogical University of Freiburg.

Brusilovsky P., Karagiannidis C. & Sampson D. (2004). Layered evaluation of adaptive learning systems. *International Journal of Continuing Engineering Education and Life-Long Learning, 14*, 402-421.

Chin, D.N. (2001). Empirical evaluation of user models and user-adapted systems. *User Modeling and User-Adapted Interaction, 11*, 181-194.

Clark, R. E. (1994). Media will never influence learning. *Educational Technology Research and Development, 42*, 21-29.

Cook, J. (2002). *Evaluating Learning Technology Resources*. LTSN Generic Centre. Retrieved April 19, 2005, from http://www.alt.ac.uk/ docs/eln014.pdf

Cotton, D. & Gresty, K. (2005). Reflecting on the think-aloud method for evaluating e-learning. *British Journal of Educational Technology, 37*, 45-54.

Cusick, A. & Ottenbacher, K. (1994). Goal attainment scaling: Continuing education evaluation tool. *Journal of Continuing Education in the Health Professions, 14,* 141-154.

Davis, F. D. (1986). *A technology acceptance model for empirically testing new end-user information systems: Theory and results*. Doctoral dissertation, Sloan School of Management, Massachusetts Institute of Technology.

Davis, F. D., Bagozzi, R. P., & Warshaw, P. R. (1989). User acceptance of computer technology: A comparison of two theoretical models. *Management Science, 35(8)*, 982-1003.

De Bra, P. (2000). Pros and cons of adaptive hypermedia in web-based education. *Journal on CyberPsychology and Behavior, 3,* 71-77.

De Bra, P., Aroyo, L., & Cristea, A. (2004). Adaptive web-based educational hypermedia. In M. Levene & A. Poulovassilis (Eds.), *Web dynamics, adaptive to change in content, size, topology and use* (pp. 387-410). Heidelberg: Springer.

De Bra, P., Smits, D., van der Sluijs, K., Cristea, A. & Hendrix, M. (2010). GRAPPLE: Personalization and adaptation in Learning Management Systems. In *Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications 2010* (pp. 3029-3038). Chesapeake, VA: AACE.

Dillon, A. (2001). User acceptance of information technology. In: W. Karwowski (Ed.), *Encyclopedia of Human Factors and Ergonomics*. London: Taylor and Francis.

Dillon, A. & Morris, M. (1996). User acceptance of new information technology – theories and models. In: M. Williams (Ed.), *Annual Review of Information Science and Technology, Vol.31* (pp. 3-32). Medford NJ: Information Today.

Doignon, J.-P., & Falmagne, J-C. (1999). *Knowledge spaces*. Berlin: Springer.

Ericsson, K. A. & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data.* London: MIT Press.

Gaiser, T.J. (2008). Online focus groups. In N. Fielding, R.M. Lee, & G. Blank (Eds.), *The SAGE handbook of online research methods*. London: SAGE Publications Ltd.

Gediga, G., Hamborg, K.-C., & Düntsch, I. (1999). The IsoMetrics Usability Inventory: An operationalisation of ISO 9241-10. *Behaviour and Information Technology, 18,* 151-164.

Gena, C. (2005). Methods and techniques for the evaluation of user-adaptive systems. *The Knowledge Engineering Review*, *20*, 1-37.

Gena, C. & Weibelzahl, S. (2007). Usability engineering for the adaptive web. In P. Brusilovsky, A. Kobsa, & W. Nejdl (Eds.), *The adaptive web: Methods and strategies of web personalisation. Lecture Notes in Computer Science*, *Vol. 4321* (pp. 720-762). Berlin: Springer.

Georgiakakis, P., Psaromiligkos, Y., & Retalis, S. (2006). The use of design patterns for evaluating personalisable web-bases systems. In S Weibelzahl, A Paramythis & J Masthoff (Eds.), *Proceedings of the Fifth Workshop on User-Centred Design and Evaluation of Adaptive Systems* (pp.440-449).

GRAPPLE – Generic Responsive Adaptive Personalised Learning Environment, http://www.grapple-project.org

Hannu, K., & Pallab, P. (2000). A comparison of concurrent and retrospective verbal protocol analysis. *American Journal of Psychology, 113(3),* 387-404.

Harmatz, M. G., & Lapuc, P. S. (1968). A technique for employing a yoked control in free operant verbal conditioning experiments. *Behavioral Research and Therapy, 6,* 483.

Harrigan, M., Kravcik, M., Steiner, C., & Wade, V. (2009). What do academic users really want from an adaptive system? In G.-J. Houben et al. (Eds.), *User modelling, adaptation, and personalization, LNCS Vol. 5535* (pp. 454-460). Berlin: Springer.

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of experimental and theoretical research. In P. A. Hancock & N. Meshkati (Eds.), *Human Mental Workload* (pp. 139-183). Amsterdam: North Holland.

Harvey, J. (Ed.). (1998). *Evaluation Cookbook.* Learning Technology Dissemination Initiative. Heriot-Watt University. Retrieved November 11, 2008, from http://www.icbl.hw.ac.uk/ltdi/cookbook/

Harvey, J., Higgison, C., & Gunn, C. (2000). Evaluation. In C.A. Higgison (Ed.) *Online Tutoring e-book.* Heriot-Watt University and The Robert Gordon University. Retrieved November 10, 2006, from http://otis.scotcit.ac.uk/onlinebook/

Herlocker, J.L., Konstan, J.A., Tervee, L.G, & Riedl, J.T. (2004). Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems, 22*, 5-53.

Hirschheim, R. (2007). Introduction to the special issue on "Quo Vadis TAM – Issues and Reflections on Technology Acceptance Research". *Journal of the Association for Information Systems, 8*, 203-205.

Höök, K. (1998). Evaluating the utility and usability of an adaptive hypermedia system. *Journal of Knowledge-based Systems, 10*, 311–319.

Höök, K. (2000). Steps to take before IUIs become real. *Journal of Interacting with Computers, 12,* 409-426.

Höver, K. & Steiner, C.M. (2009). Adaptive learning environments: A requirements analysis in business settings. *International Journal of Advanced Corporate Learning, 2*, 27-33.

Internal Organization for standardization (ISO) (1998). *ISO 9241-11: Ergonomic requirements for office work with visual display terminals (VDTs) – Part 11: Guidance on usability*. Geneva.

Jeng, J. (2005). Usability assessment of academic digital libraries: Effectiveness, efficiency, satisfaction, and learnability. *Libri, 55*, 96-121.

Karagiannidis, C. & Sampson, D. (2000). Layered evaluation of adaptive applications and services. In P. Brusilovsky, O. Stock, & C. Strapparava (Eds.), *Adaptive hypermedia and adaptive web-based systems. Lecture Notes in Computer Science, Vol. 1892* (pp. 343-346). Berlin: Springer.

Keller, J.M. (1993). *Instructional Materials Motivation Survey*. Florida State University.

Kickmeier-Rust, M.D., & Albert, D. (2010). Micro adaptivity: Protecting immersion in didactically adaptive digital educational games. *Journal of Computer Assisted Learning 26*, 95-105.

Kickmeier-Rust, M., Marte, B., Linek, S., Lalonde, T., & Albert, D. (2008). The effects of individualized feedback in digital educational games. In *Proceeding of the 2nd European Conference on Game-Based Learning (ECGBL)*. Barcelona, Spain.

Kiresuk, T. J., & Sherman, R. E. (1968). Goal attainment scaling: A general method for evaluating comprehensive community health programs. *Community Mental Health Journal, 4 (6),* 443-453.

Kiresuk, T.J., Smith, A., & Cardillo, J.E. (Eds.). (1994). *Goal attainment scaling: applications, theory, and measurement.* Hillsdale: Lawrence Erlbaum

Knapper, C. K, & Cropley, A.J. (2000). *Lifelong learning in higher education (3rd ed.).* London: Kogan Page.

Knutov, E., Bra, P.M.E. de, Pechenizkiy, M. (2009). AH 12 years later: a comprehensive survey of adaptive hypermedia methods and techniques. *New Review of Hypermedia and Multimedia, 15,* 5-38.

Kvale, S. (1996). *InterViews: An introduction to qualitative research interviewing*. Thousand Oaks, CA: Sage.

Lawless, S., Connor, A.O., & Mulwa, C. (2010). A proposal for the evaluation of adaptive personalised information retrieval. In *Proceedings of the CIRSE 2010 Workshop on Contextual Information Access, Seeking and Retrieval Evaluation.* Milton Keynes, UK.

Retrieved January 26, 2011 from http://www.scss.tcd.ie/~mulwac/cirse2010-workshop.pdf

Leutner, D. (1992). *Adaptive Lehrsysteme: Instruktionspsychologische Grundlagen und experimentelle Analysen [Adaptive Learning Systems: Instructional psychological basics and experimental analyses]*. Weinheim: Psychologische Verlags Union.

Ling, C. & Salvendy, G. (2005). Extension of heuristic evaluation method: a review and reappraisal. *Ergonomia. An International Journal of Ergonomics and Human Factors, 27,* 179-197.

López Jaquero, V., Montero, F., González, P. (2008). Quality of adaptation: User cognitive models in adaptation quality assessment. In V. López Jaquero, F. Montero, J.P. Molina, & J. Vanderdonckt (Eds.), *Computer-Aided Design of User Interfaces VI* (pp. 265-276). Berlin: Springer.

Marcus, N., Cooper, M., & Sweller, J. (1996). Understand instructions. *Journal of Educational Psychology*, *88,* 49-63.

Masthoff, J. (2006). The user as wizard: A method for early involvement in the design and evaluation of adaptive systems. In S Weibelzahl, A Paramythis & J Masthoff (Eds.), *Proceedings of the Fifth Workshop on User-Centred Design and Evaluation of Adaptive Systems* (pp. 460-469).

Morgan, D. L. (1996). Focus groups. *Annual Review of Sociology, 22*, 129-152.

Nielsen, J. (2003). *Usability 101: Introduction to Usability*. Retrieved January 24, 2008, from http://www.useit.com/alertbox/20030825.html

Nielsen, J. & Molich, R. (1990). Heuristic evaluation of user interfaces. In *Proceedings of the SIGCHI conference on Human factors in computing systems: Empowering people* (pp. 249-256). New York: ACM.

Oppermann, R. (1994). Adaptively supported adaptability. *International Journal of Human Computer Studies, 40*, 455-472.

Paas, F.G., Van Merrienboer, J.J., & Adam, J.J. (1994). Measurement of cognitive load in instructional research. *Perceptual and Motor Skills, 79*, 419-430.

Paramythis, A. & Weibelzahl, S. (2005). A decomposition model for the layered evaluation of interactive adaptive systems. In L. Ardissono, P. Brna, & A. Mitrovic (Eds.), *Proceeding of the International Conference on User Modeling*, *Lecture Notes in Artificial Intelligence*, *Vol. 3538* (pp. 438-442). Berlin: Springer.

Paramythis, A., Totter, A., & Stephanidis, C. (2001). A modular approach to the evaluation of adaptive user interfaces. In S. Weibelzahl, D. Chin & G. Weber (Eds.), *Proceedings of the Workshop on Empirical Evaluations of Adaptive Systems at the International Conference on User Modeling* (pp. 9-24). Freiburg: Pedagogical University of Freiburg.

Paramythis, A., Weibelzahl, S. & Masthoff, J. (2010). Layered evaluation of interactive adaptive systems: framework and formative methods. *User Modeling and User-Adapted Interaction, 20*, 383-453.

Reid, D.J. & Reid, F.J.M. (2005). Online focus groups. An in-depth comparison of computer-mediated and conventional focus group discussion. *International Journal of Market Research, 47,* 131-162.

Roach, A.T. & Elliott, S.N. (2005). Goal attainment scaling: An efficient and effective approach to monitoring student progress. *Teaching Exceptional Children, 37*, 8-17.

Sewell, M., & Marczuk, M. (2004). *Using Cost Analysis in Evaluation*. Tucson, AZ: University of Arizona, College of Agriculture and Life Sciences, Cooperative Extension.

Someren, W. M., Barnard, F. Y., & Sandberg, A. C. J. (1994). *The think aloud method. A practical guide to modelling cognitive process.* London: Harcourt Brace & Company.

Sweller, J., van Merrienboer, J.J.G., & Paas, F.G.W.C. (1998). Cognitive architecture and instructional design. *Educational Psychology Review, 10,* 251-296.

Totterdell, P. & Rautenbach, P. (1990). Adaptation as a problem of design. In D. Browne, P. Totterdell, & M. Norman (Eds.), *Adaptive User Interfaces*. Academic Press.

Van Velsen, L., Van der Geest, T., & Klaasen, R. (2007). Testing the usability of a personalized system: comparing the use of interviews, questionnaires and thinking-aloud. In *Proceedings of the Professional Communication Conference. IEEE International. (pp. 1-8)*.

Van Velsen, L., Van der Geest, T., Klaasen, R., & Steehouder, M. (2008). User-centered evaluation of adaptive and adaptable systems: a literature review. *The Knowledge Engineering Review, 23,* 261-281.

Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User acceptance of information technology: Toward a unified view. *MIS Quarterly, 27(3),* 425-478.

Weibelzahl, S. (2001). Evaluation of adaptive systems. In M. Bauer, P. Gmytrasiewicz, & J. Vassileva (Eds.), *User Modeling 2001, LNCS Vol. 2109* (pp. 292-294). Berlin: Springer.

Weibelzahl, S. (2003). *Evaluation of adaptive systems.* Dissertation. University of Trier, Germany.

Weibelzahl, S. (2005). Problems and pitfalls in the evaluation of adaptive systems. In S. Y. Chen & G. D. Magoulas (Eds.), *Adaptable and Adaptive Hypermedia Systems* (pp. 285-299). Hershey, PA: IRM Press.

Weibelzahl, S. & Lauer, C.U. (2001). Framework for the evaluation of adaptive CBR-systems. In U. Reimer, S. Schmitt, & I. Vollrath (Eds.), *Proceedings of the 9th German Workshop on Case-Based Reasoning* (pp. 254-263). Aachen: Shakerm.

Wilkinson, D. & Birmingham, P. (2003). *Using research instruments: A guide for researchers.* London: RoutledgeFalmer.

In: Evaluation in E-learning
Editors: Y. Psaromiligkos, A. Spyridakos et al.
ISBN: 978-1- 61942-942-0

Chapter 8

A CONTINUOUS LAYERED EVALUATION OF ADAPTIVE HYPERMEDIA SYSTEMS

***Vinicius F. C. Ramos*[1], *Geraldo B. Xexéo*[1,*] *and Paul M. E. De Bra*[2†]**
[1]Universidade Federal do Rio de Janeiro, Brazil
[2]Technische Universiteit Eindhoven, Netherlands

ABSTRACT

A hypermedia system gives too much freedom to the user to navigate through the hyperspace. An Adaptive Hypermedia System (AHS) gives personalized content, presentation, and navigation support to help user find the information she needs or complete her tasks. In this chapter we focus on AH systems with educational purpose and are interested in evaluating to which extent AHS affect the way students learn. Brusilowsky et al. proposed in 2001 a layered evaluation of AHS, which splits the AHS into layers that are separately evaluated. Ortigosa and Carro proposed in 2003 a continuous empirical evaluation of adaptive courses. Their method is used to identify possible problems by analyzing data obtained from the access log of the students, and also suggest improvements while the course is still happening. Our proposal combines the continuous empirical evaluation with the layered evaluation approach. For example, we can use the layered evaluation technique to collect data about how much time each student spent in each page of the course and the score each student obtained in a certain test to identify differences between the patterns that lead to good or bad scores. Based on this information, a continuous empirical evaluation can be triggered to suggest some pages to students that have performed poorly in the test. Besides the time spent in each page and test scores, we also suggest other aspects of a distance learning course that can be exploited by the continuous layered evaluation technique.

[*] E-mail address: vfcr,xexeo@cos.ufrj.br
[†] E-mail address: debra@win.tue.nl

1. INTRODUCTION

Distance learning has become very popular with the growth of the World Wide Web and the easy access to computers and the Internet. Nowadays, there is a plethora of distance learning platforms where instructors can create and publish their courses on the Web [11, 3, 8]. When using distance learning platforms, students are faced with quite a large amount of content, and navigating through the course in an efficient way may be a challenging task even for experienced users. In order to aid students navigate through the maze of webpages a distance learning course may contain, existing platforms employ a variety of strategies to organize their content. For example, some platforms are designed to show the course content in a sequential order, thereby providing a step-by-step set of activities that students must perform in a predetermined order. Though such strategy makes students visit the content in the order the instructor thinks is the best one, it forces *all* students to follow that ordering, thus neglecting that each student has its own needs and interests, and has a singular pace of learning.

Adaptation techniques were developed to allow a distance learning course to adapt the content to the needs of each student. The platform creates a profile for each student containing the pages visited and the activities performed so far by the student. This profile, called the user model, is used by the adaptation mechanism to suggest pages that the student should visit or activities and tests that should be performed. For instance, a strategy used by the platform *AHA!* [11] is to classify each link of a webpage as *recommended* or *non-recommended* based on the user model. Then, if a link turns out to be recommended, it is colored in a way that stands out from the text of the page and subtlety prompts the students to click on the link. Non-recommended links are colored using the same color as the text of the webpage, and therefore are disguised so that students are prone to ignore their existence. For this reason, non-recommended links are less likely to be clicked on. This strategy is known as *link hiding*.

Not only is it very hard to organize the content of a course in an effective way, but it is also hard to check whether a given course was effective in guiding students to the information they need. Evaluating the structure and organization of a distance learning platform is the main topic of this chapter. It is important to remark that our purpose is not to find errors or problems in a course, but to draw insights on the type of courses that are more effective in guiding students to find the information they need. We believe our conclusions may be used by course authors to improve the effectiveness and the navigability of their courses.

We shall focus on the area of adaptive hypermedia (AH), which was created in the 90's and employs a lot of techniques from other areas of research, such as information retrieval and artificial intelligence. AH Systems are characterized by employing the user model idea described above to adapt the course to the needs of each student [1]. Though AH systems exist in many different contexts, most of the existing ones have been developed for distance learning, such as *AHA!* [11, 12], APeLS [10], ELM-ART [25], Interbook [3] and KBS Hyperbook [20]. Existing AH systems employ different architecture and adaptive techniques, which make it very hard to develop a general methodology for evaluating their performance. In this chapter we will discuss some aspects about the evaluation of AH systems and will illustrate our approach using the platform *AHA!*.

The evaluation approach we describe here combines the techniques of layered evaluation [4, 5] and continuous empirical evaluation [21]; we name this combined approach as continuous layered evaluation (see Section 3 for a more thorough description). In order to illustrate this approach, we present a simple example. Consider a webpage of a distance learning course that contains some reading material. Assume that this page is not frequently visited by the students, but is considered by the author of the course as an important material. Therefore, it is expected that students should fail to perform some activities and tests correctly since they have not visited this important material. From this observation, the evaluation procedure should be able to identify the problem and suggest to the author of the course another page from which this material could be linked. Similarly, the evaluation procedure should be able to suggest this important material to students that have been failing to perform tests and activities correctly.

Our goal in this chapter is to point out some aspects that we observed in a distance learning course constructed using *AHA!* and discuss their impact in the continuous evaluation of AH systems. We also mention novel aspects that can be taken into account during the evaluation and discuss what their role in the evaluation process is.

The remainder of this chapter is organized as follows. We start by giving an overview of related work on AH systems and their evaluation in Section 2. In Section 3 we present the continuous layered evaluation technique, which is the essence of our approach to evaluating AH systems. In Section 4 we show the distance learning platform *AHA!*, which we use in Section 5 to illustrate our ideas for evaluating AH systems. Finally, we draw some conclusive remarks in Section 6.

2. STATE OF ART

In this section we discuss the main papers in the area of AH and evaluation of AH systems. We start by mentioning some recent surveys. Chin [9] presents a review of empirical evaluations on adaptive hypermedia, user modeling, plan recognition and user interfaces. He says that the reviewed articles are insufficient to the empirical evaluation method but he gives some insights on how to improve the quality of the evaluations, such as the rules of thumb for experimental design and the common threats to experimental validity. Velsen et al. [24] present a literature review of the evaluation methods used in adaptive systems. This survey discusses important considerations about the methods and variables used to evaluate an adaptive system. They also highlight the importance of using empirical data and found missing empirical values in some evaluations analyzed. They mention that analyzing data logs is not a trivial task and may only be useful after some preprocessing (such as data triangulation via, for example, think-aloud approach or surveys). Knutov et al. [19] review the most used systems and contrast their characteristics with the ones presented in the beginning of the 90’s, when the first adaptive systems were developed.

They try to come up with an up-to-date taxonomy of adaptation techniques, and investigate the requirements of a new reference model or architecture of adaptive systems.

We now discuss some of the most popular AH systems that were developed for educational purposes. The ELM-ART [25] is an intelligent interactive AH system to support learning programming in LISP 1. It is designed as an online intelligent learning environment that supports example-based programming. The ELM-ART was created as a hyperbook, this means that it is a course on the web with sections and subsections referenced by links with a menu annotated as a metaphor of traffic lights referring to information from the individual user model [23]. This approach of converting printed textbooks into hyperbooks, using the same technique of ELM-ART, has been developed in Interbook [3]. In the evaluation of ELM-ART, Weber and Specht [26] proved that adaptive link annotation is important to students who have previous knowledge on the discipline they are studying, while novice students benefit more when they have the "next step" link.

The Interbook is an AEH system which supports the terminology called hyperbook and has the link annotation as the main adaptive navigation technique. Link annotation is a technique where the links are indicated by status, i.e., there are several links states such as visited, ready to be learned, or not ready to be learned. The states are used to represent the suggested path to the student according to the set of pages he has already visited and the prerequisite knowledge needed to access the content in the course. Brusilovsky and Eklund [2] presented that students tended to accept the navigation advice when they were unfamiliar with a complex interface, and also argued that adaptive engine caused students to adopt less sequential paths throughout the course. Brusilovsky et al. [5] revisited this previous evaluation to propose a layered evaluation framework that offers certain benefits to the developers of AH systems with educational purpose. This is the basis of this chapter and is presented in more detail in Section 3.

The KBS Hyperbook [20] is an adaptive hyperbook system designed to be used in an introductory course of computer science. The technique of presenting information is similar to ELM-ART and Interbook, which means that it uses annotation to the section and subsections links. It differs from other systems in the fact that one of the adaptation techniques is based on a goal-driven approach and the students can choose their own goals.

We now briefly discuss some papers that deal with the specific problem of *evaluating* AH systems. The evaluation of AH systems has been studied by different authors in the literature [26, 22, 4, 5, 9, 15]. Moreover, in [27, 28, 24] problems and pitfalls of evaluating adaptive systems are discussed. All of these papers address problems, difficulties, limitations, and benefits of the evaluation process, but only few of them (e.g., [28] and [5]) do not treat the evaluation of the adaptive system "as a whole."

In [7], Calvi presents results for the evaluation of an online Italian language course. The course has an engine to enable and disable links depending on the level of knowledge of the student. Her evaluation is focused on the comparison between two groups of students, the first consisting of the students that used the adaptation engine and the second consisting of the students that used the engine without the adaptation mechanism. Although her work does not give any statistical evidence that the adaptive engine improves the quality of learning, Calvi notices that students in the first group were able to find the material they needed using a fewer number of clicks.

An empirical evaluation of AH systems can be used to verify the success of the adaptation mechanism or to find problems in the adaptive engine. Empirical studies carried out by Kaplan et al. [16] and Brusilovsky and Pesin [6] suggest that adaptive navigation support can speed up the navigation and learning of the students. Weibelzahl and Weber present in [28] the limits and the advantages of using empirical evaluation in AH systems. They notice that empirical evaluation can find errors and problems that could have remained undiscovered even when another evaluation technique had been used.

3. The Continuous Layered Evaluation Technique

In this section we present the continuous layered evaluation technique. We start explaining the layered evaluation technique, which was proposed to break the paradigm of evaluating an adaptive system as a whole [17, 28, 21, 5]. In [5], it is proposed to decompose an AH system in two distinct layers, called *user modelling* (UM) and *adaptation decision making* (ADM). They define the UM layer as the evaluation of the user model of the students. For example, the UM layer is responsible for analyzing whether a student has not visited a certain important webpage, has not understood a specific material, or has not succeeded in finding the information she needs. They define the ADM layer as the evaluation of specific aspects of interaction. It uses the UM layer to adapt the AH system to the needs of the student. For example, if the evaluation of the UM layer detects that a student has not succeeded in completing a task, a pop-up message can be triggered by the ADM layer with additional information and links that can help the student finish the task.

The other pillar of our evaluation technique is the continuous empirical evaluation introduced by Ortigosa and Carro [21]. They discussed situations in which an empirical analysis can help detect problems such as low performance (students fail to do the exercises related to the topic being studied), disorientation (students browse the course with no logical order), lack of motivation (students do not browse the course frequently), and dissatisfaction (students drop off the course at the beginning). Ortigosa and Carro also give some insights about which aspects of the course design could be the cause of these problems. For example, they mention that if students perform low in some activities of the course, then it may be the case that some information is not clearly explained in some webpage or the topics of the course are not well organized. Another point raised by Ortigosa and Carro was that unidentified dependences between topics might exist. Usually, this type of problems is very hard to identify, but Ortigosa and Carro remark that students may need to know some topics (called prerequisite) before studying others. Then, it would be natural to request that a previous visit to a prerequisite or a minimum score in the tests associated to them should be required in order to access a certain page. Then, if there are unidentified dependences in a course, this could become evident if students that visited the prerequisite get better results in the current topic than those who did not.

Not only do we consider the aspects pointed out by Ortigosa and Carro, but we also propose new aspects for the evaluation process. For example, we observe that students constantly return to the same webpages, a phenomenon that we call repetitive browsing, and that students visit advanced pages without having visited more fundamental pages or doing required activities, which we call curious browsing. These two aspects represent pathologies in the way one would want a student to visit the pages of the course. We also investigate how much time students spend in each webpage. This turns out to be an interesting empirical information regarding how effective the content of the pages are in guiding the students throughout the links of the course, especially in pages that contain a very large number of links, which could in principle lead the student not to know which link to follow.

The combination of the layered evaluation approach by Brusilovsky et al. [5] and the continuous empirical evaluation by Ortigosa and Carro [21] are the basis of our methodology. The proposed evaluation technique uses the idea of the layered evaluation, focusing on UM and ADM layers, to continuously evaluate a course. The goal of the continuous evaluation is to reveal possible problems in the course or in the way students are using the course while the course design can still be modified. Therefore, the evaluation can be used to improve the course design in real time and provide a better learning experience to the students.

Figure 1 illustrates the layers we are considering for the evaluation of *AHA!*. (We present *AHA!* in more details in Section 4.) The UM layer collects data regarding how students access the AH system (e.g., information regarding the tasks that were initiated or completed, surveys answered, and pages visited). When contrasting with other information in the UM layer, it is possible to infer some characteristics about the learning experience of the student (e.g., if the student has followed many bad links—see the definition of bad links in Section 4). This information is then used by the ADM layer to adapt the system to the needs of the user. Figure 1 also highlights how the continuous layered evaluation approach fits the decomposition of the adaptation layers. After the adaptation has taken place, the student again interacts with the system and then it is possible to evaluate whether the adaptation itself was efficient. In Section 5 we provide more information on how the proposed technique should be implemented in *AHA!*.

4. *AHA*!

The *AHA!* system is a Web-based adaptation engine and authoring tool developed at the Eindhoven University of Technology, starting in 1996. Initially *AHA!* was used only to serve as an adaptive course text for the course "Hypermedia Structures and Systems" offered to students from Dutch and Flemish universities. Later it was extended to become a general-purpose adaptive hypermedia system, available as Open Source software [14]. Since 2009 *AHA!* is being developed as a new system with a new architecture. In 2011, *AHA!* became *GALE* [29].

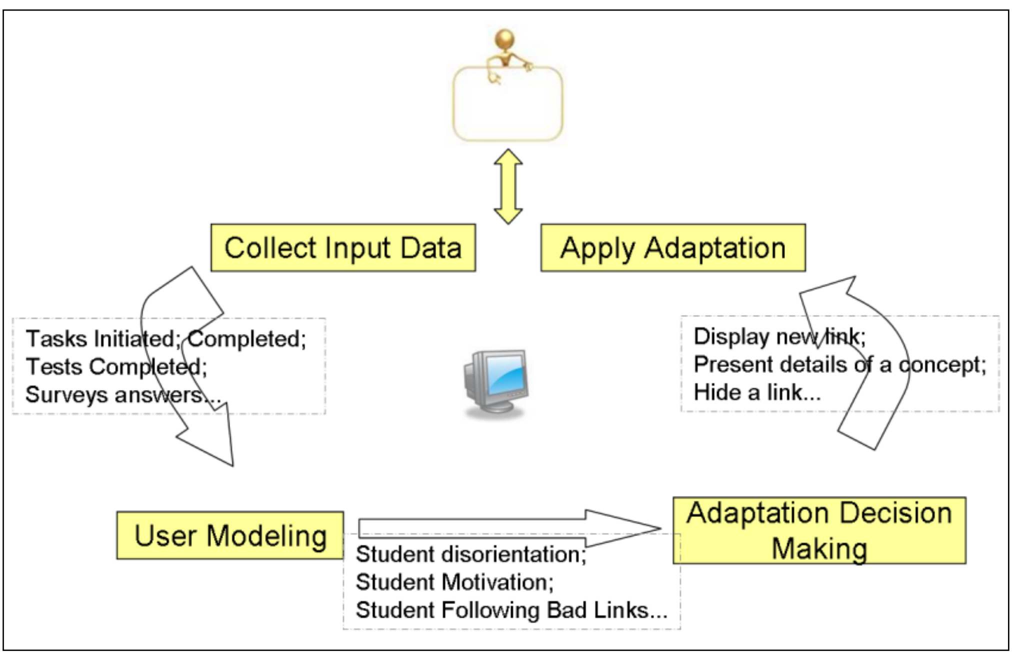


Figure 1. Adaptation Decomposing Model.

An adaptive course in *AHA!* consists of *concepts*, most of which (and in the hypertext course all of which) are connected to *pages*. The pages contain information and links to concepts. (In *AHA!* links must refer to concepts, not to pages, and the link destination page may be adaptively selected by the adaptation engine.) A page may also contain links to "external" pages on the Web. One of the approaches used in *AHA!* to adapt a course is to use different colors for the links of the pages. The author of the course defines rules to determine the conditions under which a presentation class is associated with a link. These rules are not written by hand by the author but generated by from a structure of *concept relationships*, of which the *prerequisite* is the best known type. A graphical tool, the Graph Author, makes creating this adaptation structure easy. Figure 2 illustrates the use of the Graph Author. It shows the concept relationships used for the adaptive paper [14]. The red dashed arrows are the prerequisite relationships. They indicate that enough knowledge of the source concept must be obtained before the target concept becomes recommended. "Enough" means 50% by default but for some prerequisites a different knowledge level is required. For instance, when 15% of the knowledge of "authoringtools" is reached the concepts "graphauthor" and "amt" become recommended topics. A template with "*AHA!* code" defines the adaptive behavior of the prerequisite. These templates are created by an *AHA!* expert. The course author never writes or even inspects that code.

The hypermedia course uses three presentation classes, called *bad*, *good*, and *neutral*, which have the following meaning and presentation style:

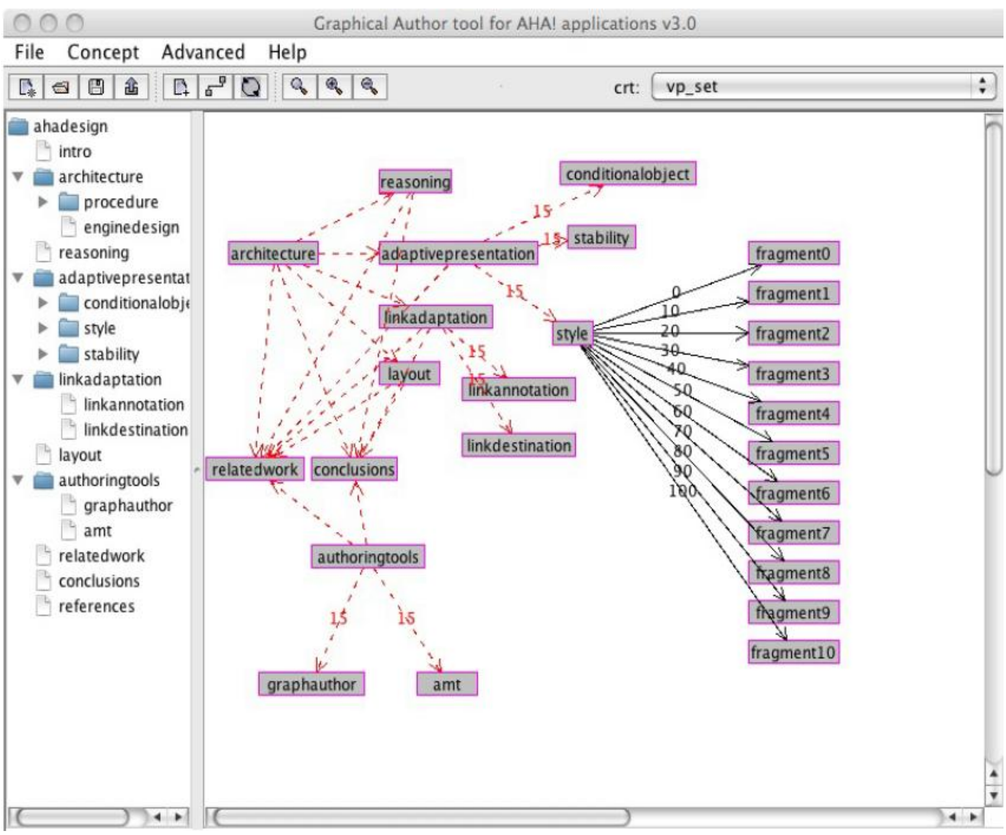


Figure 2. Graph Author tool and its concept relationships used for the adaptive paper [14].

- The *bad* links point to what we call non-recommended concepts, which means that according to the rules defined by the author, the student is expected to study something else before accessing these concepts. This “study” may consist of some reading and performing some (multiple-choice) tests in *AHA!*.

  Bad links are colored in black and are not underlined, which implies that they are indistinguishable from the textual information of the page. So, bad links are *hidden* within the text, though they are fully operational and can be clicked on at any time. It is well known in the literature as the technique of *link hiding*.
- The *good* links point to a recommended concept that the student has not yet visited after it became recommended. Good links are colored in blue.
- The *neutral* links point to a recommended concept that the student has already visited after it became recommended. Neutral links are colored in purple.

Because of the choice of adaptation and colors the hypertext course looks like a standard (non-adaptive) website with pages and blue and purple links. This presentation style gave us the opportunity to study the use of the hypertext course like a traditional non-adaptive hyper-document.

We remark that *AHA!* is capable of very different presentation styles, using automatically generated navigation aids, like different types of partial tables of content, and using different types of link annotations, including colored balls as used in ELM-ART [25] and Interbook [3]. In fact, it was show in [13] that *AHA!* can even emulate Interbook completely.

5. Continuous Layered Evaluation of *AHA!*

In this section we discuss our implementation of the continuous layered evaluation technique with a course developed in *AHA!*. We also discuss a previous evaluation of the course “Hypermedia Structures and Systems” by Ramos and De Bra [22] that was implemented in *AHA!*; to simplify the discussion we refer to this course as the *hypertext course*. The *hypertext course* we evaluated was offered during the academic year of 2007-2008 by Dutch and Flemish universities and this evaluation motivated the implementation of the evaluation technique proposed in this chapter.

We start by highlighting important aspects raised in [22]. First, it was studied the notion of hubs introduced by Kleinberg [18] in the context of Web search. Essentially hubs are webpages with a relatively large out-degree (i.e., links pointing from that webpage to other webpages). Hubs are defined to represent the webpages from which one can find a link to an “interesting” webpage, which in [18] is termed as *authority*. It was pointed out in [22] that in the context of AH systems, hubs are beneficial to decrease the depth of the course (i.e., maximum number of links that need to be followed to reach any page in the course), but on the other hand pages with very large out-degree may generate some confusion to the students, since they may not know which link should be followed. In order to verify if pages with large out-degree are indeed used as hubs in the course, the notion of *Empirical Hub Coefficient* (EHC) of a webpage X was proposed as the ratio between the number of times that students clicked on a link of X to go to a different concept and the number of times that students accessed X. Intuitively, pages with large EHC are the ones used as hubs, that is, pages with large number of links and also visited a lot to follow one of their links. The *hypertext course* has many pages with a fairly large out-degree, which as explained above could have been a problem for the author and the students of the course, since having a myriad of links in a page may generate confusion about which link the student should follow. In order to argue why the empirical hubs did not cause major problems in the *hypertext course*, the notion of *informative pages* was investigated. A page was classified as an *informative page* if students spent a relatively large amount of time when accessing the page (i.e., if students read the textual information presented in the page). Therefore, the presence of informative pages mitigates the problem inherent to pages with large out-degree, since the textual information can clarify the student about what is the best link to follow at that step. An important observation is that pages that are empirical hubs but end up not being classified as informative pages should be avoided. Note that this information can only be obtained using empirical data and manifests the importance of combining layered evaluation with continuous empirical evaluation.

Another point observed in [22] was a phenomenon that the authors called *curious browsing*; it was observed that students accessed many pages via bad links. This means that the more a student is exposed to a link (even if the link is classified as a bad link and therefore is disguised within the text of the page) the more likely it is that the student will get curious and will follow the link before it is classified as good.

We will now discuss some new aspects to be considered in the evaluation of AHS. First, the study presented here is divided into three phases: collecting the dataset, implementing the continuous layered evaluation in *AHA!* courses, and analyzing the results of the evaluation.

The first phase is already implemented and the dataset test was gathered in the *hypertext course* during the academic year 2009-2010.

The second phase is under development. The proposed evaluation will encompass the following aspects: the students' navigational paths (i.e., sequence of pages) that students follow during the learning process, the structure of the course, the improvement on the students' knowledge, the student disorientation, and the students' surveys opinion.

We presented in the beginning of this section the motivation and the evaluation of the *hypertext course* implemented in *AHA!*. As mentioned in [24], some empirical data analysis needs pre-processing and there are some features to be implemented that will require students to fill out surveys during the course. Thus, the following topics represent extra tasks and their respective problems, and are now being implemented at the new academic classes of the *hypertext course* for the academic year 2009-2010.

a. Navigation Paths Analysis

The navigation paths are used to represent the sequence of concepts visited in a course. For each identifier in the sequence it represents a concept in the course. The main idea of analyzing the navigation path is to check if students completing a test successfully visited some page before finishing the test that students failing the test did not. At this point we want to find patterns between all the students that failed the test. After getting a pattern, we shall contrast it to students that succeed the test. For illustration purpose, we refer to Figure 3. In (1), we show the navigation paths for three students that failed the test S. In (2), we show the pattern extracted from these three students (i.e., which concepts these students visited in common).

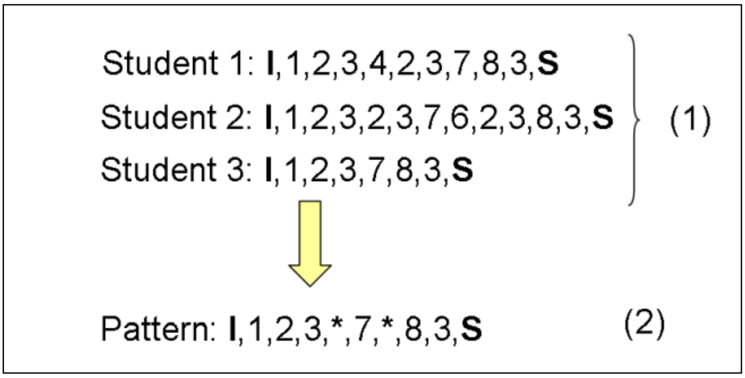


Figure 3. Navigation Paths sequence from students that fail in the test **S**. (1) represents the students' paths. (2) represents the common pattern extracted from (1), where * stands for visits to zero or more concepts.

Figure 4 presents, in (3), three students that succeeded the test S. In (4), we show the pattern extracted from the students that succeeded the test. In order to determine why students failed the test S, one needs to compare the patterns for the students failing the test (Figure 3.(2)) and for the students succeeding the test (Figure 4.(4)). In the examples of Figures 3 and 4, one can observe that the path "2,*,5,*,7" appear only in Figure 4(4). This suggests that students that have completed the test successfully have visited concept 5 before performing the test S, while students that failed the test have not visited concept 5. This gives evidence that visiting concept 5 before performing the test S is important to the students. Therefore, the evaluation mechanism can suggest the author of the course to add a prerequisite so that students can only perform test S after visiting concept 5. Alternatively, in order to corroborate the hypothesis that concept 5 is important for test S, the author of the course may add a review with the content of concept 5 before the test S.

b. One Student following Bad Links

Navigating through links in a hypertext course, which have the hiding links technique in the adaptation engine, is important to be evaluated because the hiding links technique depends on the UM. Although following *bad* links is not prohibited and do not spoil the student in the course, a large number of *bad* links being followed by the student could indicate that the UM is not capturing the knowledge of the user correctly. Therefore, the question we should answer in this evaluation is: is the user's knowledge being captured by the system, since it is changing all the time?

Our approach is simple but necessary to evaluate the UM. Once we identify that a student has been following *bad* links constantly, we present her a survey to identify her needs and discover the cause of this access via bad links. Depending on the students' answers, the UM could not be updated as the system should be. The student could have clicked in a link that is shown as a *neutral* or *good* but the UM still represent it as a bad link, or the student could have not noticed that the link is in fact a *bad* link. In the former case, the UM status is not being updated. In the latter case, the design of the course is not desirable by the authors. Both of them need the attention of the designer of the course.

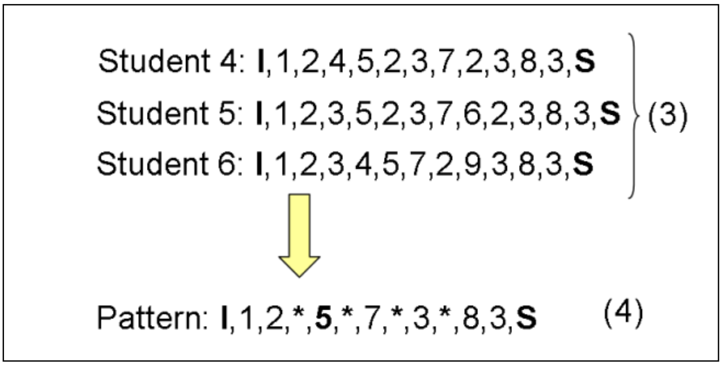


Figure 4. Navigation Paths sequence from students that succeed in the test **S**. (3) represents the students' paths. (4) represents the common pattern extracted from (3), where * stands for visits to zero or more concepts.

c. Different Students Following the Same Bad Link

We have discussed in the last topic about the case when a single student follows *bad* links constant. The current topic follows the same spirit but has a subtle (albeit important) difference. Here we want to discuss the case where *many* students follow *bad* links. Considering that *bad* links are non-recommended concepts (see Section 4), a large number of students following the same non-recommended concept raises the doubt of why this link is not recommended if many students are clicking on it. Again, this could be a problem of the UM or the design of the course. Giving another survey to the students would be sufficient to identify the needs of the students and the cause of this situation.

d. Student Motivation

In a distance learning course, regardless of whether it is adaptive or not, it is always hard to identify if the students are motivated. The author of an adaptive course needs to know when student loose stimulus; more importantly, the author of the course wants to know the cause of this fade of motivation and if the adaptation engine or the course design make the student feel confused and disoriented. We want to concentrate in two types of behaviours: a student who enters a course and spends a long time navigating through the course pages and never visits perform the tests, and a student that visits the test page but does not complete it. The rest case can be caused by many factors. For example, the student may simply not know where the test page is or may have not noticed that there is a test to be performed. Also, the student may not feel prepared to complete the test or may have decided to give up. In order to investigate whether a student is likely to fall into one of these categories, we propose to analyze the total time spent by the student in the whole course and in each concept of the course. If the student has spent a large amount of time navigating through the course and accessing again and again a large number of concepts in a small period of time, then probably the student is motivated to pursue the course, but got lost and does not know what to do. This suggests that the test needs to be relocated to a better place in the course or needs to be highlighted. Such behaviour is also likely to be observed in the case when the student does not feel prepared to complete the test. However, in this latter case, we expect the student to keep visiting all the concepts referring to that topic over and over again. In these cases we suggest to send a survey to the student.

e. Average Time Spent

The evaluation of an AH system for distance learning always concerns about usability, student performance, or students' knowledge improvement. One point that should be evaluated is the average time spent by students in a concept. We can distinguish the three presentation classes in *AHA!*: *bad*, *good* and *neutral* links, which represent the status of a concept. We associate these presentation classes to the concept status. We consider the concept status: non-recommended, recommended but not visited, and recommended and visited. These concept statuses are associated to the link status *bad*, *good* and *neutral*, respectively. For example, a *good* link represents a concept that has not yet been visited (and consequently has not yet been learned by the student). We expect that students spend a large amount of time reading a page of such a concept. On the other hand, a bad link represents a non-recommended concept, for which we expect that students are already missing some information or required knowledge to understand the concept. Hence, the average time spent by the students in such concepts may even be higher than for recommended concepts that were not visited. Finally, the *neutral* link represents an already learned concept. We then expect that the average time spent by the students in such concepts is lower than for the other concept status. It is interesting to check that the concept is already known, so that the student only needs to skim quickly through the page to find a specific piece of information of a link to be followed. Note that this idea is intrinsically related to the *informative pages* introduced in [22].

The third phase of the evaluation correspond to the analysis obtained with the data collected in the *hypertext course* offered to Dutch and Flemish universities in the academic year of 2009-2010.

CONCLUSION

In this chapter we have considered the problem of evaluating adaptive hypermedia systems with emphasis on systems for educational purposes. Our goal in evaluating adaptive educational systems is to understand to which extent adaptation affects the way students learn and how an educational system can be improved to make the learning experience more effective. We have proposed an evaluation approach that combines the techniques known as layered evaluation and continuous empirical evaluation. We consider the aspects of empirical hub coefficient, informative pages, and curious browsing obtained in a previous paper and we introduce new aspects, such as the navigational paths, the analysis of access via bad links, the student motivation, and the time spent in each page. We have illustrated our approach using the framework of the platform *AHA!*.

REFERENCES

[1] P. Brusilovsky. Methods and techniques of adaptive hypermedia. *User Modeling and User-Adapted Interaction*, 6(2-3):87–129, 1996.

[2] P. Brusilovsky and J. Eklund. A study of user model based link annotation in educational hypermedia. *Journal of Universal Computer Science*, 4(4):429–448, 1998.

[3] P. Brusilovsky, J. Eklund, and E. Schwarz. Web-based Education for All: A Tool for Development Adaptive Courseware. In *Computer Networks and ISDN Systems*, pages 1–7, 1998.

[4] P. Brusilovsky, C. Karagiannidis, and D. Sampson. The benefits of layered evaluation of adaptive applications and services. In S. Weibelzahl, D. Chin, and G. Weber, editors, *Empirical Evaluation of Adaptive Systems. Proceedings of Workshop at the Eighth International Conference on User Modeling, UM2001*, pages 1–8. 2001.

[5] P. Brusilovsky, C. Karagiannidis, and D. Sampson. Layered evaluation of adaptive learning systems. *International Journal of Continuing Engineering Education and Lifelong Learning*, 14:2004, 2004.

[6] P. Brusilovsky and L. Pesin. Adaptive navigation support in educational hypermedia: An evaluation of the ISIS-Tutor. *Journal of Computing and Information Technology*, 6:27–38, 1998.

[7] L. Calvi. Formative evaluation of adaptive CALLware: A case study. In *AH: Proceedings of the International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems*, pages 276–279, London, UK, 2000. Springer-Verlag.

[8] R. M. Carro, A. Ortigosa, E. Mart n, and J. H. Schlichter. Dynamic generation of adaptive web-based collaborative courses. In J. Favela and D. Decouchant, editors, CRIWG, volume 2806 of *Lecture Notes in Computer Science*, pages 191–198. Springer, 2003.

[9] D. N. Chin. Empirical evaluation of user models and user-adapted systems. *User Modeling and User-Adapted Interaction*, 11(1-2):181–194, 2001.

[10] O. Conlan, C. Hockemeyer, V. Wade, and D. Albert. Metadata driven approaches to facilitate adaptivity in personalized elearning systems. *Journal of the Japanese Society for Information and Systems in Education*, 1:38–45, 2003.

[11] P. De Bra, A. Aerts, B. Berden, B. De Lange, B. Rousseau, T. Santic, D. Smits, and N. Stash. *AHA!* The Adaptive Hypermedia Architecture. In *HYPERTEXT: Proceedings of the fourteenth ACM conference on Hypertext and hypermedia*, pages 81–84, New York, NY, USA, 2003. ACM.

[12] P. De Bra and L. Calvi. *AHA!* An Open Adaptive Hypermedia Architecture. *The New Review of Hypermedia and Multimedia*, 4:115–139, 1998.

[13] P. De Bra, T. Santic, and P. Brusilovsky. *AHA!* meets Interbook, and more... In *Proceedings of World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education 2003*, pages 57–64, Phoenix, Arizona, USA, 2003. AACE.

[14] P. De Bra, D. Smits, and N. Stash. The design of *AHA!* In *HYPERTEXT: Proceedings of the seventeenth conference on Hypertext and hypermedia*, pages 133–134, New York, NY, USA, 2006. ACM.

[15] C. Gena. Methods and techniques for the evaluation of user-adaptive systems. *Knowledge Engineering Review*, 20(1):1–37, 2005.

[16] C. Kaplan, J. Fenwick, and J. Chen. Adaptive hypertext navigation based on user goals and context. *User Modeling and User-Adapted Interaction*, 3(3):193- 220, 1993.

[17] C. Karagiannidis and D. G. Sampson. Layered evaluation of adaptive applications and services. In *AH'00: Proceedings of the International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems*, pages 343-346, London, UK, 2000. Springer-Verlag.

[18] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *J. ACM*, 46(5):604–632, 1999.

[19] E. Knutov, P. De Bra, and M. Pechenizkiy. Ah 12 years later: a comprehensive survey of adaptive hypermedia methods and techniques. *New Rev. Hypermedia Multimedia*, 15(1):5–38, 2009.

[20] W. Nejdl and M. Wolpers. KBS Hyperbook - A Data-Driven Information System on the Web. In *International World Wide Web Conference*, 1999.

[21] A. Ortigosa and R. M. Carro. The continuous empirical evaluation approach: evaluating adaptive web-based courses. In *UM'03: Proceedings of the 9th international conference on User modeling*, pages 163–167, Berlin, Heidelberg, 2003. Springer-Verlag.

[22] V. F. C. Ramos and P. De Bra. The influence of adaptation on hypertext structures and navigation. In *HT '10: Proceedings of the 21st ACM conference on Hypertext and hypermedia*, pages 77–82, New York, NY, USA, 2010. ACM.

[23] E. Schwarz, G. Weber, and P. Brusilovsky. World-wide intelligent textbooks. In *World Conference on Educational Telecommunications*, pages 302–307. AACE, 1996.

[24] L. van Velsen, T. van der Geest, R. Klaassen, and M. Steehouder. User-centered evaluation of adaptive and adaptable systems: a literature review. *The Knowledge Engineering Review*, 23(03):261–281, 2008.

[25] G. Weber and P. Brusilovsky. ELM-ART: An adaptive versatile system for Web-based instruction. *International Journal of Artificial Intelligence in Education*, 12:351–384, 2001.

[26] G. Weber and M. Specht. User modeling and adaptive navigation support in www-based tutoring systems. In *International Conference on User Modeling*, pages 289–300. Springer, 1997.

[27] S. Weibelzahl. Problems and pitfalls in evaluating adaptive systems. In S. Chen and G. Magoulas, editors, *Adaptable and Adaptive Hypermedia Systems*, pages 285–299. IRM Press, 2005.

[28] S. Weibelzahl and G. Weber. Advantages, opportunities, and limits of empirical evaluations: Evaluating adaptive systems. *Künstliche Intelligenz*, 3:17–20, 2002.

[29] D. Smits and P. De Bra. GALE: A highly extensible adaptive hypermedia engine. In *HT'11: Proceedings of the 22nd ACM Conference on Hypertext and Hypermedia*, pages 63–72, New York, NY, USA, 2011. ACM.

In: Evaluation in E-learning
Editors: Y. Psaromiligkos, A. Spyridakos et al.
ISBN: 978-1- 61942-942-0

Chapter 9

VALIDATING A COMPUTER-BASED TUTOR THAT PROMOTES SELF-REGULATED WRITING-TO-LEARN

Virginie Zampa[1,*] and Philippe Dessus[2,†]
[1]LIDILEM, Université Stendhal, Grenoble, France
[2]LSE, Grenoble-2 University and IUFM, Grenoble-1 University, France

ABSTRACT

In this chapter, we introduce *Apex 2.1* (i.e., Assistant for Preparing EXams), a prototype system that provides automated feedback to e-learning students regarding the summaries they wrote about the courses they attended. At first, we present some theoretical underpinnings on how the system has been designed, and then we detail *Apex 2.1* architecture. Eventually, the first results of a validation study involving three groups of stakeholders (students, teacher, administrator) are presented. The utility, acceptability and usability of the system are examined.

Keywords: Latent Semantic Analysis, Self-regulated Learning, Computer-based Feedback

1. INTRODUCTION

This chapter is focused on the presentation of *Apex 2.1* (an Assistant for Preparing EXams), a prototype system that delivers automated pieces of feedback to e-learning students on the summaries they wrote about the courses they attended. *Apex 2.1* firstly enables students to choose a subject content they want to study, from a query composed of one or more key-word(s). Then the system delivers a set of documents semantically close to this

[*] E-mail: Virginie.Zampa@u-grenoble3.fr.
[†] E-mail: Philippe.Dessus@umpf-grenoble.fr.

query. At any time the students can write out a summary of the document they have read and understood and can get an automated feedback on this summary. If they do so, students can revise their summaries and enter in a self-regulated workflow: reading-writing-feedback-revision, which could lead to a better understanding of the chosen subject content.

The remainder of this chapter is as follows. We initially introduce the theoretical underpinnings on which this system is based, as well as its architecture and the semantic analysis method (Latent Semantic Analysis) on top of which it is built. We eventually detail its first usability evaluation study upon three different categories of users.

2. Writing to Learn

In a nutshell, the design of *Apex 2.1* starts with the assumption that students should write in order to learn and the better their self-reflection (and self-assessment) on their learning is, the better their learning capability will become. Elaborating more on this assumption will lead to a better explanation on the potential use of *Apex 2.1* in an e-learning situation. According to the so-called "writing-to-learn" approach, the activity of writing yields to learning, in exploring and highlighting the relations between ideas ("strong text" view, Emig, 1977), due to very close similarities between writing and thinking. Further research (Klein, 1999) showed that the writing-to-learn approach actually shares four different views and hypotheses:

- the *spontaneous writing hypothesis*, in which writers generate de facto knowledge, without any planning or revision processes (close to the knowledge telling process, Bereiter and Scardamalia, 1987);
- the *forward search hypothesis*, in which ideas are externalized in texts and new ones are then inferred through re-reading;
- the *backward search hypothesis*, claiming that writers first have rhetorical goals from which ideas and arguments derive, and thereby writers learn;
- *the genre writing hypothesis*, suggesting that writers use their knowledge of genre structure to create and analyse relationships between elements of texts, which in turn leads to learning.

This bulk of hypotheses highlight the idea that writing is a self-regulated activity with multiple goals, whose account cannot be constrained, and that e-learning settings make this assumption more crucial. In such settings, feedback to students' learning cannot be as frequent as in presence, and students have to cope with their isolation from each other and to engage in self-regulated activities. Research has been developed on self-regulation of learning in a distance learning context, and Vovides et al. (2007) devised a model on this activity. Firstly, students work on the *object level*, in preparing their activity according to the ongoing task. They can also develop some cognitive strategies for carrying out these activities. Then, students perform a first rough assessment of their production (its adequacy, its relation to their knowledge, etc.). Thirdly, a reflection on a meta level allows them to perform a comparison between the latter comparison and the object level, often offered by artifacts (computer-based services, prompts, etc.), so as to compare their perceived level of learning with the proposed one by the artefacts. Eventually, the students can adapt their work, which in turn fuels the possible update of the object level and can be re-acted in a further loop.

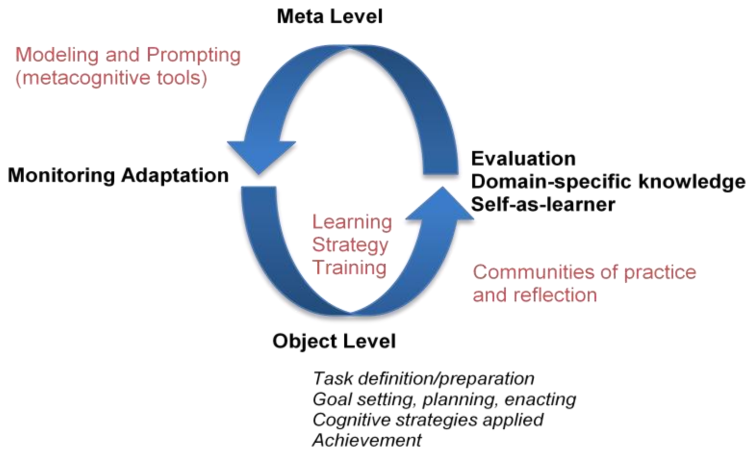


Figure 1. A Model of Self-Regulated Learning in e-learning (after Vovides, et al., 2007, p. 68).

We claim that *Apex 2.1* enables students to have a deep understanding of the course they learn without (at least immediate) support from their teachers or tutors. This software belongs to a new line of e-learning services allowing their users to be freely engaged in workflows from which they can easily move to another one and which fosters their self-regulation. It is worth noting that both teachers and tutors benefit from *Apex*'s use as well, by providing an overall view of students' activity and understanding. Thus they can focus on higher-level or more individualized student support. Since the use of this kind of services is new, our aim is to collect fine-grained evidence of their usage, and views on the way students get acquainted with them.

3. System Description

Apex 2.1 has already been subject to a first implementation (Dessus and Lemaire, 2002), but no evaluation in real-world settings has been undertaken so far. *Apex 2.1* is written in PHP and C, and built on top of Latent Semantic Analysis (Landauer and Dumais, 1997), a statistical method introduced in the next section. A running version of this service is available at http://augur.wu-wien.ac.at/apex2/Recovery/progPhp/Apex2.php.

3.1. System Architecture

The way *Apex 2.1* works is depicted in Figure 2. Two main loops (in yellow or grey) are successively carried out: the first one is a *reading loop*, allowing students to read a set of texts semantically related to a query, the second one is a *writing loop*, in which students write out a summary from the source texts previously read and understood, and can get feedback on this understanding.

Let us explicate the *reading loop*. A session typically starts when a student types a query, i.e., some key-words related to the subject domain or the theme he or she wants to be acquainted to (see Figure 3). Then *Apex 2.1* displays to the student a text to read at first, which is the closest text to the query. All the texts are presented to the student successively and are as closest to the query as possible, as indicated by LSA. The learner model is updated accordingly, indicating that a given text has been actually presented and read.

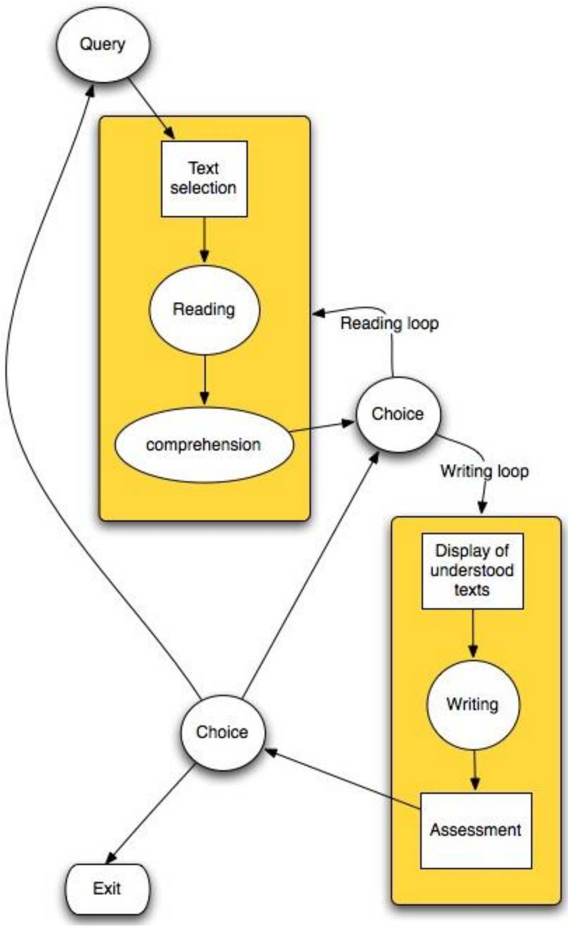


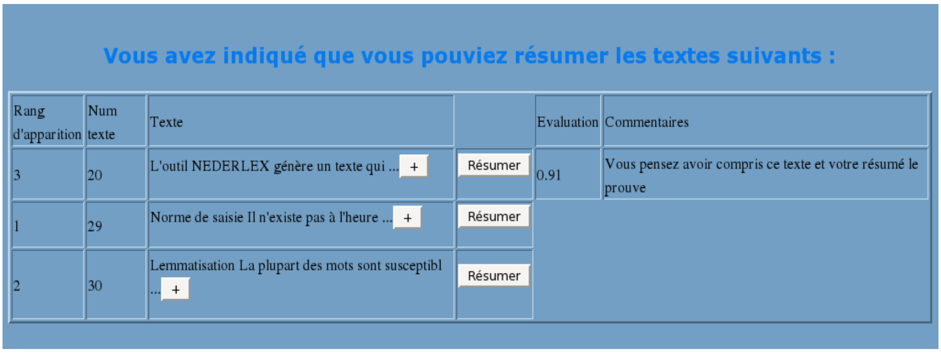
Figure 2. The way *Apex 2.1* functions. User-machine interactions.

Figure 3. A message starting the reading loop, asking the student to type some key-words: “What is the course topic you want to read?”.

Once the text has been read, the student indicates whether it is understood or not, and whether he or she is able to summarize it (see Figure 4). During this step, the learner model is updated, indicating whether the text is summarizable. *Apex 2.1* gives texts to read as long as there is no summarizable text yet *and* texts sufficiently close to the query but not delivered yet. On the contrary, *Apex 2.1* suggests the student to perform a new query.

Figure 4. The interface for reading course texts. Once the text has been read, the student can assess his or her own comprehension: “I can summarize the text” or “I cannot summarize the text”.

Once the student has understood at least one text, he or she can either read more texts on the given topic or start a *writing loop*. In the latter case, a table with the list of all the summarizable texts is displayed (Figure 5). The first column indicates the reading order of each text, the second column its ID number. The third column displays the first lines of each text read so far, which can be displayed as a whole by clicking on the + button. The “summarize” (*résumer*) button of the fourth column gives access to the writing zone (displaying the latest version of the summary, if it exists, or an empty field if not). Eventually, the last two columns respectively display the semantic proximity between the summary and the corresponding course text, as computed by LSA, and the comparison of this value and the student’s opinion about his or her own comprehension of the text (i.e., “You said you understood the course text and your summary shows it is actually the case”, « *Vous pensez avoir compris le texte et votre résumé le prouve* »).


Vous avez indiqué que vous pouviez résumer les textes suivants :

Rang d'apparition	Num texte	Texte		Evaluation	Commentaires
3	20	L'outil NEDERLEX génère un texte qui ... +	Résumer	0.91	Vous pensez avoir compris ce texte et votre résumé le prouve
1	29	Norme de saisie Il n'existe pas à l'heure ... +	Résumer		
2	30	Lemmatisation La plupart des mots sont susceptibl ... +	Résumer		

Figure 5. Overall view of the texts read so far, displayed at the beginning of the writing loop.

Apex 2.1 currently works for a learner connected individually and updates a simplified learner model, as mentioned above. This model is coded in a text file, each line representing a course text to read and four columns representing respectively: the text ID, the “already-presented” identifier (0 or 1), the “understood” identifier (0 or 1), the proximity value (i.e., the LSA-based semantic comparison between the source-text and the corresponding student’s summary). This text file is updated for every student’s move and is re-initialized for every new query.

The student, after any feedback request, can perform one of the following actions:

- revise the summary and ask for a new feedback, since it is available for modification;
- go back to the table of texts read so far (see figure 5) and start with a new summary, or modify one already existing;
- read a new course text;
- perform a new query;
- quit the application *Apex 2.1*.

3.2. Latent Semantic Analysis

Latent Semantic Analysis (LSA, Landauer and Dumais, 1997) is a well-known technique that captures semantic information in texts by uncovering word usage regularities. Extensive research has proven its efficiency in the domain of natural language processing, and more specifically for computer-based instruction—tutoring systems, interactive learning environments (Dessus, 2009). LSA represents the pieces of text to be analysed (e.g., course texts, students’ writings) in a multidimensional space. The processing steps are as follows. LSA takes as input a large set of texts and a word-paragraph matrix of co-occurrences is firstly built, then its dimension is reduced (to about 300 dimensions) by means of a statistical procedure. This reduction enables the emergence of semantic relations between words, paragraphs or texts. Due to this method two words can be considered as semantically close to each other, if they appear in *similar* contexts (i.e., sentences, paragraphs, texts), even if they actually never appeared in the same one. However, this method works best only if a sufficiently large corpus of words is processed (i.e., multi-million-word large corpus).

Apex 2.1 is built on top of LSA, which is used for providing the most adequate texts to the students (*reading loop*), and also to assess to what extent students' summaries are semantically close to the course texts they refer to (*writing loop*). Initially, we determine an arbitrary threshold value (0.6), above which the selected texts are sufficiently close to the query (*reading loop*). With regard to the *writing loop*, the same threshold has been used, and the following prompts can be displayed: "*You said you understood this text and obviously you did*" when the semantic proximity between the source text and its summary is above the threshold, and "*You said you understood this text, but obviously you didn't*" on the contrary.

4. System Validation

The validation of *Apex 2.1* is based on a corpus of sixty-six course texts (i.e. course notes or research articles) on natural language processing (NLP) and/or computer assisted language learning (CALL). These texts are homogenous regarding their length in words (an average of 650 words each). The texts provided in the reading loop were taken from this corpus. We added a larger and more general corpus to this base, so as to integrate general knowledge about language as well. This "general" corpus contains 101,123 different forms. It is formed from various texts representing several types, styles and vocabularies (articles from the daily newspaper *Le Monde*, texts written by children, texts written for children) and has been subject to validations (Denhière, Lemaire, Bellissens, and Jhean-Larose, 2007).

In order to test *Apex 2.1* in conditions as close as possible to e-learning settings, we defined three groups of participants.

- *a "user" group*, consisting of 11 masters degree or PhD students in NLP and/or CALL at our university. This group used *Apex 2.1* within the same conditions as e-learning students.
- *a "demo" group*, consisting of 28 bachelor degree students in educational sciences at our university. This group was given a demonstration of a standard learning session using *Apex 2.1* and answered a questionnaire afterwards.
- *a "teacher/administrator" group*, consisting of two persons: the manager of the pedagogical ICT department at the IUFM (i.e., "*Institut universitaire de formation des maîtres*", or Teacher Training Institute) of Grenoble university, and the manager of digital workspaces in the same institution. This group followed the same tasks as the previous one (demo).

The reason why the last two groups have been enrolled (without testing *Apex 2.1* practically) was to evaluate the degree to which such a tool is acceptable to potential groups of users and prescribers. In brief, the participants of the first two groups, as learners, performed an empirical "evaluation" task whereas the participants of the last group performed an "inspection" task, i.e., a prescriptive analysis of the possible uses of the system.

4.1. Participants Task Description

The "user" group participants were of an average age of 24.6 years old (standard deviation 4.26) and were distributed as follows: five first year master students, two second year master students and four PhD students. Before starting the experiment, they were asked to acknowledge the protocol and to sign it, and then they had to fill in a questionnaire about their course revision methods. Afterwards, they used the software by entering keywords of their own choice. They had to summarize five texts of their choice (that they had read and that they considered they understood), without any time limit. Their usage trace on the software was recorded (e.g., chosen keywords, number of searches, texts read, time taken to read each text, texts understood or not, duration of each loop, etc.). At the end of this task, they were asked to fill in a second questionnaire regarding their opinion on the software. The whole set of documents given to this group is available at http://augur.wu-wien.ac.at/apex2/ Expe/groupe-utilisateur.pdf.

The "demo" group participants (average age: 23.4 years old, SD: 4.5) had to fill in the same questionnaire as the 'user' group regarding their revision methods. Then they were shown a screencast demonstration of *Apex 2.1* and answered a second questionnaire stating their opinion about *Apex 2.1*. The questionnaire is available at http://augur.wu-wien.ac.at/apex2/Expe/groupe-demo.pdf.

The "teacher/administrator" group consisted of two persons, one aged 42, the other one aged 60. They followed a demonstration of *Apex 2.1* and answered a questionnaire available at http://augur.wu-wien.ac.at/apex2/Expe/ groupe-admin.pdf.

4.2. Hypotheses

We defined the following hypotheses, based on criteria of utility, usability and acceptability for CALL (Bétrancourt, 2007; Tricot, et al., 2003):

a) *The utility hypothesis* assumes that the use of the system induces a benefit for the user (time-wise, interest-wise, etc.). It can be split in two sub-hypotheses:

- *Query-reading hypothesis: Apex 2.1* can be used as a search engine to find texts relevant to the theme chosen by the user. The texts have been provided by teachers; they are thus appropriate and from reliable sources. Hence the system allows students to read texts both relevant and understandable to them.
- *Self-regulation/evaluation hypothesis*: Students can freely enter each loop (query-reading and writing-evaluation), which allows them to regulate their learning better; furthermore, *Apex 2.1* provides them with feedback on their validated summaries.

b) *The usability hypothesis*, which corresponds to *Apex 2.1*'s handiness, refers to browsing and the interface. The aim is to evaluate, from an ergonomic point of view, the ease with which the system can be used and with which it can help the users achieve the goals they have in mind.

c)*The acceptability hypothesis* corresponds to the "value of the intellectual representation (attitudes, opinions, etc. either positive or not) of a system, its utility and its usability[1]" (Tricot, et al., 2003, p. 396). This corresponds to a general point of view about usage demands induced by the system itself; in our case, this corresponds to knowing whether students find *Apex 2.1* easy to use and whether they would use it for their own work if they were given this opportunity.

4.3. Results

In this chapter we present the results for all three participant groups. The previous hypotheses have only been studied in depth for the "users" group.

4.3.1. The "Users" Group

As noted previously, we gathered two types of data from the "users" group: their traces (users' behaviour in terms of queries, readings and writings) and their answers to the questionnaires. The analysis of the traces has shown the following points.

Utility/Query-reading hypothesis. 7 users out of 11 only carried out only one query (see Table 1), and thus worked on the same theme for all texts, two carried out two queries, one carried out three and the last one four (average number of queries: 1.64, SD 1.03). Regarding reading, all 11 participants read between 5 and 13 texts (average 8.0, SD 3.46). However, we should note that most texts judged as "not understood" and thus not summarized were only skimmed through. Out of 33 texts, only 5 were read in more than 2 minutes and the average reading time for the 28 others was 30 seconds. It also seems that, when participants indicate whether they have understood the text and are capable of summarizing it, they give an appreciation of interest rather than an evaluation of their understanding. If the text is of some interest to them, they read it thoroughly and summarize it, otherwise they skim through the text and go to the next one. *Apex 2.1* would then be used to select texts corresponding to the user's expectations, which validates our first hypothesis.

Utility/Self-regulation-evaluation hypothesis. The following data only refer to read and summarized texts. On average, the inter-participant time taken to read a text varies between 3'24" and 14'19" (average 8'2"; SD 2'50"). This inter-participant variation is thus very high. Regarding the intra-participant variation, the standard deviation on reading time ranges from 41" to 5'52", average 2'40". All participants followed the instructions and wrote at least five summaries, taking from 1'49" to 27'40 to write them down (average writing time for each text, 7'5", SD 4'16"). The average summaries LSA-based evaluation was 0.73; only 6 out of 55 summaries got a value below the threshold (0.6) and 3 out of 6 were written by the same participant. Only one summary was rewritten. The students thus did not modify their summary after obtaining their evaluation. They did not take the feedback into account and did not try to improve them. This can probably be attributed to the fact that most feedback indicated that the summaries were correct, which was enough for them. More detailed feedback, rather than simply a correct/incorrect evaluation, would have certainly had more impact.

[1] Free translation of « la valeur de la représentation mentale (attitudes, opinions, etc. plus ou moins positives) à propos d'un système, de son utilité et de son utilisabilité » (Tricot, et al., 2003, p. p. 396).

Table 1. General data on reading activities using *Apex 2.1*

Student ID	Nb queries	Nb texts read and judged as understood	Nb texts read and judged not understood	Total Nb texts read
1	1	5	0	5
2	1	5	0	5
3	1	5	0	5
4	1	5	0	5
5	1	5	0	5
6	1	5	4	9
7	3	5	7	12
8	1	5	5	10
9	4	5	1	6
10	2	5	8	13
11	2	5	8	13

Regarding the analysis of both questionnaires, *the first questionnaire* referred mainly to using computers for course revision. Most participants mentioned they supplemented their courses with documents found on the Internet (7 out of 11 on a "weekly" or "nearly daily" basis). Furthermore, they simultaneously used pen-and-paper and computer to revise (8 out of 11 distributed between "nearly every day" and "several times a day"). Hence, the "users" participants are used to revising with their computer. This first questionnaire allowed us to verify that they are used to working on a computer and thus any problem of usability would be directly attributable to *Apex 2.1* and not to lack of expertise in browsing, note taking or computer searches. The *second questionnaire* referred to *Apex 2.1* usability and was divided in two parts.

Usability hypothesis. In the first part, named "usage difficulties" and elaborated from the NASA-TLX test (Hart and Staveland, 1988), we note that for the majority of users, *Apex 2.1*:

- does not lead to any physical pressure,
- leads to a pressure related to the experiment by itself,
- is easy to use.

In the second part, named "functions of *Apex 2.1*", we note that, for the majority of the participants:

- the texts provided corresponded to the query,
- the texts provided were suitable, precise, etc.

This confirms our second hypothesis. Finally, most participants would use *Apex 2.1* from time to time if accessible, which confirms our acceptability hypothesis.

4.3.2. The "Demo" Group

The "demo" group answered a two-part questionnaire. Regarding their use of computers to revise courses, participants supplemented their course notes with documents found on the Internet less frequently (13 out of 28 "around once a week" and 9 out of 28 "around once a month") and used more the paper-and-pen method on its own to revise (15 out of 28 "nearly every day"). Their answers on computer and paper-and-pen simultaneous usage were: 1 "never", 9 "once a month", 10 "once a week", 6 "nearly every day", 2 "several times a day". Regarding *Apex 2.1's* functions, most participants think that *Apex 2.1* could help them acquire knowledge as easily as, or even more easily than, revision methods (19 answers, 11 "more easily", 8 "as easily"). Thus, the acquired knowledge is considered to be accurate (for 15 participants out of 28) and wide (14 out of 28) with their usual method. Finally, if available, 20 out of them would use *Apex 2.1* from time to time for course revision purposes.

4.3.3. The "Teacher/Administrator" Group

Both participants in this group stated that using *Apex 2.1* could enhance the acquisition of new knowledge, compared to traditional methods (paper-pen), and furthermore that the new knowledge would be more precise. Both would suggest a frequent use of the tool to their students. They considered that *Apex 2.1* would provide students with better focus on the information required to learn course content. The administrator pointed out that the software gives students access to a wider range of information but remains easy to use.

4.3.4. Summary of Results

Carrying out this study with three groups has given us a twofold evaluation: an "inspection evaluation" with the "teacher/administrator" group and a more "empirical" evaluation with the other two, with criteria of utility, usability and acceptability. Regarding the *inspection evaluation*, both teachers/administrators consider that *Apex 2.1* fulfils all three criteria. Regarding the *empirical evaluation*, the "demo" group validates all three criteria. On the other hand, concerning the "*users"* group, the utility hypothesis regarding texts evaluation is not completely validated. When summaries are assessed correct, participants do not try to improve them, and when judged incorrect, participants did not rewrite them. However, the second part of the hypothesis has been validated, as well as both other hypotheses.

CONCLUSION

The purpose of this chapter was to describe the underlying principles and a first validation of *Apex 2.1*. The software, dedicated to exam preparation, provides different uses to the learner. Firstly, *Apex 2.1* can be used as a text database with an integrated search engine. The advantage of this use, compared to an Internet search engine, lies in the fact that the text base has been built by the teacher and therefore only contains suitable texts (which avoids wasting time in Internet searches), and furthermore only contains reliable texts (which is not necessarily the case on the Internet). The search engine, based on LSA, partially prevents from problems related to polysemy, synonymy and inflections (Landauer and Dumais, 1997).

Apex 2.1 has been designed to help learners acquire knowledge through reading and writing (the "write-to-learn" approach). The advantage is to bring a flexible approach where the user evolves according to his or her wishes. Furthermore, he or she brings feedbacks to "real" contents, rather than answers to Multiple Choice Question papers or other closed exercises.

During the validation process, we have noticed that participants select the texts they wish to read and summarize. Most texts they selected as unable to summarize are texts they have not read thoroughly. We also noticed that most participants indicate that *Apex 2.1* is easy to use and they would use it from time to time if available. This therefore supports the continuing development of this software.

This version has some limitations. The main one is lack of feedback refinement after the evaluation. Indicating to the learners that they seem to have understood the text, does not encourage them to try and improve their summaries. On the other hand, indicating that the text has not been understood without giving any suggestions is not constructive. For these reasons, in the next version currently under development (see Dessus et al., 2011), we aim at remedying these problems and allowing different kinds of feedback such as:

- coherence within the summary (so as to detect breaks in coherence);
- links between summary and source-text (so as to indicate the user sentences that might be off the point, or even those that can be reused in the summary).

This will then allow users to know on which points to work again in order to improve their summary or synthesis. The feedback would then improve both summary writing techniques (e.g., concentrating on inter-paragraph coherence), but also on the content (thanks to the emphasis on semantic links between source-texts and syntheses). Furthermore, in order to improve the appropriation of texts they have read, the learners will be able to highlight important sentences and take notes in the dedicated notepad. Finally, they will be able to keep track of the different versions of their summaries and the associated notes and comments.

This article shows that it is possible to provide e-learning students with a tool that automatically assesses some semantic properties of their written production, and thus their understanding. This tool, though imperfect yet, has been positively appreciated by three categories of potential users.

AUTHORS' NOTES

This work is partly supported by the LTfLL (*Language Technologies for Lifelong Learning*) research project, FP 7, ICT-STREP. We wish to thank Thomas Lebarbé and Lucy Garnier for translating a part of the paper, the participants of the validation study, as well as the MSH-Alpes, Grenoble for providing material support.

REFERENCES

Bereiter, C., and Scardamalia, M. (1987). *The psychology of written composition*. Illsdale: Erlbaum.

Bétrancourt, M. (2007). L'ergonomie des TICE: quelles recherches pour quels usages sur le terrain ? In B. Charlier and D. Peraya (Eds.), *Regards croisés sur la recherche en technologie de l'éducation* (pp. 77–89). Bruxelles: De Boeck.

Denhière, G., Lemaire, B., Bellissens, C., and Jhean-Larose, S. (2007). A semantic space for modeling children's semantic memory. In T. K. Landauer, D. McNamara, S. Dennis and W. Kintsch (Eds.), *Handbook of Latent Semantic Analysis* (pp. 143-165). Mahwah: Erlbaum.

Dessus, P. (2009). An overview of LSA-based systems for supporting learning and teaching. In V. Dimitrova, R. Mizoguchi, B. du Boulay and A. Graesser (Eds.), *Artificial Intelligence in Education. Building learning systems that care: From knowledge representation to affective modelling (AIED2009)* (pp. 157-164). Amsterdam: IOS Press.

Dessus, P., and Lemaire, B. (2002). Using production to assess learning: An ILE that fosters Self-Regulated Learning. In S. A. Cerri, G. Gouardères and F. Paraguaçu (Eds.), *Intelligent Tutoring Systems (ITS 2002)* (pp. 772-781). Berlin: Springer.

Dessus, P., Lemaire, B., Loiseau, M., Mandin, S., Villiot-Leclercq, E., and Zampa, V. (2011). Automated free-text assessment: Some lessons learned. *International Journal of Continuing Engineering Education and Life-Long Learning, 21*(2/3), 140-154.

Emig, J. (1977). Writing as a mode of learning. *College Composition and Communication, 28*(2), 122-128.

Hart, S. G., and Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In P. A. Hancock and N. Meshkati (Eds.), *Human Mental Workload* (pp. 139–183). Amsterdam: North-Holland.

Klein, P. D. (1999). Reopening inquiry into cognitive processes in writing-to-learn. *Educational Psychology Review, 11*(3), 203–270.

Landauer, T. K., and Dumais, S. T. (1997). A solution to Plato's problem: the Latent Semantic Analysis theory of acquisition, induction and representation of knowledge. *Psychological Review, 104*(2), 211-240.

Tricot, A., Plégat-Soutjis, F., J-F., C., Amiel, A., Lutz, G., and Morcillo, G. (2003). Utilité, utilisabilité, acceptabilité: interpréter les relations entre trois dimensions de l'évaluation des EIAH. In C. Desmoulins, P. Marquet and D. Bouhineau (Eds.), *Actes de la conférence EIAH 2003* (pp. 391–402). Paris: INRP/ATIEF.

Vovides, Y., Sanchez-Alonso, S., Mitropoulou, V., and Nickmans, G. (2007). The use of e-learning course management systems to support learning strategies and to improve self-regulated learning. *Educational Research Review, 2*(1), 64-74.

In: Evaluation in E-learning
Editors: Y. Psaromiligkos, A. Spyridakos et al.
ISBN: 978-1- 61942-942-0

Chapter 10

THE DELPHI TECHNIQUE AS A PARTICIPATORY METHODOLOGY IN DESIGN, DEVELOPMENT AND EVALUATION OF T-LEARNING: E-LEARNING EVALUATION AS A SOCIAL PRACTICE

Andreja Istenic Starcic[1,*], Kemal Alic[2,†] and Matej Zajc[2,‡]

[1]University of Ljubljana, Faculty of Civil and Geodetic Engineering;
University of Primorska, Faculty of Education, Slovenia
[2]University of Ljubljana, Faculty of Electrical Engineering, Slovenia

ABSTRACT

The Delphi technique is researched and presented as a participatory methodology in expert design, development, testing and evaluation of t-learning. It is formed from the evaluation perspective of social relations when defining educational evaluations in e-learning, illustrating it on t-learning as a constitutive part of e-learning.

Television (TV) has been maintaining its predominant broadcasting, centralised transmission and programme production applications. For this reason, its role in education and training has been lagging behind in the past decades. By introducing interactive digital TV, the space became promising for the development of its new applications, and thus, for reinforcing its role in education and training. How t-learning will respond to the challenges and take place within e-learning mainly depends on how potentials of interactive digital TV will be integrated and used in the educational arena. The Delphi technique - proven to be successful in confronting and clarifying the issues of planning and development - has led to a consensus of all the stakeholders to set up technical and pedagogical guidelines for t-learning. The methodology of the Delphi approach to the issue is described and critically assessed in the framework of a real practical case and good practice focusing on t-learning general guidelines and

* E-mail: andreja.starcic@siol.net.
† E-mail: kemalic@gmail.com.
‡ E-mail: matej.zajc@fe.uni-lj.si.

development of a t-learning course. There have been several t-learning courses developed; an in-depth analysis is presented for one of the topics for the area of ‘Development of traffic skills'.

When planning t-learning, diverse end-user groups should be considered. TV is the main information medium for groups with no or low access to computer with Internet connection. Some of these groups are having gaps in their digital literacy and therefore are more vulnerable to digital exclusion. On the other hand, young generations which are highly digitally literate are shifting the TV use into the computer use, thus providing a higher level of interactivity. At the end of our study, the main characteristics of educational evaluation as social practice are highlighted in order to identify the main roles of stakeholders and indicators to be used in e-learning evaluation.

INTRODUCTION

In our iterative process of design and development of an interactive TV learning course, the Delphi technique was used to develop shared knowledge and to design and evaluate a t-learning course. In the development process of our t-learning course participatory design methods were used, enabling us to identify the relevant knowledge and multiple experiences acquired by stakeholders. The end-users and experts represented the main groups from the fields of information communication technology, pedagogy and psychology. As identified by Hamilton, some of the main benefits of working with multiple experts are improved knowledge-base validity, consistency, completeness, accuracy, relevancy, ability to deal with more complex problems and easier identification of incorrect results (Hamilton and Breslawski, 1996).

The Delphi technique enhances an efficient communication among geographically dispersed participants (ibid.). There were 270 participants in total from ten countries engaged in our Delphi project. The outcome of the project was identification of issues in t-learning design, implementation and evaluation. Our main objective was to aggregate opinions, knowledge and experiences of a number of individuals from different cultural and geographically dispersed backgrounds focusing on teaching and learning with TV.

TV is a mass medium and the access to its content has from the user point of view never been conditioned by the quality of the end-user terminal. The TV content has been viewable either on small or large TV sets or with the transition to colour TV there was no need for the end users to change their terminal equipment. The approaches presented in the paper are based on the use of minimum-requirement end-user equipment (set top BOX - STB) and transition of the TV content to the learner via a broadcast channel.

The possibility to reach the end users necessitating no additional end-user equipment might open the door to those that are often left out by modern technologies. Furthermore, certain initiatives, such as the United Nations World Summit on the Information Society (Society, 2005) and the eEurope 2005 Plan (eEurope, 2005), are in favour of pursuing the goal of making the Information Society available to everyone, at any time and from any place (Pazos-Arias, et al., 2008). Following the main objective to cover the majority of the population (TV viewers), the IST strep ELU project (ELU) has selected the minimum-requirement end-user equipment widely available in our homes.

Our sample of participating experts and users was structured in order to seek the views and confront their opinions in a series of debates and paper questionnaires focusing on the t-learning course design, development and trial. This enabled us to identify the issues in t-learning design and evaluation mapping and to lay down a wider e-learning framework. In educational evaluation we tried to meet the needs of: Planning and decision-making; Improvement, change and innovation; Accountability; Certification and Human-resource development. The main issues identified within the Delphi process were learning objectives, teaching and learning process, learning approaches, level of interactivity, level of learner engagement and control of learning, and user interface design.

The State-of-the-Art

The educational value of TV throughout history has already been proven. The courses displayed on it range from simple educational shows (e.g. “Winky Dink and You”) on broadcast TV to more formal courses broadcast inside closed communities such as universities. With digitalisation of TV, its learning potential has gained additional attention (Bates, 2005; Bellotti, Mikulecka, Napoletano, and Rohrova, 2006; Pazos-Arias, et al., 2008; Rey-López, et al., 2008; Zuga, et al., 2007). The serviceableness of the redefined medium for t-learning purposes has already been tested. The findings suggest that interactive TV applications provide support for education and entertainment for children and young people, as well as continuous education for all (Chorianopoulos and Lekakos, 2007). We have witnessed the t-learning courses produced and used in practice (Aarreniemi-Jokipelto, 2006; Aarreniemi-Jokipelto and Tuominen, 2004; Aarreniemi-Jokipelto, Tuominen, Kalli, and Rikonen, 2005). Special purpose tools for iDTV course preparation are being produced. Though, as already stated by (Bates, 2003) that “Hardly any research has yet been conducted on how people learn with interactive TV”, there is still limited pedagogical research for early pioneering developments to draw upon to help understand how learners may best learn through this medium, despite of more than 50 years of experience of using educational broadcasting. In (Bates, 2003) it is noticed that “Interactive TV focuses on the broad concepts of what is or might be possible, rather than going into detail about specific technical and operational requirements.” Also, access to description of implementation and description of the already implemented courses is negligible. Courses that have been developed or are under their developmental phase predict that users/potential learners have at their disposal additional terminal equipment, as for instance the Personal Video Recorder (PVR) in (Rey-López, et al., 2008) or have support for the return channel (Aarreniemi-Jokipelto, 2006). One of the biggest potentials of TV is to deliver the content to a vast majority of the population. T-learning has its biggest potential in the population having no access to other forms of e-learning: “There is increasing evidence that web-based learning is not reaching the social sectors which are more reluctant to contact with the new technologies, thus leading to inequalities in the access to education and knowledge in the Information Society” (Pazos-Arias, et al., 2008). Also, broadcast video described as a "powerful medium that can provide narrative visualization and can engage multiple sensors of learners simultaneously is often neglected in today's t-learning" (Palmer, 2007).

Introduction to T-Learning (Quote a Small Introduction to T-Learning and Its Scope. What Is the Learning Paradigm Offered by T-Learning?)

From the user point of view, TV has brought many improvements. For educational purposes, the following advances show the biggest potential:

- Higher image quality – more colours and better contrast will make the reading easier. Smaller details can be included into the picture presented.
- Inclusion of data into a broadcast TV stream - data in combination with a STB (Set Top Box) computational power allows simple games, multimedia-rich teletexts, or user accessible applications.
- Access to an additional content through interaction (a return channel needed)
- Interaction user TV - users can browse the data available to them. They can either request the data through a remote channel or just interact with the data included into the broadcast stream.
- Personalization of the content – since there is more content available one user can access, the personalization of the content is necessary. Depending on the user profile, the system makes recommendations on the content to be accessed.

Even though TV is a meeting the new form, some of its characteristics remain the same. Speaking in terms of learning they can be summarized into:

- Remote control – is almost the same as it is used to be. The four added coloured buttons are not considered to be an important functional improvement. The market propositions towards more advanced remote controls have never found the way into our homes. We do not expect this to happen in the near future.
- Scheduled TV – TV has remained above all a broadcasting medium. So, the schedule has still to be followed. Special on-demand services are currently only in urban areas.

T-Learning on the Context of E-Learning

The t-learning concept should not be compared to the PC-based learning concept (or as usually referred to e-learning). Productions of the courses themselves differ, since there is a huge difference between the two concepts which can be found already at their foundations. TV is watched from a distance which normally exceeds three meters, while the PC monitor is usually placed thirty centimetres from the end-user (in our case the learner). One may argue that the TV screen is larger than the PC screen, thus the screen resolution should be considered, which is in case of TV still relatively small (the high-definition TVs are coming into our homes, but they are still rather scarce). The only means of interaction with TV is the remote control (though there are more advanced remote controls already available in the market they are rarely found in the end-user homes), whereas for PC learning the student can use a mouse and a keyboard. While watching TV is often a social event, we usually sit in front of the computer alone. The last major difference to be noted is that TV shows are based on the video content, while the PC-based content is still prevailing in a text form enriched with pictures. A more detailed analysis of the differences has already been conducted by numerous authors (Lytras, et all, 2002; Colace, et all, 2008).

METHOD

E-Learning Evaluation as a Social Practice

According to Lincoln and Guba, the aim of the evaluation is to determine the value of a t-learning course, in order to improve it and assess its impacts (Lincoln and Guba, 1986). Our evaluation was process-oriented, and consisted of a formative and summative evaluation. Our purpose was to capture the process and collect information on teaching and learning activities and characteristics (teaching and learning approaches and learning objectives related to learning outcomes). A qualitative research methodology was used to investigate the educational process. "/.../ qualitative research study things in their natural settings, attempting to make sense of or interpret the phenomena in terms of the meaning people bring to them" (Denzin and Lincoln, 1994). The information was gathered in a series of rounds from experts, authors, observers and end-users.

The selected focus groups were used to address the topics which had not been considered by our participants prior to the evaluation study, and had turned up in focusing on the t-learning of diverse learners groups. The aim was to identify new topics and the level of understanding and interpret individual actions and attitudes within a given context. The focus groups were most effective when a large volume of information was to be gathered in a short period of time (Morgan, 1998).

The evaluation process plays an important role in educational policies and practices when promoting t-learning in diverse user groups. As to the process of defining the worth or value of an educational policy and/or practice within the context of t-learning, the evaluation process involves stages of standards identification. The participatory process in which various stakeholders confront their understanding when setting up standards is crucial. Caro (Caro 1977) defines the purpose of evaluation as information and judgement.

The Delphi method has been identified as an iterative process designed to combine the expert opinion into group consensus (Lynn et al., 1998; Kenney et al., 2001). According to Moore (Moore, 1987) in the Delphi method, a heterogeneous sample is applied to provide confrontation of a spectrum of opinions. It permits collection of rich evaluation data in an iterative process demonstrating participants how their ideas and opinions are utilized (Garavalia and Gredler, 2004) in the various processes (design, evaluation, research).

The intention of the Delphi is not to provide statistically significant results. Its main value is to find out whether the consensus is generated or not. The argumentation for the extreme positions is an important part of the Delphi process bringing an important outcome. Existence of a consensus obtained by using the Delphi method does not necessarily mean that a correct answer had been found (Kenney et al., 2001).

The Delphi method involves an in-depth data collection from multiple trustful sources. Lincoln and Guba (Lincoln and Guba, 1985) describe four aspects of trustworthiness, which can be applied in a qualitative process: credibility, transferability, dependability and confirmability. How may the readers view conclusions of evaluation study based on the Delphi method within their contexts and use them in their own situations, impacts transferability as well as generalization of the findings of our study.

Utilization of the Delphi Method

A qualitative research methodology (Stake, 1994) was used to investigate the course content design, trial lesson and authoring process is in their natural environments (Denzin and Lincoln, 1994), with specific attention paid to the context (Greene, 1994).

The aim of using the Delphi was to identify the issues in the t-learning design, implementation and evaluation process with a special focus laid on learning objectives, learning processes, learning approaches, level of interactivity, level of the learner's engagement and control of the learning and end-user interface design. A score-card methodology was used to gather the data provided by the evaluation teams. The use of the score card in the evaluation process is presented in Table 1.

Unlike our questionnaires, the aim of using the Delphi method was to reach a consensus of opinions, judgements or choices (Keeny et al., 2001). The Delphi method starts with the initial open-ended questionnaire.

(a)Based on the information gathered, the coding process produces a list of value statements in which the observers assess the course contents, trial lessons and authoring process.
(b)A “weight” is assigned to each statement to provide its objective ranking statements from most to least important. The rating factor allotted to each qualitative measure is assigned with a number of representatives from each stakeholder group. The resulting rating-factor averages are used for the evaluation frame.
(c)A Score-card is created (see in Table 1).
(d)A calculation model to obtain the score-card results is designed: Score (S) X Weight Factor (WF) = Adjusted Score (AS). In the calculation process, each score (S) is multiplied by the weight factor (WF) to find the adjusted score (AS).

Table 1. Score-card use in the evaluation

Assessment area	Course content		Trial lessons		Authoring process	
	Expert scorecard	Author scorecard	Observer scorecard	User scorecard	Expert scorecard	Author scorecard
Learning objectives						
Learning processes						
Learning approaches						
Level of interactivity						
Level of the learner's engagement and control of learning						
End-user interface design						

Usually, two to four rounds often take the form of structured questionnaires to incorporate feedback to each panel member. The rounds are analysed and re-circulated in order to encourage the panel members to become more involved and motivated (Kenney et al., 2001, Walker and Selfe, 1996). In each round, the participants are given information about the responses from a previous round (Hasson et al., 2000).

A ten-step procedure for conducting the two-round Delphi was used: (1) formation of the steering committee, (2) selection of expert panels, (3) design of the first round questionnaires, (4) testing of questionnaires, (5) conduct of the first-round questionnaires among the panellists, (6) analysis of the first round responses, (7) design of second round questionnaires, (8) conduct of second round questionnaires among the panellists, (9) analysis of the second-round responses and (10) data analyses.

The procedure started by setting up a panel group identifying and consolidating the main issues to be dealt with in our questionnaires using the Likert scales. The scales were designed for different experts and end-users. The questionnaires were translated in the languages of each of the participating countries. The scale ranged from 0 = strongly disagree to 4 = strongly agree. The questions were conducted in two stages. The experts allotted weights to each questions. After the first stage, the answers were compiled according to the assigned weights, thus providing the basis for designing the second-stage questionnaires. The number of the questions was then reduced. The second stage over, the data were summarised and provided to decision makers and respondents. Besides filling-in the questionnaires, the panel members also participated in an interview enabling them to gather detailed information to be used in the t-learning course design, development and trial process.

Participants

There were altogether 270 users from ten countries participating in our Delphi. Our sample was structured from experts and users from the ICT, pedagogical and psychological fields.

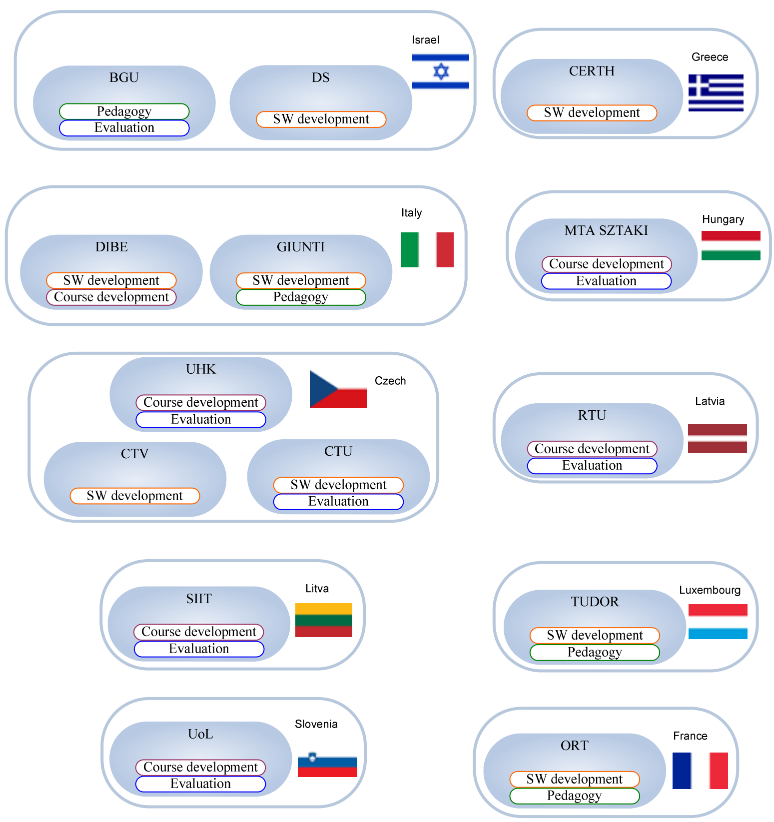


Figure 1. Main roles of the interdisciplinary multicultural teams participating in the Delphi evaluation project.

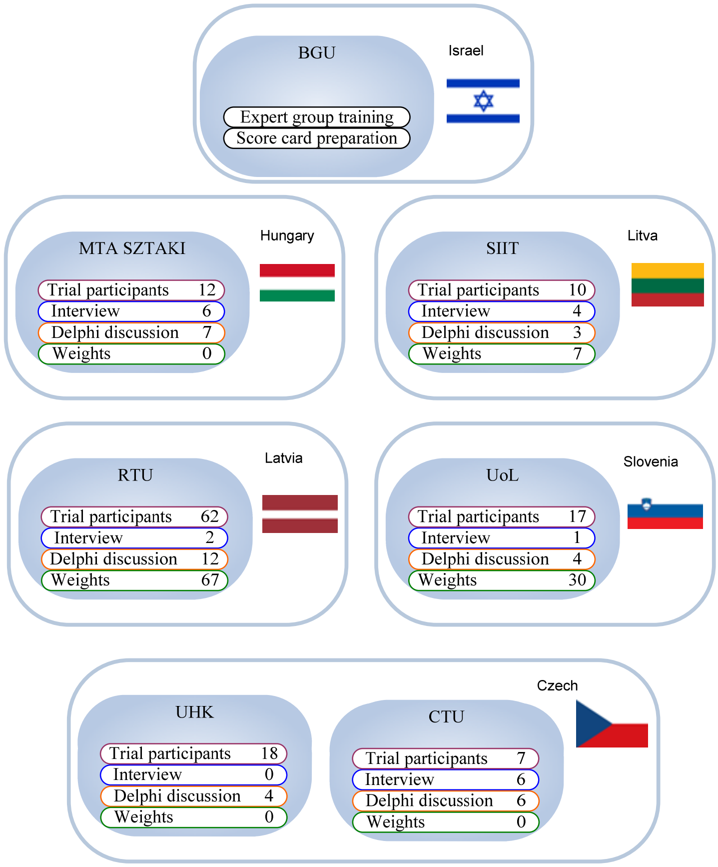


Figure 2. Project structure and the main evaluation tasks in the participating teams.

Our aim was to seek their views and confront their opinions within a series of debates and paper questionnaires about the t-learning course design, development and trial processes. The data were collected in the period from 2006 to 2008. The structure of the Delphi project participants is presented in figures 1 and 2.

Results and Discussion

Development of Traffic Skills

There are six courses developed within the project. For one of them, i. e. Development of the traffic skills course in-depth analysis is presented. The course was used as a supplementary learning material and providing additional learning material, video demonstrations, cases and examples. Powerful video means were used in the t-learning context to represent the multimedia rich content.

Using this type of the t-learning course is not mandatory for the learner. Also the learner assumed to already have some knowledge on the topic. The objective is to improve the learners' understanding of the traffic rules. The paradigm cause and consequences is studied in details. Knowledge and concepts are illustrated with real life examples.

Instructional templates were developed to address specific competence, skill and knowledge requirements of the learning process. They optimise decision-making regarding teaching actions and assigning tasks to students. The structure developed for the learning course allows for:

- explanation,
- demonstration,
- performance,
- feedback, and
- assessment.

Games, multimedia page and Audio/Video (A/V) stream are instructional elements the course is based on. They are well described in (Alic, et al. 2009; Bellotti, et al., 2006; Bellotti, et al., 2008; Vrochidis, et al., 2008).

The learning course is divided in four subsections. The composition of the first ones is generally the same:

- explanation and demonstration (the learner can follow the content presented in either the A/V stream or in the multimedia pages and presentations and
- performance (the learner is asked questions on topics introduced in the presentation part. The video is used only as a background and synchronization).

The last subsection of the learning course provides an assessment. It follows the video, is time unlimited and the learners can browse through it as long as they want.

The main story is presented in the A/V part. The interactive elements are connected to actions presented in the A/V stream. During the performance part, the A/V part is used only for synchronization. The A/V part is used for explanation and demonstration.

Games are used in the performance and assessment part to assess the learner's knowledge. According to the number of score acquired, the user's profile is built. Knowledge about the user is used in proposing on additional content accessible through a delivery subsystem, multimedia pages and presentations.

The multimedia pages and presentations are used in providing the learner with an additional content.

In the first three subsections, the A/V stream is used to explain the facts, elements and concepts. The next to follow is demonstration. There is minimally one real life example shown through the video stream. During the A/V stream, the learners may skip the narrative part and follow the content presented in the multimedia pages and presentations in which they can either explore examples related to the instructions being presented or improve their knowledge on topics closely related to the one being presented. Generally, the learners cannot use the given options but can follow the main A/V course or surf through an additional content.

The second part in the first three subsections is Practice A/V stream runs in the background and is not related to the information processed. Through games and multimedia elements (in our case only pictures), the learners are given an opportunity to practice the knowledge they need. They are offered to practice of their interest and profile. This subsection allows following each of the prepared units the learner will choose. Through the games developed in the ELU project, the learner is able to practice to recall the already known facts, elements and concepts.

The last subsection is assessment. The learners who are asked about facts, elements and concepts are given a feedback at the end of the course.

The course has the following properties that need to be briefly described in this book:

- The learning course is of highly interactive nature. Throughout the course, the users are encouraged to interact with the TV content. They are asked questions and offered different additional contents.
- Personalization of the content. Through the course the system builds a simple user profile basis of which the learner is offered an additional content (the A/V stream part of the content is the same for all the learners).
- The course combines asynchronous and synchronous learning events. A similar approach to our knowledge has never been used before. Our approach was already described in our previous work (Alic et al., 2008).
- When additional content is proposed to the learner the system takes into account the time component as well. Since the source is based on the A/V stream there are always time limitations. The course proceeds at the speed appropriate for differently skilled learners in a different way.

The Delphi process with serial rounds and follow-up actions provided detailed information. The gathered information was firstly implemented in the design structure of the existing course and secondly investigated in relation to the present and future issue of t-learning.

By using the Delphi method, we wanted to evaluate learning events. It was noted that the learners formed the learning objectives laid down in the Explanation and Demonstration events clear and as such they performed them. With the two learning events enabled them learners either to watch the A/V stream or access a certain interactive content. The different learner groups performed the learning process differently and the learning events were combined in different ways too. The learners' behaviour was observed in terms of their engagement and control of learning which was provided in different ways:

- A/V stream-based course: For the learners refusing to interact with the content, the interaction was reduced to a minimum. These learners were allowed to lean back, listen and watch the TV course almost in a traditional manner. In the test group, there were just few that interacted only in the obligatory parts.
- The majority of the learners used the option of accessing the additional content for several times in the course. They spent most of the time watching the A/V stream of the course and accessed the additional content only if the description of the content, when offered, seemed appealing/interesting to them. We noticed two types of learners' behaviour in accessing the additional content:
- Learners opened the additional content and browsed through it for a longer period of time. They were reading the additional content and watching the A/V stream at the

same time. When the situation in the A/V stream was less interesting to them, they were reading the additional content.

- Learners opened the additional content, browsed through it, closed it and followed the happening in the A/V stream.
- Some learners accessed all the available additional content. Regardless the description of the additional content, they loaded it on the screen and quickly browsed through it. It was at this point that they decided if they wanted to dedicate more time to it or not.

The learners found the Performance and Feedback events clear, too. Different learners managed to answer a different number of questions. Moreover, the learners that incorrectly answered at least one of the questions reacted differently to the offered additional content. There was no clear distinction between the learners accessing all the additional content and those not accessing it. As seen from our observations, the learners decided by themselves whether they needed additional reading or not.

An important conclusion drawn from the Performance and Feedback events is that the learners did not mind waiting for a few seconds. One of the possible situations allows the learners to finish the interactive content early and stay with the empty A/V stream for some time (in our case maximum ten seconds). The learners did not mind this kind of situations and were not bothered by exact synchronisation between the interactive content and the A/V stream.

The different learning approaches were not addressed in the desired extent due to the technological limitations. By including the A/V stream into the course, we lose flexibility in personalisation of the content and we need to follow the A/V stream timeline. On the other hand, if we do not use the A/V stream, we can build courses that are more similar to the computer-based ones, but in our opinion the video should be the centre of the t-learning course as it brings an additional value to learning.

The human computer interaction design was evaluated through learners' comments and questionnaires. We were interested in the reaction times and ability of the learners to navigate through the course with a simple remote control.

- None of the learners complained about the size of the text. They were asked to watch TV from the distance they usually take back at home. As they were settling down there was no text on the screen. Our conclusion is that learners are used to reading the subtitles that are common while watching foreign shows. We see no need for the user interface to adapt the size of letters on the screen.
- Some of the learners got frustrated after pressing the correct button on the remote control and there was no reaction of the system. Moderators encouraged these learners to continue. This is the nature of the TV set. The remote control is not a reliable device and the same button sometimes needs to be pushed more times in a row for the TV set to react.
- The learners were not used to handle the remote control. We noticed the learners had more problems with the remote control at the beginning of the course than towards its end. We assume that if the learners used their own remote control, there would be fewer difficulties for them.

- Throughout the project, we dealt with the issue of the response time of the system. As the STB is a low-cost dedicated computer, we assumed that there would be a great possibility for the learners to be distracted by the slow system. The learners were asked if the slow system had bothered them and surprisingly enough none of them complained. The loading time was not an issue for our test group either.
- The learners' experience was positive. They particularly liked the ability to extend the lifetime of the course. The main idea after watching the course was enriching the current documentaries with the new technology. Having further questions and availability to read more on topics of their interest seemed generally appealing to the learners. In the demonstrational course they missed the ability to access all the additional contents after the A/V stream had ended.

CONCLUSION

The Delphi method was used effectively sharing multiple experiences in a process of identification of relevant approaches to the t-learning design and evaluation. In the process the knowledge and experiences of numerous experts, authors, observers, and end-users from ten countries provided the basis for our course design. The wide context of ten countries was analysed with regard to different groups of end-users.

The conclusion drawn from this study shows that the utilization of the Delphi is not the best solution. Generalization of our findings depends on the audience and their comprehension of the reported interpretations and on the way they can apply ours in their contexts.

REFERENCES

Alic, K., Istenic-Starcic, A., Tasic, J. F. and Zajc., M. (2009). Conceptual interactive multimedia learning content production. *Electrotechnical Review*, 76(5), 269-274.

Alic, K., Zajc, M., Tkalcic, M., Burnik, U. and Tasic, J. F. (2008). Development of an interactive television t-learning course. In *14th IEEE Mediterranean Electrotechnical Conference,* (pp. 139-144). Ajaccio: IEEE.

Aarreniemi-Jokipelto, P. (2006). *Modelling and content production of distance learning concept for interactive digital television.* Unpublished PhD, Helsinki University of Technology, Helsinki.

Aarreniemi-Jokipelto, P., and Tuominen, J. (2004). Experiences with an Interactive Learning Environment in Digital TV. In *Proceedings of the IEEE International Conference on Advanced Learning Technologies (ICALT)*. Joensuu, Finland.

Aarreniemi-Jokipelto, P., Tuominen, J., Kalli, S., and Rikonen, T. (2005). Experimenting with Digital Television Learning Environments. In *16th EAEEIE Annual Conference on Innovation in Education for Electrical and Information Engineering*. Lappeenranta, Finland.

Bates, P. J. (2003). t-learning Study, Final Report. In (Vol. 09).

Bates, P. J. (2005). New Learning Perspectives. In (Vol. 09).

Bellotti, F., Mikulecka, J., Napoletano, L. and Rohrova, H. (2006). Designing a Constructionistic Framework for T-Learning. In *Innovative Approaches for Learning and Knowledge Sharing.* (pp. 549-554). Berlin / Heidelberg: Springer.

Bellotti, F., Vrochidis, S., Ramel, S., Tsampoulatidis, I., Lhoas, P., Pellegrino, M., Bo, G. and Kompatsiaris, I. (2008). A Technological Framework for the Authoring and Presentation of T-learning Courses. *International Journal of Emerging Technologies in Learning (iJET),* 3, 10-19.

Caro, F. G. (1977). *Readings in evaluation research* (2nd ed., pp. 3-38). New York: Sage.

Chorianopoulos, K., and Lekakos, G. (2007). Learn and Play with Interactive TV. *Computers in Entertainment (CIE), 5*(2).

Colace, F., Santo, M. D., Ritrovato, P., and Mascambruno, P. R. C. (2008). From E-Learning to T-Learning. In *3rd International Conference on Information and Communication Technologies: From Theory to Applications, 2008. ICTTA 2008.* (pp. 1-6). Damascus.

eEurope, *eEurope 2005 plan*. Available at http://ec.europa.eu/ information_society/eeurope/ 2005/index_en.htm, accessed January 2010.

Denzin, N. K. and Lincoln, Y. S. (1994). Introduction: Entering the field of Qualitative Research. In N. K. Denzin and Y. S. Lincoln (Eds.) Handbook of Qualitative Research (pp. 1-17). Thousand Oaks: Sage Publications.

ELU, *ELU project*. Available at http://www.elu-project.com/, accessed November 2009.

Garavalia, L. and Gredler, M. (2004). Teaching Evaluation through Modeling: Using Delphi Technique to Assess Problems in Academic Programs. *American Journal of Evaluation, 25(3),* 357-380.

Greene, J. C. (1994). Qualitative Program Evaluation. In N. K. Denzin and Y. S. Lincoln (Eds.) Handbook of Qualitative Research (pp. 530-544). Thousand Oaks: Sage Publications.

Hamilton, D. M. and Breslawski, S. (1996). Knowledge Acquisition for Multiple Site-related Domain Expert System: Delphi Process and Application. *Expert System With Aplications, 11(3),* 377-389.

Hasson, F., Keeney, S., McKenna, H. (2000). Research guidelines for the Delphi survey technique. *Journal of Advanced Nursing, 32 (4),* 1008-1015.

Kenney, S., Hasson, F., McKenna, H. P. (2001). A critical review of the Delphi technique as a research methodology for nursing. *International Journal of Nursing Studies*, *38*, 195-200.

Lincoln, Y. S., and Guba, E. G. (1985). *Naturalistic Inquiry*. Bewerly Hills: Sage.

Lincoln, Y. S. and Guba, E. (1986). Research, evaluation and policy analysis: heuristic and disciplined inquiry. *Review of Policy Research, 5(3),* 546-565.

Lynn, M. R., Layman, E. L., Englebardt, S. P. (1998). Nursing administration research priorities: A National Delphi study. *Journal of Nursing Administration 28(5),* 7-11.

Lytras, M., Lougos, C., Chozos, P., and Pouloudi, A. (2002). Interactive Television and e-Learning Convergence: Examining the Potential of t-Learning. In *European Conference on e-Learning*. Brunel University, UK.

Morgan, D. L. (1998). *The focus group guidebook*. Thousand Oaks, CA: Sage.

Moore, C. (1987). *Group techniques for idea building*. Newbury park, CA: Sage.

Palmer, S. (2007). An evaluation of streaming digital video resources in on- and off-campus engineering management education. *Computers and Education, 49*(2), 297-308.

Pazos-Arias, J. J., López-Nores, M., García-Duque, J., Díaz-Redondo, R. P., Blanco-Fernández, Y., Ramos-Cabrer, M., Gil-Solla, A. and Fernández-Vilas, A. (2008).

Provision of distance learning services over interactive digital tv with mhp. *Computers and Education, 50(3),* 927-949.

Pazos-Arias, J. J., López-Nores, M., García-Duque, J., Díaz-Redondo, R. P., Blanco-Fernández, Y., Ramos-Cabrer, M., Gil-Solla, A. and Fernández-Vilas, A. (2008). Provision of Distance Learning Services over Interactive Digital TV with MHP. *Computers and Education,* 50, 927-949.

Rey-López, M., Díaz-Redondo, R. P., Fernández-Vilas, A., Pazos-Arias, J. J., López-Nores, M., García-Duque, J., Gil-Solla, A., and Ramos-Cabrer, M. (2008). T-MAESTRO and its authoring tool: using adaptation to integrate entertainment into personalized t-learning. *Multimedia Tools and Applications, 40*(3), 409-451.

Stake, R. (1994). Case Studies. In N. K. Denzin and Y. S. Lincoln (Eds.) *Handbook of Qualitative Research* (pp. 236-247). Thousand Oaks: Sage Publications.

Society, *World Summit on the Information Society*. Available at http://www.itu.int/wsis/index.html, accessed January 2010.

Vrochidis, S., Bellotti, F., Bo, G., Napoletano, L. and Kompatsiaris, I. (2008). T-Learning Technologies. In *Encyclopedia of Information Science and Technology* Hershey. New York: Information Science Reference.

Walker, A.M., Selfe, J. (1996). The Delphi method: a useful tool for the allied health researcher. *British Journal of Therapy and Rehabilitation 3 (12),* 677-681.

Zuga, B., Ozolina, A., Kapenieks, A., Slaidins, I., Jirgensons, M., Strazds, A., and Stale, G. (2007). Towards a T-Learning Content and Usability Testing Environment. In *Annual Proceedings of Vidzeme University College, "ICTE in Regional Development"* (pp. 87-95). Valmiera, Latvia.

INDEX

#

A

B

C

D

E

F

G

H

I

J

K

L

M

N

O

P

Q

R

S

T

U

V

W

Y