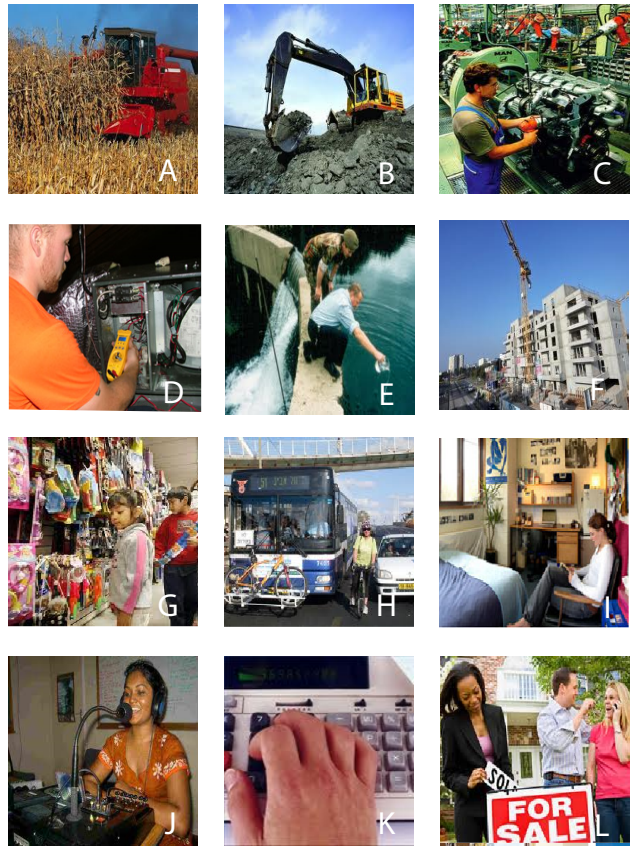
Proceedings of the

1st European Workshop on HCI Design and Evaluation

the influence of domain on Human Computer Interaction design and evaluation

Limassol, Cyprus - April 8th, 2011



Georgios Christou

Panayiotis Zaphiris

Effie Lai-Chong Law

(editors)

International Standard Industrial Classification of All Economic Activities

A - Agriculture, forestry and fishing
B - Mining and quarrying
C - Manufacturing
D - Electricity, gas, steam and air conditioning supply
E - Water supply; sewerage, waste management and remediation activities
F - Construction
G - Wholesale and retail trade; repair of motor vehicles and motorcycles
H - Transportation and storage
I - Accommodation and food service activities
J - Information and communication
K - Financial and insurance activities
L - Real estate activities
M - Professional, scientific and technical activities
N - Administrative and support service activities
O - Public administration and defence
P - Education
Q - Human health and social work activities
R - Arts, entertainment and recreation
S - Other service activities
T - Activities of households as employers
U - Activities of extraterritorial organizations and bodies

Publisher: IRIT Press, Toulouse, France
Proceedings of the 1st European Workshop on HCI Design and Evaluation
Georgios Christou, Panayiotis Zaphiris & Effie Lai-Chong Law (eds.)
102 pages. 2011

ISBN :978-2-917490-13-6 EAN : 9782917490136

Introduction and Acknowledgements

These are the proceedings of the 1st European Workshop on HCI Design and Evaluation. The broad theme of the workshop is the influence of domain on HCI design and evaluation. We have tried to include a variety of different domains in which HCI design and evaluation plays an important role, and we have had many submissions with topics ranging from E-Learning to Web and Information Technology, and from Safety Critical and Emergency domains to Industrial and Small to Medium Business Software domains. We even had submissions for User Evaluation Methodologies. For a detailed list of the themes covered in the workshop, take a look at the table of contents of the proceedings.

Naturally, as in any event, a long list of people and organizations are involved to create something that runs smoothly and works well. Such is the case here as well. Thus, as chairs, we owe a big thanks to the many people who have made this workshop possible. We firstly thank the two organizations that have made the workshop possible, ESF COST Action IC0904 – Towards the Integration of Transectorial IT Design and Evaluation, and SIGCHI Cyprus. We thank of course our session chairs, Andri Ioannou, Jan Gulliksen, and Ebba Thora Hvannberg. We thank our invited speakers, Alistair Sutcliffe and Davide Rocchesso. We thank our sponsors, European Science Foundation, Cyprus Tourism Organization, Cyprus University of Technology, University of Leicester, and European University Cyprus for giving us the support required for the smooth hosting of this workshop. A special thank you to Marco Winkler who designed the front and back cover of the proceedings, and more! Finally we thank our program committee. The program committee is the following:

The chairs,

Georgios Christou, European University Cyprus
Panayiotis Zaphiris, Cyprus University of Technology
Effie Lai-Chong Law, University of Leicester

Table of Contents

Guiding design and evaluation of e-learning systems

Carmelo Ardito
Università di Bari
via Orabona, 4
70125 Bari, Italy
ardito@di.uniba.it

Maria Francesca Costabile
Università di Bari
via Orabona, 4
70125 Bari, Italy
costabile@di.uniba.it

Rosa Lanzilotti
Università di Bari
via Orabona, 4
70125 Bari, Italy
lanzilotti@di.uniba.it

1st European Workshop on HCI Design and Evaluation: The influence of domains. April 8, 2011. Limassol, Cyprus

Abstract
E-learning is becoming very important in fields where access to learning materials needs to be brought about effectively and efficiently. Our experience in dealing with e-learning systems of different types has shown that several aspects influence the design and the evaluation of such systems, primarily the educational aspect, since the main purpose of such systems is to achieve educational goals. We are well aware that specific conceptual models and practical protocols are needed, capable of considering all the aspects that influence design and evaluations of e-learning systems. As a consequence, in the last few years we have developed the *TICS (Technology, Interaction, Content, Services)* framework and the eLSE (*e-Learning Systematic Evaluation*) methodology that, taking into account the most important aspects of e-learning system quality, aim at guiding designers and evaluators.

Keywords
E-learning system design, e-learning system evaluation, usability evaluation techniques

Introduction
In the e-learning context, a major challenge for designers is to develop software tools and environments that can engage novice learners and support their learning even at a distance. Towards this

end, there should be a synergy between the learning process and the learner's interaction with the software. As for any interactive system, usability is a primary requirement. If an e-learning system is not usable, the learner spends more time learning how to use the software rather than learning the contents. Besides being usable, an e-learning system addresses other challenges, e.g., it must be effective in meeting pedagogical objectives. System evaluation should thus integrate an assessment of the educational aspects.

Despite the large number of e-learning systems now available, one of the barriers to successful deployment of technology-based learning is the lack of high quality systems tailored to the needs of individual users and groups. Various definitions and frameworks for the quality of e-learning systems are reported in literature, but the identified solutions appear as a short blanket, able to cover only some of the multiple aspects that characterize the complexity of the problem [2, 3, 5].

Quality of e-learning systems has been primarily considered from the point of view of education experts, focusing on the educational content with little attention to how it is presented. With respect to this point of view, we emphasize the interaction dimension and, specifically, the interaction between the user (teacher or learner) and the overall system, not only its content (the learning materials). On the basis of our experience, we believe that, in the e-learning context, proper conceptual models, as well as design and evaluation techniques and methodologies, have to be considered in order to foster a user-system interaction that facilitates the learning process. It is not possible to neglect the influence of the user interface on the learners' activities [4, 10]. Many e-learning systems provide various functionalities and services, but it is often difficult, if not impossible, for users to find and use them in an effective and efficient way. Thus, we have proposed a definition of e-learning systems quality, which highlights the importance of the user interface in designing and/or evaluating the overall quality of e-learning systems [6]. If designers are fully aware of the high value of the user interface, they will not neglect this aspect of e-learning systems and will create products of better quality, which will contribute to the success of technology-based learning.

Looking at e-learning systems from the point of view of Human-Computer Interaction (HCI) experts, we have to acknowledge the great influence of the educational objectives of such systems in their design and evaluation. With the intent of balancing the above points of view in design and evaluation of e-learning systems, we have developed the TICS framework and the eLSE methodology described in the next sections.

TICS framework

By considering the literature on e-learning systems quality, the experience of HCI experts and the results of observing real users interacting with e-learning systems, a new framework, called *TICS (Technology, Interaction, Content, Services)*, has been developed. It addresses the most important aspects to be considered when an e-learning system is designed or evaluated. TICS primarily focuses on the quality dimensions recommended by the Italian Ministry of Education in [9], integrating them with dimensions that refer to the interaction aspect that, as we have discussed above, plays a crucial role in the fruition of the e-learning material. In the following the TICS aspects, and their

specific dimensions, as shown in Figure 1, are described.

Technology
The technology aspect focuses on the dimension *Hypermedial Technology*, since this is the technology on which e-learning systems are usually based. This aspect refers to the technological problems that can obstruct or make difficult the use of the e-learning system, namely compatibility with different operating systems; performance; monitor resolutions; hardware and software for system accessibility.

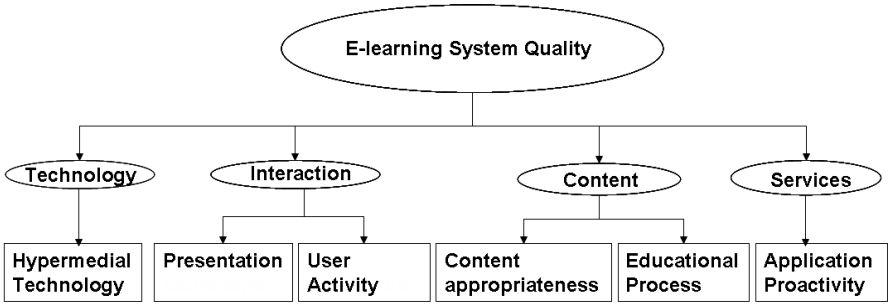


figure 1 *Quality aspects of e-learning systems*

Interaction
The explicit inclusion of the interaction aspect is one of the novelty of our approach with respect to the quality dimensions usually addressed in the e-learning literature. We believe it is crucial for technology-based learning. This aspect involves *Presentation* of the educational material and of the provided services and tools, and *User Activity* performed during the interaction with the system. In particular, user errors should be prevented as much as possible. If an error occurs, the e-learning system should provide appropriate support to manage it.

Content
The content aspect refers to the *Appropriateness* of the educational material that must be learned and to the *Educational Process* that involves the person learning through the system. Appropriateness and quality of the educational material could be achieved through an accurate learner-centered design [12]. This aspect also refers to the way the material is taught and to the capability of the e-learning system to propose study activities to the learner, who should also be free to autonomously choose his/her learning path (e.g. alternating moments of study, exercises, and verifications). Besides, the learner should have the possibility to go more deeply into the course content and the system should provide concrete, real, and helpful examples to facilitate understanding of the educational content.

Services
The services aspect refers to the *Application Proactivity*, which involves the tools that facilitate and support the user during the navigation through the system and the fruition of the educational material. It is necessary to provide users with communication tools, auto-evaluation tools, help services, search engines, references, scaffolding, and so on. Ease of learning and ease of use of such tools allow users to concentrate their efforts on the learning paths without being required to spend too much time trying to understand the way the e-learning system works.

Some authors already stated that the user interface must be designed and evaluated on the basis of well

defined criteria and methods, specific for e-learning [13]. The TICS framework has been developed in order to further emphasize that e-learning systems have to consider several aspects that influence their design and evaluation. TICS dimensions and corresponding guidelines should guide designers to focus on what is required for creating good quality systems. On the basis of TICS, we have also developed an evaluation methodology, described in the next section.

eLSE Methodology

eLSE (e-Learning Systematic Evaluation) methodology proposes to evaluate e-learning systems with the support of appropriate evaluation patterns [6, 8]. Here we briefly summarize the eLSE methodology; its goal is to increase the reliability and the effectiveness of e-learning system evaluation by proposing a structured and systematic approach to it, which considers all the important aspects that influence the quality of such systems, as addressed by the TICS framework.

eLSE suggests analyzing an application along specific dimensions (those highlighted by TICS), which address the appropriateness of design with respect to the peculiar nature and purposes of e-learning systems. eLSE proposes an inspection technique based on the use of *evaluation patterns* that are specifically defined for e-learning systems. For this reason, it is called *Patter- Based inspection* or (PB inspection).

For each TICS specific dimension (i.e. Hypermedial Technology, Presentation, User Activity, Educational Process, Content Appropriateness, Educational Process, Application Proactivity), general usability principles are identified and then decomposed into finer-grained criteria. By considering the literature on e-learning, results of users studies, and the experience of HCI experts, a number of specific guidelines have been identified and associated to these criteria, to be taken into account during the design phase. Then, a set of evaluation patterns, addressing these guidelines, has been defined [6, 8].

An evaluation pattern is a description of what an evaluator has to do when inspecting an application. Evaluation patterns guide the evaluator's activities by describing which elements of the application to look for, and which actions the evaluators must perform in order to analyse such elements. In this way, even novice evaluators, with lack of expertise in usability and/or application domain, are able to come out with more complete and precise results.

Evaluation patterns make possible to maximize the reuse of evaluators' know-how, by capturing usability inspection expertise, and by expressing it in a precise and understandable form, so that it can be easily reproduced, communicated, and exploited. They therefore allow evaluators to take "... advantage of any of the efforts done in previous works, to reduce the effort needed to achieve a new one" [11]. Evaluation patterns are precisely formulated by means of a template that provides a consistent format. More details on the library of evaluation patterns defined for evaluating e-learning systems can be found in [1].

eLSE methodology organizes the evaluation process activities into a *preparatory phase* and an *execution phase*. The preparatory phase is performed only once for each analysis dimension; its purpose is to create a conceptual framework that will be used to carry out actual evaluations. The output of the preparatory phase

can be easily shared among different evaluators, or different evaluation laboratories that have similar interests and evaluate such applications from similar points of view. The preparatory phase consists of the identification of guidelines to be considered for the given dimensions and the definition of a library of evaluation patterns. The execution phase is performed every time a specific application must be evaluated. It mainly consists of a PB inspection, performed by some evaluators. If needed, the inspection can be followed by user testing sessions, involving real users. At the end of each evaluation session, the evaluators must provide designers and developers with a proper evaluation feedback.

eLSE methodology also proposes an important distinction between the e-learning platform and the educational modules. The e-learning platform is the software environment that usually offers a number of integrated tools and services for teaching, learning, communicating, and managing learning material. The educational modules, also called learning objects, are the specific learning material provided through the platform. Design guidelines and evaluation patterns defined for the platform differ from those ones defined for e-learning modules, since different features and criteria need to be considered [1].

The eLSE methodology can be easily adapted to different types of applications in different domains, by identifying specific analysis dimensions and the proper set of evaluation patterns.

Validating the PB inspection

The main novelty of the eLSE methodology is the PB inspection. This technique, particularly suited for novice evaluators, has been empirically validated through a controlled experiment, which compared PB inspection to heuristic evaluation and user testing [7]. The study involved 73 novice evaluators and 25 end users, who evaluated an e-learning application using one of the three techniques. The comparison metric was defined along six major dimensions, covering concepts of classical test theory and pragmatic aspects of usability evaluation. The study showed that evaluation patterns, capitalizing on the reuse of expert evaluators know-how, provide a systematic framework which reduces reliance on individual skills, increases inter-rater reliability and output standardization, permits the discovery of a larger set of different problems and decreases evaluation cost.

Acknowledgements

The research reported in this paper has been supported by the following grants: COST Action 294 MAUSE (Towards the MAturation of Information Technology USability Evaluation), COST Action IC 0904 TWINTIDE (Towards the Integration of Transectorial IT Design and Evaluation), the Italian MIUR project L4ALL (Learning For All).

Citations

[1] Ardito, C., Costabile, F., Marsico, M. D., Lanzilotti, R., Levialdi, S., Roselli, T. and Rossano, V. 2006. An approach to usability evaluation of e-learning applications. Univers. Access Inf. Soc., 4, 3 (2006), 270-283.

[2] Chua, B. B. and Dyson, L. E. 2004. Applying the ISO 9126 Model to the Evaluation of an e-Learning System In Proceedings of the ASCILITE 2004 (Perth, Australia, 2004), 184-190.

[3] Dmytro, R. 2004. Distance Courses Quality: A Learner's View. In Proceedings of the IEEE International

Conference on Advanced Learning Technologies (ICALT'04) (Joensuu, Finland, 2004), 1080-1081.

[4] Ewing, J. and Miller, D. 2002. A framework for evaluating computer supported collaborative learning. Educational Technology & Society, 5, 1 (2002), 112-118.

[5] Herrington, A., Herrington, J., Oliver, R., Stoney, S. and Willis, J. 2001. Quality Guidelines for Online Courses: The Development of an Instrument to Audit Online Units. In Proceedings of the ASCILITE 2001, (Melbourne, Australia 2001), 263-270.

[6] Lanzilotti, R. A holistic approach to designing and evaluating e-learning quality: usability and educational effectiveness. PhD Thesis, University of Bari, Bari, Italy, 2006.

[7] Lanzilotti, R., Ardito, C., Costabile, M. F. and De Angeli, A. 2011. Do patterns help novice evaluators? A comparative study. Int. J. Hum.-Comput. Stud., 69, 1-2 (2011), 52-69.

[8] Lanzilotti, R., Ardito, C., Costabile, M. F. and De Angeli, A. 2006. eLSE Methodology: a Systematic Approach to the e-Learning Systems Evaluation. Educational Technology &Society, 9, 4 (2006), 42-53.

[9] Montedoro, C. and Infante, V. Linee Guida per la Valutazione di Qualità del Software didattico nell'e-Learning. ISFOL, I libri del Fondo Sociale Europeo Roma, 2003.

[10] Moore, M. 1989. Editorial: Three types of interaction. American Journal of Distance Education, 3, 2 (1989), 1-7.

[11] Nanard, M., Nanard, J. and Kahn, P. 1998. Pushing reuse in hypermedia design: golden rules, design patterns and constructive templates. In Proceedings of (Pittsburgh, Pennsylvania, United States, 1998). ACM, 11-20.

[12] Quintana, C., Carra, A., Krajcik, J. and Elliot, S. Learner-Centred Design: Reflections and New Directions,. In Human-Computer Interaction in the New Millennium. (New York, 2001). ACM Press, Addison-Wesley.

[13] Squires, D. and Preece, J. 1996. Usability and learning: Evaluating the potential of educational software. Computers & Education, 27, 1 (1996), 15-22.

User Experience in Educational Virtual Worlds

Panagiotis Zaharias
Computer Science Department, University of Cyprus
CY-1678 Nicosia, Cyprus
zaharias@cs.ucy.ac.cy

Marios Belk
Computer Science Department, University of Cyprus
CY-1678 Nicosia, Cyprus
belk@cs.ucy.ac.cy

Panagiotis Germanakos
Institute of Informatics and Telematics - CNR
56124, Pisa, Italy
Computer Science Department, University of Cyprus
CY-1678 Nicosia, Cyprus
pgerman@cs.ucy.ac.cy

George Samaras
Computer Science Department, University of Cyprus
CY-1678 Nicosia, Cyprus
cssamara@cs.ucy.ac.cy

CHI 2011, May 7ï12, 2011, Vancouver, BC, Canada.
ACM 978-1-4503-0268-5/11/05.

Abstract
This paper explores user experience (UX) in educational Virtual Worlds. A two-month user experience evaluation was conducted where 37 learners used an educational Virtual World for a problem-based learning task. Two user experience measurements were performed at two distinct phases so as to assess usersô perceptions of the Virtual World environment over time. Results reveal that all UX qualities were rated lower at the second measurement mainly due to the many technical problems that users faced throughout the study. However, only the decrease in pragmatic quality was found to be statistical significant. To this end, results indicate that the environment needs usability improvement in terms of efficiency and effectiveness and technical stability so as to provide a better user experience.

Keywords
User Experience, Virtual Worlds, Problem-based Learning, Education, AttrakDiff2.

ACM Classification Keywords
H5.2. User Interfaces: Evaluation/Methodology.

General Terms
Design, Human Factors

Introduction

The evolution of Virtual Worlds (VWs) (e.g., Second Life) has extended research and development on learning technologies and distance learning. Virtual Worlds have a great impact on technology enhanced learning [9], providing prospective learners unprecedented ways of collaboration and feelings of immersion and social presence. Despite the attention they have received as new learning environments there is a lack of validated design principles and guidelines for developing learning spaces within VWs. Accordingly, user experience evaluation studies are scarce. Very few studies have been reported in the literature [16, 14] while most of the relevant research is still anecdotal.

In this empirical study we conducted a user experience (UX) evaluation where learners used an educational VW for two months for a problem-based learning task. Two UX measurements were performed at two distinct phases so as to assess usersô perceptions of the VW environment over time.

The remainder of the paper is organized as follows: Section 2 discusses the related work. In Section 3 and Section 4 we present the methodology and results of this empirical investigation, respectively. Finally, Section 5 concludes the paper with a discussion on the experimental results.

Background and Related Work

In this section we present related research work on User Experience and Virtual Worlds for education, both of which lie at the foundation of our study.

User Experience

User Experience (UX) studies the feelings and thoughts of an individual about a product (e.g., interactive system). UX is dynamic, because it changes over time as the circumstances change [4]. Being a multi-dimensional and complicated area a universal definition has not been agreed to date. Nevertheless, most of the definitions given to UX [20, 2] agree that UX focuses on the hedonic and affective aspects of human-computer interaction (HCI), but it also includes a personôs perceptions of the practical aspects such as utility, ease of use and efficiency of a system.

Effective HCI design and evaluation involves two important qualities: i) usability (i.e., traditional HCI), and ii) hedonic, beauty and affective [2]. Based on Jordan [3], the latter complements traditional HCI qualities (i.e., pragmatic) by suggesting a fixed hierarchical structure of qualities that contribute to positive experience. That is, a product has to provide functional and usability qualities before hedonic aspects can take effect. In contrast to Jordan, Karapanos et al. [4] assume the importance of these different qualities to vary with several contextual factors, i.e., individual differences, type of product, situation the product is used, and change of experience over time.

Recent research on HCI extends traditional task-based analysis and evaluation (e.g., usability evaluation), but rather focuses on hedonic and affective (e.g., surprise, diversion, intimacy) aspects of HCI design and evaluation. Mandic and Kerne [5] demonstrated an addition to email (called faMiliar), which visualizes ñrhythms in social engagementsò (p. 1617) and builds on intimacy as a core construct. In laboratory studies of usersô preferences of Media player skins, Tractinsky and

Lavie [6] and Tractinsky and Zmiri [7] showed the choice of personalised user interfaces (i.e., media player skins) to be driven by usability, aesthetic and symbolic considerations. Desmet et al. [1] demonstrated how affect could become a design goal. Zhang and Li [8] found the perceived affective quality of a course management system to be an antecedent of its perceived usability, usefulness and the intention to use. Germanakos et al. [19] have shown that usersô cognitive and emotional characteristics have significant impact in the adaptation and personalization process of web environments by increasing usability and satisfaction during navigation and learning process.

Regarding UX evaluation, one of the most influential models is the one proposed by Hassenzahl [22]; according to this model each interactive product has a pragmatic (related to usability) and hedonic quality that contribute to the UX. Based on this model a very well known and widely used instrument has been developed, the AttrakDiff, which has been employed in our empirical study (version AttrakDiff2). It is composed of four main constructs with seven anchor scales (total 28 items). The constructs are [21, 22]: Pragmatic Quality (PQ), which is related to traditional usability issues (such as effectiveness, efficiency, learnability, etc.); Hedonic Quality Stimulation (HQS), which is about personal growth of the user and the need to improve personal skills and knowledge; Hedonic Quality Identification (HQI), which focuses on the human need to be perceived by others in a particular way; and Attraction (ATT) which is about the global appeal of an interactive system or product.

3D learning spaces and educational Virtual Worlds

Virtual Worlds (VWs) or MUVEs (Multi User Virtual Environments) are computer-generated (not exclusively in 3D or animated graphics) environments, in which multiple users navigate, interact and communicate having a form of embodied representation [17].

This study considers VWs as learning and educational environments. VWs quickly captured the attention of the educational community as a highly engaging medium and a prospective learning environment that supports synchronous and asynchronous collaboration and user (learner) immersion in realistic or imaginary environments. There are numerous distinguished characteristics in VWs that can transform and enhance the quality of learning and educational activities. As Dalgarno and Lee [9] argue, VWs provide five important learning affordances as they facilitate learning tasks that: *"…lead to the development of enhanced spatial knowledge representation of the explored domain" (p.18); "…would be impractical or impossible to undertake in the real world" (p.19); "…lead to increased intrinsic motivation and engagement" (p.20); "…lead to improved transfer of knowledge and skills to real situations through contextualisation of learning" (p.21); "…lead to richer and/or more effective collaborative learning than is possible with 2-D alternatives" (p.23).*

A strand of research has investigated issues related to social and motivational aspects of educational VWs [23, 24]. Another strand of research has focused on the use of VWs as collaborative and problem-based learning environments [13, 18]. Other researchers have focused on interaction with technological features of VWs, e.g., multisensory representation can help learnersô interest,

fun, ability to navigate [12, 11], interaction with data gloves [10], etc.

Methodology

A total of 40 students (57% male, 43% female, age varying from 20-24) from the University of Cyprus participated in a two-month problem-based learning study. To undertake the learning tasks and activities, a learning space was built in Second Life where the users had to use in-world collaboration tools and techniques (i.e., live text and voice chatting, forums for comments, etc.). Most of the users did not have any previous experience with such a VW. Before starting the empirical study, a series of introductory tutorials on Second Life were conducted (i.e., how to configure an avatar, how to create objects, etc.). During the first week groups of 3-4 students were assigned to design and develop interfaces for several interactive systems (e.g., Realtorôs Agency, Online Game Shop, Universityôs Management System, Smart Home Management System, etc.) utilizing Second Life.

The interface design of each system was based on the Logical User-Centered Interactive Design (LUCID) methodology. Some indicative tasks the students had to undertake throughout the design/development cycle were: i) literature review on similar systems, ii) determine the typical users of the system, iii) analyse the interfaceôs design, following the Hierarchical Task Analysis (HTA) methodology, iv) design and develop the systemôs interface prototypes, v) design usersô navigation model.

Users were asked to evaluate their experiences using and collaborating through the VW at two distinct times by employing an online version of AttakDiff2; the first was at the end of the second week and the second at the end of the eighth and last week (henceforth time 0 and time 1, respectively).

Results

A paired sample T-test was performed in order to compare the user experience evaluations at the two different times. The analysis showed that difference in mean scores for the time 0 and time 1 was significant in PQ (mean difference=.436, t=2.651, p=.012).

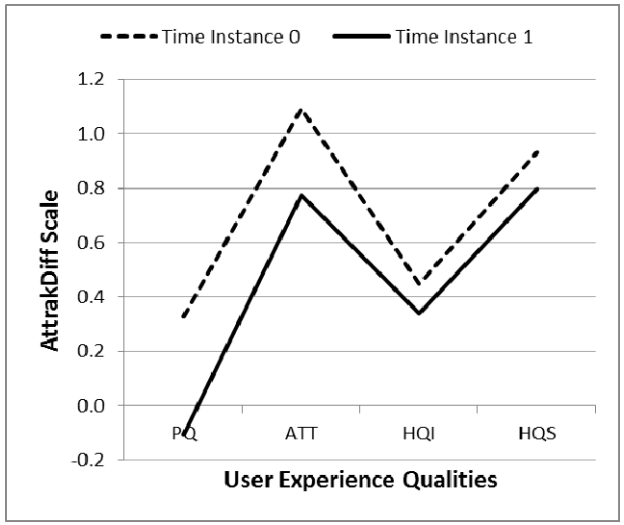


figure 1. Overall Results

Differences found in other scales were not significant: ATT (mean difference=.320, t=1.966, p=.057), HQI (mean difference=.108, t=1.004, p=.322) and HQS (mean difference=.135, t=1.456, p=.154).

Discussion

According to the results, all the evaluations at time 0 were higher than the evaluations at time 1 (figure 1). Nevertheless, only the difference in Pragmatic Quality (PQ) was found to be statistical significant. Several technical problems emerged during this 2 month period which caused service unavailability and sometimes led to loss of usersô objects and artifacts. That was the main reason for usersô frustration and irritation. In addition some users kept complaining about usability issues of the environment. To this end, PQ which refers to usability and functional aspects of interaction was rated quite low at both times of evaluation. Such finding calls for improvement of usability design mainly in terms of efficiency and effectiveness (focused on objectsô manipulation). It is also remarkable that according to the analysis all the other UX qualities such as Attraction (ATT), Stimulation (HQS) and Identification (HQI) were rated lower at time 1 (end of the two-month period). It is very likely that usersô frustration because of the several technical problems they faced was getting bigger as the time was passing and this led to a lower perception of HQS, HQI and ATT. Especially for the attraction, a preliminary interpretation can provide an explanation because of the low rating of PQ. This is in accordance with prior research [21] that has found that pragmatic quality influences attraction. In any case results reveal that there is a lot of room for usability improvements of such an educational environment so as to provide better UX.

Two main limitations have been identified regarding this study: a) a larger sample of users is needed in order to be more confident with the results of this investigation, and b) several technical problems emerged during the study; if there were a better operation of the server it is very likely that results and overall evaluation would be much different.

In conclusion, this empirical work performed two UX measurements so as to assess usersô perceptions of an educational Virtual World. The relevant research is in its infancy and further empirical studies are needed to investigate UX issues in such environments. This work focused on measurements at two distinct times. A future research prospect is to employ methods and tools [4] that can assess UX perceptions over time in order to shed light on this complex, dynamic and unexplored phenomenon.

References

[1] Desmet, P. M. A., Overbeeke, C. J. and Tax, S. J. E. T. Designing products with added emotional value: development and application of an approach for research through design. *Design Journal, 4* (2001), 32-47.

[2] Hassenzahl, M. and Tractinsky, N. User experience - a research agenda. *Behaviour & Information Technology 25: 2* (2006), 91-97.

[3] Jordan, P. W. Designing Pleasurable Products: An Introduction to New Human Factors. *Taylor & Francis* (2000).

[4] Karapanos, E., Hassenzahl, M. and Martens, J., User experience over time. *Ext. Abstracts CHI 2008*, ACM Press (2008), 3561-3566.

[5] Mandic, M. and Kerne, A. Using intimacy, chronology and zooming to visualize rhythms in email experience. *Ext. Abstracts CHI 2005*, ACM Press (2005), 1617-20.

[6] Tractinsky, N. and Lavie, T. Aesthetic and usability considerations in userôs choice of personal media

players. In *Proc. British HCI 2002*, ACM Press (2002), 70-73.

[7] Tractinsky, N. and Smiri, D. Exploring attributes of skins as potential antecedents of emotion in HCI. *Aesthetic Computing*, MIT Press (2006), 805-421.

[8] Zhang, P. and Li, N. Love at first sight or sustained effect? The role of perceived affective quality on user's cognitive reactions to information technology. In *Proc. ICIS 2004*, (2004), 283-95.

[9] Dalgarno, B. and Lee, M. J. W. What are the learning affordances of 3-D virtual environments? *British Journal of Educational Technology 41*, 1 (2010), 10-32.

[10] Mikropoulos, T. Presence: a unique characteristic in educational virtual environments. *Virtual Reality 10*, 3 (2006), 197-206

[11] Minogue, J., Gail Jones, M., Broadwell, B. and Oppewall, T. The impact of haptic augmentation on middle school students' conceptions of the animal cell. *Virtual Reality 10*, 3 (2006), 293-305.

[12] Zhigeng, P., Adrian, C., Wolfgang, M., Abdennour, E.R., Makoto, S., Xiangning, L., Jun, M., Katsuhito, A. and Masaharu, I. A haptic virtual environment for molecular chemistry education. *Transactions on Edutainment I 5080*, Springer (2008), 28-39.

[13] Nelson, B. C. and Ketelhut, D. J. Exploring embedded guidance and self-efficacy in educational multi-user virtual environments. *Computer-Supported Collaborative Learning 3,* 4 (2008), 413-427.

[14] Hasler, B. S., Buecheler, T. and Pfeifer, R. Collaborative work in 3D virtual environments: A research agenda and operational framework. *Online Communities 5621*, Springer (2009), 23-32.

[15] Karapanos E., Zimmerman J., Forlizzi J. and Martens J.B. User Experience Over Time: An Initial Framework, In *Proc. CHI 2009*, ACM Press (2009), 729-738.

[16] Minocha, S. and Reeves, A. J. Interaction Design and Usability of Learning Spaces in 3D Multi-user Virtual Worlds*. Information and Communication Technology*, Springer (2010), 157-167.

[17] Omale N., Hung W., Luetkehans L. and Cooke-Plagwitz J. Learning in 3-D multiuser virtual environments: Exploring the use of unique 3-D attributes for online problem-based learning. *British Journal of Educational Technology 40*, 3 (2009), 480-495.

[18] Vosinakis, S., Koutsabasis, P. and Zaharias, P. An Exploratory Study of Problem-Based Learning in Virtual Worlds. To appear in *Proc. of 3rd Conference in Games and Virtual Worlds for Serious Applications*, (2011).

[19] Germanakos, P., Tsianos, N., Lekkas, Z., Mourlas, C., Belk, M. and Samaras, G. Towards an Adaptive and Personalized Web Interaction using Human Factors, *Advances in Semantic Media Adaptation and Personalization*, Taylor & Francis (2009), 247-282.

[20] Forlizzi, J. and Battarbee, K. Understanding experience in interactive systems. In *Proc. DIS 2004*, ACM Press (2004), 261.

[21] Schrepp, M., Held, T. and Laugwitz, B. The influence of hedonic quality on the attractiveness of user interfaces of business management software. *Interacting with Computers 18*, 5 (2006), 1055-1069.

[22] Hassenzahl, M. The interplay of beauty, goodness, and usability in interactive products, *Human-Computer Interaction*, 19 (2004), 319-349.

[23] Barab, S. A., Thomas, M, Dodge, T., Carteaux, R. and Tuzun, H. Making learning fun: Quest Atlantis, a game without guns. *Educational Technology Research and Development 53*, 1 (2005), 86-108.

[24] Lim, C. P., Nonis, D. and Hedberg, J. Gaming in a 3D multi-user virtual environment (MUVE): Engaging students in Science lessons. *British Journal of Educational Technology 37*, 2 (2006), 211-231.

Towards a hierarchical model for the perceived quality of an augmented reality platform

Costin Pribeanu
ICI Bucharest
8-10 Bd. Maresal Averescu.
Bucureşti, 011455, Romania
pribeanu@ici.ro

Alexandru Balog
ICI Bucharest
8-10 Bd. Maresal Averescu.
Bucureşti, 011455, Romania
alexb@ici.ro

COST Workshop 2011, April 8, 2011, Limassol, Cyprus.
ACM 978-1-4503-0268-5/11/05.

Abstract
In this paper we explore a hierarchical model for the perceived quality of an augmented reality (AR) teaching platform that is based on three core components: perceived ease of use, perceived enjoyment and perceived usefulness. The model is capturing several key quality attributes for an AR based educational system.

Keywords
Usability evaluation, perceived quality, augmented reality, e-learning, hierarchical models

ACM Classification Keywords
H5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous. See [3] for help using the ACM Classification system.

General Terms
Guides, instructions, author's kit, conference publication

Introduction
Educational systems based on desktop augmented reality (AR) technologies are integrating real life objects

into computer environments. Touching and holding real objects is an appealing user experience that is increasing the students' motivation to learn and could better support active and collaborative learning [9, 11]. As AR technologies become more wide-spread, there is an increasing interest in their usability. Designing for usability is not an easy task in emerging AR technologies where the focus is mainly on designing and implementing new interaction paradigms [5, 7].

The quality of an e-learning platform should satisfy three kinds of requirements: usefulness for the learning process, ease of use, and increased motivation for the learner. Quality could be seen as general concept that is manifested by specific dimensions. As such it could be modeled as a higher-order factor. In general, the higher-order factors have proven to be successful in increasing theoretical parsimony and reducing model complexity [6, 12].

In this paper we explore a hierarchical model for the perceived quality of a learning scenario for Biology implemented onto an augmented reality teaching platform (ARTP). This educational platform was developed in the framework of the ARiSE (Augmented Reality for School Environments) research project. A "3D process visualization" paradigm was implemented that was targeted to enhance students' understanding and motivation to learn the human digestive system.

Firstly we will present a hierarchical model that is more suitable to integrate three core components of the perceived quality: perceived ease of use (PEOU), perceived enjoyment (PE) and perceived usefulness (PU). Then we will explore the relation between the perceived quality and the intention to use ARTP for learning.

Figure 1. Students testing the Biology scenario in 2008

Perceived quality of educational platforms

ISO standard 9126:2001 defines quality in use as the extent to which specified users accomplish specified goals with effectiveness, productivity, security and satisfaction in a given context of use [10]. Quality of educational platforms could be discussed from different perspectives (learner, teacher or school).

Perceived quality of e-learning systems is an important research topic in recent years. Understanding extrinsic and intrinsic factors that are influencing the acceptance of new technologies helps designers to focus on critical aspects. There are several influencing factors and many ways to model the quality among which the most frequently used are the perceived ease of use (related to usability), the perceived enjoyment (related to user experience) and perceived usefulness [1, 4, 13, 14].

In a previous work [3] we developed a measurement model for ARTP that was grounded in the technology acceptance models (TAM) theory in order to explain the causal relations between various factors influencing the intention to use. Although the model was useful to test various hypotheses, the variance explained was small, which suggests looking for an alternative model. This study extends the research by specifying the perceived quality as a hierarchical, reflective construct and examining its impact on the intention to use.

Case study

ARTP is a desktop AR environment: users are looking to a see-through screen where virtual images are superimposed over the perceived image of real objects

placed on a table [15]. The real object is a flat torso of the human digestive system that is used by two users staying face to face (see Figure 1). A pointing device having a colored ball on the end of a stick and a remote controller Wii Nintendo as handler has been used as interaction tool that serves for three types of interaction: pointing on a real object, selection of a virtual object and selection of a menu item.

The model specification and estimation is done on a sample collected in 2008. A total of 139 students 8th grade participated. Each of them tested the Biology scenario by performing four tasks: The participants have been assigned 4 tasks: a demo program explaining the absorption / decomposition process of food and three exercises: (1) asking to indicate the organs of the digestive system, (2) asking to indicate the nutrients absorbed / decomposed in each organ, and (3) asking to indicate the organs where a nutrient is absorbed / decomposed. After testing they filled in a questionnaire (five points Likert scale). Scale reliability was 0.94 which is acceptable (Cronbach's alpha).

We hypothesized that the perceived quality of the Biology scenario is associated with two extrinsic motivational factors (PEOU and PU) and an intrinsic motivational factor (PE).

We specified the construct "Quality of ARTP" (Q-ARTP-B) as a second-order, hierarchical reflective model which comprises three first-order reflective constructs: perceived ease of use (PEOU), perceived enjoyment (PE), and perceived usefulness (PU). Q-ARTP-B represents a general concept that is manifested by specific dimensions measuring the perceived quality of the application implementing the Biology scenario. In this respect, Q-ARTP-B is a superordinate construct [6].

We analyzed the data collected for the Biology scenario for outliers and normality, 9 cases were eliminated, thus resulting a final sample of 130 observations. Then we carried on an Exploratory Factor Analysis (EFA) using SPSS 18. Principal axis factoring with Promax rotation was used to extract factors. Eigenvalue and scree plot were used to determine number of factors extracted. The initial result revealed a three-factor solution, which accounted for 51% of variance. To achieve a more meaningful and interpretable solution, it was necessary to delete some items with low loadings or those loaded on more than one factor [8]. The final results with these three factors accounted for 69.38% of variance.

Table 1. Constructs and items

Perceived ease of use	
O1	Observing through the screen is clear
O2	Understanding how to operate with ARTP is easy
O3	Superposition between projection an the real object is clear
O4	Reading the information on the screen is easy
Perceived usefulness	
U1	Using ARTP helps to understand the lesson more quickly
U2	After using ARTP I will get better results at tests
U3	After using ARTP I will know more on this topic
Perceived enjoyment	
E1	The system makes learning more interesting
E2	Performing exercises is captivating
E3	Overall I like learning with this system
E4	Overall I find the system interesting
Intention to use	
I1	I would like to have this system in school
I2	I intend to use this system for learning
I3	I will recommend to other colleagues to use ARTP

From a learner's point of view the ease of use is likely to be influenced by the devices used to perform specific tasks. In an AR setting, the real objects used are also influencing both extrinsic and intrinsic motivational so is worth to identify particular ease of use aspects that are critical for a given AR application.

The results of EFA highlight three items related to the quality of visual perception as perceived by users: observing through the screen, superposition between projection and the real object, and reading the information on the screen. The descriptive statistics also showed that the two items got the lowest mean value (M=3.76, SD=1.07 for O1 and M=3.62, SD=1.02 for O2) while most of the items were rated over 4.00.

Then we applied Confirmatory Factor Analysis (CFA) using Structural Equation Modeling (SEM) with AMOS 17. A two-step approach [2] was adopted for examining the measurement model and following the structural model.

The results of the first-order factor model (the factors were free to correlate) did not provide a satisfactory result with a chi-square value of 68.60 (df=41), which was significant at the $p<.05$ level. Other fit indices revealed an acceptable fit (the chi-squared per degree of freedom was 1.67, SRMR=0.043, TLI=0.946, CFI=0.960, RMSEA=.072 with 90% CI 0.04-0.10 and pclose=.113). All of the items have statistically significant relationships with their factors. All loadings were above 0.67 and a good majority of them above 0.70. The first-order factors are highly correlated: 0.68 (PEOU, PU), 0.86 (PU, PE), 0.71 (PEOU, PE) thus showing evidence for convergent validity.

As expected, the fit indices of the second-order factor model were identically with those of the first-order model. The degree of explained variance (R^2 values) of hierarchical model is reflected in its components, that is, PEOU (.56), PU (.82), and PE (.90), indicating acceptable reliability for all factors. The Composite Reliability (CR) exceeded the recommended level of 0.70. Also, Average Variance Extracted (AVE) exceeded 0.50 (Table 2). All the path coefficients from Q-ARTP-B to its components are significant at $p<0.001$.

Note that the loadings of the first-order factors to the second-order factor are all above 0.75 and that the t-values were significant ($p<0.001$).

PEOU has a significant contribution to Q-ARTP-B ($\gamma=0.75$). Critical quality attributes for this scenario are related to the clarity of the visual perception and the ease of understanding how to operate with ARTP.

PU has a significant contribution to Q-ARTP-B ($\gamma=0.91$), showing that ARTP is useful to understand the lesson more quickly, to get better results at tests and to learn more on this topic.

The strong relationship between first-order factors and the second-order factor indicates a convergent validity of the posited second-order model. As such, the second-order model is very effective in representing the data. Both the magnitude of path coefficients and variance explained by the model showed higher values than those previously obtained [3].

Table 2. Convergent validity

Construct	CR	AVE
PEOU	0.800	0.501
PU	0.821	0.606
PE	0.874	0.636

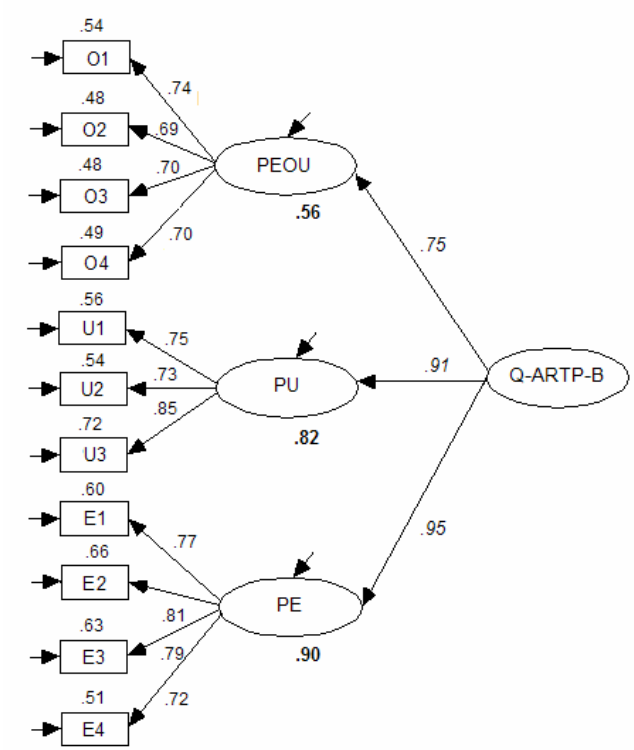


Figure 2. Hierarchical model of the perceived quality

We postulated a structural model where *Intention to use* comprising three items was specified as the dependent variable and the second-order factor (i.e. Q-ARTP-B) was hypothesized to affect it.

The model fit meets the recommended cut-off values, demonstrating adequate fit between the hypothesized second-order model and the observed data. The chi-square per degree of freedom was 1.670. Both TLI (0.942) and CFI(0.953) are above 0.90 and SRMR=0.048, RMSEA=.072 with 90% CI .049-.094 and pclose=.058. The *t*-value associated with the relationship between Q-ARTP-B and Intention to use (β=0.89*)* was 6.67. The model explains 79% of the variance in the intention to use.

Discussion and conclusion

There are some inherent limitations in our study. The sample used in this study is small and the data was collected with an evaluation instrument that was intended for another research model. As shown in a previous work [3] there are also several limitations of the evaluation instrument due to the novelty of the platform (neither specific questionnaires nor similar models available), the target user population (young and not happy to answer a long questionnaire), and the mixed character of the e-learning system (utilitarian and hedonic).

However, the hierarchical model proved to be a better alternative to predict the intention to use (79% vs. 20%) and is a first step in modeling the perceived quality of an AR-based learning scenario. This study confirmed the previous finding that PE had a slightly higher influence than PU on the intention to use.

The fact that this study is targeting a specific learning scenario implemented onto ARTP is both a weakness and strength. On the one hand, the generalizability is not possible, not even at ARTP level. On the other hand it enables a better understanding of strengths and weaknesses of each learning scenario in terms of ease of use, usefulness and enjoyment.

In the next future we intend to develop a similar model for the Chemistry scenario and comparatively analyze the results.

Acknowledgements

This work was supported from the European research project FP6 027039 and the national research project TEHSIN 503/2009.

References

[1] Alonso-Tapia, J., Pardo, A. (2006) Assessment of learning environment motivational quality from the point of view of secondary and high school learners. *Learning and Instruction 16*, 295-309.

[2] Anderson, J.C., Gerbing, D.W. (1988). Structural Equation Modeling in Practice: A Review and Recommended Two-Step Approach. *Psychological Bulletin*, vol. 103, no. 3, pp. 411-423.

[3] Balog, A., Pribeanu, C. (2010) The Role of Perceived Enjoyment in the Students' Acceptance of an Augmented Reality Teaching Platform: a Structural Equation Modelling Approach . *Studies in Informatics and Control, 19 (3)*, 319-330.

[4] Davis, F.D (1989) Perceived usefulness, perceived easy of use, and user acceptance of information technology. *MIS Quaterly 13*, 319-340.

[5] Dunser, A., Grasset, R., Billinghurst, M. (2008) A survey of evaluation techniques used in augmented reality studies. *Proceedings of SIGGRAPH Asia*, ACM, 1-27.

[6] Edwards, J. (2001) Multidimensional constructs in organizational behavior research: An integrative analytical framework. *Organizational Research Methods 4(2)*, 144-192.

[7] Gabbard, J., Swann, E. (2008) Usability engineering for augmented reality: Employing user-based studies to inform design. *IEEE Transactions on Visualization and Computer Graphics 14(3)*, 513-525.

[8] Hair, J.F., Black, W.C., Babin, B.J., Anderson, R.E., Tatham, R.L. (2006). Multivariate Data Analysis. 6th ed., Prentice Hall, 2006.

[9] Huang, H.M., Rauch, U., Liaw, S.S. (2010) Investigating learners' attitude towards virtual reality learning environments: based on a constructivist approach, *Computers & Education, 55*, 1171-1182

[10] ISO 9126-1:2001 *Software Engineering - Software product quality. Part 1: Quality Model*.

[11] Krauss, M., Riege, K., Winter, M., Pemberton, L. (2009) Remote Hands-On Experience: Distributed Collaboration with Augmented Reality. *Proceedings of. EC-TEL 2009, LNCS 5794*, Springer, 226-239.

[12] MacKenzie, S.B., Podsakoff, P.M., Jarvis, C.B. (2005). The Problem of Measurement Model Misspecification in Behavioral and Organizational Research and Some Recommended Solutions. *Journal of Applied Psychology*, 90(4), 710-730.

[13] Sanchez-Franko, M. (2009) WebCT – The quasimoderating effect of perceived affective quality of an extending Technology Acceptance model. *Computers & Education 54*, 37-46.

[14] Teo, T. (2010) Development and validation of the e-learning acceptance measure, *Internet and Higher Education 13*, 148-152

[15] Wind, J., Riege, K., Bogen M. (2007) Spinnstube®: A Seated Augmented Reality Display System, Virtual Environments: *Proceedings of IPT-EGVE – EG/ACM Symposium*, 17-23.

The Fast Lane: Rapid Document Triage Using an Eye-tracker.

Fernando Loizides
Centre for HCI Design
School of Informatics
City University London
Fernando.Loizides.1@city.ac.uk

George Buchanan
Centre for HCI Design
School of Informatics
City University London
George.Buchanan.1@city.ac.uk

Abstract
Document triage is the process by which an information seeker goes through a set of documents and decides their relevance to his or her information need. In this paper we examine this process with a particular focus on how a limited time frame (simulating fast decision making) affects behavior and relevance judgments. We present results using an eye-tracking methodology, showing document features viewed by information seekers during a time constrained document triage process.

Keywords
Document Triage, Eye-tracking, Information Seeking

ACM Classification Keywords
H.1.2 User/Machine Systems

General Terms
Human Factors, Experimentation.

Introduction and Related Work
Document triage takes place after the automatic information retrieval process, and before a relevance judgment occurs. Currently, there has been limited research of the manual human behavior into this area as well as a lack of specialized software tools. Conceptualization of information seeking is mostly limited to the process in its entirety (e.g Marchionini's model for information seeking [6]).

Our current work has shown areas of importance during document triage to the information seeker [2, 4]. This research corresponds to, and complements earlier work which looked at how document features influence relevance decisions [7]. As further research[1] is performed, focusing directly on the document triage process, it is becoming clearer that the overall pattern

[1] This refers to already published work but also includes current work by the authors which is not yet presented.

of behavior in document triage is more generalized rather than task oriented. This is a positive insight for software developers and designers. By deciphering the inner workings of document triage, software tools can be effectively created to assist this, so far, mostly unaided process. Currently, there have been limited attempts at creating specialized tools for supporting the document triage process [3, 5]. In order for this process to be better understood it is vital to identify and test the limitations that may constrain the users. One such constraint that has previously been looked at is the effects of display configurations in the document triage process [8]. Another constraint has been the reading and organizing of document sets [1].

In this paper we explore a natural constraint to information seekers when performing document triage. Document triage is by nature a fast part of the information seeking process. This fact, coupled with individual user time constraints, often leads to relevance decisions being made quickly. Evidence of low tolerance for manual searching can be seen in how information seekers rarely triage past the second page in an on-line search results list [9]. Users would rather modify their query term, thus relying on the automatic process to do the work. Furthermore, it was seen in [2, 4] that information seekers can be biased against long documents and documents with a bad structure, which hinders their locating information quickly enough. It is also commonly known that users do not spend much time (a few seconds at most) scrutinizing a website before making a decision of whether or not it is useful. This is ported to the academic document triage field in which users often make a relevance judgment early on, but may continue to triage further for reassurance [4].

The work presented here explores document triage when performed at a fast pace. We attempt to answer the following research questions under the constraint of time:

1. What parts of a document do information seekers feel are vital, important and unnecessary to look at during a fast search?

2. At what point in the document triage process, when given a whole document, are information seekers consistent with their relevance ratings?

3. How much would information seekers deviate from their regular document triage behavioral pattern given a limited time-frame?

Study Description

We chose 20 participants, 7 female and 13 male, for the study. The first 2 participants were chosen as pilot study subjects while the remaining 18 participated in the main experiment. All participants were at postgraduate level or above in an HCI (Human Computer Interaction) related discipline. Their ages ranged from 21 to 50 years of age. The criteria for the selection of appropriate participants was their familiarity with academic documents; namely to have read academic documents before and are therefore accustomed to the general format. Further to this, the participants needed to have a minimum understanding and experience in using electronic document readers and a basic foundation in HCI. Special care was taken however, in order that the participants were not familiar with the specific documents presented to them in the study.

We briefed the participants on the definition of the term *document triage*. We did not give out information on how users are expected to act during document triage that may bias the behavior of our participants. A pre-study questionnaire was given to the participants containing general questions such as age, experience in using document readers, documents triaged on average per day and time spent on the process per day. The second part of the questionnaire asked the participant to give a subjective rating of the importance they would give to different document elements with respect to the document triage process. The elements in question were: Main Title, Abstract, Introduction, Plain Text, Conclusion, References, Figures and Emphasized Text (This includes bold, italic, underlined text and bullet points). Participants were also asked which medium between electronic and physical they preferred and to what extent, when performing document triage.

In the main study, participants were given two information needs. For the first information need, they were given six documents to triage and place a relevance score from zero to ten (zero being extremely irrelevant and ten being extremely relevant to the information need). The second information need included ten documents for the participants to triage in the same fashion. Participants were not given an open-ended time frame but were assigned a specific time limit per document. A within-subjects design allowed for participants to perform triage in one of three time limits (30, 60 and 90 seconds). These results could then also be cross examined with the data presented in [4] where the same documents were used but participants were permitted an open ended time frame. There is one limitation to this study however. In the original study presented in [4] no eye-tracker was used. Thus, we can only compare the participants' relevance scores directly. The documents chosen to be scrutinized by the participants were all rated by 3 experts as highly relevant and selected specifically for their distinct features (such as length, structural clarity and Main title relevance).

The documents were presented one at a time on a 19 inch screen in Portable Document Format (PDF). The participants' eye gaze was tracked using the non-intrusive Tobii x50 eye-tracker. The Tobii x50 eye-tracker allows 0.5-0.7 degree accuracy, a spatial resolution of 0.35 degrees, a drift of < 1 degree and a frame rate of 50 Hz. The participants' freedom of head movement was 30 x 15 x 20 cm at 60 cm from tracker. The participants' screen was recorded using BB Flashback Screen Recorder Standard Edition. A dictionary was supplied for the clarification of unknown words and questions could be asked at any point in the study provided they were related to clarifying participant goals. Guidance regarding the relevance of the documents themselves however was not given.

Following the main study a semi-structured interview was performed in order to gain qualitative feedback from the participants. Parts of the study were played back to the participants and they were asked to explain their thoughts during interesting or confusing behaviors to the investigator. This approach was deemed more appropriate than interrupting participants during their triage process. Participants were rewarded with £10 (GBP) for their time.

Results

In this section we begin by presenting the pre-study questionnaire feedback which shows natural subjective bias of the users before initiating document triage towards certain document features. We then continue to present the actual behaviour and participants' navigational patterns and visual focus, explaining which document features were viewed most and those which were scarcely or not viewed. We also look at the participant relevance ratings for the three groups and compare the results with our open-time frame study [4]. Finally, we summarize and discuss the most distinct findings.

***Navigational Patterns*.** When users perform document triage, they will navigate through a document following a specific pattern. These patterns were recorded and reported on in earlier work [4]. When users triage while under pressure from a time constraint, they are found to replicate some of these patterns also.

The most common navigational behavior was that of 'step up navigation'. This linear skimming of the document and stopping at areas of interest was performed by all participants in all three groups 92% of the time. The remaining 8% of the time, participants would produce a hybrid behavior between 'Flatline navigation' and 'Begin and End navigation'. 'Flatline navigation' is a term used to describe no scrolling beyond the initial page of a document. 'Begin and End navigation' indicates that a user looks at the initial page and scrolls to the conclusions section only.

	Viewing area	Group1 (30)	Group2 (60)	Group3 (90)
Initial Page	10.8	42.7	27	46.7
Main Title	1.1	10.8	5.5	9.5
Abstract	1.2	16.5	15.1	21.1
Introduction	2.3	8.1	5.9	7.1
Headings	2.1	15.3	8.1	8.2
Emphasised Text	1.9	2.2	2.8	4.6
Images	6.2	4.9	6.8	10.5
Diagrams and Tables	2.3	12.5	4.6	6.3
Conclusions Section	1.6	6.1	5	7.5

Figure 1: Average document area per feature Vs. Average percentage of visual attention between Groups.

User Attention. The areas of participants' visual focus between the three groups were analyzed and reported on (See Figure 1). Participants were allowed to examine every document for a limited amount of time; up to 5 seconds extra was allowed after the limit was reached for the decision to be orally given to the investigator. Participants with a 30 second limit, 60 second limit and 90 second limit will be referred to as Group 1, Group 2 and Group 3 respectively. According to our study document features accounted for 18.8% of the total document area. Participants focused on these features for 68.5% of the time. This fact, including scrolling time, suggests the remaining plain text area was skimmed, but not scrutinized, therefore unlikely to be of benefit to the participants.

Document Relevance Ratings. Figure 2 shows the average rating and standard deviation for the three groups in our study, as well as the average rating from a fourth group taken from [4]. This fourth group underwent the same document triage process on the same documents but was permitted an open-ended time frame. As we can see from the results, there is no

significance between time taken to triage documents and relevance scores.

	Group 1	Group 2	Group 3	Group 4
Average score	5.1	5.82	6.02	5.99
St.Dev	2.41	2.42	2.15	2.2

Figure 2: Average ratings and standard deviation for Group 1 (30 secs), Group 2 (60 secs), Group 3 (90 secs) and Group 4 (open time frame).

Discussion

It was the intention of this paper to report on two primary findings. First, we researched the behavior (in terms of navigational and viewing areas) of users, while performing document triage with a limited time constraint. Second, we compared and contrasted the relevance judgments made between our participants' groups and studies from our previous work.

Behavior. We saw behaviors from all three participants groups which coincide with findings from previous work. One example is the visual importance of the initial page [2, 4]. The navigational patterns used by our participants, replicated those performed by users with no time limit. We did however, discover an overwhelming bias in the number of participants choosing to use the 'step up' navigation rather than any other. A contributing factor to this behavior is the attraction of headings to users' focus during skimming. Headings receive increased attention when an information seeker performs within-document triage with a short time constraint. This is especially noticed from our 30 second group. Headings, Diagrams and Figures, as well as the conclusions section receive the most visual attention for the information seeker, relative to their viewing area on a document.

Relevance Judgments. Interestingly, we saw no significant change in the average relevance ratings given by participants between any of the three groups presented here. The same statement holds when comparing the results with users from a previous group. This group of participants performed triage on the same set of documents with no time restrictions [4].

Overall, we have identified that time is not a largely contributing factor in the relevance decision process. Furthermore, we have identified (and cross referenced our results with previous studies) document parts which are used to make the relevance decisions. These findings can also be used as heuristic guidelines into informing supporting software designs. We can inform document triage software by targeting visually important document features and bypass features which do not contribute substantially towards a user's decision process.

Future Work

As more data is gathered regarding document triage, we see patterns evolving in information seekers' behaviors. These patterns, being generic, can be standardized into models for librarians, software developers and human computer interaction scientists to use in understanding, conceptualizing and facilitating the document triage process further. A prototype software tool is currently under development which aims to provide rapid triage support to information seekers in searching through a document set effectively.

Conclusions

In this paper we have explored the effects of limited time on document triage in a process we label *rapid triage*. We assigned an information need to participants and gave them a set amount of documents to triage in three different time frames. By comparing their subjective feedback with their actual behavior, using an eye-tracker, we were able to distinguish areas of interest and level of importance for several features on a document empirically. We were able to determine that time does *not* produce a significant factor in the document relevance ratings of information seekers. We have also produced further questions for research and progressed in the formulation of a standardized general model for the document triage process.

Bibliography

[1] R. Badi, S. Bae, J. M. Moore, K. Meintanis, A. Zacchi, H. Hsieh, F. Shipman, and C. C. Marshall. Recognizing user interest and document value from reading and organizing activities in document triage. In IUI '06: Proceedings of the 11th international conference on Intelligent user interfaces, pages 218{225, New York, NY, USA, 2006. ACM.

[2] G. Buchanan and F. Loizides. Investigating document triage on paper and electronic media. In Proceedings of the European Conference on Reasearch and Advanced Technology for Digital Libraries, number 35, pages 416{427, 2007.

[3] J. David, W. William, and C. B. Schroh David, Proulx Pascale. Information triage with trist. In International Conference on Intelligence Analysis, 2005.

[4] F. Loizides and G. Buchanan. An empirical study of user navigation during document triage. In Proceedings of the European Conference on Reasearch and Advanced Technology for Digital Libraries, pages 138{149, 2009.

[5] F. Loizides and G. R. Buchanan. The myth of _nd: user behaviour and attitudes towards the basic search feature. In JCDL '08: Proceedings of the 8th ACM/IEEE-CS joint conference on Digital libraries, pages 48{51, New York, NY, USA, 2008. ACM.

[6] G. Marchionini. Information Seeking in Electronic Environments. Cambridge University Press, 1995.

[7] T. Saracevic. Comparative e_ect of titles, abstracts and full texts on relevance judgments. In Proceedings of the American Society for Information Science, number 6, pages 293{299, 1969.

[8] B. Soonil, B. Rajiv, M. Konstantinos, M. J. Michael, Z. Anna, H. Haowei, M. C. C., S. F. M., C. M. Francesca, and P. Fabio. E_ects of display con_gurations on document triage. INTERACT 2005 : human-computer interaction, (3585):130{143, 2005.

[9] A. Spink, B. J. Jansen, D. Wolfram, and T. Saracevic. From e-sex to e-commerce: Web search changes. Computer, 35(3):107{109, 2002.

Evaluating the Usability of Data Entry Web Pages for E-Government Sites

Jan Stage
Aalborg University
Department of Computer Science
Selma Lagerlºfs Vej 300
DK-9220 Aalborg East, Denmark
jans@cs.aau.dk

CHI 2011, May 7ï12, 2011, Vancouver, BC, Canada.
ACM 978-1-4503-0268-5/11/05.

Abstract
Usability is a key quality factor for interactive web pages. In this paper, we present experiences from a usability evaluation of web pages for data entry on e-government websites. In the evaluation, we started focusing on usability in the traditional meaning where it is defined in relation to the direct user. Yet the usability for the user of the data that were entered turned out to be much more important. In our case, there was a difference of focus in usability for this secondary user compared to the primary or direct user. We discuss the implications of these experiences for usability evaluation in the e-government domain.

Keywords
Usability evaluation, interaction style, form fill-in, experimental comparison, E-Government

ACM Classification Keywords
H.5.2 [**Information Interfaces and Presentation**]: User Interfaces: *interaction styles, evaluation/ methodology*.

General Terms
Evaluation

Introduction

The World Wide Web is often associated primarily with search and retrieval of information. Yet a growing number of interactive websites also involve data entry.

Electronic delivery of public services, known as e-government, is an emerging application area. There are many design challenges in developing e-government websites, including meeting government legal requirements and serving a varied user population [2]. For many e-government websites, data entry is a key element in the user interaction. The user group of typical E-Government services is a broad and multi-faceted population that uses the system rarely. This makes form fill-in particularly useful as a key interaction style [3].

This paper reports from a usability evaluation where we compared two basically different implementations of form fill-in for data entry on an E-Government website. The comparison was based on use of a specific electronic form for data entry on a website with a broad population of users, where the majority is novices. The website is typical of e-government systems. In the following section, we describe the method of the experimental comparison. Then we provide the results. In the next section, we discuss the results and the difference between primary and secondary users. Finally, we provide the conclusion.

Method

In this section, we describe the experimental comparison of two systems for data entry on an e-government website.

System

The web pages we evaluated are used to apply for a construction permit at a municipality. A Danish software company has developed 2 different systems (denoted ñAò and ñBò) to support this application process, based on a pdf-form and a wizard, respectively.

System A is a form-based system designed using a PDF form with text fields, checkboxes and radio buttons to collect information about the planned construction work. System B is developed in Flex and the user interface is designed using a stepwise wizard with 11 steps.

Setting

The usability evaluation was conducted in a state-of-the-art usability laboratory, with a test participant was sitting at a table using systems A and B. A test monitor was sitting next to the participant during the entire session. A data logger was in a control room noting task completion times.

Participants

Ten participants were selected for this study, all of which were potential end users of the systems. The participants were required to have some experience with ñdo it yourselfò (DIY) work in order to be selected for the experiment. Participantsô DIY experiences varied where 2 had only painted their homes and 8 had either restored parts or all of their homes. Participants varied in age with an average of 39.9 years (SD = 13.1) and four were trade- or craftsmen and 6 had an academic background, e.g. from architecture and structural engineering.

Procedure
The test monitor started by introducing the participants to the systems, the purpose of the evaluation and to the evaluation procedure, including the tasks they would be solving with the system. The tasks were given to the test participantôs one by one and always in the same order. As we were focusing on the effectiveness and efficiency, i.e. completion rates and time on task, the test monitor did not ask participants to think aloud during task solving. There were 2 tasks, see Table 1.

Task No.	Task
1	Apply for construction of a 24 m^2 garage.
2	Apply for a 54 m^2 addition to the house containing extra rooms.

Table 1: Summary of the tasks scenarios.

The experiment was designed to be within subjects such that all 10 participants completed tasks 1 and 2 using systems A and B. The tasks are henceforth referred to as A1, A2, B1 and B2. In order to reduce the impact of learning effects on our results, half of the participants solved the tasks using system A first followed by B and the other half used B first followed by A. Participants were given 30 minutes per system to solve the tasks, and participants were interrupted in their task solving if they were not done.

Data Collection
Data was collected through video recordings from the usability laboratory. The video material is a picture in picture setup showing the desktop feed from the pc and the face of the test participants. Notes were also taken by the data logger and test monitor during the test sessions and the post test interviews.

Data Analysis
In order to make a quantitative comparison of the effectiveness of systems A and B, we rated the forms created by the participants. A specialist from a Danish municipality made a rating scheme containing quality ratings ranging from 1 to 3. Ratings were defined from the perspective of employees at the municipalities and how much time they would need to spend on correcting potential errors. Thus ratings were defined as follows:

1. Less serious: The employee would spend less than 5 minutes correcting the error.

2. Serious: The employee would spend between 5 to 10 minutes correcting errors.

3. Very serious: The employee would wither spend more than 10 minutes correcting errors, must contact the applicant or return the application form in order to correct the errors.

Potential errors in input fields of systems A and B were rated by the specialist using the above scheme, e.g. a minor spelling error in an address would be rated as 1 whereas a missing signature would be rated as 3 and so forth.

Field by field each application form was then walked through where ratings were given according to the above scheme. Each time we encountered an error with e.g. a missing signature we would withdraw 3 ñpointsò corresponding to the severity rating. The maximum rating for each application form was 29 (no errors) and

the lowest was 0 (an empty application or one full of errors).

Results

This section presents first the quantitative and then the qualitative results of our study.

Task Completion Time

Table 2 provides an overview of task completion times based on tasks A1, A2, B1 and B2 where the black cell indicates that the participant ran out of time before starting the task and the grey cells indicate that participants started the tasks but did not have time to finish.

The total average completion time when using system B is 20:30 mm:ss (sd = 07:02) and for A this number is 18:14 (sd = 03:23). The highest total completion time for system A is 24:46 and 30:15 for B while the lowest totals are 13:22 for A and 13:30 for B. Thus results show that participants had a higher average completion time by using system B compared to A.

Participants 1,4,5,6 and 9 used system A first then B and from this we see that the total times for system A generally is higher than that of system B. The opposite result is observed for participants 2, 3, 7, 8, and 10 who used system B before A as completion times are higher for B in most cases.

Form Ratings

Table 3 shows the ratings of the application forms based on tasks A1, A2, B1 and B2 where the black and grey cells indicate incomplete forms as mentioned previously.

	System A (mm:ss)			System B (mm:ss)		
Partici-pant	**A1**	**A2**	**Total**	**B1**	**B2**	**Total**
1	9:32	10:58	20:30	7:16	9:25	16:41
2	8:40	6:31	15:11	20:40	6:30	27:10
3	14:20	10:26	24:46	30:15	-----	30:15
4	7:15	6:40	13:55	7:10	6:20	13:30
5	11:01	7:52	18:53	7:59	5:40	13:39
6	10:39	4:04	14:43	6:30	5:37	12:07
7	11:27	11:37	23:04	14:44	6:31	21:15
8	7:46	5:36	13:22	12:44	5:54	18:38
9	5:58	9:59	15:57	8:31	8:58	17:29
10	17:56	4:01	21:57	20:01	6:54	26:55
Average time	10:27	07:46	**18:14**	14:19	06:52	**20:30**
Highest time	17:56	11:37	**24:46**	20:40	09:25	**30:15**
Lowest time	05:58	04:04	**13:22**	07:10	05:37	**13:30**

Table 2: Task completion times. Black cells indicate that the participant newer started on the task. Grey cells indicate that participants started the task but never completed it within time.

From the table it can be seen that application forms filled out using system B generally results in higher ratings than those created with system A. One exception, however, is participant 6, whose application forms scored 24 in A2 and 22 in B2.

By pooling A1, A2 and B1, B2 a t-test reveals a highly significant difference between application ratings of systems A (nA = 19) and B (nB = 17) where p = 0.0005.

	System A		System B	
Participant	**A1**	**A2**	**B1**	**B2**
1	16	20	24	24
2	22	21	27	13
3	7	12	14	
4	17	21	25	25
5	22	22	23	25
6	18	24	26	22
7	22	22	25	23
8	19	16	23	24
9	14	23	26	23
10	17	8	20	18
Avg. rating	**17.4**	**18.9**	**23.6**	**21.9**
Highest rating	**22**	**24**	**27**	**25**
Lowest rating	**7**	**8**	**14**	**13**

Table 3: Form ratings. Black cells indicate that the participant newer started on the task. Grey cells indicate that participants started the task but never completed it within time.

Higher ratings when using system B is also the result when looking at participants 2, 3 and 10, who are the 3 oldest with ages 61, 63 and 46 respectively.

Distribution of Rating Types

Table 4 shows a distribution of the rating types 1, 2 and 3 according to systems A and B and participants. Rating types are labeled as R1, R2 and R3 in the table.

Forms created through use of system A were given a total of 26 ratings being less serious, 34 serious and 34 very serious. In the case of system B there were given 19 less serious, 33 serious and 2 very serious ratings. This result shows that the number of very serious ratings was reduced considerably when participants used system B while the less serious and serious remains comparable.

	System A			System B		
Participant	**R1**	**R2**	**R3**	**R1**	**R2**	**R3**
1	8	4	2	6	2	0
2	1	4	2	2	0	0
3	7	4	8	3	3	1
4	2	3	4	0	4	0
5	0	4	2	0	5	0
6	0	5	2	2	4	0
7	0	4	2	0	5	0
8	3	1	6	3	4	0
9	3	3	4	1	4	0
10	2	2	2	2	2	1
Total	**26**	**34**	**34**	**19**	**33**	**2**

Table 4: Rating types. Number of ratings given per application form per participant. R1 = Less Serious, R2 = Serious and R3 = Very Serious.

It is also noteworthy that the 2 very serious errors in the forms from system B are produced by participants 3 and 10, which are among the oldest participants (63 and 46 years of age).

Discussion

Our results show that there is no significant difference in task completion times when comparing system A and B. This came as a surprise when we started analyzing the results. We had expected that system B would be more efficient to use for the participants. Through a review of the task solving processes of the participants, we realized that the reason for this was the browsing process that some participants went through while solving the tasks. Thus system B made them browse through more information related to the application.

In many e-government websites, the primary users are citizens who request a service that is delivered by the authority that has the website. These citizens are typically novice users that use the service quite rarely. In our case, most house owners only make a couple of building applications in their life. With this pattern of use, the traditional concept of efficiency is not very relevant.

To better understand the concepts of usability and efficiency in this domain, we have to introduce the users of the information that is entered by the primary users. This group can be denoted as secondary users of the data entry system as they work with the resulting information [1]. Efficiency is much more important for these secondary users who are the experts working for the authority. Our findings show that form ratings between the two systems differed significantly and that B forms scored higher ratings. The better ratings is primarily caused by a considerable reduction of very serious errors, as the less serious and serious remained comparable to forms created in system A, see **Table 4**. In practice this means that employees at the municipality would spend an average of at least 51 minutes correcting errors in forms created by system A and at least 18.5 minutes in average when system B is applied. Thus, the difference between A and B are not only significant in a statistical sense but also in a practical, as B may save a lot of time for employees at the municipalities.

Conclusion

We have presented an experimental comparison of the efficiency and effectiveness of two systems for data entry on an e-government website. The results show that system B was not more efficient than system A for the primary users, but for the secondary users, there was a remarkable difference.

This experiment illustrates that for a domain like e-government, we need to include secondary users in our understanding of usability.

References

[1] Eason, K. Information Technology and Organisational Change, Taylor and Francis (1987).

[2] Hourcade, J. P. and Fox, J. E. Designing public government web sites. In CHI '05 Extended Abstracts on Human Factors in Computing Systems, ACM Press (2005), 2039-2040.

[3] Shneiderman, B. Designing the User Interface, 3rd ed. Addison-Wesley, 1998.

Adaptive web site design based on fuzzy user profiles, usability rules and design patterns

Dimitrios K. Kardaras
Athens University of Economics and Business.
76, Patission Str.
Athens, GR10434 GREECE
kardaras@aueb.gr

Xenia J. Mamakou
Athens University of Economics and Business.
76, Patission Str.
Athens, GR10434 GREECE
xenia@aueb.gr

Bill Karakostas
HCI Design Center
City University
Northampton Square
London, EC1V 0HB, UK
billk@soi.city.ac.uk

Abstract
This paper introduces an approach to automatic design of web sites, based on user profiles, web usability heuristics and web design knowledge. The methodology draws on ontologies that define the user, service and web design domains, and utilizes fuzzy relations composition, in order to capture the complex interdependencies among user preferences, web site usability criteria and web design choices, to automatically construct the layout and context of web pages.

Keywords
Ontologies, adaptive websites, fuzzy relations.

ACM Classification Keywords
H5.2. [Information interfaces and presentation (e.g., HCI)]: User Interfaces--- User interface management systems (UIMS), User-centered design; I.5.1. [Pattern recognition]: Models---Fuzzy sets.

General Terms
Design, Human factors, Management.

Introduction
The term *adaptive Web site* refers to sites that automatically adapt their design and interface,

according to a user's profile. The adaptation process of data-intensive web sites operates on three main components: content, navigation, and presentation [5]. It is important to design adaptive web sites, since a poorly designed web site may result in a loss of 50 per cent of potential sales, due to users being unable to find what they want, and a loss of 40 per cent of potential repeat visits, due to initial negative experience [4], [11].

Domain knowledge and ontologies

Only by capturing domain knowledge can automatic web site adaptation be realized. Domain knowledge that characterizes the context as a source of information [7], needs to be captured and formalized as ontologies. Such ontologies define the context factors relevant to the user, his or her environment and situation.

Adaptive websites

Today, many online businesses are adaptive to each user or user group preferences and profiles. Approaches to adaptation vary as some of them introduce new content, adapt the structure of links and change the presentation of page elements. Various adaptation techniques are used, depending on the component that is targeted, the degree of prior knowledge about the user (user profile) and the method used to elicit knowledge dynamically, i.e. as the user interacts with the web site [9].

Knowledge mining approaches such as Bayesian networks and decision [6] are used to identify patterns in web site usage and use such patterns to assist users navigate between web pages and/or rearrange pages based on their relevance to the user or according to the user's expectation of where the page should be.

Collaborating filtering relies not only on a single user, but on groups of users with similar profiles/interests to select and customize content. One such approach, based on declarative rules, relies on both profile and configuration knowledge [5].

Adaptation can be carried either on the server side or the client side. Server side adaptation implies that knowledge is stored on the web site, and adaptation takes place on the server, prior to delivering content to the user. Client-side adaptation approaches employ agents that are responsible for collecting user information and interests and customize the site accordingly [1]. Attentive agent-based information systems do not intrude in the user's ongoing activity, but work on the background and gather information about the user's behavior, model the user and provide content that is relevant but not critical. Attentive systems agents may for example monitor eye gaze to determine what the user is reading and look for relevant information. The results of their inquiry are displayed discreetly on the side of the page.

The effectiveness of adaptation approaches in general, can be improved by using formalized knowledge of the domain, context and even of the adaptation process, in terms of ontologies [3]. The most important advantage of using an ontological description of such knowledge in languages such as OWL-S is that it can be augmented, enriched and synthesized using suitable reasoning mechanisms, and also exploited in reasoning on context ontologies. To overcome the problems of partial and

imprecise context information, fuzzy knowledge representation approaches can be used.

Composition of Fuzzy Relations

Fuzzy relation is an extension of relations in conventional set theory. Classical relations are given by sets, while fuzzy relations are given by fuzzy sets [10]. Like classical relations, fuzzy relations also map elements of one universe, say X, to those of another universe, say Y, through the Cartesian product of the two universes. However, the fuzzy relation possesses stronger expressive power while relating x and y [2], since the relation between ordered pairs of the two universes is measured with a membership function expressing various "degrees" of strength of the relation on the unit interval [0,1]. Hence, a fuzzy relation $\tilde{R}$ is a mapping from the Cartesian space X x Y to the interval [0,1], where the strength of the mapping is expressed by the membership function of the relation for ordered pairs from the two universes, of $\mu_R(x,y)$ [10]. Composition of fuzzy relations can be defined as following [8]. Suppose two fuzzy relations $\tilde{R}$ and $\tilde{S}$ are defined on sets X, Y and Z. That is, $\tilde{R}$ is a fuzzy relation on the Cartesian space XxY and $\tilde{S}$ is a fuzzy relation on YxZ. The composition $\tilde{R} \circ \tilde{S} = \tilde{T}$ is expressed by the fuzzy relation of XxZ and this composition is defined as: $\mu_T(x,z)=\mathrm{Max}[\mathrm{Min}(\mu_R(x,y),\mu_S(y,z))]$. If the relations $\tilde{R}$ and $\tilde{S}$ are represented by matrices M_R and M_S, the matrix $M_{R\circ S}$ is obtained from the product of M_R and M_S which is: $M_{R\circ S}=M_R \circ M_S$

The proposed approach

The generation process of adaptive web sites assumes the availability of ontologies that capture the domain knowledge and the modeling of their interrelationships using fuzzy relations. The composition of fuzzy relation provides the inference mechanism for adapting web design alternatives to the user types. The proposed approach consists of the following steps:

1) Construct the User Ontology that specifies the user styles (T={type-1, type-2, … type-n}) in terms, shown as u_i's in figure 1, that represent the user cognition, the demographics such as age, sex, income, attitude towards innovation, ability to articulate needs and priorities, educational level etc. User types are associated with the perceived for each type importance of services and service characteristics.

2) Develop the Service Ontology that represents the characteristics (s_{ci}'s in figure 1) that constitute the service described in the web site, and captures knowledge that pertains to the contribution of each characteristic to service quality. For example, in the case of a hotel web service, these characteristics could be price, view, room amenities, beach etc.

3) Construct the Web design Ontology, which specifies the web design features (w_{Fi}'s in figure 1) and represent web usability heuristics. Design features can be, for example, text, images, videos etc., which are associated with size, location on the web page, font types, etc.

4) Define fuzzy relations $\tilde{U}$, $\tilde{R}$ and $\tilde{P}$ that associate the Ontologies:

$\tilde{U}$ is the fuzzy relation "defines style" associating each user characteristics (u_i) with each user types (T),

where: $\tilde{U} = u_i$ xT , $\forall\ u_i$, T

$\tilde{R}$ is the fuzzy relation "requires" which shows how strongly each service characteristic (s_{ci}) is required by each user types (T), where: $\tilde{R}$ =Tx s_{ci} , $\forall$ T, s_{ci}

$\tilde{P}$ is the fuzzy relation "is suitable", which indicates the most suitable presentation of each service characteristic (s_{ci}) by comparing and contrasting each alternative web design features (w_{Fi}), where:

$\tilde{P} = s_{ci} \times w_{Fi}$, $\forall\ s_{ci}$, w_{Fi}

5) Infer the web design adaptation alternative that is most suitable for a given domain instance (user type, service and web characteristics). The adaptive web design decisions are made by the following fuzzy compositions: $\tilde{AW} = \tilde{T}$ o ($\tilde{R}$ o $\tilde{P}$), where $\tilde{T}$ is the fuzzy relation "closer to type (i)" and indicates how close a considered user is to each of the user types of the set T of user types.

The composition takes into account the degree to which a user fits a user profile, and then adapts the design of the web page, according to the user requirements for service and the suitability of alternative web design features.

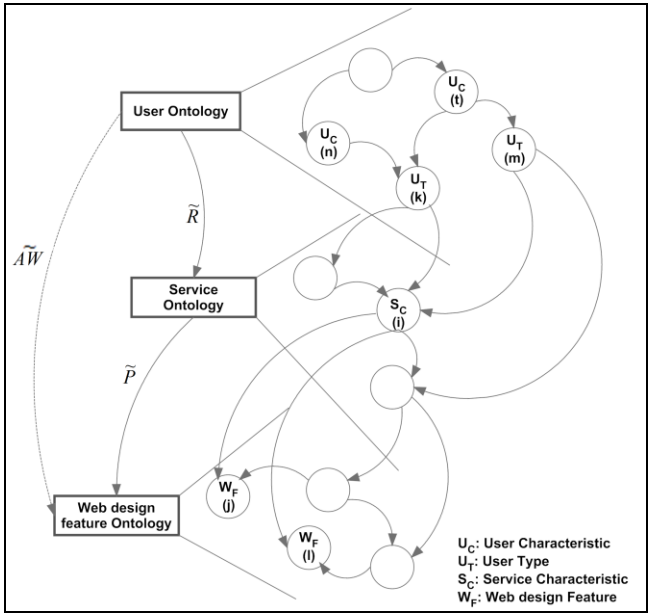


Figure 1. Integrating Ontologies with Fuzzy Relations for web adaptation

Illustrative example

The following example illustrates the proposed methodology. Let us consider three different user types (type1, type2 and type3). In the case of a hotel web site, showing hotel features such as price, location and amenities, and using web page elements such as text, image and video we have:

T={type1,type2,type3}, S_i={low,medium,high} and W_j={low,medium,high}, where i is the specific service feature and j is the specific web design element. As a result, the corresponding fuzzy relations R_i and $P_{i,j}$ for i=price and j=text will be:

$$\tilde{R}_{price} = \begin{array}{c} \\ type1 \\ type2 \\ type3 \end{array} \begin{array}{c} \begin{array}{ccc} low & med & high \end{array} \\ \begin{bmatrix} 0.2 & 0.4 & 0.1 \\ 0.7 & 0.8 & 0.4 \\ 0.5 & 0.3 & 0.8 \end{bmatrix} \end{array}$$

which shows how strongly the service feature 'price' is required by each of the three different customer types. For example, customer type 1 is price sensitive with a low degree (0.2), medium (0.4) and high (0.1).

$$\tilde{P}_{price,text} = \begin{array}{c} \\ low \\ med \\ high \end{array} \begin{array}{c} \begin{array}{ccc} low & med & high \end{array} \\ \begin{bmatrix} 0.3 & 0.2 & 0.1 \\ 0.5 & 0.4 & 0.4 \\ 0.5 & 0.3 & 0.2 \end{bmatrix} \end{array}$$

which indicates the suitability of presenting price with text, taking into consideration the importance for the hotel characteristic price. For example, if price is considered to be a low requirement, that means that it should be presented by text 0.3 low, 0.2 medium and 0.1 high (where *low*, *medium* and *high* for text indicate the font size). Other web page elements and properties such as size, location etc. can be represented in a similar manner. Then, the resulting fuzzy composition which relates elements of user types (T) to elements of text (W_{text}) will be found by the type:

$$\begin{aligned} &\mu_{R\circ P}(type1, low) \\ &= \max\left(min(0.2,0.3), min(0.4,0.5), min(0.1,0.5)\right) \\ &= max(0.2,0.4,0.1) = 0.4 \end{aligned}$$

Consequently, the fuzzy composition $\tilde{R} \circ \tilde{P}$, is:

$$\tilde{R}_{price} \circ \tilde{P}_{price,text} = \begin{array}{c} \\ type1 \\ type2 \\ type3 \end{array} \begin{array}{c} \begin{array}{ccc} low & med & high \end{array} \\ \begin{bmatrix} 0.4 & 0.4 & 0.4 \\ 0.5 & 0.4 & 0.4 \\ 0.5 & 0.3 & 0.3 \end{bmatrix} \end{array}$$

which shows that a type 3 user should view the price presented by text 0.5-low, 0.3-medium and 0.3-high. Thus, price when represented as text on the web page, should probably be shown using small size fonts for this type of customer.

Consider now a customer with a profile that is closer to a type 2 user:

$$\tilde{T} = \begin{array}{c} \begin{array}{ccc} type1 & type2 & type3 \end{array} \\ \begin{bmatrix} 0.4 & 0.8 & 0.2 \end{bmatrix} \end{array}$$

The decision for web adaptation comes from the composition $\tilde{T}$ o ($\tilde{R}$ o $\tilde{P}$), which is:

$$\tilde{AW} = \tilde{T} \circ (\tilde{R}_{price} \circ \tilde{P}_{price,text}) = \begin{array}{c} \begin{array}{ccc} low & med & high \end{array} \\ \begin{bmatrix} 0.5 & 0.4 & 0.4 \end{bmatrix} \end{array}$$

The composition shows that for the specific customer, the service price, when presented by text, should probably be presented by small fonts. The same composition is calculated for each service characteristic, web feature, and user type, until the most suitable design is chosen among all alternatives. Thus, customers having different profiles will view different web pages that dynamically adapt to their service requirements and corresponding design choices.

Conclusions

The use of fuzzy knowledge representation allows web page designs to be dynamically adapted to user profiles. The proposed approach suggests that the integration of ontologies and fuzzy relations provides the means to adapt web design along a continuum that is defined by the user profile and preferences as well as the suitability of web design alternatives. In the near future we are planning to validate the above approach by applying it to the domain of tourism web sites, by carrying out a number of experiments involving several user groups. The profiles of participants will be captured and codified using the ontologies described above. Then, the users will be asked to evaluate the usability of the automatically adapted web pages. The results of such experiments will allow us to assess the effectiveness of the proposed approach.

References

[1] Abedin, B. and Sohrabi, B. A Web Designer Agent, Based on Usage Mining Online Behavior of Visitors. *World Academy of Science, Engineering and Technology 6,* (2005), 41-45.

[2] Bodjadziev, G. and Bojadziev, M. Fuzzy Logic for Business, Finance, and Management. *Advances in Fuzzy Systems – Applications and Theory*, 12, World Scientific, Singapore, (1997).

[3] Buriano, L., Marchetti, M., Carmagnola, F., Cena, F., Gena, C. and Torre, I. The Role of Ontologies in Context-Aware Recommender Systems. In *Proc. of 7th International Conference on Mobile Data Management* (MDM 2006), (2006), 80-80.

[4] Cunliffe, D. Developing usable Web sites – a review and model. *Internet Research*, 10(4), (2000), 295 – 308.

[5] De Virgilio, R., Torlone, R., and Houben, G.J. A Rule-Based Approach to Content Delivery Adaptation in Web Information Systems. *Proceedings of the 7th International Conference on Mobile Data Management* (MDM'06), Nara, Japan, (2006), 21-24.

[6] Ha, S.H. Digital Content Recommender on the Internet. *IEEE Intelligent Systems* 21(2), (2006), 70-77.

[7] Hervas, R and Barvo, J. (2011), Towards the ubiquitous visualization: Adaptive user-interfaces based on the Semantic Web, *Interacting with Computers*, 23, 40-56.

[8] Lee, K. First Course on Fuzzy Theory and Applications. *Advances in Soft Computing*, Springer Berlin / Heidelberg, (2005).

[9] Paik, H-Y. and Benatallah B. Building Adaptive E-Catalog Communities Based on User Interaction Patterns. *IEEE Intelligent Systems*, 17(6), (2002), 44-52.

[10] Ross, T. J. *Fuzzy logic with engineering applications*, John Wiley and Sons, (2004).

[11] Van der Merwe R. and Bekker J. A framework and methodology for evaluating e-commerce web sites. *Internet Research: Electronic Networking Applications and Policy*, 13(5), (2003), 330–341.

Usability Evaluation: Commonalities and Discrepancies in Games and Safety Critical Systems.

Philippe Palanque
IRIT, Groupe ICS
118 Route de Narbonne
Toulouse, 31062, France
palanque@irit.fr

Regina Bernhaupt
IRIT, Groupe ICS
118 Route de Narbonne
Toulouse, 31062, France
regina.bernhaupt@irit.fr

Marco Winckler
IRIT, Groupe ICS
118 Route de Narbonne
Toulouse, 31062, France
winckler@irit.fr

Abstract

This paper presents the basic principles of widely agreed upon practice for usability evaluation in two different domains: Games and Safety Critical Interactive Systems. The user interfaces (and their associated interaction techniques) proposed in the applications deployed in these two domains are usually very different while the objects and functions to be manipulated might be very similar. For instance in a flight simulator game a rear view of the aircraft is usually proposed (see Figure 1), this rear view is also proposed in flight simulators (see Figure 2) while in a real cockpit such representation is not possible (see cockpit capture Figure 7). This is surprising as the key point in flight simulator games is to propose an as-realistic-as-possible cockpit interface.

Based on an overview on the development processes and the respective usability evaluation methods used in these two domains, a comparison of the approaches showing commonalities and discrepancies is presented. Concluding remarks focus on how these two domains could fruitfully cooperate, by adopting and enhancing methods used in the two domains. Additionally, such comparison could provide foundations for trans-

Figure 1. Screen shot from piloting game (outside view of the aircraft)

Figure 2. A rear view of the VRS Hornet on Flight Simulator

sectorial usability evaluation methods identifying both ground for common shared tools and techniques and also identifying idiosyncrasies.

Keywords
Games, safety-critical systems, usability evaluation, development processes.

ACM Classification Keywords
H5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

General Terms
Human-Factors, Reliability.

Introduction
Since early work on software quality from McCall in 1977 [8], usability has been explicitly identified (even within the software engineering community) as a key factor for the quality of software. However, for the evaluation of usability of any interactive system it is important to take into consideration the application domain as the application domains can heavily influence how the usability is evaluated. For instance, when evaluating usability in games it is important to take into account that games typically are not task based and that playing a game is (by definition [4]) non-productive and oriented towards entertainment. On the other hand, in domains where safety plays a more important role than usability, this preponderant factor influences usability evaluation and the interpretation of results: detected usability problems in a given user interface might not be resolved by improving the user interface, but by training the user to simply learn how to use the problematic interface in a safe and error-free manner.

Goal of this research is to compare the domains of safety-critical systems and games in terms of usability evaluation methods. We argue that both domains can benefit from each other by incorporating the knowledge available in the other domain into their usability evaluation approaches.

First, the two domains are described, giving an overview on the development processes used, the applied usability evaluation methods (including their need of technical infrastructure). Next the approaches in the domains are compared showing commonalities and discrepancies. Concluding remarks highlight potentialities for the two domains to learn from each other. Lastly, we outline a draft research agenda for a more global approach on trans-sectorial usability evaluation methods.

Games
For the development of games there is a common agreement in the community that successful games rely on an iterative development approach. Usability evaluation is an important aspect in games development: if a game is not usable (e.g. the interaction technology does not allow easy learning how to play the game), a game is typically not successful. There is a broad range of game development processes used, the majority of them having special evaluation phases (including usability evaluation).

Games Development Processes
When looking at the development of games, standard software engineering approaches like the waterfall-model or Boehm's spiral model have been described as useful for game development. In the majority of cases a successful game development company does not

Figure 3. Test bench for UAVs command and control system (Copyright © 1999-2011 by Aero-News Network, Inc. All rights reserved)

Figure 4. Microsoft playtest lab

Figure 5. Games usability testing in the field

release any information about the software development process used. Novak [9]reports on a mixture of concepts and methods that support iterative development as well as flexibility in the development process.

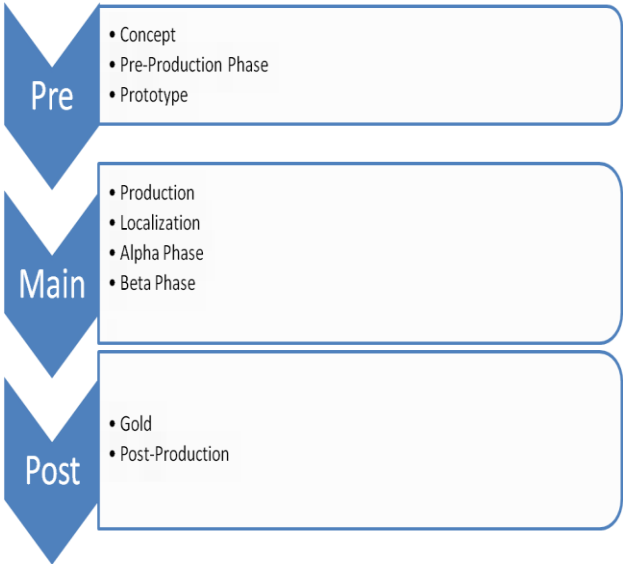


Figure 6 an example of game development process model (showing the 3 main phases)

Figure 6 presents an abstract view of a game development process. We have not represented the iterations that are of typical HCI practice (in the pre-phase), typical software engineering practice such as presented in the V model (in the main-phase) and closer to the agile processes (in the post-phase).

The following phases typically guide a game development process [9]:

(1) **Concept**: This phase is dedicated to the initial game idea and is devoted to producing a first concept document describing the game. The development team in this phase is typically small (e.g. consisting of designer, programmer, artist and producer).
(2) **Pre-Production Phase**: This phase includes the development of art style guides, production plans and first description of the game design and the technical design document.
(3) **Prototype**: Goal of this phase is a first working piece of software allowing to demonstrate key characteristics of the game and enabling to understand basic concepts related to the general user experience of the game ("Is the game fun to play?").
(4) **Production**: The production phase can range from few weeks development to years of programming. This phase can be structured additionally, following approaches like an increment to completion approach, a cascade approach or an "iterative- until you drop" approach [6].
(5) **Localization**: an important phase for games that will be delivered to different markets (countries) is the localization phase. In this phase game-play can be adjusted to suit the tastes of the market, to allow language translation and modifications due to local regulatory authorities.
(6) **Alpha-Phase**: This is the phase when a game is playable from start to finish, allowing different evaluation methods to be applied to better understand aspects like fun, playability and user experience.
(7) **Beta-Phase**: Main goal during this phase is normally to fix bugs. In terms of user experience in this phase lots of fine-tuning is necessary to improve the overall user experience. The beta-phase includes steps like certification or submission (the hardware-manufacturer of the proprietary platform will test the

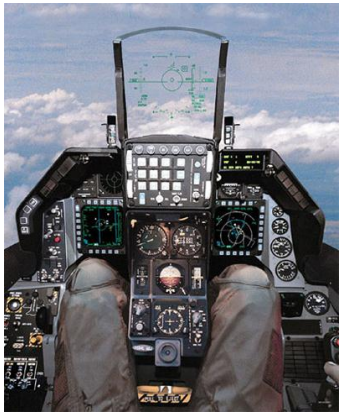


Figure 7. F-16 A/B fighter cockpit – inside view of the aircraft

game).
(8) **Gold**: In this phase the game is sent to be manufactured.
(9) **Post-Production**: In this phase subsequent versions of the game may be released (including patches and updated) and allows to improve the user experience of the game.

Games development has some additional special milestone (depending on the product developed). Especially the release candidate milestone is important for games that are produced for special consoles, as the release candidate is then tested and evaluated by the game console manufacturer.

Usability Evaluation in Games Development Processes

METHODS

For the usability evaluation there is a broad range of methods used, that are applied in the different phases of the game development. At the early stages of the development (concept phase) these methods are: focus groups, interviews, informal play-testing, questionnaires; at later stages, especially during the production phase these are play testing (including bio-metrical measurements), (semi-structured) interviews, observation, video coding, quantitative comparisons of gamers behaviours, questionnaires focusing on users attitudes or experiences or heuristic evaluation (including the usage of heuristics that focus additionally on the playability of the game).

Two usability evaluation methods from the entertainment and games domain have become a focus of attention in other areas (1) play testing and (2) beta-testing. Play tests are used during all phases of a game development, with the majority of usability and play testing performed once the alpha phase is accomplished until the beta-phase is finished. (Alpha is the milestone in games development were the game can be played the first time from the beginning to the end, still missing some branches in the game that are not finalized, or graphics that are not ready; Beta is milestone in the game development when no further developments and changes are made for the game play, and the only goal is to get rid of bugs). Beta-testing is a form of game-play evaluation that also includes usability evaluation. During Beta tests the game is distributed to (a selected community of) end users that are playing the game. These users (sometimes thousands of users) provide feedback in terms of usability (e.g. mapping of the input to the various game elements) as well as in terms of game-play (allowing for example the fine-tuning of game-rules based on player-death logging).

EQUIPMENT

Usability evaluation of games is conducted both in the lab (Figure 4) and in the field (Figure 5). To enable the evaluation of usability at early development stages prototypes are developed and used. To understand how usability and overall user experience and fun are related usability evaluation for games development today rely on sophisticated usability labs, including bio-physiological measurements, eye-tracking or motion sensors.

Safety-Critical Interactive Systems Development Processes

In safety critical systems several quality factors deeply influence the development process such as reliability, fault-tolerance or security. However, usability could/should be seen as a critical factor especially when

work such as [10] report that 80%-90% of accidents in industry are attributed to human error.

One key element of the development process of safety critical systems is the certification phase which involves an external body who is in charge of assessing the process that has been followed during the development of the system. Figure 8 presents such a process in which the last phase precisely concerns certification activities. It is important to note that, however, such phase is more prominent when the system involves "normal" citizens.

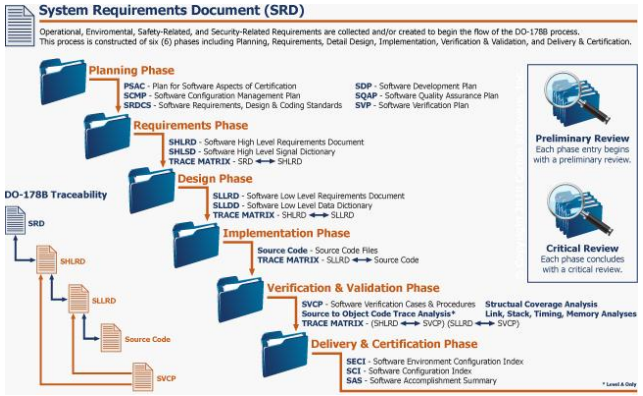


Figure 8. DO 178B [1] standard (no usability – no users)

Another typical aspect of the safety critical domain is the fact that standards processes are available and should be carefully followed. Figure 8 is defined in the DO 178-B standard from [1] while Figure 9 is from another standard IEC 61511 dealing specifically with safety issues [5].

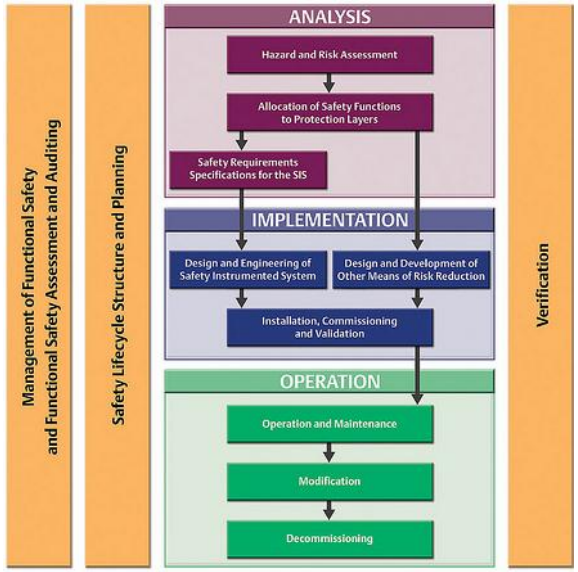


Figure 9. IEC 61511 [5] Process Safety (operations are part of the process)

Usability Evaluation in Safety-Critical Interactive Systems

Usability evaluation in safety critical systems is typically conducted as part of the development processes presented above. However, it is usually not an explicit phase as usability concerns are taken into account in more or less every phase.

From the scientific perspective, usability evaluation is scarcely described or even reported. From our experience, for instance in the case of civil aircrafts cockpit designs, usability and operability concerns are dealt with by means of expert users (called test pilots) who are involved from the beginning (architectural

design of the cockpit) to the very late final phases (design of the training of airlines pilots). This could be compared to participatory design approaches.

Using a simulator as a test-bed for design and testing of concepts and interfaces can be seen as a specific equipment or supporting usability studies. For instance, while designing the elements of a head-up display, scenarios can be defined and tested to assess the impact of a display freeze of the head-up display on the operations (i.e. the pilot's tasks).

To summarize a broad range of usability evaluation methods are used in this area but they typically do not refer and relate to the standard method descriptions advocated in the field of HCI.

Equipment for Usability Evaluation
As for games, usability evaluation for safety critical systems is conducted both in the lab and in the field. But given the necessary focus on safety and reliability, evaluations in the lab are based on the extensive usage of very complex and realistic simulators (e.g. force feedback systems to simulate the behavior of physical equipment in the cockpit, vibrations, 3D movements of the cabin …). Figure 3 presents a realistic simulator for a UAV (unmanned aerial vehicle) control and command system embedding a physical model of the aircraft, weather, gravity ….

When usability evaluation is carried out in the field, very specific equipment has to be designed and embedded to gather information about the operations and the pilot. As for games, they can collect biometric measures about the crew but in addition, they have to be very robust to variations as the evaluation has to be as close as possible of the real operations. For instance, eye tracker devices must be able to track eyes movements both in the dark (while landing at night) and while facing sun (while departing East in the morning or landing West in the evening) [1].

Lastly, usability testing in safety critical systems exhibits the following constraints:

- Expensive users (retired pilot, current pilots from airlines, test pilots working for the manufacturer),
- Expensive real tests for a new designs/systems (50k€ one hour flight test to be included in the budget),
- Complex tests scenarios that are usually based both on normal and abnormal conditions (previous incidents/accidents, emergency situations, partial failure of systems …).

Summary of Similarities and Discrepancies

Usability is usually evaluated according to the time spent in training (costly training is perceived as poor design of the UI).

Games - Process	**Safety-Critical**
Development Process	
Process Model: iterative	Process Model: waterfall or V model
Specialized development phases include localization,	Specialized development phase include certification

internationalization, gold	
Special verification phase by game console manufacturer	Special customization phase of procedures by the airline
System should be of type walk-up and use (or with a very simple practice scenario included)	System is designed together with training material. Similarity with respect to other interfaces is relevant (to support cross-qualification of pilots).
Usability Evaluation Methods	
Usability can be less important than user experience	Safety and reliability are more important than usability (User experience is not an issue).
Usability evaluation is combined with user experience evaluation (e.g. playability tests)	Usability (performance and efficiency) is prominent for the definition and validation of operations
Beta-testing as a unique way to gather user feedback before the software is released. This generates a huge amount of data for improvements (e.g. balancing).	Incident and accident reporting activities constitute longitudinal studies and inform future designs and re-designs.

Concluding Remarks

Looking at the specificities of the two domains it is surprising to find a set of similarities: both domains share the approach that usability as a factor of the interaction with the system might be less important than other software quality factors. Both domains have a set of additional development phases and usability evaluation is typically quite heavy in terms of infrastructure. Being interested in the domain dependent adaptation of usability evaluation, the domain of safety-critical system and the domain of games are promising candidates to understand how usability evaluation methods must be adapted to better fit the domain.

We believe that safety critical systems and games represent the two extremes of a continuum of domains where usability evaluation methods have to be customized and adapted to fit their specificities.

In a nutshell and to provide a second adaptation of Landsburg's words as stated in [2]: "Usability wants us to die rich; UX wants us to die happy", we propose a new vision of it by saying “Usability wants us to die fit; UX wants us to die happy; Safety want us to die still alive”.

Acknowledgements

This work is partly funded by Airbus under the contract CIFRE PBO D08028747-788/2008, R&T CNES (National Space Studies Center) Tortuga R-S08/BS-0003-029 and ruwido.

References

[1] Dehais, F., Causse M. & Pastor J. Toward the definition of a pilot’s physiological state vector through

oculometry: a preliminary study in real flight conditions. In proceedings of HCI Aero 2010, ACM Digital Library.

[2] DO-178B (1992) Software Considerations in Airbone Systems and Equipment Certification. RTCA Inc, EUROCAE, December

[3] Hassenzahl M., Law E & Hvannberg E.T. (2006) User Experience – Towards a unified view. In proceedings of "User Experience – Towards a Unified View: Second International COST294-MAUSE Open Workshop", NordiCHI 2006. P. 1-3.

[4] Huizinga J (1950) Homo Ludens: A Study of the Play-Element in Culture. Beacon Press, Boston, MA

[5] IEC 61511 Functional safety - Safety instrumented systems for the process industry sector. Beuth Verlag GmbH, Berlin, 2003.

[6] Irish D (2005) The Game Producer's Handbook. Thomson Course Technology PRT.

[7] Landsburg S. (1993) "The Armchair Economist: Economics and Everyday Life", The Free Press, 241 p.

[8] Mc Call J (1977) Factors in software quality. (Ed.) General Electric.

[9] Novak J (2008) Game Development Essentials. Delmar Cengage Learning.

[10] Salminen S. And T. Tallberg, Human errors in fatal and serious occupational accidents in Finland, Ergonomics 39 (1996) (7), pp. 980–988

Project COPE: Development and Evaluation of a Technology Concept in Emergency Response Domain

Leena Norros
VTT Technical Research Centre of Finland
P.O. Box 1000
02044 VTT, Finland
Leena.Norros@vtt.fi

Paula Savioja
VTT Technical Research Centre of Finland
P.O. Box 1000
02044 VTT, Finland
Paula.Savioja@vtt.fi

Marja Liinasuo
VTT Technical Research Centre of Finland
P.O. Box 1000
02044 VTT, Finland
Marja.Liinasuo@vtt.fi

Abstract
In this paper we describe what kind of usability evaluation and development methods we have used in concept development and evaluation work in emergency response domain. The work described in this paper was carried out in EU FP7 funded project COPE (Common Operational Picture Exploitation). The aim was to create a technology concept and related work practices to support common operational picture (COP) in on-line emergency response work. Various usability and human factors methods were used in the course of the project and a specific field research method labeled parallel augmented exercise was created to enable studying safety critical work in live settings. In this paper we address the issue of conducting usability work in the very unique domain of emergency response.

Keywords
Concept development, methods, emergency response work

ACM Classification Keywords
H5.2. Information interfaces and presentation (e.g., HCI): User Interfaces.

Introduction

The starting point for the EU project COPE (Common Operational Picture Exploitation; see http://cope.vtt.fi/) was to develop technological concepts that support and enable common operational picture (COP) among a multi-agency emergency response professionals. At the same time it was acknowledged that any development of a technical tool has an effect on the work practices within the activity system (task-artifact cycle). Consequently, a further goal for development work was identified: the development of emergency response practices and concepts of operation (ConOps). In addition, one of the starting points for the whole project was that a special emphasis was to be laid on carrying out the development work in a user- and usage-centered manner and ensuring high level end user involvement throughout the project.

In this paper we give a brief introduction, first about the research process as a whole and secondly about the usability development and evaluation methods used in the project. More detailed descriptions of the methods can be found e.g. in [7] [6, 8]

Background

From a methodological point of view the COPE project required innovative ways of designing the research. This is due to the specific nature of the emergency domain. The work carried out by fire fighters, paramedics, and their leaders is extremely safety critical, very procedural but at the same time dynamic in situ, and it has developed within a strong culture of professional practice which is not very open to interventions concerning new tools or ways of working.

The VTT human factors group has developed a work analysis method called core task analysis [5] to be used for analyzing and developing work activities in different safety critical domains. The idea in core task analysis is to identify *the core* of a particular activity, i.e., the result-critical content of the work that is defined by the objectives of the work and the constraints of the domain (environment). Main information gathering methods in core task analysis are interviews, observations, and analytical considerations concerning the nature of the domain. In the analysis phase of core task analysis the information from the ways in which the professionals describe their work is combined with the observed practice of conducting the work and embedded in the characteristics of the domain. Based on this, *core task demands* of a particular work can be identified. Core task demands denote the ways of working which enable reaching the objectives of the work in the dynamic, complex and uncertain environment.

Early on in the COPE project, creating common operational picture (COP) was identified to be one of the key core task demands that must be fulfilled in order to reach the objectives of emergency response (ER). What is the way COP is constructed and exploited in ER, what is the way COP brings added value to ER and, above all, how the added value of the technology concept on COP can be evaluated, are the basic questions with which the human factors group in COPE project was faced with and towards which the methodological choices had to be tuned.

Flow of HF research in COPE

During the course of the project we tried to avoid focusing on the possibilities of technology as the main

driver for the development of tools in Emergency Response. The challenge is to balance between the needs of the practice and the technology push. The whole project was designed in a manner (Figure 1) that made it possible for the technology development to conduct fast prototyping and small scale technical tests while at the same time the human factors group was gathering data and analyzing the end user activity. The products of technology development, early prototypes and concept descriptions were used as design interventions in the human factors work to enable early feedback from the users.

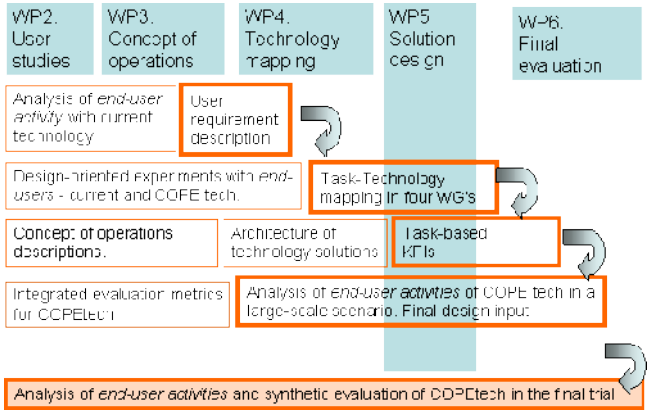


Figure 1. Organization of the COPE project in work packages. Work packages are depicted in blue and each row below the blue boxes represent a phase in the project. Time progresses from left to right in each row, starting from the top row and proceeding to the row below. The boxes with strengthened outline represent the output in each phase (row) that is used in the next phase (row).

Figure 1 demonstrates the organization and flow of research and development in COPE. WP2 concentrated on user studies and the demands of different individual personnel in emergency response. WP3 considered emergency response from the processes and ConOps point of view. As WP5 was the technology development work package, WP4 acted as an intermediate playground for mapping the understanding of end user activity resulting from the work of WPs 2&3 to technological innovations created in WP5. Finally WP6 was the evaluation work package the aim of which was to organize the evaluation of the operational effect of the developed technological solutions and concept of operations.

The project started by letting the individual work packages act rather independently (first row in Figure 1). During this phase several important HF research steps took place: Human Factors White Paper was written to define the scope of the COPE project with regard to the entire emergency management, and to provide an initial conceptualization of the concept Common Operational Picture (COP). Deliverable ñUse Case Descriptions and a Human Factors Engineering Frameworkò was written. It describes first responder work on the basis of previous literature. A proposal for so called initial use cases was made and a plan for empirical data gathering concerning end user activities proposed. Deliverable ñEnd User Requirementsò was written based on empirical user studies (see later this paper). At this stage it became clear that more interaction was needed among the Human Factors group and the engineering design. Design-oriented experiments and four technology mapping working groups were established to increase the interaction and joint design work (second row of the Figure 1). Next, following steps were achieved: Deliverable ñHF-Based Design Inputs to COPE technology ï Conceptual and

Empirical Considerations of Common Operational Pictureò was produced. This deliverable includes detailed analysis of end user interviews. Moreover it reports in detail the design study accomplished by one of the technology mapping groups at the Emergency Services College (ESC) in Kuopio. This deliverable also makes a first proposal for the methodology to be used in the Human Factors evaluation of the COPE technology. The work was summarized in Deliverable ñThe Technology Mapping Workshop Outputsò. An important phase in the design process was when the human factors ïoriented work for defining the concept of operations was found to be restricted if it only would focus on the singular technologies that the individual working groups had been working with. Hence, the need for a comprehensive architecture was discovered and the architecture was articulated within the technology mapping activity and the work for defining the quality criteria for a good system could begin (Key Performance Indicators) (third row in Figure 1): Deliverable ñScenario Descriptions from User Perspective ï Key Performance Indicators and Scenario Requirementsò was written. As the next step the human factors group focused on further analyzing the data from available end-user interviews and design experiments to formulate a set of comprehensive human factors evaluation metrics. This was used in two large scale exercises with end-users and developed further. (The fourth row in Figure 1): The results of end-user studies in design contexts were used to improve and finalize the COPE technology concept.

The final demonstration took place in September 2010 in the form of a large-scale final exercise. About 100 end-users and experts participated in the exercise and comprehensive data acquisition was organized by the human factors group concerning the activities of the personnel and their experience of the technologies (last row of Figure 1). Deliverable ñCOPE Technology Enabled Capacity for First Responderò summarizes all the empirical results acquired of the usage of COPE technologies during the project. For synthesizing the results verification and validation evaluations were accomplished and reported.

Methods used in human factors work

Methods of gathering data about the end user activity

Typically we would start familiarizing with a new domain by conducting e.g. Contextual Inquiry (CI) [1] type of studies. But in the domain of fire and rescue we cannot interfere with the actual activity to the extent which would be necessary for CI. The reasons are obvious and numerous. E.g. we cannot even go close enough to the fire fighters acting in a situation to ask questions and explanations and really to understand what is going on. Another method to familiarize with a domain is to gather the task descriptions and procedures that have already been created. This was done to a certain extent in COPE but the data gathered this way does not reflect all aspects of the work, as even though procedures are followed quite systematically they still leave room for individual ways of acting for the fire fighter crews.

INTERVIEW METHOD

Our interview method was based on the Critical Decision Method [3] to elicit personal, lived experiences from fire fighters, and to ground discussion in real events (versus hypothetical scenarios that will often elicit only opinions and generalizations). The interview methodology can be described in terms of ósweepsô or iterations of the same incident experienced by the fire

fighter, focusing on different levels and types of information each time [2]. The CDM was adapted with a final sweep through the incident where the interviewee had the opportunity to envision the use of various technologies in the context of the event.

OBSERVATION METHOD

Since it would have been too difficult to observe professional firefighters in real-life tasks, we observed firefighter students taking their practical examinations. In the exam a crew of fire fighter paramedics performed four small-scale emergency exercises in natural settings. Each crew was assigned the tasks of roof fire, cellar fire, car-accident and ventilation of an industry hall. At each incident site there was an instructor who gave a short briefing about the situation, observed the crewsû performance and evaluated it. The comments of the instructor were also considered as data concerning emergency response work.

PARALLEL AUGMENTED EXERCISE

A wider scale human factors evaluation took place in the first field test in the form of a parallel augmented exercise. In this design study the evaluation of the technology concept was connected to a live accident exercise conducted at ESC. Students performed in the exercise as usual but experienced fire fighters, forming the ñparallel groupò, used the developed technology and envisioned what it would be like to work in the on-going accident with such technology.

Methods for evaluating usability of the COPE concept

A longitudinal evaluation approach was developed for the project. Accordingly, the evaluation took place stepwise and assessment methods were adapted to the maturity of the technology concept. Hence, in the earlier design phase, the evaluation was integrated into the technology mapping process. Design workshops were organized which each focused on singular applications.

VERIFICATION

The evaluation took two basic forms, i.e. verification and validation. In the verification earlier project documents were used as reference to test the fulfillment of requirements. The requirements had been defined on the basis of analysis of emergency response activity. In the verification the method was a walkthrough accomplished during the final trial. Walkthrough was arranged in the form of interdependent tasks that some fire fighters performed with the new technology. Thinking aloud was encouraged and after the performance, a brief interview about the experience and opinions was performed to each fire fighter separately.

VALIDATION

IN validation the focus was on testing whether the entire COPE technology concept would support emergency responders cognitive work demands, especially those connected to creation of Common Operational Picture. A new method was used that is labeled the Usability Case [4]. It provides a systematic reasoning tool and reference for gathering data of the technology under design, and for testing its usability in the targeted work. Following this method, a model was created that defined the intrinsic work demands that are needed to maintain COP, i.e. sense making, coordination and maintaining common ground. Next, the concept level requirements for the entire technology were defined. Thereafter, the singular technological functionalities of the COPE applications

were ordered in four main concept solutions. Then connections were identified concerning which singular technologies would fulfill each identified concept solution, how each concept solution would fulfill the concept requirements, and how the concept requirements would support the intrinsic cognitive demands of COP. On the basis of this reasoning, claims about the COPE technology were created. Testing of the fulfillment of these claims was the content of the validation evaluation. All user experience and behavioral data collected during the project in different evaluation phases was pooled in a database and used as evidence to test whether the claims about the advantages of the COPE technology for creating COP could be supported. Also negative evidence was registered.

Discussion

Above we have described what kind of design process and evaluation methods were used in project COPE. Throughout the project, the requirement of evaluating technology and working practices in a highly safety-critical work but without interfering it created challenges that, we consider, were overcome by the methods used. The design process was participatory with high involvement of end users in several design experiments. The co-operation of technology development and human factos was organized through technology mapping. The interview method CDM provided the interviewees an opportunity to reflect on the naturalistic decision making processes of real work. Observation was conducted in training situations which gave the instructor an opportunity to explain the work to the researchers. Evaluation was conducted through parallel augmented exercise which ensured zero interference of the activity but enabled professionals reflection of promisingness of the new concept. Usability Case was used for organizing all the human factors data and identifying the validity of the technology concept.

Citations

[1] Beyer, H. and Holtzblatt, K. Contextual Design: Defining Customer-Centered Systems. Morgan Kaufman, San Francisco, 1998.

[2] Crandall, B., Klein, G. A. and Hoffman, R. R. Working Minds. A Practitionerôs Guide to Cognitive Task Analysis. MIT Press, London, 2007.

[3] Klein, G. A., Calderwood, R. and MacGregor, D. Critical decision method for eliciting knowledge. IEEE Transactions on Systems, Man and Cybernetics, 19, 3 (1989), 462.

[4] Liinasuo, M. and Norros, L. Usability Case - integrating usability evaluations in design. In Proc. Mause workshop on Downstream Utility, 2007.

[5] Norros, L. Acting under Uncertainty. the Core-Task Analysis in Ecological Study of Work. VTT, Espoo, 2004.

[6] Norros, L., Hutton, R., Grommes, P., Colford, N., Liinasuo, M. and Savioja, P. Analysis of work demands of multi-agency emergency response activity for developing information support systems. In proc. ECCE 2009

[7] Norros, L., Liinasuo, M. and Hutton, R. Evaluating the promsingness of new technological tools for safety critical work. Interacting with Computers, (in press).

[8] Valkonen, P. and Liinasuo, M. Role playing with fire fighters: using a worst case scenario and verbal re-enactment in the role play. In Proc. NordiCHI 2010 (Reykjavik, Iceland). ACM, New York, NY, USA, 2010, 805-808.

Groupware Evaluation: An overview

Irene-Angelica Chounta
HCI Group, Dept. of Electrical
and Computer Engineering
University of Patras, 26500 Rio
Patras, Greece
houren@upatras.gr

Nikolaos Avouris
HCI Group, Dept. of Electrical
and Computer Engineering
University of Patras, 26500 Rio
Patras, Greece
avouris@upatras.gr

Abstract
This paper discusses current trends in the field of evaluation of groupware applications. We review existing evaluation methodologies in CSCW, present the experience gained from case studies in the field and work in progress. It outlines the importance of using traditional single user evaluation methods in combination with evaluation of groupware specific issues, like communication and awareness mechanisms as well as the importance of context of use in the process.

Keywords
evaluation, groupware applications, collaboration

ACM Classification Keywords
H5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

General Terms
Evaluation, groupware, collaboration

Introduction
Collaborative applications (or groupware) are software systems that support people involved in a common task to achieve their goals. These systems have been developed in the broader field of Computer Supported Cooperative Work (CSCW) that describes different kinds of activities that are carried out by a group that uses information and communication technology. A similar area of study is that of Computer Supported Collaborative Learning (CSCL) that puts special emphasis on the learning outcome of the collaborative activity. The two areas share many technological and theoretical background issues. The technology in this context usually supports communication and coordination of the activity, as well as providing structure for the activity, stores record of action and intermediate and final states of the activity outcome. An activity of this nature, in general, involves two or more people who work together sharing a common objective. Evaluation of groupware applications is an open challenge for the past years and the subject of many studies. For years, research has focused on the

various aspects of collaboration that should be supported through groupware applications in order to turn collaborative activities into a success. Evaluation experts claim that until a general framework or set of guidelines is established, we should focus in evaluating the various factors of collaborative systems separately [1]. Groupware evaluation itself involves the assessment of various aspects. The objective of evaluation may be the collaborative experience, the group activity outcome, the mechanics of group interactions or communication means used, or simply the groupware application used. The objective of the evaluation study should become explicit as it is often difficult to distinguish these different aspects. For example evaluation of the technology may be done through study of a typical case of its use and often cannot be distinguished from it. One problem that has been put forward, is to define commonly acceptable methods and tools for evaluation of groupware applications. Such methods need to involve a combination of various complementary approaches that take in consideration the typical context of use of the groupware application. In the following sections we provide a brief review of the field of groupware evaluation, the experience we have gained from various evaluation studies and some ideas on the perspective of this field.

Groupware evaluation: a brief review

In general, there are some principles that groupware systems are expected to support such as, communication between partners, the establishment of shared understanding and goals, support for awareness of others' actions. In this fashion, we can define the groupware systems as the combination of technology software (applications that offer common resources, fig. 1) and hardware devices (e.g. microphones, cameras etc) that promote collaboration among a group in order to achieve a shared goal. It is common in today's complex environments, a groupware application to contain both synchronous and asynchronous communication components, synchronous and asynchronous content sharing (e.g. file sharing or joint editors) and content management components. The scenarios of typical use of such systems are extended and complex and involve usually distributed groups of people and multiple contexts of use.

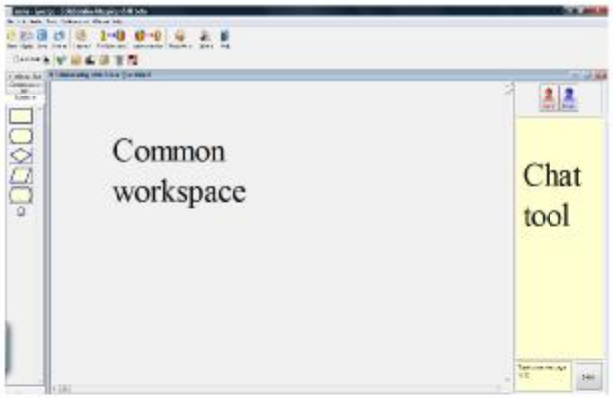


Figure 1. Example of a synchronous groupware application, it typically includes a communication tool (Chat) and a shared work space, including various awareness mechanisms.

While for single-user applications there is already an extensive corpus of knowledge relating to evaluation methods and tools, groupware evaluation methods are still in evolution. Originally some of the most well known groupware evaluation methods derive from HCI methodologies for single user evaluation as classified by McGrath in Table 1 [12]. These methods have been adapted for the collaborative context. Such methods

are: Expert based methods like use of heuristics [2,3], User testing, usually in labs [16], survey of user opinions through interviews & questionnaires, focus groups, etc, and ethnographic methods, like contextual inquiry. Among these techniques, the most prominent one involves ***user studies in the lab***. For groupware systems however, it is particularly difficult to create a group in a laboratory that will reflect the real case scenarios [17].

Table 1: Evaluation Types (from [12])

		Manipulation	
		Rigorous	**Minimal / None**
Setting	**Naturalistic**	Field Experiment	Field Study, Case Study
	Controlled	Laboratory Experiment	Exploratory

Contextual inquiry in the field may be also used for groupware evaluation, however the distributed nature of such applications constrains the evaluation process necessitating high amount of resources, while it is particularly difficult to reconstruct the user experience, as this necessitates processing of large amounts of data. These restrictions led to the development of discount methods and frameworks that are based on human information processing models. We summarize some of these frameworks:

Mechanics of collaboration describe the basic operations that appear during group work towards a common goal. Collaboration, including communication and awareness aspects, is analyzed and mapped through small-scale actions described by mechanics of collaboration [8]. Frameworks such as ***Collaboration Usability Analysis (CUA)*** and ***Groupware Heuristic Evaluation*** were based on the mechanics of collaboration. Collaboration Usability Analysis (CUA) is based on a hierarchical task model for group work than involves field studies and incorporates several modified task analysis techniques. It provides high and low-level representation of collaborative activity and the interactions that take place within it [18].

Groupware Task Analysis (GTA) [20] is another analysis method and framework for the design of groupware that focuses on the triplet people, work, situation. Hierarchical task analysis in combination with human information processing models, such us keystroke level model, has also been used in groupware evaluation either in order to decompose group work to the level of individual subtasks [13] or to be integrated and adapted in collaborative settings [6].

In addition, concept-specific frameworks of study have been proposed, focusing on specific aspects of collaborative work such as awareness, communication and coordination. These frameworks put emphasis on participatory design and evaluation [19], extending data logging methods for collaborative systems [10], as well as methods for modeling and mapping awareness within a collaborative setting [14] and breakdown analysis [9].

Lessons learned from evaluation studies

In this section we briefly discuss a number of recent groupware evaluation studies that we conducted. These are indicative of the current trends in evaluation methods of this field. The presented studies are part of a more general framework that aims to provide guidelines and rules towards successful collaborative group activities. During these studies traditional HCI methods for single user application evaluation were combined with CSCW evaluation methodologies, while special emphasis was given to the effectiveness of alternative awareness mechanisms. The studies also involved qualitative analysis of video to assess the collaboration activity. We present two of these studies.

The first study [5] involved Heuristic Evaluation of a web-based argumentation tool used by communities of practice. For the evaluation of the application, heuristic evaluation for single-user interface was combined with heuristics for CSCW based on the mechanics of collaboration. Groupware Heuristic Evaluation [2] provides a set of heuristics for Shared-Workspace groupware that can be used by a number of experts for the evaluation of groupware. The groupware application supported both synchronous and asynchronous collaboration and was studied for both settings. Overall fifty (50) evaluators took part in the study. It was clear that in evaluation of groupware applications there are various aspects that need to be studied and evaluated separately but taken into consideration for an overall assessment. Expert based inspection methods should be combined with user observation methods, while we need not only focus on the collaborative functionality but also on the user interface design issues. This has been confirmed by studies that revealed that most of the issues occurring in a collaborative session are due to inconsistencies of the interface rather than communication and awareness problems [5].

The second study [4] involved study of synchronous collaboration for problem-solving and examined allocation of attentional resources during different collaborative sessions. Three dyads' practice was monitored and analyzed. The dyads were formed in order to study different group dynamics. One dyad consisted of participants of similar knowledge background and equally motivated towards collaboration, a second dyad consisted of users of different knowledge background where the inexperienced user was monitored by the eye tracker and the third dyad was also formed by users of different background but the experienced user was monitored. The logfiles of the collaborative activity where combined with the logfiles of the eyetracker in order to analyze the interplay between task, awareness mechanisms and collaborative practice. This study revealed that awareness and communication failures are often interpreted as unwillingness towards collaboration or gradual lost of interest in a collaborative activity. The lack of adaptive awareness mechanisms that help users to set priorities towards the common goal instead of just informing about who, when and where is active, leads users to withdraw from the joint activity. They remain visible in the common workspaces, are aware of the actions of their partners but take no actual role in the collaborative activity. [4]

Another objective is to define evaluation metrics for groupware applications. Quality of collaboration has been correlated to automated metrics taken from logfiles [11]. In the same way, it is expected that data of eye movements and activity on the common

resources of groupware applications are able to provide us with useful information concerning the usability and quality of collaborative systems.

Conclusions and further research

Grudin in his article ñWhy CSCW applications fail: problems in the design and evaluation of organizational interfacesò [7], outlines a number of key issues in groupware design and evaluation. These concerns are still valid, despite that they have been stated over twenty years ago. Collaborative activities are difficult to model and analyze. The outcome of collaborative activities rely on many factors such as the quality of collaboration, the context of the activity and the tools that mediate the activity. The complexity of collaboration setting makes it difficult to identify the source of observed problems. So for instance tools that may support efficiently groups for playing a game may not support students during a learning activity, or social networking tools may not be satisfactory as argumentation system. So there is still a long way to go in establishing a framework for evaluation of groupware applications, that takes all these issues in consideration and can be used in a wide range of collaborative applications. On one hand, as discussed in this paper, there is already a substantial experience in the community of research and practice and reports of good practice exist that indicate the need to use a combination of evaluation methodologies, ones relating to single-user interfaces and more specific methods relating to groupware characteristics. However an open issue remains how to include the context in which each application is used and the quality of the outcome and how to take them into account in the evaluation process.

References

[1] Ackerman, M., Halverson, C., Erickson, T., & Kellogg, W. Resources, co-evolution and artifacts: Theory in CSCW, UK: Springer, (2008).

[2] Baker, K., Greenberg, S., Gutwin,C. Heuristic Evaluation of Groupware based on the Mechanics of Collaboration. Lectures in Computer Science, Engineering for Human-Computer Interaction, (2001), 123-139.

[3] Baker, K., Greenberg, S., Gutwin,C. Empirical development of a heuristic evaluation methodology for shared workspace groupware, Proceedings of the ACM conference on Computer supported cooperative work, (2002), 96 ï 105.

[4] Chounta I.A., Avouris N. Study of the effect of awareness on synchronous collaborative problem-solving. In 2nd International Conference on Intelligent Networking and Collaborative systems Proceedings ï INCoS, (2010), 153-160.

[5] Chounta, I.A., Avouris, N. Heuristic Evaluation of an Argumentation Tool used by Communities of Practice. In EC-TEL Workshop: TEL-CoPs'09: 3rd International Workshop on Building Technology Enhanced Learning solutions for Communities of Practice, (2009).

[6] Ferreira, A., Antunes, P. A Technique for Evaluating Shared Workspaces Efficiency, in Computer Supported Cooperative Work in Design III, Lecture Notes in Computer Science, 4402, Springer, (2007), 82-91.

[7] Grudin, J., Why groupware applications fail: Problems in design and evaluation, Office: Technology and People, 4(3), (1988), 245-264.

[8] Gutwin, C., Greenberg, S. The Mechanics of Collaboration: Developing Low Cost Usability Evaluation Methods for Shared Workspaces, IEEE 9th International

Workshops on Enabling Technologies: Infrastructure for Collaborative Enterprises (WET ICE'00), (2000), 98.

[9] Hartswood, M., Procter, R. Design guidelines for dealing with breakdowns and repairs in collaborative work settings. In International Journal of Human-Computer Studies, 53(1), (2000), 91-120.

[10] Helms, J., Neale, D.C., Isenhour, P.L. and Carroll, J.M. Data Logging: Higher-Level Capturing and Multi-Level Abstracting of User Activities. In ***Proceedings of the 40th annual meeting of the Human Factors and Ergonomics Society, (2000)***.

[11] Kahrimanis, G., Chounta, I.A., Avouris, N. Study of correlations between logfile-based metrics of interaction and the quality of synchronous collaboration, 9th International Conference on the Design of Cooperative Systems, Workshop on Analysing the quality of collaboration, (2010). Published in International Reports on Socio-Informatics (IRSI), 7(1), 2010

[12] McGrath, J.E. Methodology matters: doing research in the behavioral and social sciences. In ***Human-computer interaction***, Ronald M. Baecker, Jonathan Grudin, William A. S. Buxton, and Saul Greenberg (Eds.). Morgan Kaufmann Publishers Inc., 1995, 152-169.

[13] Min, D., Koo, S., Chung, Y.H., Kim, B. Distributed GOMS: an extension of GOMS to group task. In ***Systems, Man, and Cybernetics, IEEE SMC '99 Conference Proceedings,*** 5, (1999), 720-725.

[14] Neale, D.C., Carroll, J.M, Rosson, M.B. Evaluating computer-supported cooperative work: models and frameworks. In ***Proceedings of the 2004 ACM conference on Computer supported cooperative work*** (CSCW '04). ACM, (2004), 112-121.

[15] Olson, G. M., and Olson, J. S. Technology support for collaborative workgroups. In G. M. Olson, J. Smith, and T. Malone (Eds.), Coordination theory and collaboration technology. Hillsdale, NJ: Lawrence Erlbaum Associates, (2001), 559-584.

[16] Olson, G. M., and Olson, J. S. User centered design of collaboration technology. Journal of Organizational Computing, 1, (1991), 61-83.

[17] Pinelle,D., Gutwin, C. A review of groupware evaluations. WETICE ô00: 9th International Workshop on Enabling Technologies, IEEE Computer Society, (2000), 86ï91.

[18] Pinelle,D., Gutwin, C., Greenberg. S. Task analysis for groupware usability evaluation:Modeling shared-workspace tasks with the mechanics of collaboration. ACM Transactions on Computer-Human Interaction, (2003), 10(4):281ï311.

[19] Ross, S., Ramage, M., Rogers, Y. PETRA: participatory evaluation through redesign and analysis. In Interacting with Computers, 7(4), (1995), 335-360.

[20] Van der Veer, G.C, Lenting, B.,F., Bergevoet, B.A.J. GTA: Groupware task analysis -- Modeling complexity, Acta Psychologica, 91(3), Usage of Modern Technology by Experts and Non-professionals, (1996), 297-322.

A comparative evaluation of mouse, stylus and finger input in shape tracing

Stanislaw Zabramski
Department of Informatics and Media
Ekonomikum, Kyrkogårdsg. 10
751 20 Uppsala, Sweden
stanislaw.zabramski@im.uu.se

Dimitrios Gkouskos
Department of Informatics and Media
Ekonomikum, Kyrkogårdsg. 10
751 20 Uppsala, Sweden
dimitrios.gkouskos@im.uu.se

Mats Lind
Department of Informatics and Media
Ekonomikum, Kyrkogårdsg. 10
751 20 Uppsala, Sweden
mats.lind@im.uu.se

CHI 2011, May 7–12, 2011, Vancouver, BC, Canada.
ACM 978-1-4503-0268-5/11/05.

Abstract
A comparison of the aggregated performance measurements for three input methods in a line-tracing task is presented. Results indicate that users perform best using touch input, both in tasks with and without visual feedback, therefore we recommend touch input as the preferred input method for simple drawing tasks.

Keywords
Mouse, stylus, touch, freehand, tracing, comparison.

ACM Classification Keywords
H.5.2 User Interfaces: Evaluation/methodology.

Introduction
Research on input methods and their influence on human input has been focused mainly on the performance aspects. Many input devices have been tested on their effectiveness in pointing, dragging, crossing and path steering navigation tasks and this knowledge is used for different analyses and comparisons [7, 4, 6]. These standard navigational tasks became the subjects of mathematical modeling. Fitts' Law is a proven method that models linear pointing and clicking tasks but it appears to be not as well suited for modeling two dimensional tasks [5]. This has been supplemented by the Steering Law, which is a

more suitable predictive model for investigating two dimensional navigation tasks by considering them as a constrained motion within predefined tunnels of error [1].

The recent popularity of hand held devices equipped with displays capable of sensing multiple ways of human-input created new opportunities for creative users. Any surface-based human input can be broken down to a time-series of 2D coordinates. Therefore we can use the analogy of line tracing to describe the output of the continuous user's action that takes place e.g. on a touch-sensitive surface.

Line tracing can be an example of a task which might be negatively influenced by the low accuracy of the input method used for drawing lines but also by any kind of constraints imposed on the user. However, we have been unable to find a model specifically designed for unconstrained drawing in freehand input with initially unpredictable user error. Therefore, we decided to investigate the shape-based approach to assess the potential of small screens for unconstrained free-hand tracing tasks.

In order to compare the three most commonly used input devices, namely mouse, pen and touch input, in a drawing task, we performed an experiment with 16 participants.

In the experiment, the participants had to complete a tracing task of tracing a simple, random shape with all three input methods. The purpose of this experiment is to find out if there is a clear difference in performance when performing the tracing task between the three input-devices that are being investigated.

Method

In our study, users were asked to trace over, in one stroke, a greyed-out shape that was displayed on the screen using one of the three input methods in question in a randomized order. Additionally, we controlled the visual feedback of tracing that imitated drawing with black or invisible ink. We have also measured the accuracy and time it took the participants to perform the task.

The shape that was given to the participants was designed with use of a modified version of Method 4 described by Attneave [2]. We created asymmetrical non-sense contour shapes that did not bear any resemblance to well-known shapes. The modification of Attneave's method was limited to making the shapes consisting of at least two of each kind of perceptually meaningful properties like: convex corners, concave corners, straight line segments, and curve line segments. These segments of the shape did not cross at any point and their parameters like length or angle were randomized.

There are multiple factors that can describe the differences between two shapes: general shape, translation, rotation, and scale [3]. In order to calculate the user's accuracy of each trial of the tracing task, we decided to calculate the error score based only on the pixel-wise difference in general shape between the shape given and the shape that each participant produced.

We extracted multiple pixel-based values from each generated shape and used the following equation to calculate error scores for each task:

$$error = \sqrt{\left(\frac{CP}{DP} - 1\right)^2 + \left(\frac{DP}{TP} - 1\right)^2} \qquad (1)$$

Where:

- CP is the number of common pixels between the participant-generated shape and the original shape
- DP is the total number of the participant-generated pixels
- TP is the total number of pixels of the original shape.

The error value is represented by the length of line from points A to B as seen on the plot below:

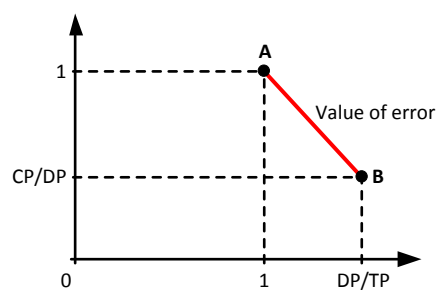


figure 1: Error measurement plot

Point A (1,1) is a perfect score with zero error where CP=DP=TP. Point B represents a user's score.

While performing such a tracing task it is theoretically possible to achieve maximum accuracy. This would mean that a user has traced over a shape perfectly and created the same amount of pixels in the exact same position as the original shape that was presented.

Experiment design

The experiment has a mixed design. The input methods have a within subjects design and were assigned in a randomized order for counterbalancing. The visual feedback has a between subjects design. 8 randomly selected participants performed the task with visual feedback and 8 without it. An HP Touchsmart TM2 Tablet PC with a 12.1 inch screen and a resolution of 1280*800, with stylus and finger input, as well as a Logitech basic optical mouse were used. The HP TM2 was used in tablet mode with the stylus and finger input, parallel to the desk, and in laptop mode with the mouse. Morae version 3 was used in order to acquire time data from the trials. Participants had to fill in a pre-test questionnaire and were offered a minute long introductory session for the stylus and touch input in MS Paint. Then, they were presented with the shape and instructed to "trace over the shape in one stroke, starting from the top right corner". Afterwards they were asked to fill in a post-test questionnaire regarding their preferences and opinions for the input devices that were tested. The following figures display the shape that was presented to the participants along with a high scoring and a low scoring participant attempt.

Results

First the shape error data were analyzed for deviations from normality by means of the Kolgomorov-Smirnov test. A liberal decision criterion of 0.1 was used in order not to use parametric test unduly. However, none of these tests showed any such deviations and the further analyses were performed using ANOVA using a decision criterion of 0.05. The ANOVA for shape error revealed neither significant differences between the input methods ($F(2,28)=0.876$, $p=0.427$), the visualization of feedback conditions ($F(1,14)=0.26$, $p= 0.61$), nor

any interaction between these factors ($F(2,28)=0.317$, $p=0.731$). The grand mean error value was 0.741.

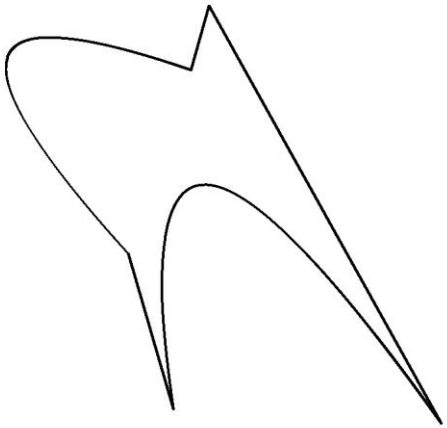


figure 2: Original shape

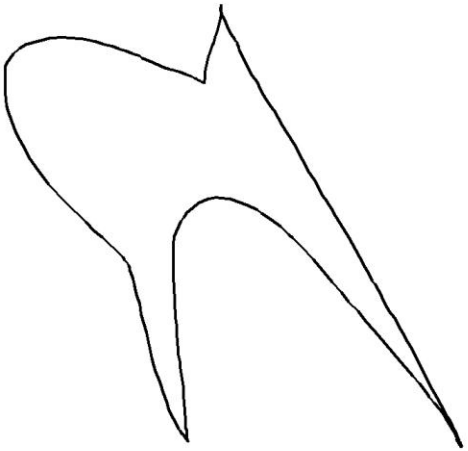


figure 3: High scoring user generated shape

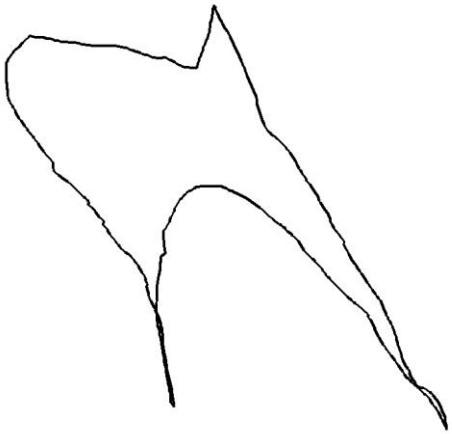


figure 4: Low scoring user generated shape

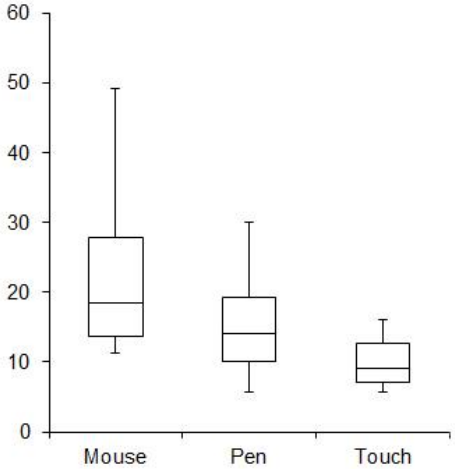


Figure 5: Box Plot of Times Measured

Then we turned to the timing data. The ANOVA was performed on the logarithm of the time in seconds because reaction type data are known to be non-normally distributed otherwise. There was no effect of visualization of feedback ($F(1,14)=0.0613$, $p=0.808$) nor any interaction between this factor and input method ($F(2,28)=0.275$, $p=0.761$). However, there was a clear difference between the input methods ($F(2,28)=49.535$, $p<0.0000001$). The mean times for each input device were mouse=23.00s, pen=15.10s and touch =9.81s.

Conclusion

Results show that all input methods have comparable error scores for the shape that was used. There are, however, large differences between time scores for each input method. Touch outperforms mouse by a factor of 2.3 and pen by a factor of 1.54.

Discussion

The precise line-tracing task might be representative of multiple tasks related from creative graphics design and free-hand drawing to complex linear selections of multiple graphical elements. Therefore the results of our study show that for at least moderately complex drawing tasks touch input is much more efficient than pen or mouse (what was also confirmed in tasks with no visual feedback) and might be the preferred input method in graphics design applications. This was supported by the qualitative post-test data that indicates that our users preferred touch input with pen input second and mouse input last. More of complex shapes need to be investigated, as it is expected that the results may vary from the one produced for this experiment. Finally posture must be considered as well, when using touch input, especially since there might be additional muscle strain due to the lack of proper support for the user's arms during the use of touch devices.

Acknowledgements

We thank The Swedish IT-User Centre (NITA) for partial financing of this study.

References

[1] ACCOT, J. AND ZHAI, S. 1997. Beyond Fitts' law: models for trajectory-based HCI tasks. Proceedings of the SIGCHI conference on Human factors in computing systems, ACM.

[2] ATTNEAVE, F. AND ARNOULT, M.D. 1956. The quantitative study of shape and pattern perception. Psychological Bulletin 53, 452-471, APA.

[3] COSTA, L.D.F. AND CESAR-JR, R.M. 2001. Shape analysis and classification: theory and practice. CRC.

[4] FORLINES, C., WIGDOR, D., SHEN, C. AND BALAKRISHNAN, R. 2007. Direct-touch vs. mouse input for tabletop displays. Proceedings of the SIGCHI conference on Human factors in computing systems, p.647-656, ACM.

[5] MACKENZIE, I.S. AND BUXTON, W. 1992. Extending Fitts' law to two-dimensional tasks. Proceedings of the SIGCHI conference on Human factors in computing systems, ACM.

[6] MACKENZIE, I.S., SELLEN, A. AND BUXTON, W.A.S. 1991. A comparison of input devices in element pointing and dragging tasks. Proceedings of the SIGCHI conference on Human factors in computing systems: Reaching through technology, ACM.

[7] SASANGOHAR, F., MACKENZIE, I.S. AND SCOTT, S.D. 2009. Evaluation of Mouse and Touch Input for a Tabletop Display Using Fitts' Reciprocal Tapping Task, Proceedings of the Human Factors and Ergonomics Society Annual Meeting, Volume 53, Number 12, p.839-843, Human Factors and Ergonomics Society.

Usability evaluation in exclusive domains: How to access domain knowledge.

Asbjørn Følstad
SINTEF
Pb 124, Blindern
0314 Oslo, Norway
asf@sintef.no

Abstract
The concept of domain exclusiveness is introduced to differentiate between domains with respect to ease of domain knowledge access. Whereas traditional usability evaluations are the method of choice for non-exclusive domains, exclusive domains may require evaluation methods that draw on the domain knowledge held by domain experts, such as users experienced in the domain. Group-based expert walkthrough and Cooperative usability testing are reviewed as examples of evaluation methods that facilitate access to such domain knowledge.

Keywords
Usability evaluation, domain exclusiveness, domain knowledge, group-based expert walkthrough, cooperative usability testing.

ACM Classification Keywords
H5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

1st European Workshop on HCI Design and Evaluation, Limassol, Cyprus, April 8, 2011.

Introduction

Usability evaluation has been established as an indispensible part of interactive systems development. Methods for usability testing and usability inspections are being used across an ever growing spread of domains, that is, types of contexts in which the interactive systems are to be implemented. The application of usability evaluation methods across domains has sensitized the HCI community to the need to match the evaluation methods with the domain in which they are used, as is seen from the theme of the current workshop [1].

Adequate application of usability evaluation methods depends on in-depth knowledge of the domain of the interactive system. In an evaluation, usability is examined relative to defined users and goals, as well as the social, environmental and organizational context [7]. Domain knowledge is needed to verify the relevance of scenarios and tasks for usability testing and walkthroughs, identify breaches in usability through inspections, and analyze data from usability testing. Without sufficient domain knowledge, the quality of the usability evaluation will suffer.

However, the availability of domain knowledge varies greatly between domains. For some domains, such as eCommerce and eGovernment, domain knowledge is abundantly available. For other domains, as for example medical care at hospital wards and medical emergency response, easily accessible domain knowledge is scarce.

The availability of domain knowledge within a given domain may be a critical for the adequacy of a given usability evaluation method; in particular for domains of which domain knowledge is exclusive, that is, not easily available to the outsider. To argue for this position, the concept of ***domain exclusiveness*** is proposed. This concept is then used to demarcate the domains which may not be adequately treated with traditional usability evaluation methods. Finally, two evaluation methods are reviewed that may enable us to overcome the challenges associated with traditional usability evaluation methods in exclusive domains. Please note that the presented argument is work in progress; it is the hope of the author that this paper can generate the discussion needed to develop it further.

Domain exclusiveness as a function of specialization and required training

Domain exclusiveness refers to the unavailability of domain knowledge to the outsiders of a specific domain and is a function of the domainûs (a) level of specialization and (b) level of required training. In the highly exclusive domains of medical care at hospital wards and medical emergency response domain knowledge depends on a particular professional and occupational status (high level of specialization), and several years of education (high level of training). Such domains may be called ***specialist domains***.

Domains that are highly specialized but require only limited training, as for example professional sales and parking enforcement, may be called ***limited training domains***.

Domains that require extensive training but are encountered in a wide spread of contexts, such as general office work and project leadership, may be called ***generalist domains***.

Finally, domains that are general and require little training, such as eCommerce and eGovernment customership, may be called ***popular domains***. The four domain categories are mapped out in figure 1.

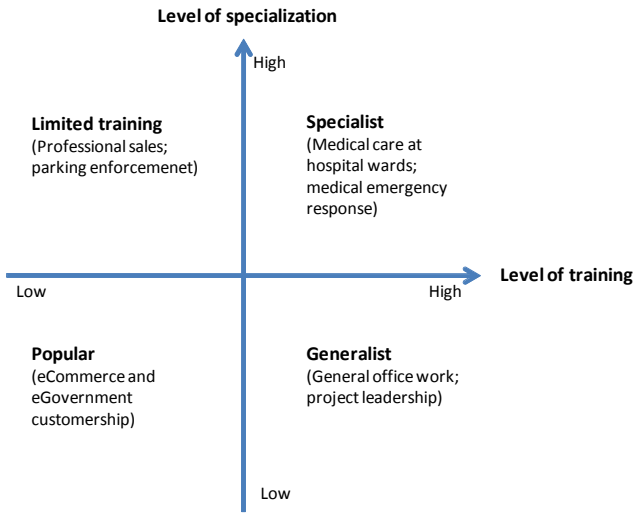


figure 1. Domain exclusiveness mapped as a function of the levels of specialization and training in the domain.

Usability evaluation in domains with low levels of specialization

Popular domains represent little challenge to usability evaluation methods. The usability experts conducting the evaluations are likely to have access to sufficient domain knowledge on basis of existing literature, including elaborate usability guidelines, as well as detailed market research. Also, as the potential user groups for these domains include any person with an internet access, the usability experts will most likely have hands-on experience as user of range of systems similar to that under evaluation. In consequence, traditional usability evaluation methods are adequate. Usability testing can be conducted in line with renowned textbooks [2;10], where the usability expert is the main interpreter of the evaluation results and the user is the object of observation. Usability inspections can be conducted with method such as heuristic evaluation [9] or cognitive walkthrough [12] where the usability expert is the main judge of usability on basis of general usability knowledge and knowledge of relevant user groups.

Similarly, generalist domains may be argued to represent no great challenge in usability evaluations, even though these domains require high levels of training and therefore are more exclusive than popular domains. Usability experts are likely to be skilled in generalist domains, as for example general office work and project leadership, through their education and work experience. Also, since general purpose systems for office work support have been an important category of interactive systems for more than 20 years, a fair amount of the existing HCI literature, implicitly or explicitly, target generalist domains[1]. In consequence, it may be held that usability evaluations in generalist domains also may be adequately conducted with traditional evaluation methods.

Usability evaluation in domains with high levels of specialization

In domains with high levels of specialization, domain knowledge is typically not easily accessible to usability

[1] The prevalence of office work support systems in the literature on usability evaluation is for example seen in the title of the ISO standard on usability [7].

experts. To acquire adequate domain knowledge, usability experts may have to engage in extensive context research or training, which may represent a prohibitive overhead due to resource limitations in the development project. Further, the elusiveness of specialized domain knowledge may make its elicitation and description fundamentally difficult [11].

In consequence, the use of traditional usability evaluation methods in domains with high levels of specialization is problematic. The validity of the evaluation result depends on access to adequate domain knowledge, and the domain knowledge is elusive. For limited training domains it may be held that this does not represent a fundamental problem as usability experts may be trained in the domain with moderate effort. For specialist domains the problem remains since providing usability experts with domain training is typically not an option.

Accessing domain knowledge: Usability evaluations with domain experts

A key resource to domain knowledge in exclusive domains is domain experts. Users experienced in an exclusive domain will typically hold domain expertise, as for example skilled workers [8]. Traditional usability evaluation methods do not include mechanisms to draw on the domain knowledge of domain experts, but the HCI literature do include usability evaluation methods adapted to draw on this resource. Two of these are Cooperative usability testing and the inspection method Group-based expert walkthrough.

Group-based expert walkthrough

Group-based expert walkthrough [4] is a usability inspection method particularly designed to allow domain experts to participate as evaluators, in spite of their lack of training in usability. The inspection is lead by a usability expert, taking the evaluators step by step through scenarios of use to identify breaches in usability. The inspection procedure is rather rigid to provide the necessary scaffolding for meaningful domain expert participation.

In an empirical study investigating the performance of domain experts as usability inspectors [3], groups of domain experts evaluated domain-specific solutions for (a) medical personnel at hospital wards, (b) parking wardens, and (c) politicians and political advisors. The three domains all hold fairly high levels of specialization. Parking enforcement may be seen as a limited training domain, political work and medical care at hospital wards may be seen as specialist domains.

The inspection method was group-based expert walkthrough. For each group of domain experts a similar inspection was conducted by a group of usability experts. The performance metric of the study was impact on subsequent development.

Three interesting findings were made: First, the overlap in findings between the evaluator groups was small; only 12 percent of the 229 identified usability issues were identified by both domain experts and usability experts. Second, most of the usability issues identified by the domain experts (59 %) were specific to the domain (as opposed to domain-independent), whereas the comparable proportion for the usability experts was only 15 percent. Third, the impact of the domain expertsô findings was substantial; 54 percent of the usability issues identified by domain experts had led to changes in the interactive systems three months after

the evaluation, as opposed to 27 percent for the usability experts [3].

Cooperative usability testing

Cooperative usability testing [5] is a usability testing method involving the participating users in interpretation and analysis, thereby providing access to the domain knowledge of the participants. The method is structured in interchanging interaction and interpretation phases, where the test leader and the participating user in the interpretation phase walk through the userôs task completion in the preceding interaction phase, discussing and interpreting their observations and experiences. Thus, the interpretation phases may be seen as cooperative augmentations of the traditional usability test.

In an empirical study of cooperative usability testing [6], medical emergency personnel and mobile sales personnel participated in evaluations of specialized interactive systems to support their respective work contexts. Both domains are high in specialization. Medical emergency response may be seen as a specialist domain whereas mobile professional sales may be seen as a limited training domain. As in the study on group-based expert walkthrough, the performance metric was impact in the subsequent development process. Three interesting findings were made: First, the interpretation phases generated a substantial proportion of usability issues not previously identified in the interaction phases. Second, the usability issues of the interpretation phases had the same impact on the subsequent development process as the usability issues of the interaction phases. Third, whereas the usability issues identified in the interaction phases typically were related to interaction design or static design (characteristics of visual layout or wording), the interpretation phases generated a wider spread of usability issues including needed information, needed functionality, and requirements for use and content [6].

Conclusion and future work

Exclusive domains represent a challenge to traditional usability evaluation methods, in particular domains with high levels of specialization. However, evaluation methods drawing on the domain knowledge of domain experts, such as experienced workers, may alleviate this challenge.

Empirical studies of group-based expert walkthrough and cooperative usability inspection indicate two potential benefits of drawing on domain expertsô knowledge: First, improved identification of certain categories of usability issue, as for example domain-specific issues. Second, improved impact on the subsequent development process; possibly caused by the perceived relevance of the usability issues identified by the domain experts.

Existing studies on usability evaluation methods drawing on the domain knowledge of domain experts only provide early evidence on the potential benefits of these methods. Even so, it seems reasonable to suggest the possibility that such usability methods, as for example group-based expert walkthrough and cooperative usability testing, may help us to overcome the challenges related to usability evaluation in exclusive domains.

Whereas traditional usability evaluation methods should be the methods of choice for non-exclusive domains,

usability experts may be well advised to consider alternative evaluation methods for exclusive domains in order to benefit from the domain knowledge of domain experts.

Acknowledgements

This paper was written with the support of the research projects R2D2 networks and RECORD, both supported by the VERDIKT programme of the Norwegian research council.

Citations

[1] Christou, G., Zaphiris, P., Law, E.L-C. ***1st European Workshop on HCI Design and Evaluation: The influence of domains.*** http://sites.google.com/site/ehcide/call-for-papers.

[2] Dumas, J.S., Redish, J.C. ***A Practical Guide to Usability Testing***. Intellect Books, Bristol, UK, 1999.

[3] Fßlstad, A. Work-domain experts as evaluators: usability inspection of domain-specific work support systems. ***International Journal of Human–Computer Interaction 22***, 3 (2007), 217ï245.

[4] Fßlstad, A. Group-based expert walkthrough. ***Proc. 3rd. COST294-MAUSE International Workshop*** (2007), 58ï60. http://141.115.28.2/cost294/upload/522.pdf.

[5] FrßkjÞr, E. and HornbÞk, K. Cooperative usability testing: complementing usability tests with user-supported interpretation sessions. ***Proc. CHI 2005***, ACM Press (2005), 1383ï1386.

[6] Fßlstad, A. and HornbÞk, K. Work-domain knowledge in usability evaluation: Experiences with Cooperative Usability Testing, ***The Journal of Systems and Software, 83*** (2010), 2019-2030.

[7] ISO, 9241-11. ***Ergonomic requirements for office work with visual display terminals (VDT)s—part 11 guidance on usability***, ISO/IEC 9241-11, 1998 (E).

[8] Kensing, F. and Munk-Madsen, A. PD: structure in the toolbox. ***Communications of the ACM 36***, 4 (1993), 78ï85.

[9] Nielsen, J. Finding usability problems through heuristic evaluation. ***Proc. CHI 1992***. ACM Press (1992), 373ï380.

[10] Rubin, J. and Chisnell, D. ***Handbook of Usability Testing: How to Plan, Design, and Conduct Effective Tests, second ed***. Wiley Publishing, Indianapolis, 2008.

[11] Suchman, L.A. Making work visible. ***Communications of the ACM 38***, 9 (1995), 56ï64.

[12] Wharton, C., Rieman, J., Lewis, C., and Polson, P. The Cognitive Walkthrough: A Practitioners Guide. In Nielsen, J., Mack., R.L. (Eds.) ***Usability Inspection Methods***, p. 105-140. John Wiley, 1994.

What do you mean? What do I mean? A novel application of repertory grid at the user interface.

Andrew Saxon
Birmingham City University,
Institute of Art and Design,
Corporation Street,
Birmingham, UK B4 7DX
andy.saxon@bcu.ac.uk

Shane Walker
Birmingham City University,
Institute of Art and Design,
Corporation Street,
Birmingham, UK B4 7DX
shane.walker@bcu.ac.uk

David Prytherch
Birmingham City University,
Institute of Art and Design,
Corporation Street,
Birmingham, UK B4 7DX
david.prytherch@bcu.ac.uk

Abstract
This paper focuses on the adoption and adaptation of methodologies drawn from research in psychology for the evaluation of user response as a manifestation of the mental processes of perception and cognition. We present a robust alternative conceptualization of an evaluative methodology, which allows the surfacing of views, feelings and opinions of individual users producing a richer, more informative texture for user centered evaluation of software or product. We will present our methodology so that readers can firstly, review the method as a theoretical exercise and secondly, creatively adapt and apply the method to their own research or practical context.

Keywords
User-centered design, Usability, Motivation Systems Theory, Repertory Grid, Human Computer Interaction, User Experience, User satisfaction, Usability evaluation, User interface, Evaluation methodology, Motivation, Emotion, HCI, Methodology adaptation.

ACM Classification Keywords
H. INFORMATION SYSTEMS
H5. INFORMATION INTERFACES & PRESENTATIONS
H5.2 User Interfaces
Subject Descriptor: Evaluation/Methodology

General Terms
User-centered design, Usability, Motivation Systems Theory, Repertory Grid, Human Computer Interaction, User Experience, User satisfaction, Usability evaluation, User interface, Evaluation methodology, Motivation, Emotion, HCI, Methodology adaptation.

Introduction

Repertory Grid has previously been used to permit a qualitative comparison between different pages of a website or between a new website, the previous website and its competitors. The technique has also been extensively used in comparative product evaluation. The method detailed here permits a direct comparison between a websiteôs design objectives and its user experience. A similar adaptation could potentially be applied to the testing of any designed item, where a direct understanding of the success of a design approach is sought by comparing the stated design aspirations with the userôs actual experience.

We suggest that looking to other domains for potentially adaptable methodologies to provide quantitative user data may often provide a fruitful approach. We like the repertory grid because it is extremely adaptable and *not* tied to one domain, so specific domain influences are largely irrelevant.

The philosophy underlying our methodology is fundamentally concerned with eliminating interviewer bias to the greatest extent possible whilst focusing on the userôs own perceptions in his/her own terms. This is not done reductively by iterating draft questionnaire items, but constructively by simply asking users what is important to them. Our approach, illuminating the general by close examination of the specific, (paraphrasing Denscombe) [1] lends weight to theories of participatory design and amplifies the userôs voice in the software design cycle in ways that are authentic, genuine and transparent.

The ubiquitous nature of information technology today means that the computer is no longer just a tool for those who are compelled to use it, or have to learn to use it, as was the case in the 1980s. Interfaces, in particular on the Internet, must appeal to a broad base of users with varying levels of skill and ability, and should work first time to ensure the user is not óput offô the experience. Aesthetic considerations (often referred to as beauty) may also be considered significant in this context, (Hartmann, Sutcliffe & De Angeli) [2] but have only attracted significant interest from researchers in the last few years. (Hassenzahl) [3]. Norman [4] suggests that attractive things generate positive moods that impact directly on improving thought processes and therefore task performance. Hassenzahl [3] suggests the process is more complex, and that a clear link to task performance is not proven. However, there is agreement that aesthetics satisfies basic human needs and therefore can impact on motivation. (Tractinsky & Hassenzahl) [5].

In order to assess how far visual design techniques applied to the user interface can harmonize with psychological needs for optimal performance on specific tasks and attainment of goals, we must base questions

on a fundamental understanding of key influencing variables of the interaction process, together with a clear knowledge of their relative importance to the individual user. In perceptual terms, interactive computer systems are not just representations of knowledge, but interactive *experiences* that should seek to fully exploit the userôs senses and emotions, developing new ways to deliver effective communication.

Adherents of the quantitative approach, upon recognizing that a userôs subjective response to the software under evaluation can be an important aspect of evaluation, will often add a criterion that aims to deal with the subjective dimension of how much the user *liked* using the software. This is illustrated in Nielsenôs [6] model of the attributes of system acceptability, which explores and charts the notion of usability through consideration of such matters as: learnability, efficiency, memorability, user error-rate, and satisfaction. In so doing, they are acknowledging that no matter how well it performs in the other categories, if the software and its interface are not liked by the user, it may not actually be used much, and may even be discarded. On the World Wide Web, it is accepted that the competition is just one click away, so the importance of gaining a clear understanding of usersô experience with web software cannot be overstated.

Our method is derived from Kellyôs [7] Repertory Grid Technique. This method uses a highly qualitative approach to the surfacing of a userôs experience using his/her own frame of reference. The technique is highly amenable to customization by the experimenter, to suit the particular needs of his/her investigation, and details on customization carried out by the authors are described. This paper is based on earlier work detailed in Saxon, Walker & Prytherch [8]

Repertory Grid Technique

The Repertory Grid Technique was developed from Kellyôs [7] Personal Construct Theory. This theory had, at its starting point the belief that people are scientists. By this, it is meant that each of us has constructed our own point of view about the world (a theory), our own opinions on what will happen in a given situation (a hypothesis) and that in our day-to-day behaviour, we continually test our views against our reality (we conduct experiments) and, based upon our day-to-day experiences (the experimental data,) we modify our world-view (our theory), forming new opinions on how the world works (a revised hypothesis).

According to Kelly [7], as these constructs represent a set of individual qualitative judgments about the world, it is preferable to think of them as being scalar, having two contrasting poles. Kelly [7] observed that our judgments are usually bipolar, rather than unipolar. The concept of *good* can only exist as the opposite of *bad;* the concept of *wet* can only exist as the opposite to *dry*. Therefore, when an individual offers a judgment, the question *'as opposed to what?'* needs always to be asked in order to further clarify the individualôs meaning and the context. For example, the term *good* may mean *skilful* as opposed to *novice* when discussing sporting ability, or it may mean *virtuous* as opposed to *evil* when discussing the moral character of another individual. The exact meaning of a word in a given context cannot therefore be assumed. Even where the opposite pole is thought obvious by the

experimenter, it should still be sought in order to eliminate observer bias. Repertory Grid Technique sets out to measure subjective aspects of a participantôs experience using a highly qualitative approach whose methods are amenable to customization by the experimenter.

By far the most common method of construct elicitation is Kellyôs [7] triadic method, where three elements are shown to the interviewee, who is then asked to specify some way in which two of these elements are alike and thereby different from the third. The reply, representing the likeness pole of the bipolar pair, is then recorded. Then, the interviewee is asked in what way the third element differs from the other two. This reply, representing the contrast pole in the bipolar pair is also recorded. The process is then repeated using as many triads of elements as the interviewer thinks appropriate.

Another less common, yet equally effective method according to Ryle and Lunghi [9] is the dyadic method, where only two elements are shown to the interviewee, so that the two construct poles can be elicited more directly. The dyadic method is felt to be simpler in operation than the triadic method, without sacrificing either validity or reliability. Fransella and Bannister [10] also state that:

ñKelly based his triad method for eliciting constructs on his theory as to how constructs are first formed. But since one is eliciting constructs already established in the personôs repertoire there is no reason why three elements need be usedéThere is nothing sacrosanct about the triad. It is equally reasonable to use two elements for elicitationéò

Epting et.al. [11] evaluated the two different formal methods proposed by Kelly [7] for identifying the contrast pole during the Repertory Grid interview. These are called the *opposite* method and the *difference* method, the latter, over time, becoming the more popular of the two. Using this method, after the likeness pole has been identified, as stated above, the interviewee is asked to specify some way in which two of the elements in the triad are alike and thereby different from the third.

Using the opposite method, after the likeness pole has been identified, as stated above, the interviewee is asked to specify what he/she considers to be the opposite of the characteristic given for the likeness pole. Epting et al. [11] found that the opposite method produced a greater number of bipolar constructs than the difference method when using the Repertory Grid procedure. Curtis et al. [12] also point out that the opposite method produces fewer óbent constructsô than the difference method. However they go on to note that the opposite method also has shortcomings; namely that the opposite method creates contrast poles that can be more extreme than the difference method, and that the opposite method can produce constructs that do not apply readily to their elements.

For reasons of operational simplicity without loss of reliability or validity, and the need in this study for several bipolar constructs to be elicited during the interviews for each element, we used dyadic elicitation of constructs, and the opposite method of contrast pole elicitation. Laddering (Fransella and Bannister [10]) was used as needed to more finely tune the contrast poles to the target domain.

Niemeyer et al. [13] offer a third method of construct elicitation; the *contrast* method; which may provide fruitful refinement.

Applications of the Repertory Grid Technique to software evaluation

When considering Repertory Grid Technique as the vehicle for gathering qualitative assessments of the user experience, as opposed to quantitative assessments of software usability, exemplars can be seen in the literature that demonstrate a precedent, and which may be shown to clarify the appropriateness of the theory for use in this context.

Baber [14] provides an example of the use of Repertory Grid Theory in comparative product evaluation, suggesting that further use could be made of the technique in the early stages of product design, when compiling a list of desirable design characteristics. Baber (p. 164) [14] also envisages its use: ñéas a means of defining usersô conceptions of usability, perhaps through a comparison of a range of products which users are encouraged to consider in terms of functionality, ease of use, etc.ò

Boy [15] presents the Group Elicitation Method (GEM), which seeks to facilitate the mutual understanding of terms and ideas among experts who have been gathered together in order to provide solutions to design or usability issues. Boy (p. 33) [15] states: ñAlthough we have used GEM to design new systems, we have observed that it would be useful for evaluating existing systems and suggesting alternatives.ò

Both Baber [14] and Boy [15] appear to have recognised a symmetry between initial product design and user testing that deploys an individualôs personal construct system as a kind of yardstick against which to evaluate a product, either at the beginning or at the end of the design process.

Verlinden and Coenders (p. 143) [16] describing their approach to the use of Repertory Grid for the evaluation of a website state: ñMost usability techniques are quantitative and measure the performance, retrieval times, success times/failure ratesé Although [a quantitative method] might provide some information on experiences with respect to websites, it exposes a number of shortcomings.ò

Repertory Grid is offered as it ñéfacilitates an objective approach of capturing subjective aspects of web usability.ò (Verlinden & Coenders, p. 144) [16]

The approach they use is comparative, but in this case the comparison is made between different pages of the website under evaluation, one against the other.

Hassenzahl and Trautmann (p. 167) [17] set out to evaluate the ñholistic overall impression [or] ôcharacterô of a web site.ò The method used was to employ Repertory Grid Technique to compare a new website design created for a German online bank, with the old design, along with the other six prominent online banking sites available in Germany. The experiment produced useful data, enabling a comparative evaluation of the eight websites to be conducted, facilitating a view on the success of the new site design in terms of its usersô experience. (Hassenzahl & Trautmann) [17]

Hassenzahl, Beu, & Burmester [18] discuss the evolved needs of software product evaluation during the ó00ôs, stating that the industry-wide focus on usability engineering over the last 30 years must now be extended to acknowledge contemporary usersô needs. Hassenzahl et al. [18] perform a selective review of existing measurement methods and tools. Repertory Grid is cited as a valuable tool in this regard.

The method they describe presents interviewees with a randomly drawn triad from the software product set occupying the design space of interest to the evaluators. The interviewee is then asked the classic question: in what way are two of these three products similar to each other, but different from the third? In this way, the personal construct systems of the interviewees are used as a means of communicating responses to the products under evaluation.

Hassenzahl, Wessler, & Hamborg [19], seeking ways to better capture design requirements in the early stages of software design, introduce the Structured Hierarchical Interview for Requirement Analysis. (SHIRA).

As well as being useful in the early stages of designing, SHIRA has (as yet unexplored) potential for use as an evaluative tool. Hassenzahl, Beu, & Burmester , (p. 7) [18] note that: ñSHIRA is especially suited to gather information at early stages of the design process for interactive systems. However, it might also be possible to evaluate software at a later stage regarding how it fits the userôs expectations.ò Further, there appears to be a good ófitô with the use of Personal Construct Theory in utilising an interviweeôs own terms of reference in any evaluation.

Tan and Tung [20] set out to investigate website designersô notions of what makes an óeffectiveô website. They used Repertory Grid Technique to ascertain a clear set of design principles, which were derived from constructs elicited from a panel of 20 designers, who evaluated six high-traffic websites. Tan and Tung (p. 65) [20] note that this approach helped to avoid the use of predetermined methods that may limit the scope of the study such as: ñ pre-structured questionnaires to collect data...[and] é scripted actions that govern the way participants walkthrough a website.ò

In a similar way to Boy [15] then, Tan and Tung [20] used Repertory Grid technique to achieve a consensus view drawn from expert consultations.

Fallman and Waterworth [21] sought to couple Repertory Grid Technique with empirical data analysis methods in the evaluation of user experience and affective responses to artefacts. They used seven digital artefacts, including a digital camera, a PDA and a mobile phone. These artefacts were used as elements, and constructs were elicited from 18 participants. Although the focus of the work appears to have been on the use of empirical methods, the choice of such a wide variety of elements may have caused the elicitation process to be rather diffuse, as evidenced by the example constructs listed. This may have been exacerbated by the relatively short artefact familiarisation time provided to participants by the experimenters. However they conclude that as: ñthe data are meaningful to the participant, not to the experimenter é the data that is found within a participantôs own repertory grid has not already been influenced by ð i.e. interpreted in the light of ð the

researcherôs pet theory.ò (Fallman & Waterworth, p. 4) [21].

The preceding exemplars demonstrate that Repertory Grid Technique may be used successfully in the evaluation of designed artefacts in general, and websites and software products in particular.

Applying the Repertory Grid Technique

As has been shown above, Repertory Grid Technique has strong possibilities for application where a comparative view is needed of how one software product fares against its competitors in the eyes of its users. Where a software product must be evaluated in a stand-alone manner however, a novel approach is needed. Optimally, the required approach retains the many advantages offered by Repertory Grid Technique, whilst facilitating the stand-alone evaluation of the software user experience.

Our approach is described below: In brief, we use an initial card-sorting activity to identify precise themes for enquiry, after which an individualised questionnaire is created that can later be used to evaluate the test software. This questionnaire is made up of themes that were identified by the participant as important to him/her self. The specific questions used in the questionnaire were conceived and worded by the participant, thus ensuring that the terms of reference used fell within his/her range of convenience.

The specific method is now explained, step-by-step.

Interviewees are sought from members of the target audience for the software under evaluation. The interviews are conducted singly. Interviewees are screened by questionnaire as to their IT skills.

The interviewee is welcomed to the experimental venue. We use an office, provided with a round table, chairs, a desk and a computer connected to the Internet, running the software under evaluation. At the table, the interviewee is introduced to the aim and purpose of the experiment, and any initial questions are answered.

A list of the softwareôs design objectives is offered to the interviewee. These are written on index cards, one objective per card. The interviewee is asked to sort the cards, ranking them in order of importance according to his/her own opinion, by laying them out in a line on the table with ómost importantô at one end and óleast importantô at the other. Any questions asked by the interviewee are answered.

The final card ranking arrived at by the interviewee is noted and recorded for future reference.

The five cards ranked as most important are selected by the interviewer and the design objectives written on them are noted. These objectives correspond to the *elements* in a Repertory Grid interview.

For each design objective (element), the interviewee is then asked to think of a characteristic that the software under evaluation would need, in order to fulfil that design objective in such a way that their experience of using the software would be positive. The resulting statement is written down on the Personal Statements Sheet form. Then the interviewee is asked to think of the opposite to the already stated characteristic, with

the prompt ñéas opposed to what?ò This statement is also written down on the Personal Statements Sheet form. The two characteristics stated: one desirable, the other its opposite in the interviewee ôs own terms represent his/her personal construct relating to that design objective. The bipolar construct is written down as anchors, one at each end of a seven-point ratings scale.

This step is repeated, to produce a total of three bipolar constructs for each of the five elements, making 15 in total. The interviewer must take great care to avoid influencing the interviewee ôs choice of words at all times. Careful use of repetition and reflecting during the interview will help the interviewee to state clearly what they mean, where difficulty is experienced.

All 15 constructs are written out to create a personal Ratings Form, in effect, creating a ólankô user experience questionnaire that is completely personalised to the interviewee, being based on elements, that he/she has chosen from a large pool of design objectives, using dyadic personal constructs elicited by the opposite method from these elements.

The interviewer and the interviewee then move to the desk with the computer, where the interviewer briefly introduces the software under evaluation. Any questions asked are answered.

Each interviewee is invited to use the software under evaluation to achieve the same real-world task. During the process, any questions asked are answered.

When the interviewee states that he/she had finished, he/she is invited to complete the personal Ratings Form, scoring the software under evaluation against his/her previously elicited constructs.

The interviewee is asked to share any other comments or views that have not been captured by the personal Ratings Form, and these are noted.

The interviewee is thanked, and the experiment is concluded.

The data gathered are then ready for analysis.

A Repertory Grid interview session as described above takes approximately two hours from beginning to end. We used 10 participants, in separate sessions. All interviews must be conducted in the same manner, preferably using the same venue.

Three methods are suggested for analysis of the data gathered, and these are described below:

Ranking of Design Objectives by Importance

As each participant had ranked all of the design objective cards in order of importance, the information may be tabulated in order to indicate an overall importance ranking as expressed by the whole group. This ranking may then be examined. On several occasions, this has produced surprising results where design objectives ranked ómost importantô by the system designers were low-ranked by interviewees. The converse is also seen, where design objectives ranked óleast importantô by the system designers were high-ranked by interviewees.

Tabulation of each participant's personal ratings with mean of overall scores for each participant

This method tabulates scores taken from the Personal Ratings Sheets drawn up for each participant. The five design objectives selected by the participant as most important (the Repertory Grid elements) and the scores for each of the three sets of paired, bipolar statements associated with each design objective (the Repertory Grid constructs) are presented, summed and a mean score is calculated. The data are thus readily available for an óeyeballô test. Robson [22], Clegg [23], prior to further analysis.

Personal Ratings Sheet Discussion

Participantsô Personal Ratings Sheets are reviewed and discussed. This discussion presents the individual subjective judgements made by each participant about their user experience of the software under review. Standard content analysis methods are useful here. Due to the Repertory Grid methodology used in the experiment, the judgements are stated using participantsô own range of convenience and in their own terms of reference. These dimensions of judgement were hitherto hidden from view.

Summary of Repertory Grid Technique applied to software evaluation.

After all 15 personal constructs discussed above have been elicited, each participant is invited to use the software under review to perform the same real-world task. Once the participant has finished this task, he/she is invited to complete a Personal Ratings Sheet. This takes the form of a questionnaire, which uses the personal constructs elicited earlier in the session as questionnaire items. Each questionnaire item uses the personal construct ólikenessô and ócontrastô poles as anchors, separated by a seven-point ratings scale (1-7) with 7 representing greatest agreement with the ólikenessô pole and 1 representing greatest agreement with the ócontrastô pole.

The questionnaire items have therefore been elicited from the participant before using the software to perform any tasks. Participants must have no prior experience of using the software beforehand. Whilst participants are using the software, the experimenter transfers the personal constructs to the Personal Ratings Sheet, creating a questionnaire óblankô. The questionnaire is administered after the user task is finished.

Summary and conclusions

Valuable data on the user experience is available if novel methods of investigation are applied. Many traditional HCI evaluation models have not been sufficiently sensitive to elicit data on user behaviour, which has significant implications for software design. However, the need to understand more about the covert judgments made by users is now critical for the success of software applications.

An important outcome has been the development of a framework for surfacing these affective judgments in a structured manner. The technique described here could be extended and applied to the design development of software capable of delivering a highly positive user experience.

0

Data collected using this method are well structured, and easily accessible to later content analysis. The initial óeyeball testô of the data immediately highlights areas of interest. The data are, as can be expected, very rich and highly value laden, but the values seen are those of the respective interviewees whose covert judgments regarding the software under evaluation have been surfaced and recorded. Overall, the protocol is simple to deploy in the experimental setting.

Future Plans

Although this case study describes the evaluation of a web application, we believe that this technique can be extended to facilitate new product evaluation.

Work is now in progress on the development of a Qualitative User Experience Satisfaction Tool (QUEST) which will be deeply rooted in our stated philosophy of ñask the userò and will incorporate key aspects of the approach described here.

Acknowledgements

The authors wish to thank IGI Global for permission to extract from our previous publication: Saxon, A. Walker, S. & Prytherch, D. (2009) 'Whose questionnaire is it, anyway?ô. In T. Spiliotopoulos, T. Papadopoulou, P. Martakos, D. & Kouroupetroglou, G. (eds.) Integrating Usability Engineering for Designing the Web Experience: Methodologies and Principles, IGI Global, Pennsylvania.

Citations

[1] Denscombe, M. (2007). *The Good Research Guide: For Small-scale Social Research Projects*: McGraw-Hill International.

[2] Hartmann, J., Sutcliffe, A., & De Angeli, A. (2007). *'Investigating attractiveness in web user interfaces*. Paper presented at the SIGCHI Conference on Human Factors in Computing Systems, San Jose, California.

[3] Hassenzahl, M. (2008). Aesthetics in interactive products: Correlates and consequences of beauty. *In:* Schifferstein, H. N. & Hekkert, P. (eds.) *Product Experience* (pp. 287-302). Elsevier.

[4] Norman, D. A. (2004). *Emotional design: Why we love (or hate) everyday things,* New York, Basic Books.

[5] Tractinsky, N. & Hassenzahl, M. (2005). Arguing for Aesthetics in Human-Computer Interaction. *i-com,* 4**,** 66-68.

[6] Nielsen, J. (1993). *Usability Engineering*. San Francisco: Morgan Kaufmann (Academic Press).

[7] Kelly, G. (1955). *The psychology of personal constructs* (Vol. 1 & 2). New York, NY: Norton.

[8] Saxon, A. Walker, S. & Prytherch, D. (2009) 'Whose questionnaire is it, anyway?ô. In T. Spiliotopoulos, T. Papadopoulou, P. Martakos, D. & Kouroupetroglou, G. (eds.) Integrating Usability Engineering for Designing the Web Experience: Methodologies and Principles, IGI Global, Pennsylvania.

[9] Ryle, A., & Lunghi, M. (1970). The dyad grid: a modification of repertory grid technique. *British Journal of Psychology*, *117*.

[10] Fransella, F., & Bannister, D. (1977). *A manual for repertory grid technique*. London: Academic Press.

[11] Epting, F., Suchman, D., & Nickeson, C. (1971). An evaluation of elicitation procedures for personal constructs. *British Journal of Psychology*, *62*(4), 513-517.

[12] Curtis, A., Wells, T., Lowry, P. B., & Higbee, T. (2008). *An Overview and Tutorial of the Repertory Grid Technique in Information Systems Research.* Communications of the Association for Information Systems (CAIS), vol. 23 (3), pp. 37-62.

1

[13] Neimeyer, G. J., Bowman, J. Z., Saferstein, J. (2005) *The Effects of Elicitation Techniques on Repertory Grid Outcomes: Difference, Opposite, and Contrast Methods.* Journal of Constructivist Psychology, vol. 18 (3), pp. 237-252.

[14] Baber, C. (1996). Repertory grid theory and its application to product evaluation. In Jordan, P., Thomas, B., Weerdmeester, B. & Mclelland, I. (Eds.), *Usability evaluation in industry* (pp 157-165). London: Taylor and Francis.

[15] Boy, G. (1997). The group elicitation method for participatory design and usability testing. *Interactions of the ACM*, *4*(2).

[16] Verlinden, J., & Coenders, M. (2000). *Qualitative usability measurement of websites by employing the repertory grid technique*. Paper presented at the CHI 2000: extended abstracts, The Hague, The Netherlands.

[17] Hassenzahl, M., & Trautmann, T. (2001). *Analysis of web sites with the repertory grid technique*. Paper presented at the CHI 2001: extended abstracts: interactive poster sessions, New York, N.Y.

[18] Hassenzahl, M., Beu, A., & Burmester, M. (2001a). Engineering Joy. *IEEE Software*, *18*(1).

[19] Hassenzahl, M., Wessler, R., & Hamborg, K. (2001b). *Exploring and understanding product qualities that users desire*. Paper presented at the Joint AFIHM-BCS conference on Human-Computer Interaction IHM-HCI'2001, Toulouse, France.

[20] Tan, F. B., & Tung, L. L. (2003). *Exploring website evaluation criteria using the repertory grid technique: A web designers' perspective.* Paper presented at the Second annual workshop on HCI research in MIS, Seattle: Washington.

[21] Fallman, D., & Waterworth, J. A. (2005). *Dealing with User Experience and Affective Evaluation in HCI Design: A Repertory Grid Approach*. Paper presented at the CHI Workshop on Evaluating Affective Interfaces: Innovative Approaches.

[22] Robson, C. (1993). *Real World Research*. Oxford: Blackwell.

[23] Clegg, F. (1990). *Simple Statistics*. Cambridge: Cambridge University Press.

An Integrated Approach Towards the Construction of an HCI Methodological Framework

Tasos Spiliotopoulos
Department of Mathematics & Engineering
University of Madeira
9000-390 Funchal, Portugal
tasos@m-iti.org

Ian Oakley
Madeira Interactive Technologies Institute
University of Madeira
9000-390 Funchal, Portugal
ian@uma.pt

Abstract
We present a methodological framework aiming at the support of HCI practitioners and researchers in selecting and applying the most appropriate combination of HCI methods for particular problems. We highlight the need for a clear and effective overview of methods and provide further discussion on possible extensions that can support recent trends and needs, such as the focus on specific application domains.

Keywords
User experience, usability, methods, framework, HCI, domains.

ACM Classification Keywords
H5.2. Information interfaces and presentation: User Interfaces - evaluation/methodology.

General Terms
Human factors.

Introduction and background
The use of human-centered design models and methodologies facilitates software development processes. However a plethora of approaches exists and it can be challenging for developers to appropriately match tools to problems. One way to address this issue is via methodological frameworks,

which facilitate the development process by aiding selection of the most appropriate design methods according to the characteristics of a particular project.

The contribution of this paper is the presentation of a novel framework for systematically categorizing and evaluating HCI methods. It is anticipated that this framework will ease the process of selecting appropriate HCI methods for particular design and evaluation tasks. An overview of the framework is provided, its structure and qualities are discussed and directions for its future development are proposed. The discussion highlights areas of improvement for similar efforts, including providing effective overview of methods, and extensions that support the study of recent trends and needs in HCI, such as the shift of emphasis to user experience (UX) and the focus on specific application domains.

Overview of methodological approaches

An apparent way of categorizing HCI methods can be based on the development stage that the method is applied to, be it analysis, design, implementation, formative evaluation, or summative evaluation. However, such a categorization does not provide any direct appreciation of the different kinds of results and insights provided by a method. Nor does it highlight the resources required, or the fact that some methods can be effectively used in more than one development phase. Fitzpatrick and Dix proposed an alternative schema for categorizing HCI methods according to their strategic application [3]. In this approach, four strategies are proposed, based on the resources at human and system level (i.e. real or representative users, and real or representative system) thereby creating a 2 by 2 matrix with top level categories of real world, virtual engineering, soft modeling and hard review. However in this framework, methods are then classified at a second level based on their type and the way they are used, resulting ultimately in a usage analysis table. The result of this process is that the top-level categorization does not reflect the underlying goal of classifying the methods and, as such, does not offer useful insights. A third approach involves categorizing methods by the type of participants featuring in the UX evaluation [6]. This leads to a top-level breakdown into categories of lab tests, field studies, online surveys, and expert evaluations without actual users. Another recent approach classifies usability evaluation methods into data gathering & modeling methods, user interactions evaluation methods, and collaborative methods [4].

An HCI methodological framework

The first step in the development of the methodological framework presented in this paper was to gather a comprehensive corpus of HCI methods. Overall, 41 HCI methods were sourced from the literature and studied, analyzed and systematically described (for the complete list, see [8]). The set of methods included both traditional usability methods and those that take into account experiential aspects of a system.

In the framework described in this paper, a top-level categorization of the methods is achieved according to the way that they are used, resulting in four categories: *inquiry*, *prototyping*, *inspection* and *testing*, with 14, 8, 12 and 7 methods in each category, respectively.

Inquiry methods refer to the study of the use of an existing system or the characteristics of a system under construction by querying users and other stakeholders. Methods of this type typically require a considerable number of participants and, since they are based on

participants' opinions and feelings about a system, provide subjective results. Examples include questionnaires, interviews and focus groups. Prototyping methods involve the construction of prototypes, which enable the study of the characteristics of a system at an early stage in the development process. Prototypes can be further classified according to their fidelity to the final system, level of detail, and scope, into high or low fidelity, and horizontal or vertical prototypes. Examples of prototyping methods include paper prototyping and storyboards. Inspection methods involve the close examination of a system by one or more HCI professionals. Typical inspections can be based on heuristics or guidelines, and can be driven by scenarios (i.e. walkthroughs). In testing methods, experiments on the system or a prototype are conducted with the participation of users. Typical examples include think-aloud protocol, wizard-of-Oz, and retrospective testing.

This categorization was selected as it highlights both the usage of a method and the type of resources that are required. In general, methods in different groups were found to exhibit high complementarity and low overlap in their results, focus and required resources. For example, inquiry methods tend to provide subjective results, as opposed to testing methods, whereas inspection methods do not require the direct participation of users, as inquiry and testing methods do.

The key to the methodological framework is the comparative evaluation of the HCI methods, which is based on a set of measurable criteria. This comparative evaluation, in combination with a good understanding of the workings of each method from their description and analysis, and a good apprehension of the needs of the system under development, can facilitate the selection of the most suitable combination of methods for each project.

The criteria for the comparative evaluation were selected so that they also enhance the general overview of the available methods. The HCI methods and evaluation parameters are tabulated in order to enable quick and effective overview and comparison. An excerpt of this table, comprising only 4 of the 41 methods studied is depicted in Table 1. The evaluation parameters are explained below.

- Type of method: This refers to the classification of methods as inquiry, prototyping, inspection or testing.
- Life-cycle stage: One or more of the following: requirements analysis, design, implementation, formative evaluation, and summative evaluation.
- Type of results obtained (quantitative - qualitative). This is a particularly important parameter, since methods that provide different types of results usually exhibit high complementarity and low overlap, thereby leading to a more efficient development process. Quantitative results are easily analyzed statistically, presented and interpreted in reports, and can be used for comparison of products or ideas in a straightforward way. On the other hand, qualitative results are not easily documented, but can provide important insights that can be easily missed in a quantitative approach.
- Bias (subjective - objective results). The results derived from the application of a method may be influenced to a significant extent by a personal predilection in part of a user or a usability expert.

Table 1. Excerpt from the HCI Methodological Framework depicting 4 of the 41 methods studied. The tabular presentation of the methods enables a quick and effective overview and comparison. The brief, descriptive analysis of the main strengths and weaknesses provides further insights aiding the selection of the most appropriate combination of methods.

Name	Focus groups	Contextual interviews	Heuristic walkthrough	Automatic logging of use
Type	inquiry	inquiry	inspection	testing
Dev. phase	req. analysis, design, formative evaluation	req. analysis	design, formative and summative evaluation	formative and summative evaluation
Type of results	qualitative	qualitative	qualitative, quantitative	quantitative
Bias	subjective	subjective	objective	objective
Cost	low (recruitment of participants)	medium (recruitment of participants during work time, trip to the workplace)	low	low (recording and logging equipment, data analysis)
HCI experts	yes	yes	yes	no
No. of users	6-9	5-10	-	- (no users recruited specifically)
Level of detail	high, low	high, low	high, low	low
Immediacy	yes	yes	yes	no
Location	lab, work place	work place	lab	lab
Intrusiveness	yes	yes	no	no
Strengths	The users' preferences and ideas come from spontaneous reactions. Group dynamics come into effect. Easily repeatable. Can focus on specific issues.	Takes into account the context of use of a system. Can focus on specific issues and aspects of the system in detail. Most effective for exploring an application domain.	Inexpensive, flexible, structured, quick and repeatable evaluation method. Can be applied on low fidelity prototypes. Easily documented results. Can focus on specific parts of a system.	Demonstrates how a system is really used. Allows data collection from a large number of real users, in an automatic and systematic way. Allows for a longitudinal approach to studying users' behavior. Provides easily documented results.
Weaknesses	Results are subjective and not easily documented or analyzed. The presence of a group moderator is imperative to keep the group on track and make sure that participants do not influence each other.	Results are subjective and not easily documented or analyzed. Employee participants may be biased.	Heuristics limit significantly the scope of the evaluation. A walkthrough covers only a small part of the system. Inherent bias due to the selection of tasks/scenarios to be evaluated.	Provides answers to *how* people use a system, but not *why*. Does not take into account experiential aspects of the use of the system. The results require statistical analysis. Caution is needed as in not to breach users' privacy.

This is something that should be taken into account in the interpretation of the results and the selection of methods.

- Cost. Includes items such as required equipment (e.g. a usability lab), prototype development costs, user recruitment costs and the cost of usability experts (if required).
- Need for recruiting HCI experts. Boolean parameter referring to whether the method requires HCI experts for correct execution.

- Number of users. A typical number of users that are needed to participate in the activities described in the method.
- Level of detail of the results (high - low). The results derived from the application of a method, may be low-level (e.g. relating to icon size or positioning) or high-level (e.g. general impressions of a system).
- Immediacy of results. Whether the method yields results immediately, or if further analysis (e.g. statistics) is required.
- Location. The site where a method's activities take place (e.g. lab, field, workplace).
- Intrusiveness. A user's behavior may be influenced by the presence of an observer, interviewer or recording equipment. This criterion highlights the extent to which the method is intrusive.
- Strengths. The strengths and main advantages of each method are described briefly using natural language.
- Weaknesses. The weaknesses and main disadvantages of each method are described briefly using natural language.

Discussion

This framework integrates characteristics from similar efforts [3,6,7] to provide a complete and comprehensive catalogue and comparative tool.

The main advantages of the framework and the points of differentiation from similar attempts are enumerated below:

- **Systematic**. A lucid and eloquent overview of methods is achieved by categorizing and positioning them in a single table.
- **Critical Review**. Inclusion of a descriptive analysis of the main advantages and disadvantages (strengths and weaknesses) of each method. This brief overview will effectively assist HCI professionals in incorporating and applying a method in their projects.
- **Comprehensive**. The framework covers 41 HCI methods in total, embracing aspects of system development from initial conception to final testing. This compares well to previous attempts (e.g. [6,7]).
- **Extensible**. The framework is template-based, so it can be updated by appending new methods, and allows for the revision of the characteristics of each method in a collaborative way from the HCI community. New parameters and possible ways of categorization can be included to address emerging needs of the HCI community, such as the shift to UX or the focus on a specific domain.

The main disadvantages and limitations of the framework are as follows:

- **Subjective**. The characteristics and the values for each method's parameters have been elicited either from the relevant literature, or from the personal experience of a small number of HCI researchers and practitioners.
- **Non-experiential**. The framework does not explicitly address the experiential aspects of interaction with a system.

Suggested enhancements

In order to address the shortcomings we propose a number of enhancements to the framework. These ideas can also prove to be useful enhancements to similar efforts undertaken in this area.

First, an interactive online version of the methodological matrix should be developed. This should make use of visualization elements and techniques to simplify the overview and review of methods and support the decision-making process. Drawing from examples of information visualization used in decision making (e.g. [1]), visual elements such as color, saturation, shape, size, texture and orientation can be used to convey information. Interactivity can support the decision making process by dynamically altering the visibility and visual elements in the online version of the framework. An online interactive version of the framework will further enhance the clear overview of the methods, which is a problem in large and detailed collections of methods (e.g. [4]), and is also expected to contribute to further disseminating the methods to the community of HCI practitioners, researchers, students, software developers and other stakeholders.

Second, an interactive online version of the matrix can be enhanced to support participation from HCI researchers and practitioners. This should yield similar benefits, in terms of exposure to a wide range of opinions, as those frameworks derived from survey data (e.g. [2,5,6,7]). However, integration into an online platform will be a more streamlined and efficient way, which is expected to contribute to the dissemination of the results. Of course, this is expected to be one step further than mere collaborative editing, as done in wikis, for example. The data from many participants, including some basic profile information, can be analyzed statistically and allow the dynamic provision of different views of the framework. For instance, it will be possible to find which methods can be used when one does not have access to users, or which methods are most applicable to a specific application domain.

Acknowledgements
The work reported in this paper is supported by FCT grant SFRH/BD/65908/2009.

References
[1] Bautista, J. and Carenini, G. An integrated task-based framework for the design and evaluation of visualizations to support preferential choice. AVI ’06, ACM Press (2006).

[2] Chilana, P.K., Ko, A.J., Wobbrock, J.O., Grossman, T., and Fitzmaurice, G. Post-Deployment Usability: A Survey of Current Practices. CHI 2011, (2011).

[3] Fitzpatrick, R. and Dix, A. A Process for Appraising Commercial Usability Evaluation Methods. HCII ‘99, (1999).

[4] Law, E. L-C., Scapin, D., Cockton, G., Springett, M., Stary, C., and Winckler, M. (eds.) Maturation of Usability Evaluation Methods: Retrospect and Prospect (Final Reports of COST294-MAUSE Working Groups), IRIT Press, France (2009).

[5] Law, E. L-C., Roto, V., and Hassenzahl, M., Vermeeren, A.P.O.S. Understanding, Scoping and Defining User eXperience: A Survey Approach. CHI 2009, ACM (2009), 719-728.

[6] Roto, V., Obrist, M., and Väänänen-Vainio-Mattila, K. User Experience Evaluation Methods in Academic and Industrial Contexts. In workshop on User Experience Evaluation Methods, (2009).

[7] Vredenburg, K., Mao, J.-Y., Smith, P.W., and Carey, T. A survey of user-centered design practice. CHI 2002, ACM (2002), 471-478.

[8] Σπηλιωτόπουλος, Α. Μεθοδολογικό πλαίσιο για την ανάπτυξη εύχρηστων υπολογιστικών συστημάτων. MSc thesis, University of Athens, Greece, 2007.

Domain Values and Method Transferability: an Initial Framework

Gilbert Cockton
School of Design,
Northumbria University,
City Campus East,
Newcastle upon Tyne,
NE1 8ST, UK
gilbert.cockton@northumbria.ac.uk

CHI 2010, April 10–15, 2010, Atlanta, Georgia, USA.
ACM 978-1-60558-930-5/10/04.

Abstract
The transferability of design and evaluation methods between domains is an interesting research focus that needs to be given firmer foundations. This paper presents an initial framework that can generate useful conjectures about factors that (dis)favour transferability of methods.

Keywords
Design Methods, Evaluation Methods, Design Situations, Design Resources, Meta-Principles.

Introduction
The TwinTide COST Action (www.twintide.org) has a focus on the transferability of design and evaluation methodologies between application sectors and domains, with an aim of developing cross-sectoral methodologies. *Transferability* is a physical metaphor that involves the movement of one entity from one place to another. For example, footballers transfer between clubs, students transfer between courses and universities, money is transferred between bank accounts, and files are transferred between computers. In all of these examples, something is transferred whole from one place to another.

Two questions follow for the transferability of method(ologie)s between domains:

- Is a method(olog)y transfer whole or partial?
- How are the places in the transfer characterised?

To answer the first question, we need to know what a method(ology) comprises. To answer the second, we need to be able to characterise application domains or sectors in a way that is relevant to the transferability of method(ologie)s. For example, the UK's creative industries grew at 6% p.a. from 1997 to 2005, when they accounted for 7% of GVA. While both are key economic attributes, it is not clear that they have any automatic relevance for design method transferability. We could consider further well established sector attributes, and still struggle to identify relevant ones.

To avoid a fruitless domain-centred search for relevant attributes, a method-led approach is developed to identify potential interactions between method attributes and domains. These interactions give rise to conjectures that can be explored in research programmes. Domains are thus explored through the lens of methods, rather than vice-versa, where it is not clear where to focus, especially with so many alternative classifications of economic sectors, e.g.,

- Primary, secondary, tertiary, quaternary, quinary.
- Agriculture, Mining, Construction, Manufacturing, Transport/Utilities, Wholesale, Retail, FIRE (Finance, Insurance, Real Estate), Services

In short, if we want to reach an understanding of cross-sectoral transfer, we may never get there if we start with the nature of domains. A more tractable approach starts from the nature of methods.

This paper considers methods rather than methodologies, since the former comprise the latter, and we need to identify the parts of a methodology that can be transferred, and with what costs. The next section outlines a framework for a method-led analysis (instead of a less promising domain-centred analysis).

A Framework for Transferability

The opening conjecture is that a method is compatible with domains with which its demands are compatible. A demand is any requirement for a method that may not, or cannot, be met in a domain. For example, in healthcare, medical interventions (especially the use of medicines) need to be approved by national review bodies. Thus co-design approaches that let patients design medical interventions cannot always be used in healthcare. Transferability would depend on the nature of patient ideas, which cannot be predicted in advance. Some suggestions would require formal review, but not others (but may still require doctors' consent).

A method-led framework for transferability thus hinges on the nature of method demands. It is possible to distinguish between demands on the basis of their origins in the *focus*, *roles* and *resources* of design work.

Design Situations and the Foci of Design Work

HCI has advocated *user-centred design*. If all design was user-centred, then the term would add nothing to design. Not all design is focused on users. Even within user-centred design, we cannot always focus on users. For example, dependencies on, and availability of, the back button for web site navigation must be carefully

considered. Issues and options here are primarily technical, and have little if any bearing on user needs and preferences. Also, reliability and avoiding bias often guide decisions on evaluations. These are technical concerns within a human sciences context, not user issues. An example of a genuinely user-centred design choice would be avoiding use of the word 'abort' as a command name (e.g., on a GUI button) or in the text of a prompt (e.g., in a modal dialog box), due to its potential to cause distress to various social groups. From these examples, we can see that there are three distinct *types of design choice* here:

- Choices about users and other *beneficiaries* (e.g., pregnant women should not be asked to abort now)
- Choices about *artefacts* (e.g., disabling back button)
- Choices about *evaluations* (e.g., Likert scale sizes)

Even at this very high level of abstraction, we can anticipate method demands that are incompatible with domains. For example, incompatibilities can arise if a type of design choice is not valued within a domain. All domains must value choices about artefacts, since choices of form, features, content, capability etc. must occur in all design settings (otherwise nothing gets designed). However, design settings within traditional applied arts or engineering may not value choices about beneficiaries. Indeed, this is *why* user-centred design came into existence, i.e., as a methodology that was claimed to be superior to the designer-centred applied arts and technology-centred engineering design. Methods that commit considerable resources to understanding users on the basis of primary research (as opposed to stating an unstudied audience) create demands that are not valued when, e.g., designing a bridge or a poster for a play. Conversely, in engineering design, precise actionable verifiable specifications and requirements are more highly valued than in user-centred design. Engineering methods for requirements specification address a type of choice that receives limited attention in user-centred design, i.e., design *purpose*. Choices of design purpose can also be supported by design *briefs* for the applied arts, although without emphasis on precision and verification. However, no HCI methods support option generation, option communication and option choice for design purpose (apart from some experimental worth-focused approaches [3,4]).

One way to capture the different types of design choice that are valued in different domains is through the concept of an *Abstract Design Situation* (ADS [2]). An ADS abstracts over concrete settings to isolate the types of design choices that are made, whether simple (e.g., choice of evaluation) or complex (e.g., aligning evaluation with design purpose and beneficiaries). We can distinguish four ADS classes:

- *Applied Arts*: explicit choices about artefacts and evaluations, co-ordinated through reflection/critique
- *Engineering*: explicit choices about artefacts, purpose and evaluations, aligned via requirements
- *User-Centred*: explicit choices about artefacts, beneficiaries and evaluations, aligned via rationales
- *Design for Human Outcomes*: explicit choices about artefacts, beneficiaries and evaluations, aligned with explicit choices of purpose (e.g., worth focus)

Each type of explicit design choice (simple or complex) creates demands that may not currently be valued in a domain or sector. We can refine our initial conjecture:

(1) a method is transferrable to domains that sufficiently value and prioritise its associated design choice types.

However, this does not imply that design methods associated with 'unwelcome' choice types cannot transfer. This will only be so if the key influencers within a domain or sector cannot be persuaded to alter their perspectives. Such changes of heart are not uncommon. For example, in the UK, some areas of public service design have recently embraced user-centred co-design approaches. Similarly, the CHI-MED project in the UK (www.chi-med.ac.uk) aims to enable the adoption of user-centred and safety critical methods for medical device development. Given this, transferability is unlikely to be a property of method(ologie)s alone, but instead also results from interactions between human communities.

Methods give rise to demands through their association with types of design choices. Further demands are revealed as we drop down a level of abstraction.

Meta-Principles and the Roles of Design Work
Methods rarely fully support the making of design choices. Instead, different methods perform different *roles* in relation to making choices. To be able to choose, designers need a 'menu' of well expressed realistic options. Some methods thus focus on the *generation* of options, others on their *communication*, and others on their *credibility*. Methods thus need to be combined to support high quality choices. For example, personas are primarily a resource for *expressing* potential *beneficiaries*. Not all will be primary personas, so choices have to be made. However, for personas to be *credible*, they need to be grounded in field data. Also, for a set of personas to be comprehensive, field research methods need to be *receptive* to a broad range of possible stakeholders.

Design and evaluation methods thus have different *roles* relative to different types of design *choices* (e.g., choices of beneficiaries). Roles correspond to different *types of design work*, which in turn are instances of *meta-principles* for designing [1]:

- *receptiveness*: openness to options
- *expressivity*: clarity of options
- *credibility*: realism of options

Note that a method often only adequately performs one primary role, and thus needs to be combined with other methods to ensure that a wide range of realistic options are clearly expressed to enable high quality design choices. Methods typically also often support one type of design choice, e.g., expressivity for beneficiaries, credibility of evaluations, or receptiveness to creative opportunities for digital products or services.

As with design choice types, the values required from a meta-principle may make inappropriate demands for some domains. For example, receptiveness to creative opportunities for artefact is more strongly favoured in applied art settings than in engineering design settings, and even less so in user-centred design settings. Also, rich expressivity is favoured more in applied arts and user-centred settings than in engineering, which values precision over inspiration. Despite the scientific foundations of user-centred design, standards of credibility tend to be higher in engineering design. applied arts traditions value designers' judgements over objective credibility.

A method's transferability thus also depends on the work that it supports. We can add a second conjecture:

(2) a method is transferrable to domains that sufficiently value and prioritise it role in design work for its associated choice types.

If associated choice types are not valued in a domain, then neither is the work supported by a method.

Receptiveness, expressivity and credibility are three *simple* meta-principles that associate with a single design choice type. Complex meta-principles involve multiple design choice types, but are mostly associated with design settings that design for human outcomes, where four choice types need to be co-ordinated. In engineering design, co-ordination tends to be achieved via verifiable requirements, and in user-centred design via design rationale methods (albeit often tacitly and inconsistently). For reasons of space, the fit between domain values and the three complex meta-principles of committedness, improvability and inclusiveness [1] is not considered in this paper.

Methods thus give rise to demands through their association with types of both design choices and design work that may not be valued in a domain. Further demands are again revealed as we drop down a further level of abstraction to look inside methods.

Inside Methods: Design and Evaluation Resources
Methods support design work, but they are not themselves design work. A gap between generalised methods and actual work practices has to be filled with project specific resources. For example, for user testing, participant recruitment criteria, test briefings, instructions, activities and debriefings must all be decided upon within specific projects. There are no generic criteria here. Every user test setting is unique.

Given the loose coupling between methods as ideas and work as a reality, the word *approach* better indicates that much work is needed in response to choosing some approach. Workable methods only result when resources for an approach have been complemented, parameterised and/or applied in a specific design setting. The effort required to get from an outline approach to a practical method may not be feasible for some domains. We can thus add a third conjecture:

(3) an approach is transferrable to domains that can afford the effort required to produce a practical high quality method when using the approach's resources.

The nature of the resources that comprise approaches is an open question, but examples of *resource types* for design and evaluation methods include [8]:

(i) *scoping* resources that indicate the extent of applicability, e.g., *specialised heuristics* for computer games [5]
(ii) *expressive* resources that communicate the output of a method via specifications, reports, templates and formats etc., e.g., The ISO *Common Industry Format* (CIF) for evaluation reports [6]
(iii) *instrumentation* resources that collect issues and measures for evaluations, e.g, *physiological measurement instruments* [7]

Such resources give rise to further demands in addition to those associated with design work and choice types.

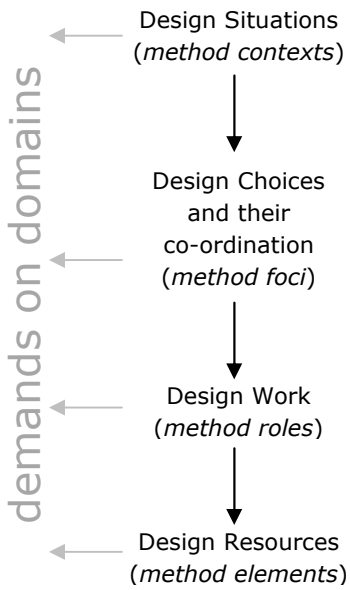


Figure 1. Method-led Outline Framework for Transferability

For example, domains that value scientific validation approaches may find heuristics unacceptable. Similarly, domains that value expert judgement may see no value in CIFs. Also, domains where user experience is not a primary consideration are unlikely to value physiological measurement instruments. We can thus add a fourth conjecture:

(4) an approach is transferrable to domains where its supplied resources are valued

Further conjectures can be derived from more detailed analysis of the interactions between design choice types, design work types, and design and evaluation resources, but there is not space here to extend the analysis. Clearly, there is more to method transferability between domains than the four conjectures derived above, but these do offer a basis for more systematic derivation of testable hypotheses.

Conclusions

A compact method-led framework (Figure 1) of choice types, work types, and resources within approaches has more promise as a basis for investigating the bases for method transferability than an approach based on demarcating domain types and their attributes. This paper has focused on presenting the framework and deriving conjectures as a contribution to the new research focus on method transferability. Space has not allowed a thorough and systematic defence of the framework. Instead, examples have been used throughout to indicate how the framework could be grounded in realistic HCI examples.

A method-led framework indicates that when considering domains, the focus should be on values within the domain, in terms of what is prized, favoured and shunned, and the worth relations resulting from what can be invested to realised prized and favoured value within a domain. However, domains are not static. As with all human structures, values can change. Method(ologie)s that are not immediately transferrable now could become transferrable by gaining support from key influencers, e.g., those responsible for legislation for the safety of medical devices, or the fairness of computer-administered examinations.

References

1. Cockton, G. 2009. Getting There: Six Meta-Principles and Interaction Design, *Proc. CHI 2009*, 2223-2232.
2. Cockton, G. 2010. Design situations and methodological innovation in interaction design. *Proc CHI EA 2010*. ACM, 2745-2754.
3. Cockton, G. Kirk, D., Sellen, A. and Banks, R. 2009. Evolving and Augmenting Worth Mapping for Family Archives. *Proc. HCI 2009*, 329-338, BCS.
4. Cockton, G. Kujala, S., Nurkka, P. and Hölttä, T. 2009. Supporting Worth Mapping with Sentence Completion. *Proc. INTERACT 2009*. II: 566-581.
5. Desurvire, H., Caplan, M. and Toth J.A. 2004. Using heuristics to Evaluate the Playability of Games. *Proc. CHI 2004*, ACM, 1509-1512.
6. ISO/IEC. 2006. *IS25062 Software engineering -- Software product Quality Requirements and Evaluation (SQuaRE) -- Common Industry Format (CIF) for usability test reports*, ISO Geneva.
7. Mandryk R.L. 2008. Physiological Measures for Game Evaluation. *Game Usability: Advice from the Experts for Advancing the Player Experience*. Eds K. Isbister and N. Shaffer, Morgan Kaufmann.
8. Woolrych, A. Hornbæk, K. Frøkjær, E. and Cockton, G. 2011. Ingredients and Meals Rather Than Recipes: a Proposal for Research that Does Not Treat Usability Evaluation Methods as Indivisible Wholes, to appear in *Int. J of HCI*.

Hidden in the Text: Exploring User Experience for the Automotive Domain

Marianna Obrist
Christian Doppler Laboratory
for "Contextual Interfaces"
HCI & Usability Unit,
ICT&S Center
University of Salzburg
Sigmund-Haffner-Gasse 18
5020 Salzburg, Austria
marianna.obrist@sbg.ac.at

Daniela Wurhofer
Christian Doppler Laboratory for
"Contextual Interfaces"
HCI & Usability Unit,
ICT&S Center
University of Salzburg
Sigmund-Haffner-Gasse 18
5020 Salzburg, Austria
elke.beck@sbg.ac.at

Elke Beck
Christian Doppler Laboratory for
"Contextual Interfaces"
HCI & Usability Unit,
ICT&S Center
University of Salzburg
Sigmund-Haffner-Gasse 18
5020 Salzburg, Austria
daniela.wurhofer@sbg.ac.at

Manfred Tscheligi
Christian Doppler Laboratory for
"Contextual Interfaces"
HCI & Usability Unit,
ICT&S Center
University of Salzburg
Sigmund-Haffner-Gasse 18
5020 Salzburg, Austria
manfred.tscheligi@sbg.ac.at

Abstract
Within the HCI community there is a need for rich descriptions of User Experience (UX) from a real life perspective in order to guide future experience-centered design processes. Most UX insights are revealed in lab or field studies where people are directly involved in the data collection process. Within this paper, we make an initial step towards getting a richer understanding of car owners' experiences by investigating already available, written experience reports in a selected online car forum with the means of qualitative text analysis. We want to show the potential of qualitative analysis for getting less biased insights on UX. Following a grounded theory approach four main categories are identified and interpreted from the reports. They are further grouped into four car experience types and described in more detail with regard to their main characteristics and linked with relevant UX components to guide future experience-centered design processes of in-car interfaces.

Keywords
User Experience, Experience Reports, Text Analysis, non-reactive Research Method, Automotive Context

ACM Classification Keywords
H5.2. Information interfaces and presentation (e.g., HCI): User Interfaces, *Evaluation/methodology*.

General Terms
Design, Human Factors

Introduction & Motivation

According to Wright and McCarthy [8], people want to share significant experiences with others, and others often expect to hear people's stories of such experiences. This process might be thought of as „recounting" or „storytelling" (see also [1][8]). Thus, communicated, remembered experiences and the way they are reported in texts can be considered as relevant sources for exploring UX. However, such stories can never catch the totality of peoples' lived experience as they are always edited versions.

We aimed at getting a deeper understanding of car owners' remembered experiences as a initial step within a large seven-year research project on UX in the automotive domain. The central focus of our work is on how people describe experiences with their cars and where these descriptions are embedded in and linked to. By investigating written experience reports in online car fora with the means of qualitative text analysis, we made an initial step towards getting a richer understanding of car owners' experiences. The study results build the basis for the second step, which aims at identifying relevant components of UX for in-car interfaces, that should guide the experience prototyping, design and UX evaluation processes.

Related Work

The presence of a researcher in a specific context can easily disturb and influence the users' experiences. Thus, several self-reporting methods such as probing, experience-sampling (ESM), or Experience Reports [5] have been introduced in HCI. All these methods provide participants with tools and materials for collecting and reflecting their experiences for themselves. Although the collected data have proved to be valuable, the participants are still aware of being part of a study and may be influenced by this fact. One way to reduce this bias and reveal more realistic data are long-term studies [4]. Another way is to apply non-reactive, unobtrusive methods (e.g., content analysis) [1]. When applying such methods, the person is not aware of the object of a research project and no specific action is requested on his/her side. The challenge is to achieve a differentiated understanding of UX by combining data from reactive methods with insights from non-reactive methods, such as peoples' descriptions of experiences (e.g., reviews, narrations [8] etc.) initiated and produced by themselves.

Stories and narrations are a vehicle that people use to condense and remember experiences, and to communicate them to an audience [3]. This happens extensively through web blog entries, discussion fora, product test reports, experience reports, etc. As Blythe and Cairns [1] showed in their study, such material provides a rich resource to inform research and design, but also requires new methods. Qualitative research is useful to explore, discover, understand, and describe phenomena that have already been identified but are not well understood yet, such as the UX concept [6].

figure 1. Car Forum selected for the analysis "Autoplenum.de" (http://www.autoplenum.de/)

Text analysis process with three successive types of theoretical coding:

1) open coding
2) axial coding
3) selective coding

Four car experience types

Text Analysis of Car Forum Entries

A qualitative analysis was conducted for uncovering peoples' car experiences based on written experience reports collected from a public online car forum. We are aware that experience descriptions in car fora only illuminate a facet of the whole experience of a car owner (without linkage to a particular car type), but we are confident that the selected approach represents an initial step towards understanding UX in the automotive context. The selected car forum "Autoplenum.de" is an appropriate source for data because it provides written descriptions of personal experiences by car owners, and serves as a place for communicating these experiences. The newest entries from the car forum were selected (using theoretical sampling) and analyzed by two coders together using theoretical coding in the tradition of the grounded theory [9]. Theoretical saturation was reached after the coding of 21 experience stories (meaning that new data did not require the creation of new categories). The average text length was 337 words; all stories were from different authors.

Text analysis by Strauss is chosen as an appropriate method to analyze narrations about remembered experiences. Our central research question was: *How do people experience their car?* In order to investigate this question, we focused on the descriptions in the car forum. We believe that remembered experiences and the way they are reported in texts provide a first foundation for formulating car experience types. By the use of text analysis, we identified categories in the text and the relations between these categories. The basic activity in the analysis was the coding of the data, i.e. isolating different concepts in the text, with three successive types of coding: 1) open coding, 2) axial coding, and 3) selective coding. This text analysis process was applied in this initial exploratory study as part of a seven-year research project on Contextual Interfaces and UX research, amongst others, focusing on the automotive domain. Based on the coding process, the following four categories were identified:

1. *Attributed qualities:* characteristics car owners attribute to their car (e.g., fuel-efficient, fast).
2. *System characteristics:* car related features mentioned by car owners (e.g., gas consumption, engine performance).
3. *Needs and motives:* arguments car owners use to describe their needs and motives for owning a car (e.g., practicability, social belonging).
4. *Car perceptions:* experiences describing the car owners' main focus and values toward their car (e.g., car as tool, car as status symbol).

By using a matrix, the above revealed categories were further linked to each other using the analysis software NVivo[1]. Thus, we identified commonalities and differences between the sub-categories. In a next interpretation step, four main car experience types were extracted based on these initial analysis steps.

Initial Car Experience Types

The four revealed car experience types are presented below, summarizing the initial analysis results from the text analysis based on the selected cases.

Type 1: Car as Tool

This type mainly sees the car as a **"means to an end",** using it for the purpose of transport. Reasons for owning a car are to facilitate everyday activities and routines. Thus, **practicability** and **suitability for everyday life** is in the foreground. Attributed

[1] NVivo website: http://www.qsrinternational.com/

Type 1: Car as Tool

Type 2: Car as Fascination Object

Type 3: Car for Self-Appreciation

Type 4: Car as Frustration Object

characteristics were for example *spacious, comfortable, adjustable, easy to use, well-arranged, family-friendly*. **Safety** and **reliability** were expressed to be essential, revealed in attributions like *reliable, safe, stable, and robust*. Within the experience reports **positive attributes** (e.g., *fast, powerful, smooth, appealing, sophisticated, praiseworthy, ideal, impressive, convenient, fuel-efficient*) and **negative attributes** (e.g., *need to get used to, fuel-inefficient, poorly equipped, old-fashioned/outdated, loud, faulty*) were mentioned at the same time. Overall, the car descriptions can be characterized as objective descriptions, listing both positive and negative attributes in a rather neutral tone. In sum, people belonging to this type are **satisfied with their cars,** they are aware that they do not have high-end cars but take negative attributes into account. However, the positive attributes outweigh the negative ones.

Type 2: Car as Fascination Object
For this car experience type, the car represents a positive entity. Interacting with the car induces positive feelings or memories, indicated by **predominant positive associations** in the texts. Motives found for this type are nostalgia (thinking of the good old times) or identity management (upgrading one's identity by driving a rare/scarce car). When reading such car descriptions, one has the feeling that the author is a **fan of his own car**, praising it and writing about the car in a very positive way. **Exaggerated affirmations** and rather **car-untypical attributions** can be found in the texts, such as *top class, amazing, awesome, genial, impressive, and incomparable*. Negative aspects of the car are only mentioned marginally or are not mentioned at all. Positive associated feelings with the car are the main focus within these reports (e.g., *robust, snappy, smooth, melodious)*. Overall, the car owners seem to have a **special relationship to their car**, often describing it as something *reliable* they are *proud of*.

Type 3: Car for Self-Appreciation
For this third type, the car represents a **means to enhance the persons' social status/identity** and to **facilitate social affiliation**. Moreover, having an enjoyable and pleasurable experience while driving the car **(enjoyment of driving)** can be found throughout these reports, and becomes a repeated key element in the persons' experiences. Another characteristic of this type is the desire to be outstanding or to **belong to a special social group**. This seems to be achieved by owning a special and unique car. The authors often mention attributes such as *rare, expensive, and authentic* for underlining the uniqueness of their car. The car seems to represent a **fixed point in their lives,** cars being stable in their value for the owner and driving the car is needed by the owner (addictive tendencies in the examined reports). Overall, the car seems to affect the personality of the car owner in a way that is considered as positive and valuable for their life and personal/social identification.

Type 4: Car as Frustration Object
This type differs from the previous three types mainly by characterizing the car as an object of frustration. **Only negative characteristics** are attributed to the car (e.g., car *causes costs, deficient, expensive, fuel-inefficient*); no positive attributes can be found in the forum reports. By listing negative aspects of the car, the car owner seems to communicate his/her frustration to the rest of the world. A clear motive for owning a car cannot be identified. Reports of this kind are highly emotional and subjective, revealing the author's disappointment and anger about the car.

Car Experience Types	Main Characteristics	User Experience Components
Type 1: Car as Tool	Practicability Suitability for everyday life Safety Reliability Flexibility	(Perceived) Usability Perceived Safety/Reliability Feeling of Security
Type 2: Car as Fascination Object	Nostalgia Non-technical attributions (very human-like associations) Emotional attachment	Aesthetics Emotions (positive emotions) Pride
Type 3: Car for Self-Appreciation	Enjoyment of driving Status Symbol Social belonging Identification/ Self-Expression	Fun/Enjoyment Pleasure Reputation User values
Type 4: Car as Frustration Object	Disappointment Anger Frustration	Emotions (negative emotions)

table 1. Overview on the four main identified car experience types with their main characteristics and linkages to UX components.

An overview on all four identified car experience types and a summary of their main characteristics can be found in table 1. The main characteristics are linked to relevant UX components, which we identified in an extensive previous literature review.

Conclusions and Future Work

Our primary goal was to identify relevant categories of "car experiences" from the written car experience reports by means of text analysis and theoretical coding following a grounded theory approach. By trying to understand how people describe their experiences with their car, main attributed qualities, system characteristics, needs and motives, and peoples' perceptions of their car were identified and further interpreted, and grouped into four main car experience types. It is essential for designers to understand these different types. When we know what people remember and recount from their experiences to others, then we can reveal domain-appropriate design recommendations and experience types later on as well as develop, for instance, measuring tools for exploring UX in consideration of the context in which the experiences occur. Currently, for instance, we develop a first version of a car experience questionnaire based on the attributes revealed from the text analysis. This questionnaire will be further extended with more data sets and validated in follow-up user studies in the

automotive context to collect further insights on car experience types using additional data sources as well as considering different car types and owners' experiences. In addition, the grouped car experience types can be further developed and used as an inspirational tool for experience-centered design (similar to personas, however, with a special focus on relevant UX components to be considered in the design process). Moreover, we intend to explore our approach towards other domain areas.

Although the text analysis used in the presented qualitative study helps to gather rich descriptions of UX, it has also limitations. Text analyses focus on written reports of remembered experiences produced by people and thus, the analysis misses the experiences not remembered by people or not communicated in the texts. As stated at the beginning of the paper, it is essential to link such data from non-reactive studies with reactive data collection methods. We are planning further user studies using narrative methods (e.g., explicitation interviews [2]) to extend our car experience type descriptions.

Acknowledgements

The financial support by the Federal Ministry of Economy, Family and Youth and the National Foundation for Research, Technology and Development and by AUDIO MOBIL Elektronik GmbH is gratefully acknowledged (Christian Doppler Laboratory for "Contextual Interfaces").

References

[1] Blythe, M. and Cairns, P. 2009. Critical methods and user generated content: the iPhone on YouTube. In *CHI '09.* ACM, NY, USA, 1467-1476.

[2] Cahour, B. 2008. Discomfort, affects and coping strategies in driving activity. In *Proc.* ECCE '08, ACM, NY, USA, Article 22, 7 pages.

[3] Forlizzi, J. and Ford, S. 2000. The building blocks of experience: an early framework for interaction designers. In *Proc DIS '00*, Daniel Boyarski and Wendy A. Kellogg (Eds.). ACM, NY, USA, 419-423.

[4] Karapanos, E., Zimmerman, J., Forlizzi, J. and Martens, J.B. 2009. UX over time: an initial framework. In *Proc. CHI '09*. ACM, NY, USA, 729-738.

[5] Korhonen, H., Arrasvuori, J. and Väänänen-Vainio-Mattila, K. 2010. Let users tell the story: evaluating user experience with experience reports. In *Proc. CHI EA '10*. ACM, NY, 4051-4056.

[6] Kuutti, K. 2010. Where are the Ionians of user experience research?. In *Proc. NordiCHI '10*. ACM, NY, USA, 715-718.

[7] Law, E. L., Roto, V., Hassenzahl, M., Vermeeren, A. P., and Kort, J. 2009. Understanding, scoping and defining user experience: a survey approach. In *Proc. CHI '09*. ACM, NY, 719-728.

[8] Schrammel, J., Geven, A., Leitner, M., & Tscheligi, M. (2008). Using narration to recall and analyse UX and emotions evoked by today's technology. In P. M. A. Desmet, M. A. K. Karlsson & J. van Erp (Eds.), Cambridge Scholars Press.

[9] Wright, P., and McCarthy J. (2010). *Experience-Centered Design: Designers, Users, and Communities in Dialogue.* Synthesis Lectures in Human-Centered Informatics, ed. John Carroll, no. 8, San Rafael, CA: Morgan & Claypool Publishers.

[10] Strauss, A. and Corbin, J. (1990). Basics of qualitative research: Grounded theory procedures and techniques. Sage Publications.

UCD practice in SMEs – an explorative study in Flanders

Jeroen Vanattenhoven
Centre for User Experience
Research – KULeuven / IBBT
Parkstraat 45 bus 3605
3000 Leuven, Belgium
jeroen.vanattenhoven@soc.kuleuv
en.be

Andrea Wilkinson
Social Spaces, MAD-Faculty
Campus C-Mine
C-mine 5
3600 Genk, Belgium
andrea.wilkinson@khlim.be

David Geerts
Centre for User Experience
Research – KULeuven / IBBT
Parkstraat 45 bus 3605
3000 Leuven, Belgium
david.geerts@soc.kuleuven.be

CHI 2011, May 7–12, 2011, Vancouver, BC, Canada.
ACM 978-1-4503-0268-5/11/05.

Abstract
In this paper we report the results of an exploratory study on the contextual factors of User-Centered Design (UCD) practice in small and medium enterprises (SMEs) in Flanders. To achieve this, we carried out a qualitative mapping study with five organizations: two SMEs, one research department in a large company, and two consultancy firms. This specific selection was made to gain insight from different perspectives on the issue. An important characteristic of UCD practice in SMEs is uncertainty on the part of the client company. This relates to what has been called "wicked problems" in literature, and what differentiates industry from academia. In order to manage uncertainty before and during projects, the organizations had to spend a lot of effort in communication with the client; in these communications visual material was preferred over textual documents.

Keywords
User-Centered Design, design practice, industry, user-experience, context, communication

ACM Classification Keywords
H.5.2 User Interfaces: User-Centered Design

General Terms
Human Factors

Introduction
The academic community has developed many different methods that can be used in UCD. An important issue with these methods is that more often than not, these methods are created from an academic viewpoint, and are difficult to implement in industry practice [3][4][6][8]. It is therefore essential to understand how UCD is used in practice in order to develop or adapt methods that work in an industry context. The main difference between the academic (scientific) context, and the industry context is that the former can invest a significant amount of time and effort into controlling every aspect of their scientific studies, whereas the latter are confronted with several constraints such as limited budgets, short timeframes etc [3][6]. The latter situation has been described as "wicked problems" [1][5]. Whereas [2][8] focus on usability evaluation methods, this paper investigates how SMEs appropriate aspects of UCD in general into their industry context for product improvement and/or innovation.

This study was carried out in the Usewell research project, which is aimed at stimulating UCD practice in Flemish SMEs. To achieve that goal we aim to develop a toolkit for SMEs starting out with UCD. The explorative study in this paper forms the first step the project: we want to understand UCD practice in SMEs from an industry point of view, in Flanders. The research question is how contextual factors in which these SMEs operate daily, influence and shape their application of UCD principles and methods. The insights obtained from this study form essential input for the design of the roadmap and toolkit.

The paper is structured as follows: first, we describe the method we used, followed by the organizations that took part in our study. Then, we discuss the main results of the study. Finally, we formulate the main conclusions.

Method
To gather data about the organizations involved, we used an informal interview/information-gathering technique called 'mapping'. Instead of asking participants to present their experiences and past projects in a standard powerpoint presentation or document, we, along with the participants, used the mapping session to map the trajectory of the players in the project: who was involved, the role he/she played, the various stages of the project, timeline, etc.

The benefit of mapping is that it shares, records and documents insight into a particular organization's working process and allows the moderator and the participants to have conversations about the project and its key components instead of following a linear structure. Using a limited icon set to map the conversation and experience, these icons act as touch points that allow researchers to gather and analyze information across organizations. In this study the mapping technique was mainly used to stimulate the conversation; the icon sets, the layers of transparent paper, and their annotations, form an external representation which helps participants in dynamic problem solving [7].

In the following instances, the mapping kit used was loosely based on the existing Map-It toolkit (www.map-it.be). The icon set for these mapping sessions included the following key icons: duration/phases, people/group roles, opportunity, achievement, tool, problem, method, role-change and documentation, lock and bomb.

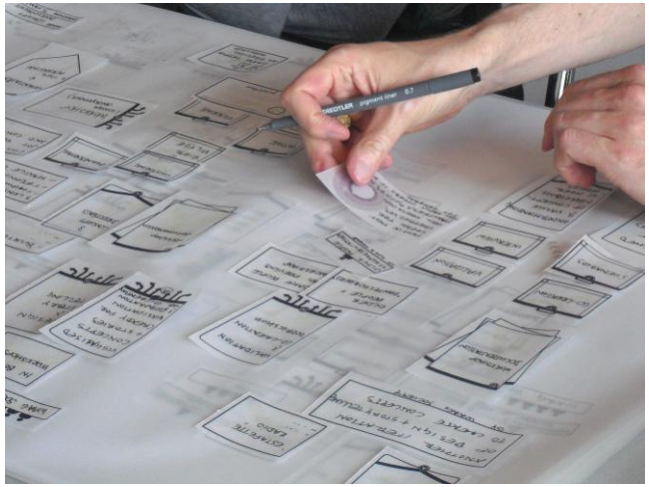


figure 1. Mapping a project using the toolkit

Mapping allows for collective group reflection on a given project. Mapping gives respondents a chance to 'lock' elements that were/are successful (allowing them to discuss why they are successful and to consider how it can be implemented in other projects) and 'bomb' aspects that should be reviewed before being implemented again (aspects that did not go well). Instead of mapping a project from left to right, the mapping technique also uses transparency and layering, so that it is possible to visualize how aspects of the project are 'built upon' instead of seeing a project as a linear succession.

Part workshop and part conversation, mapping requires a moderator to lead the mapping activity and participants from the organization that are/were actively involved on the project being mapped. In our study there were one or two moderators and one researcher observing and taking notes. In three cases the mapping was carried out with one person on behalf of the organization, in one case there were two people, and in the last case, there were four people involved.

For data analysis we relied on the researcher's notes made during the mapping study with each organization. These notes are the researcher's summaries made of participants' arguments when they placed 'bombs' and 'locks' on the mapping of their project; the notes describe the contextual difficulties encountered by the organizations, and the respective ways of dealing with these difficulties.

Participants

In order to gain insight from different perspectives, we selected two SMEs (the main focus of the project), two UCD consultancy firms, and one research department inside a larger company. All organizations (or specific teams) involved less than 10 people. The perspective of a UCD consultancy is relevant since their main activities involve helping out other companies without UCD knowledge or experience. Involving a research department of larger company helped us to get a view on the role of UCD in innovation. In sum, obtaining insights from different viewpoints seems advantageous given the explorative nature of the study.

The different organizations involved will now be described by relying on the descriptions they gave of themselves during the mapping study.

The first organization (which we will name C1 in this paper), is an SME that started as a university spin-off. Their main focus is creating software solutions for the social economy with the goal of improving the lives of people with disabilities. They operate in Belgium.

The second organization (C2) is also an SME working on information and audiovisual technologies for a broad, diverse range of clients in the educational, cultural and training sectors, in Belgium as well as abroad.

The third organization (C3) is a UCD and innovation consultancy firm offering the whole range of UCD activities and innovation, for public and private sector clients in Belgium and abroad.

The fourth organization (C4) is also a UCD and innovation consultancy (the only Dutch partner in the project), carrying out design research and service design, for a broad and diverse range of Dutch and international clients, in the public and private sector.

The fifth organization (C5) is a research and development (R&D) department inside a larger, international technology company, which is involved in "everything that goes on in the network". The role of the R&D department is to create "disruptive innovations" in the domain of visual communications for the larger company.

Results

Uncertainty

An important theme recurring in all conversations with the organizations is uncertainty on the part of their clients or other stakeholders, or on the part of the organization that is working with the client.

One form of uncertainty on the part of the organizations occurs before the start of the project when decision-makers have to be convinced to go ahead with the project. Oftentimes such a decision-maker is not the same person as the one a company talks to about the scope of the project on the part of the client (C2, C3, and C4). Furthermore, a lot of new funding can be at stake for the organization when trying to acquire new projects; it is therefore a delicate balancing act in how much effort to spend in this pre-project phase (C1 and C2).

A second form of uncertainty occurs during the projects. This uncertainty is the changing project scope. In that regard, C4 mentioned that they created a detailed plan for themselves to anticipate on possible outcomes of each phase of the project. At the same time they argued that this detailed plan was always changed significantly at the end of project. Nevertheless, they still considered the exercise of making a detailed plan and evaluating possible directions a project could take as essential.

Communication

In order to manage the above mentioned uncertainties, good communication is vital. The question is: what does good communication actually entail?

All organizations considered visual design documents (personas, user interface sketches, logical schemes…) absolutely necessary. Textual design documents were considered less powerful in order to communicate ideas to and discuss them with the client. In the project discussed with C5 videos were created in order to show a design concept in which they showed people using their application concept in scenarios. The team members found that the video conveyed their ideas better than written personas and scenarios, when they showed them inside their larger organization. The downside was the time that was needed to make these video productions. C4 also used video in order to show design concepts, or to illustrate certain aspects of their ethnographic findings. One of the tools C1 extensively used during client meetings was a whiteboard which allowed them to show, map and draw schemes and plans for discussion with the client. C3 mentioned that using mockups and prototypes they created made their client realize and appreciate the work that is needed to generate these prototypes. Due to severe time constraints C2 used screenshots (very easy, fast and cheap to make) of work-in-progress to communicate the state of the project and discuss its direction.

This does not mean that textual documents were never used; but their use was mainly for documentation i.e. to look up certain decisions (and/or their rationale) made during the project afterwards.

The above statements show us that the organizations decided in each situation what UCD method was the best course of action (based on the available time and budget), and adapted it to their needs.

PERSUASION, INVOLVEMENT

There are different kinds of communication that occur during the course of a project. Which kinds were considered important by the organizations?

Before the start of an actual project, C3 argued that persuasion is important in order to get the client on board. C3 used storytelling to make a compelling case based on their prior cases with other clients to convince the client of the outcome and benefits the proposed UCD approach. There was no standard way of doing this; based on what they knew of the client, they decided on the spot what and how to show their clients what they considered to be valuable approaches.

During projects all organizations met with the client on a regular basis in order to steer the project. C4 considered it very important to involve the client and different stakeholders directly in the project. Stakeholders were invited to workshops in which they would analyze the ethnographic research (maps, photographs, quotes…) results together. The results would not be completely analyzed before going into the workshop. This way they would avoid going into these workshop with their minds made up, and increase the involvement of the client themselves by having them work with the material.

Communication was considered an essential project activity during the discussion of the ‘locks’ and ‘bombs’. To illustrate its importance, C4 stated that half of the project work consist of communication with the client.

Conclusions

The main theme that came forward by analyzing the results of the mapping study was uncertainty on behalf

of the client which is consistent with what in literature is described as practice (as opposed to academic context) and industry [2][3][6][8]. Before a project could start, the organization had to convince the client that the proposed UCD approached justified the investment. During the project the organizations showed how they mainly used visual means of communication. To conclude, communication was a very important topic that came forward during the discussion of the 'locks' and 'bombs' on the mappings.

The main implication of these results for the design of our roadmap and toolkit is that in addition to providing useful UCD content (methods, principles), a great deal of attention will have to be paid to communicating this content to SMEs without UCD experience. We can hereby think of use cases which illustrate how a certain UCD method was adapted to the specific company context, and what the value was of the applied approach for the client.

In future work, we will perform a more thorough analysis based on the complete transcriptions of the interviews (as opposed to only using researcher's notes), an analysis of the mapping artifacts, and a comparison of these two data sources.

Acknowledgements

We would like to thank FlandersInShape for their coordination of the project and fruitful discussions relating to the issue of how to promote UCD in Flanders. Furthermore, we would like to thank all involved organizations in the project, that helped us during the mapping study by providing access to their experiences in UCD practice.

References

[1] Buchanan, R. Wicked problems in design thinking. *Design Issues 8*, 2 (1992), 5–21.

[2] Furniss, D., Blandford, A., and Curzon, P. Usability evaluation methods in practice: understanding the context in which they are embedded. *Proceedings of the 14th European conference on Cognitive ergonomics: invent! explore!*, (2007), 253–256.

[3] Kolko, J. On academic knowledge production. *Interactions 17*, 5 (2010), 80–80.

[4] Norman, D.A. The Research-Practice Gap: The Need for Translational Developers. *Interactions 17*, 4 (2010), 9–12.

[5] Rittel, H.W. and Webber, M.M. Dilemmas in a general theory of planning. *Policy sciences 4*, 2 (1973), 155–169.

[6] Stolterman, E. The nature of design practice and implications for interaction design research. *International Journal of Design*. 2, 1 (2008), 55–65.

[7] Suwa, M. and Tversky, B. External representations contribute to the dynamic construction of ideas. *Diagrammatic Representation and Inference* (2002), 149–160.

[8] Wixon, D. Evaluating usability methods: why the current literature fails the practitioner. *Interactions 10*, 4 (2003), 28–34.

Cyprus University of Technology

European University Cyprus

SIGCHI Cyprus

http://twintide.org/

IC0904 twinTide

Towards the integration of transectional IT design and evaluation

European Science Foundation provides and manages the scientific and technical secretariat for COST

COST is supported by the EU RTD Framework Programme

ISBN : 978-2-917490-13-6 EAN : 9782917490136