Proceedings of the International Workshop on

eaningful easures :

Valid Useful User Experience Measurement (VUUM)

Reykjavik, Iceland • June 18th 2008

Effie L-C. Law
Nigel Bevan
Georgios Christou
Mark Springett
& Marta Lárusdóttir
(Editors)

http://cost294.org/

Proceedings of the International Workshop on Meaningful Measures: Valid Useful User Experience Measurement (VUUM)

Editors: Effie L-C. Law, Nigel Bevan, Georgios Christou, Mark Springett, and Marta Lárusdóttir

Publisher: Institute of Research in Informatics of Toulouse (IRIT) - Toulouse, France

ISBN: 978-2-917490-02-0

Acknowledgements

First of all, we are very grateful to the local organizers – **University of Iceland** and **Reykjavik University**, especially **Marta Lárusdóttir** and **Ebba Hvannberg**, who have strongly supported us to hold our 5th COST294-MAUSE Open Workshop **"Meaningful Measures: Valid Useful User Experience Measurement (VUUM)"** (http://www.cost294.org/vuum). Thanks must also go to the authors of the workshop's papers, whose contributions serve as rich sources of stimulation and inspiration to explore the issues of interest from multiple perspectives. The quality of the contributions could further be ensured and improved with the generous help of the program committee members (Table 1). Their effective and efficient review works are highly appreciated. Besides, we are grateful to our Dissemination Team, **Marco Winckler** and **Philippe Palanque**, for designing, printing and transporting the printed proceedings to the workshop venue.

Table 1: List of the reviewers of the 5th COST294-MAUSE workshop VUUM, 18th June 2008

Name	Affiliation	Country
Bevan, Nigel	Professional Usability Service	UK
Christou, Georgios	European University Cyprus	Cyprus
Cockton, Gilbert	University of Sunderland	UK
Gulliksen, Jan	Uppsala University	Sweden
Hassenzahl, Marc	University of Koblenz-Landau	Germany
Hornbaek, Kasper	University of Copenhagen	Denmark
Hvannberg, Ebba	University of Iceland	Iceland
Jokela, Timo	Joticon	Finland
Larusdottir, Marta	Reykjavik University	Iceland
Law, Effie	ETH Zurich/University of Leicester	Switzerland/UK
Springett, Mark	Middlesex University	UK

Last but not least, we express gratitude to our sponsor – COST (European Cooperation in the field of Scientific and Technical Research; http://cost.cordis.lu/src/home.cfm). The COST Office operated by the European Science Foundation (ESF) provides scientific, financial and administrative support to COST Actions. Specifically, the COST Action 294 (http://www.cost294.org), which is also known as MAUSE, was officially launched in January 2005. The ultimate goal of COST294-MAUSE is to bring more science to bear on Usability Evaluation Methods (UEM) development, evaluation, and comparison, aiming for results that can be transferred to industry and educators, thus leading to increased competitiveness of European industry and benefit to the public. The current Workshop is the second open workshop implemented under the auspices of COST294-MAUSE. As with other past and forthcoming events of COST294-MAUSE, we aim to provide the participants with enlightening environments to further deepen and broaden their expertise and experiences in the area of usability.

Table of Contents

Meaningful Measures: Valid Useful User Experience Measurement (VUUM)

Preface

Effie L-C. Law & the VUUM Program Committee[1]
COST294-MAUSE (http://www.cost294.org)

ABSTRACT
In this Preface we first describe the motives underlying the workshop VUUM. Next, we provide a short summary of each of the sixteen accepted submissions, which are grouped into five categories, namely: Overviews, Validity, Comparisons, Commercial relevance, and UX in Particular Application Contexts. Correspondingly, some questions for further discussion are raised.

Author Keywords
VUUM, Usability measurements, User experience, Meaningfulness, Validity, Usefulness

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

BACKGROUND

"To measure is to know"
"If you cannot measure it, you cannot improve it"
(Lord Kelvin, a.k.a. Sir William Thomson, n.d.)

VUUM2008, June 18, 2008, Reykjavik, Iceland.

Lord Kelvin's dictum on measurement is frequently quoted to justify the quantification of theoretical concepts in physical sciences, computer science, engineering and social sciences [1]. Much HCI evaluation aspires to its scientific philosophy. However, while some HCI researchers and practitioners are strongly convinced about the need for measurement, others are ambivalent about the role of numerical values in providing useful, valid and meaningful assessments and understanding of complex interactions between humans and machines. Some go further and deny the measurability of affective states such as love, beauty, happiness, and frustration. Strictly, one can measure (almost) anything in some arbitrary way. The compelling concern, however, is whether the measure is *meaningful*, *useful* and *valid* to reflect the state or nature of the object or event in question.

What is measurement actually? One key definition is "the assignment of numbers to objects or events in accordance with a rule of some sort" ([9], p.384); a process that is seen as essential to the empirical determination of functional relations. Alternatively, measurement can be defined as "an *observation* that reduces an *uncertainty* expressed as a quantity" (our italics, [5]). In usability evaluation, the uncertainty is the quality-in-use; the observation can objectively be taken by usability professionals (e.g. task completion time) or subjectively by users (e.g. self-perceived duration) [2]. The debate on objective vs. subjective measurements animates many HCI discussions. In dispute are not only which type of measure is more appropriate, but also whether and how they are related and under which conditions [4]. More important perhaps however, is the question of how to interpret measurements

[1] The list is too long to be presented on the title page: Special contributions from **Nigel Bevan, Gilbert Cockton**, and **Georgios Christou** in preparing the call for papers and in reviewing the submissions together with other Program Committee members: **Mark Springett, Marta Lárusdóttir, Ebba Hvannberg, Kasper Hornbæk, Marc Hassenzahl, Jan Gulliksen**, and **Timo Jokela**.

taken and use them to support improvement of an interaction design.

Usability manifests as quality in design, in interaction and in value [8], with diverse measures from many methods and instruments [3]. The challenge is how to select appropriate measures to address the particularities of an evaluation context. The definitions above strongly support the necessity and utility of usability measures, which should provide data either for improving the system under scrutiny (i.e. formative evaluation), and/or for comparing different versions of a system or assessing whether user requirements have been achieved (i.e., summative evaluation). However, both the construct validity and predictive power of some usability measures are of particular concern. For instance, a count of user errors cannot accurately reflect the quality of interaction, nor can it well predict the actual adoption of a product/service, because there are inconsistent ways to categorize and quantify errors [4].

Whereas some qualities of intangible interactions/products may be considered as *non-measurable*, there are *un-measured* but tangible qualities such as affordances and constraints of interfaces. Some researchers argue that they are unmeasured *not* because they have nothing to do with usability, but because no suitable measure that translates them well into usability outcomes exists. Furthermore, there is a substantial philosophical literature on *qualia* [6, 10], which are not even directly detectable, never mind measurable (cf. physicists struggling with the problem of sub-particle physics). It is intriguing to consider indirect measurement and inference of qualia. Besides, we should consider alternative approaches from the fine arts where there are not systematic measures, but critical assessments of artefacts.

Most importantly, all sorts of measurements should be rooted in sound theories, usability measures are no exception. Otherwise, they are just numbers, as remarked by Thomas S. Kuhn [7]:

> "*The route from theory or law to measurement can almost never be traveled backwards. Numbers gathered without some knowledge of the regularity to be expected almost never speak for themselves. Almost certainly they remain just numbers*." (*p.44)*

GOAL & OBJECTIVES

The overall goal of the workshop VUUM is to understand challenges relating to measures of usability and user experience (UX), and to identify effective practical responses to these challenges. The following objectives are addressed:

- To gather evidence of the contextual bases for meaningful and useful usability and UX measures;
- To identify validity and reliability concerns for specific usability measures;
- To identify practical strategies for selecting appropriate usability measures and instruments that meet contextual requirements, including commercial contexts;
- To explore the notion of *qualia* from the philosophical perspective and its practical implications for usability engineering;
- To identify whether non-measurable properties of usability/UX exist, and propose alternative critical methods and techniques for their assessment;
- To extend the range of measures to currently tangible but unmeasured and under-used physical and other properties, such as affordances and constraints inherent in interfaces;
- To review and analyse the theoretical frameworks underlying different usability measures;
- To examine the applicability of existing usability measures to new interaction styles;

CATEORIZATION OF SUBMISSIONS

Sixteen accepted submissions, which are authored by experienced usability and UX researchers and practitioners, cover the aforementioned objectives to various extents. Each of the submissions has been peer reviewed by at least two members of the Program Committee. They are categorized into five major categories: Overview, Validity, Comparisons, Commercial relevance, and UX in particular application contexts. Subsequently, each submission is briefly described.

Category 1: Overviews

This category covers ISO standards, organizational stakeholder enquiry approach, systems usability approach, and worth-centred approach.

- **Nigel Bevan**: *Classifying and Selecting Usability Measures*
  The paper refers to a set of old and new standards related to usability, accessibility and user experience. The author attempts to integrate the interpretations of usability and UX extracted from several related standards. It is deemed challenging to articulate the intricate inter-relationships between the standards. Nonetheless, it is relevant to know how the ISO addresses the UX definition and measures.

- **Pekka Ketola & Virpi Roto**: *Towards Practical UX Evaluation Methods*
  The paper presents the findings of a survey conducted with personnel in different units and levels of Nokia about UX measurement needs. The authors characterize UX in two interesting ways, as a longitudinal relationship that continues post-use, and as a set of intangible benefits both for end user and for the organization. The paper also contributes to an undervalued body of knowledge about the needs and perspectives of different roles within organizations.

- **Leena Norros & Paula Savioja**: *User Experience in the Systems Usability Approach*
  The paper discusses user experience in work-related settings and in the operation of complex systems under the conceptual framework of systems usability and activity theory. It highly advocates contextual analysis and more holistic perspectives. The authors present in this short paper some strong arguments for observations and interviews as the main data source (i.e. informal documentation methods).

- **Gilbert Cockton**: *What Worth Measuring is*
  The paper presents a very useful and insightful overview of the state-of-the-art of worth-centred design (WCD), and some contentious claims about when and why emotion should be measured (i.e. "measuring emotions has to be related to their role"). The paper raises a number of fundamentally important issues and describes a useful approach to understanding UX in context. It is convincing argued that worth centredness places user-experience design and measurement meaningfully in context.

Discussion:
What is actually new in the alternative approaches to usability and UX measures?

Category 2: Validity

This category covers the issues on psychometric properties of scales, the notion of qualia, the utility of usability measures in industry, and the practical use of a newly developed scale.

- **William Green, Greg Dunn & Jettie Hoonhout:** *Developing the Scale Adoption Framework for Evaluation (SAFE)*
  In this paper the authors argue for the use of established scales for subjective measurements. This paper presents a good account of psychometric issues for scale development and choice. It summarizes the approach taken in psychology and provides a framework that is relevant for HCI.

- **Mark Springett**: *Assessing User Experiences Within Interaction: Experience as a Qualitative State and Experience as a Causal Event*
  The paper presents a nice discussion on the notion of qualia, which is addressed in the literature of UX, albeit to a limited extent. It also attempts to link the persuasiveness of the evidence with first person felt states (i.e. qualia) based on some case studies. The author puts forward an interesting argument that the 'soft' discipline of user-experience evaluation is not significantly softer than more traditional measurement of HCI phenomena.

- **Jan Gulliksen, Asa Cajander & Elina Eriksson**: *Only Figures Matter? – If Measuring Usability and User Experience in Practice is Insanity or a Necessity*
  The paper presents two case studies to illustrate how usability measures have been used in industry. The authors put forward some contentious arguments about the necessity and utility of usability measures. The controversies lie in the particularities of organizational goals, which strongly influence whether and how usability measures are used.

- **Jonheidur Isleifsdottir & Marta Lárusdóttir***:* M*easuring the User Experience of a Task Oriented Software*
  The paper uses AttrakDiff2 to probe user experience with a business tool. AttrakDiff2 is new, ambitious and requires examination. The authors administer it before and after think-aloud test. The timing of UX measurement is an issue to explore.

Discussion: Shall we fix the constructs to be measured or fix the measuring tools/methods to render usability and UX measures more meaningful? How is the meaningfulness of a usability measure determined by contextual factors?

Category 3: Comparisons

This category addresses the issue of subjective vs. objective measures, field vs. lab-based settings, and triangulation of data from multiple resources.

- **Anja B. Naumann & Ina Wechsung:** *Developing Usability Methods for Multimodal Systems: The Use of Subjective and Objective Measures*
  The paper addresses the controversial and yet unsettled issue about the relationship between the subjective and objective usability/UX measures. In addition, it looks into the variable interaction modality (mono vs. multi) when examining the extensibility of the existing usability evaluation methods. Another interesting point is to contrast within-data-types correlations with between-data-types correlations.

- **Carmelo Ardito et al**.: *Combining Quantitative and Qualitative Data for Measuring User Experience of an Educational Game*
  This paper demonstrates the applications of multiple measurement techniques and methods to evaluate the complex user experiences engendered by playing the educational game called Explore! The work reported interestingly revealed aspects of authentic field evaluation and its advantages over contrived evaluation settings.

- **Arnold Vermeeren et al**.: *Comparing UX Measurements: a Case Study*
  The paper reports different approaches to capturing data on the usage of a P2P software application. The authors

present some interesting empirical studies – field tests, lab tests and expert reviews. The authors demonstrate how to triangulate different types of data from longitudinal and field studies.

Discussion: Under which conditions are certain types of usability and UX measures correlated?

Category 4: Commercial Relevance

This category addresses the practical applications of emerging usability and UX evaluation methods in industry.

- **Mats Hellman & Kari Rönkkö:** *Is User Experience supported effectively in existing software development processes?*
  This paper addresses a very valid issue of monitoring UX into the software development process with a case study of a mobile phone company. The authors illustrate how the UX quality can be integrated into the traditional usability framework. They also point out the definitional problem of UX and the tradeoff between different usability metrics such as efficiency and satisfaction.

- **Timo Jokela**: *A Two-Level Approach for Determining Measurable Usability Targets*
  The paper makes a succinct and important contribution by distinguishing between business-relevant strategic usability goals and detailed operational usability goals. Different usability measures are required for each of this two-tier usability targets.

- **Kaisa Väänänen-Vainio-Mattila, Virpi Roto, & Marc Hassenzahl et al**.: *Exploring UX Measurement Needs*
  The paper summarizes several proposed methods for UX evaluation that have been addressed at a CHI'08 workshop *UX Evaluation Methods in Product Development* (UXEM). The useful and interesting outcome of that workshop is a list of requirements for practical UX evaluation methods.

Discussion: How can the industrial partners be convinced the usefulness and validity of alternative evaluation methods for usability and UX? Do they tend to believe in adapted traditional methods or in entirely new methods that might bring in some magic?

Category 5: UX in Particular Application Contexts

This category addresses three specific contexts: mobile applications, children's technology uses, and emergent migratory user interfaces.

- **Nikolaos Avouris, Georgios Fiotakis & Dimitrios Raptis***: On Measuring Usability of Mobile Applications*
  This paper addresses some interesting views on measuring the usability of mobile applications. It presents a comprehensive literature review as well as a nice summary of the current practice as reported in several related papers of CHI 2008. This material gives a good ground for discussing what usability measurements for mobile applications are selected from a range of choices.

- **Janet C. Read**: *Is what you see what you get? Children, Technology and the Fun Toolkit*
  The paper addresses the validity and reliability of the UX measures with children. The paper describes some interesting empirical studies on surveying children's prediction and experience in using different types of technologies.

- **Fabio Paterno, Carmen Santoro & Antonio Scorica:** *Evaluating Migratory User Interfaces*
  The paper describes an empirical evaluation study of migratory user interfaces that allow users to change device and continue the task from the point where they left. A new evaluation paradigm is clearly described along with key usability and UX issues specific to it. A useful list of identified issues relevant to migratory interfaces is identified.

  Discussion: What can we learn from the process of adapting the existing usability and UX evaluation methods or creating new ones to address context-specific characteristics?

CONCLUDING REMARKS

The workshop VUUM brings together a group of experienced HCI researchers and practitioners to explore a long-standing research problem – the meaningfulness of measurable constructs and the measurability of non-measurable ones. One may argue that basically everything can be measured, but some things may be more "measurable" than the others; how to estimate the threshold of measurability remains unclear. The sixteen interesting submissions in this volume touch upon the basic issue of the *formal-empirical dichotomy* [9]. Many arguments can be boiled down to the fundamental problem that our understanding of how people think and feel is still rather limited, which is essentially inferred from people's behaviours. Psycho-physiological and neuro-psychological data seem promising, but the issue of calibration and integration is a big hurdle to overcome. Nonetheless, we are convinced about the value, meaningfulness and usefulness of this research endeavour.

REFERENCES

1. Bulmer, M. (2001). Social measurement: What stands in its way? *Social Research, 62*(2).

2. Czerwinski, M., Horvitz, E., & Cutrell, E.(2001) Subjective Duration Assessment: An Implicit Probe for Software Usability? *Proc. IHM-HCI 2001*, 167-170

3. Hornbæk, K .(2006). Current Practice in Measuring Usability: Challenges to Usability Studies and Research, *International Journal of Human-Computer Studies*, *64*, 79-102.

4. Hornbæk, K., & Law, E. L-C. (2007). Meta-analysis of correlations among usability measures. In *Proc. CHI 2007*, San Jose, USA.

5. Hubbard, D. (2007). *How to measure anything: Finding the value of intangibles in business*. John Wiley & Sons.

6. Kerkow, D. (2007). Don't have to know what it is like to be a bit to build a radar reflector – Functionalism in UX. In E. Law, A. Vermeeren, M. Hassenzahl, & M. Blythe (Eds.), *Proceedings of the Workshop "Towards a UX Manifesto"*, 3rd September 2007, Lancaster, UK. Online at: http://www.cost294.org

7. Kuhn,T.S. (1961). The function of measurement in modern physical science. In H. Woolf (Ed.), *Quantification: A History of the Meaning of Measurement in the Natural and Social Sciences (pp.31-63).* Indianapolis: Bobbs-Merrill Co.

8. Law, E., Hvannberg, E., & Cockton, G. (2008) (Eds.). *Maturing usability: Quality in software, interaction and software*. Springer

9. Stevens, S.S. (1958). Measurement and man. *Science, 127*(3295), 383-389.

10. van Gulick, R. (2007). Functionalism and qualia. In M. Velmans & S. Schneider (Eds.) *The Blackwell Companion to Consciousness*. Blackwell.

Developing Usability Methods for Multimodal Systems: The Use of Subjective and Objective Measures

Anja B. Naumann
Deutsche Telekom Laboratories, TU Berlin
Ernst-Reuter-Platz 7, 10587 Berlin, Germany
anja.naumann@telekom.de
+4930-8353-58466

Ina Wechsung
Deutsche Telekom Laboratories, TU Berlin
Ernst-Reuter-Platz 7, 10587 Berlin, Germany
ina.wechsung@telekom.de
+4930-8353-58329

ABSTRACT
In the present paper different types of data (log-data and data from questionnaires) assessing usability parameters of two multimodal and one unimodal system were compared. The participants (N=21) performed several tasks with each system and were afterwards asked to rate the system by filling out different questionnaires. Log-data was recorded during the whole test sessions. The results show that since the questionnaire ratings differed considerably from each other, questionnaires designed for unimodal interfaces should not be used as the only data source when evaluating multimodal systems. The correlations between task duration and questionnaire ratings indicate that subjective measures of efficiency via questionnaires do not necessarily match with the objective data.

Author Keywords
Evaluation methods, multimodal interfaces

ACM Classification Keywords
H.5.2 [Information Interfaces and Presentation]: User Interfaces – theory and methods, interaction styles, haptic I/O, voice I/O, graphical user interfaces (GUI).

INTRODUCTION
In the recent years, an emerging interest in multimodal interfaces has become noticeable. But up to now, there is no standardized method for usability evaluation of multimodal systems. In particular, it is not clear if established methods covering unimodal systems provide valid and reliable results also for multimodal systems.

In an earlier paper we already showed that standardized questionnaires are not completely applicable for usability evaluation of multimodal systems, since there was only little agreement between the different questionnaires [13].

VUUM2008, June 18, 2008, Reykjavik, Iceland.

Thus the questionnaires seem to measure different constructs. Therefore further validation is needed. The purpose of this paper is first to analyze which questionnaire relates most to objective data and second to investigate if the results are consistent to our earlier findings.

RELATED WORK
The HCI literature provides a wide range of methods to measure usability. Most of them were developed to evaluate unimodal graphical user interfaces. Parameters used as usability measures include subjective data, often assessed through questionnaires, and objective data like for example log-files containing task duration or performance data. Since all these parameters are measuring at least roughly the same concept, namely usability, high correlations among them should be observable.

A meta-analysis conducted by Nielsen and Levy [10] showed that performance and predicted preference are indeed correlated. Similar results were reported by Sauro and Kundlund [12]. They found negative correlations between satisfaction (subjective data) and time, errors (objective data) and a positive correlation between satisfaction (subjective data) and completion (objective data).

However, several studies reported opposing findings: Krämer and Nitschke [8] showed that user ratings of multimodal interfaces are not affected by increased intuitivity and efficiency. Moeller [9] could not find a correlation between task duration and user judgments when evaluating speech dialogue systems. Also Frokjaer and colleagues [2] could not find correlations between user ratings and efficiency. Results from a meta-analysis by Hornbaek and Lai Chong-Law [5] showed that the user's experience of the interaction (subjective data) and objective data differ considerably from each other or show even negative correlations.

In view of the studies mentioned above, it seems necessary to use both kinds of data in usability evaluation to obtain reliable results. Thus developing methods for usability evaluation of multimodal systems can only be done by validating within these data types (e.g. validation across questionnaires) but also between these data types (e.g. comparing objective and subjective results).

METHOD

Participants and Material

Twenty-one German-speaking individuals (11 male, 10 female) between the age of 19 and 69 (M = 31.24) took part in the study. All users participated in return for a book token. Due to technical problems log-data was missing from three participants. Regarding the task duration three further cases were identified as outliers and therefore excluded from analyses including task duration.

The multimodal systems adopted for the test were a PDA (Fujitsu-Siemens Pocket LOOX T830) and a tablet PC (Samsung Q1-Pro 900 Casomii). Both systems could be operated via voice control as well as via graphical user interface with touch screen. Additionally, the PDA could be operated via motion control. Furthermore, an unimodal system (a conventional PC controllable with mouse and keyboard) was used as control condition. The application MediaScout, a media recommender system, was the same for all systems.

Procedure

The users performed five different types of tasks: seven navigation tasks, six tasks where checkboxes had to be marked or unmarked, four tasks where an option from a drop-down list had to be selected, three tasks where a button had to be pressed, and one task where a phone number had to be entered. The questionnaires used were the AttrakDiff questionnaire [3], the System Usability Scale (SUS) [1], the Software Usability Measurement Inventory (SUMI) [7], the SASSI questionnaire [4] and a self-constructed questionnaire covering overall ratings and preferences. SUMI, SASSI and AttrakDiff were used in their original form. The SUS was adapted for voice control by replacing the word "system" by "voice control". The order of the questionnaires was randomized. With the help of the questionnaires designed or adapted to cover speech based applications (SASSI and SUS) ratings were only collected for the two multimodal systems (PDA and tablet PC).

Each test session took approximately three hours. The procedure is shown in Figure 1. Each participant performed the tasks with each system. Participants were verbally instructed to perform the tasks with a given modality. This was repeated for every modality supported by that specific system. After that, the tasks were presented again and the participants could freely choose the interaction modality. Finally, they were asked to fill out the questionnaires in order to to rate the previously tested system. This procedure was repeated for each of the three systems. In order to balance fatigue and learning effects the sequence of the systems was randomized. After the third system, a final questionnaire regarding the overall impressions and preferences had to be filled out by the participants.

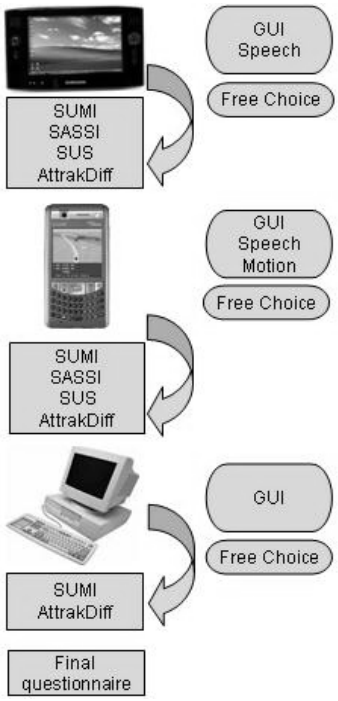

Figure 1. Example of the procedure. Order of systems, modalities and questionnaires was randomized.

During the whole experimental session log-data and psycho-physiological data were recorded. Task duration as a measure of efficiency was assessed with the log-files and was, for each system, averaged over all tasks.

For the current paper only data from the test block in which the users could freely choose modalities was analyzed. Since the participants were already familiar with the systems and all modalities, it can be assumed that they have used the modality they preferred most.

The scales and subscales for each questionnaire were calculated according to the instructions in the specific handbook [1,3,4,11]. All questionnaire items which were negatively poled were recoded so that higher values indicate better ratings.

The level of significance was set at $p<.05$.

RESULTS

Task Duration – Objective Measures

In a first step we compared the systems regarding the task duration. There was a significant difference between all three systems ($F(2,28)=32.64$, $p=.000$; $part.eta^2=.70$). Participants solved the tasks fastest when using the unimodal PC. The most time to solve the tasks was needed with the PDA. Figure 2 visualizes these results.

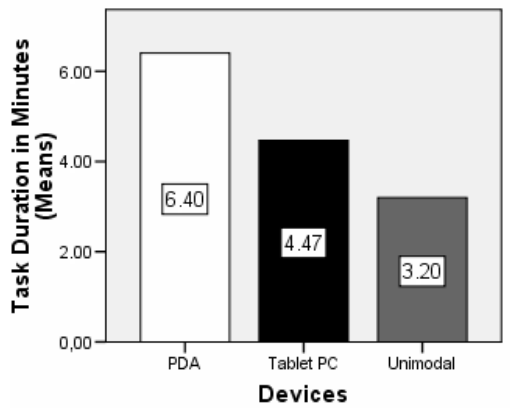

Figure 2. Task duration for all systems.

Questionnaire Ratings -Subjective Measures

Ratings on Scales Measuring Efficiency

Task duration is used as an objective measure for efficiency and should therefore be related to questionnaire scales assessing perceived efficiency. Thus the following section presents the results of the subscales developed with the purpose to measure subjective judgments of efficiency.

The results of task duration are contradictory to the SUMI questionnaire but match the results from the AttrakDiff. According to the SUMI efficiency subscale the PDA is most efficient and the unimodal PC least efficient ($F(2,40)=6.19$, $p=.005$, part.eta²=.236).

Results on the AttrakDiff in contrast agree with the objective task duration data: Best ratings on the AttrakDiff pragmatic scale (the scale measuring efficiency according to the AttrakDiff authors) got the unimodal system, worst ratings were given for the PDA ($F(2,38)=16.80$, $p=.000$, part.eta²=.469).

Scale	System	Mean	SD
SUMI Efficiency (Min.=10/Max.=30)	Tablet PC	19.00	3.39
	PDA	19.90	3.48
	Unimodal	16.67	3.15
AttrakDiff Pragmatic Qualities (Min.=-3/Max.=3)	Tablet PC	.91	.84
	PDA	.01	.88
	Unimodal	1.34	.61
SASSI Speed (Min.=0/Max.=4)	Tablet PC	1.64	.50
	PDA	1.60	.46

Table 1. Ratings on questionnaire subscales measuring efficiency.

The SASSI speed scale revealed no significant difference ($t(20)=.40$, $p=.69$, Cohen's $d= 0.09$) but the absolute rating was higher, which means better, for the tablet PC. The detailed results are given in Table 1.

Ratings on Scales Measuring Global Usability

According to ISO 9241 [6] efficiency is one of the three main factors determining usability. So we expected global questionnaire scales to be affected by task duration, as it is an efficiency measure.

On the SUMI global scale the PDA was rated best and the unimodal PC worst ($F(2,40) = 6.56$, $p = .003$; partial eta² = .247). The AttrakDiff scale attractiveness pointed to the tablet PC as the most attractive system ($F(2,38)=4.04$, $p=.026$, part.eta²=.175). Regarding SUS and SASSI ratings were only assessed for the systems supporting voice control. No differences but a medium effect could be shown for the SASSI ($t(20)=1.95$, $p=.059$, $d=0.6$). Like on the AttrakDiff the tablet PC was rated better than the PDA. The SUS ratings for the PDA and the tablet PC did not differ ($t(20)=1.23$, $p=.232$, $d=0.25$) but again the absolute ratings were better for the tablet PC. Table 2 presents the detailed results.

Scale	System	Mean	SD
SUMI Global (Min.=10/Max.=50)	Tablet PC	40.19	7.87
	PDA	45.29	10.14
	Unimodal	38.04	7.06
AttrakDiff Attractiveness (Min.=-3/Max.=3)	Tablet PC	.99	1.04
	PDA	.34	.88
	Unimodal	.74	.66
SASSI Global (Min.=0/Max.=4)	Tablet PC	2.25	.52
	PDA	1.96	.48
SUS Global (Min.=0/Max.=100)	Tablet PC	53.93	16.60
	PDA	50.12	13.38

Table 2. Ratings on questionnaire scales measuring global usability.

Correlations between Subjective and Objective Measures.

In a further step the efficiency measures from the different questionnaires were correlated with task duration. All questionnaire ratings and task durations were transformed into ranks for each participant. Concerning the questionnaire ratings, rank one was assigned to the system with the best rating. For task duration, the system with the shortest task duration got rank one. Thus positive correlations show concordance between these objective (task duration) and subjective measures (questionnaire ratings).

Correlations between Scales Measuring Efficiency and Task Duration

The ranks based on SUMI efficiency correlated negatively with task duration. A positive correlation could be observed between the ranks based on the AttrakDiff pragmatic scale and task duration. Also the rank transformed SASSI speed scale was positively correlated with task duration. Table 3 shows the detailed results.

Ranks based on	Task Duration
SUMI Efficiency (N=45)	-.577**
AttrakDiff Pragmatic Qualities (N=45)	.529**
SASSI Speed (N=30)	.324*

Table 3. Correlations (Kendall's tau-b) between task duration and subscales measuring efficiency (p<.01; *p<.05).**

Correlations between Scales Measuring Global Usability and Task Duration

Regarding the scales assessing global usability the SUMI showed a negative correlation with task duration. All other global scales were not significantly correlated with task duration (see Table 4).

Ranks based on	Task Duration
SUMI Global (N=45)	-.418**
AttrakDiff Attractiveness (N=45)	.019
SASSI Global (N=30)	.230
SUS Global (N=30)	.264

Table 4. Correlations (Kendall's tau-b) between task duration and scales measuring global usability (*p<.05).

DISCUSSION

The subjective data (questionnaire ratings) and the objective data (task duration) results showed concordances only to a limited extend. The questionnaire ratings most inconsistent to the results of all other questionnaires as well as to the task duration data were the ratings of the SUMI. The SUMI results showed correlations in the wrong direction for efficiency and for the global scale. That means that the longer the task duration the better was the rating on the SUMI. The AttrakDiff pragmatic scale in contrast showed the highest agreement with the task duration data. Thus this scale measures the construct it was developed for. A similar conclusion can be drawn from the SASSI results: Ratings on the speed scale matched the task duration data. Regarding the global scales, all questionnaires except the SUMI showed no significant correlation with task duration. According to these results, the global usability is hardly affected by the systems efficiency.

The current results are in line with our previous findings [13]. Again, the AttrakDiff and the SASSI showed the most concordance. A possible explanation could be that the kind of rating scale used in the AttrakDiff, the semantic differential, is applicable to all systems. It uses no direct questions but pairs of bipolar adjectives, which are not linked to special functions of a system. The SASSI uses direct questions but was specifically developed for the evaluation of voice control systems and may therefore be more suitable for multimodal systems including voice control than questionnaires developed for GUI based systems. Furthermore all SUMI questions were included although some of them are only appropriate for the evaluation of market-ready interfaces. These inappropriate questions may have affected the SUMI results.

In summary, some of the questionnaires showed high correlations in the right direction with the objective data. Other questionnaires' ratings stood in contrast or no relation with the objective data. Thus the contradictory findings [5,8,9,10,12] regarding the correlation between objective and subjective data may partly be caused by questionnaires with a lack of construct validity. Therefore the method assessing subjective data should be chosen carefully. Furthermore a reliable, valid and more specific questionnaire especially for multimodal interfaces is desirable. In view of the results reported, the AttrakDiff provides a proper basis for this.

REFERENCES

1. Brooke, J. SUS: A 'quick and dirty' usability scale. In Jordan P.W., Thomas B., Weerdmeester B.A., McClelland I. L. (Eds.) *Usability Evaluation in Industry*, pp. 189-194. Taylor & Francis, London, 1996.
2. Frøkjær, E., Hertzum, M., and Hornbæk, K. Measuring usability: are effectiveness, efficiency, and satisfaction really correlated? In *Proc. CHI 2000*. ACM Press, (2000), 345-352.
3. Hassenzahl, M., Burmester, M. and Koller, F. AttrakDiff: Ein Fragebogen zur Messung wahrgenommener hedonischer und pragmatischer Qualität [A questionnaire for measuring perceived hedonic and pragmatic quality]. In Ziegler J., Szwillus G. (Eds.) *Mensch & Computer 2003. Interaktion in Bewegung*, B.G. Teubner, Stuttgart (2003), 187-196.
4. Hone, K.S. and Graham, R. Towards a tool for the subjective assessment of speech system interfaces

(SASSI). *Natural Language Engineering*, *6*, 3/4 (2000) 287-305.

5. Hornbæk, K. and Law, E.L. Meta-analysis of correlations among usability measures. In *Proc. CHI 2007*. ACM Press (2007), 617-626.
6. ISO 9241-9 Ergonomic Requirements for Office Work with Visual Display Terminals, Nonkeyboard Input Device Requirements, Draft International Standard, International Organization for Standardization (1998).
7. Kirakowski, J. and Corbett, M. SUMI: The software usability measurement inventory. *British Journal of Educational Technology*, *24*, 3 (1993), 210-212.
8. Krämer, N.C. & Nitschke, J. Ausgabemodalitäten im Vergleich: Verändern sie das Eingabeverhalten der Benutzer? [Output modalities by comparison: Do they change the input behaviour of users?] In R. Marzi, V. Karavezyris, H.-H. Erbe & K.-P. Timpe (Eds..), *Bedienen und Verstehen. 4. Berliner Werkstatt Mensch-Maschine-Systeme*. Düsseldorf: VDI-Verlag (2002), 231-248.
9. Möller, S. Messung und Vorhersage der Effizienz bei der Interaktion mit Sprachdialogdiensten [Measuring and predicting efficiency for the interaction with speech dialogue systems], In S. Langer & W. Scholl (Eds.), *Fortschritte der Akustik - DAGA 2006.* DEGA, Berlin (2006), 463-464.
10. Nielsen, J. & Levy, J. Measuring usability: Preference vs. performance. *Communications of the ACM,* 37, 4 (1994), 66-75.
11. Porteous, M., Jurek, K. and Corbett, M. *SUMI: User Handbook*. Human Factors Research Group, University of Cork, Ireland, 1993.
12. Sauro, J. and Kindlund, E. A method to standardize usability metrics into a single score. In *Proc. CHI 2005*. ACM Press (2005), 401-409.
13. Wechsung I. and Naumann, A. Established Usability Evaluation Methods for Multimodal Systems: A Comparison of Standardized Usability Questionnaires. In *Proc. PIT 08*. Heidelberg: Springer (in press).

Classifying and Selecting UX and Usability Measures

Nigel Bevan
Professional Usability Services
12 King Edwards Gardens, London W3 9RG, UK
mail@nigelbevan.com
www.nigelbevan.com

ABSTRACT

There are many different types of measures of usability and user experience (UX). The overall goal of usability from a user perspective is to obtain acceptable effectiveness, efficiency and satisfaction (Bevan, 1999, ISO 9241-11). This paper summarises the purposes of measurement (summative or formative), and the measures of usability that can be taken at the user interface level and at the system level. The paper suggests that the concept of usability at the system level can be broadened to include learnability, accessibility and safety, which contribute to the overall user experience. UX can be measured as the user's satisfaction with achieving pragmatic and hedonic goals, and pleasure.

WHY MEASURE UX/USABILITY?

The most common reasons for measuring usability in product development are to obtain a more complete understanding of users' needs and to improve the product in order to provide a better user experience.

But it is also important to establish criteria for UX/usability goals at an early stage of design, and to use summative measures to evaluate whether these have been achieved during development.

Summative Measures

Summative evaluation can be used to establish a baseline, make comparisons between products, or to assess whether usability requirements have been achieved. For this purpose, the measures need to be sufficiently valid and reliable to enable meaningful conclusions to be drawn from the comparisons. One prerequisite is that the measures are taken from an adequate sample of typical users carrying out representative tasks in a realistic context of use. Any comparative figures should be accompanied by a statistical assessment of whether the results may have been obtained by chance.

VUUM2008, June 18, 2008, Reykjavik, Iceland.

For example, the test method for everyday products in ISO 20282-2 points out that to obtain 95% confidence that 80% of users could successfully complete a task would for example require 28 out of 30 users tested to be successful. If 4 out of 5 users in a usability test were successful, even if the testing protocol was perfect there is 20% chance that the success rate for a large sample of users might only be 51%.

Although summative measures are most commonly obtained from user performance and satisfaction, summative data can also be obtained from hedonic questionnaires (e.g. Hassenzahl et al., 2003; Lavie and Tractinsky, 2004) or from expert evaluation, such as the degree of conformance with usability guidelines (see for example Jokela, et al, 2006).

Formative Measures

Formative evaluation can be used to identify UX/usability problems, to obtain a better understanding of user needs and to refine requirements. The main data from formative evaluation is qualitative. When formative evaluation is carried out relatively informally with small numbers of users, it does not generate reliable data from user performance and satisfaction.

However some measures of the product obtained by formative evaluation, either with users or by an expert, such as the number of problems identified, may be useful, although they should be subject to statistical assessment if they are to be interpreted.

In practice, even when the main purpose of an evaluation is summative, it is usual to collect formative information to provide design feedback at the same time.

WHAT MEASURES SHOULD BE USED?

There are two types of UX/usability measures: those that measure the result of using the whole system (usability in use) and measures of the quality of the user interface (interface usability).

SYSTEM USABILITY

ISO 9241-11 (1998) defines usability as:

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

and ISO 9241-171 (2008) defines accessibility as:

> usability of a product, service, environment or facility by people with the widest range of capabilities

These definitions mean that for a product to be usable and accessible users should be able to use a product or web site to achieve their goals in an acceptable amount of time, and be satisfied with the results. ISO/IEC standards for software quality refer to this broad view of usability as "quality in use", as it is the user's overall experience of the quality of the product (Bevan, 1999). This is a black-box view of usability: what is achieved, rather than how.

The new draft ISO standard ISO/IEC CD 25010.2 (2008) proposes a more comprehensive breakdown of quality in use into *usability in use* (which corresponds to the ISO 9241-11 definition of usability as effectiveness, efficiency and satisfaction); *flexibility in use* (which is a measure of the extent to which the product is usable in all potential contexts of use, including accessibility); and *safety* (which is concerned with minimising undesirable consequences):

Quality in use

- Usability in use
  - Effectiveness in use
  - Productivity in use
  - Satisfaction in use
    - Likability (satisfaction with pragmatic goals)
    - Pleasure (satisfaction with hedonic goals)
    - Comfort (physical satisfaction)
    - Trust (satisfaction with security)
- Flexibility in use
  - Context conformity in use
  - Context extendibility in use
  - Accessibility in use
- Safety
  - Operator health and safety
  - Public health and safety
  - Environmental harm in use
  - Commercial damage in use

Usability in use is similar to the ISO 9241-11 definition of usability:

- Effectiveness*: "accuracy and completeness." Error-free completion of tasks is important in both business and consumer applications.*
- Efficiency: *"resources expended." How quickly a user can perform work is critical for business productivity.*
- Satisfaction: *the extent to which expectations are met. Satisfaction is a success factor for any products with discretionary use; it's essential for maintaining workforce motivation.*

Usability in use also explicitly identifies the need for a product to be usable in the specified contexts of use:

- Context conformity: *the extent to which usability in use meets requirements in all the required contexts of use.*

Flexibility in use: the extent to which the product is usable in all potential contexts of use:

- Context conformity in use: *the degree to which usability in use meets requirements in all the intended contexts of use.*
- Context extendibility in use: *the degree of usability in use in contexts beyond those initially intended.*
- Accessibility in use: *the degree of usability in use for users with specified disabilities.*

Safety: acceptable levels of risk of harm to people, business, data, software, property or the environment in the intended contexts of use.

Safety is concerned with the potential adverse consequences of not meeting the goals. For instance in Cockton's (2008) example of designing a van hire system, from a business perspective, what are the potential consequences of:

- Not offering exactly the type of van preferred by a potential user group?
- The user mistakenly making a booking for the wrong dates or wrong type of vehicle?
- The booking process taking longer than with competitor systems?

For a consumer product or game, what are the potential adverse consequences of a lack of pleasurable emotional reactions or of achievement of other hedonic goals?

SYSTEM USABILITY MEASURES

Usability in use and flexibility in use are measured by effectiveness (task goal completion), efficiency (resources used) and satisfaction. The relative importance of these measures depends on the purpose for which the product is being used (for example in some personal situations, resources may not be important).

Table 1 illustrates how the measures of effectiveness, resources, safety and satisfaction can be selected to

measure quality in use from the perspective of different stakeholders.

From an organisational perspective, quality in use and usability in use is about achievement of task goals. But for the end user there are not only pragmatic task-related "do" goals, but also hedonic "be" goals (Carver & Scheier, 1998). For the end user, effectiveness and efficiency are the do goals, and stimulation, identification, evocation and pleasure are the be goals.

Additional derived user performance measures (Bevan, 2006) include:

- Partial goal achievement. *In some cases goals may be only partially achieved, producing useful but suboptimal results.*
- Relative user efficiency. *How long a user takes in comparison with an expert.*
- Productivity**.** *Completion rate divided by task time, which gives a classical measure of productivity.*

Stakeholder:	**End User** *Usability*	**Usage Organisation** *Cost-effectiveness*	**Technical support** *Maintenance*
Goal: **Characteristic**	Personal goals	Task goals	Support goals
System effectiveness	User effectiveness	Task effectiveness	Support effectiveness
System resources	Productivity (time)	Cost efficiency (money)	Support cost
Safety	Risk to user (health and safety)	Commercial risk	System failure or corruption
Stakeholder satisfaction	Hedonic and pragmatic satisfaction	Management satisfaction	Support satisfaction

Table 1. Stakeholder perspectives of quality in use

User Satisfaction Measures

User satisfaction can be measured by the extent to which users have achieved their pragmatic and hedonic goals. ISO/IEC CD 25010.2 suggests the following types of measure:

- Likability: *the extent to which the user is satisfied with their perceived achievement of pragmatic goals, including acceptable perceived results of use and consequences of use.*
- Pleasure: *the extent to which the user is satisfied with their perceived achievement of hedonic goals of stimulation, identification and evocation (Hassenzahl, 2003) and associated emotional responses (Norman's (2004) visceral category).*
- Comfort: *the extent to which the user is satisfied with physical comfort.*
- Trust: *the extent to which the user is satisfied that the product will behave as intended.*

Satisfaction is most often measured using a questionnaire. Psychometrically designed questionnaires will give more reliable results than ad hoc questionnaires (Hornbaek, 2006).

Safety and Risk Measures

There are no simple measures of safety. Historical measures can be obtained for the frequency of health and safety, environmental harm and security failures. A product can be tested in situations that might be expected to increase risks. Or risks can be estimated in advance.

Evaluation of Data from Usage of an Existing System

Measures of effectiveness, efficiency and satisfaction can also be obtained from usage of an existing system.

Web Metrics

Web-based logs contain potentially useful data that can be used to evaluate usability by providing data such as entrance and exit pages, frequency of particular paths through the site, and the extent to which search is successful. (Burton and Walther, 2001), although it is very difficult to track individual user behaviour (Groves, 2007) without some form of page-tagging combined with pop-up questions when the system is being used, so that the results can be related to particular user groups and tasks.

Application Instrumentation

Data points can be built into code that "count" when an event occurs (for example in Microsoft Office (Harris, 2005)). This could be the frequency with which commands are used or the number of times a sequence results in a particular type of error. The data is sent anonymously to the development organization. This real-world data from large populations can help guide future design decisions.

Satisfaction Surveys

Satisfaction questionnaires distributed to a sample of existing users provide an economical way of obtaining feedback on the usability of an existing product or system.

USER INTERFACE USABILITY

The broad quality in use perspective contrasts with the narrower interpretation of usability as the attributes of the user interface that makes the product easy to use. This is consistent with one of the views of usability in HCI, for example in Nielsen's (1993) breakdown where a product can be usable, even if it has no utility (Figure 1).

System acceptability
- Social acceptability
- Practical acceptability
  - Cost
  - Compatibility
  - Reliability
  - Usefulness
    - Utility
    - Usability

Figure 1. Nielsen's categorisation of usability

User interface usability is a pre-requisite for system usability.

Expert-based Methods

Expert evaluation relies on the expertise of the evaluator, and may involve walking through user tasks or assessing conformance to UX/usability guidelines or heuristics.

Measures that can be obtained from expert evaluation include:

- *Number of violations of guidelines or heuristics.*
- *Number of problems identified.*
- *Percentage of interface elements conforming to a particular guideline.*
- *Whether the interface conforms to detailed requirements (for example the number of clicks required to achieve specific goals).*

If the measures are sufficiently reliable, they can be used to track usability during development.

Automated Evaluation Methods

There are some automated tools (such as WebSAT and LIFT) that automatically test for conformance with basic usability and accessibility rules. Although the measures obtained are useful for screening for basic problems, they only test a very limited scope of usability issues (Ivory & Hearst, 2001).

MEASURING UX, USABILITY AND ACCESSIBILITY

Usability is variously interpreted as good user interface design (ISO 9126-1), an easy to use product (e.g. Cockton, 2004), good user performance (e.g. Väänänen-Vainio-Mattila et al, 2008), good user performance and satisfaction (e.g. ISO 9241-11), or good user performance and user experience (e.g. ISO 9241-210).

Accessibility may refer to product capabilities ("technical accessibility") or a product usable by people with disabilities (e.g. ISO 9241-171).

UX has even more interpretations. ISO CD 9241-210 defines user experience as:

> all aspects of the user's experience when interacting with the product, service, environment or facility.

This definition can be related to different interpretations of UX:

- ***UX attributes*** *such as aesthetics, designed into the product to create a good user experience.*
- *The user's pragmatic and hedonic* ***UX goals*** *(individual criteria for user experience) (Hassenzahl, 2003).*
- *The actual* ***user experience*** *when using the product (this is difficult to measure directly).*
- *The measurable* ***UX consequences*** *of using the product: pleasure, and satisfaction with achieving pragmatic and hedonic goals.*

Table 2 shows how measures of system usability and UX are dependent on product attributes that support different aspects of user experience. In Table 2 the columns are the quality characteristics that contribute to the overall user experience, with the associated product attributes needed to achieve these qualities.

The users' goals may be pragmatic (to be effective and efficient), and/or hedonic (stimulation, identification and/or evocation).

Although UX is primarily about the actual experience of usage, this is difficult to measure directly. The measurable consequences are the user's performance, satisfaction with achieving pragmatic and hedonic goals, and pleasure.

User performance and satisfaction is determined by qualities including attractiveness, functionality and interface usability. Other quality characteristics will also be relevant in determining whether the product is learnable, accessible, and safe in use.

Pleasure will be obtained from both achieving goals, and as a direct visceral reaction to attractive appearance (Norman, 2004).

<table>
<tr><td>Quality characteristic</td><td>UX</td><td>Functionality</td><td>User interface usability</td><td>Learnability</td><td>Accessibility</td><td>Safety</td></tr>
<tr><td>Product attributes</td><td>Aesthetic attributes</td><td>Appropriate functions</td><td>Good UI design (easy to use)</td><td>Learnability attributes</td><td>Technical accessibility</td><td>Safe and secure design</td></tr>
<tr><td>UX pragmat-ic do goals</td><td colspan="6">To be effective and efficient</td></tr>
<tr><td>UX hedonic be goals</td><td colspan="6">Stimulation, identification and evocation</td></tr>
<tr><td>UX: actual experience</td><td>Visceral</td><td colspan="5">Experience of interaction</td></tr>
<tr><td>Usability (= performance in use measures)</td><td colspan="2">Effectiveness and Productivity in use:
effective task completion and efficient use of time</td><td>Learnability in use: effective and efficient to learn</td><td>Accessibility in use: effective and efficient with disabilities</td><td>Safety in use: occurrence of unintended consequences</td></tr>
<tr><td rowspan="2">Measures of UX consequences</td><td colspan="5">Satisfaction in use:
satisfaction with achieving pragmatic and hedonic goals</td></tr>
<tr><td>Pleasure</td><td colspan="4">Likability and Comfort</td><td>Trust</td></tr>
</table>

Table 2. Factors contributing to system usability and UX

WHAT SHOULD BE MEASURED?

In a systems development environment, UX/usability measures need to be prioritised:

1. At a high level, whose stakeholder goals are the main concern (e.g. users, staff or managers)?
2. What aspects of effectiveness, efficiency, satisfaction, flexibility, accessibility and safety are most important for these stakeholders?
3. What are the risks if the goals for effectiveness, efficiency, satisfaction, flexibility, accessibility and safety are not achieved in the intended contexts of use?
4. Which of these UX/system usability measures are important enough to validate using user-based testing and/or questionnaires, and how should the users, tasks and measures be selected?
5. Are baseline measures needed to establish requirements? (Whiteside et al, 1998)
6. Which aspects of interface usability can be measured during development by expert evaluation to help develop a product that achieves the UX/system usability goals for the important stakeholders in the important contexts of use?
7. How can UX/usability be monitored during use?

CONCLUSIONS

Discussion of UX and selection of appropriate UX measures would be simplified if the different perspectives on UX were identified and distinguished. The current interpretations of "UX" are even more diverse than those of "usability".

This paper proposes a common framework for classifying usability and UX measures, showing how they relate to broader issues of effectiveness, efficiency, satisfaction, , accessibility and safety. It is anticipated that the framework could to be elaborated to incorporate new conceptual distinctions as they emerge.

Understanding how different aspects of user experience relate to usability, accessibility, and broader conceptions of quality in use, will help in the selection of appropriate measures.

REFERENCES

1. Bevan, N. (1999) Quality in use: meeting user needs for quality, *Journal of Systems and Software*, **49**(1), pp 89-96.
2. Bevan, N. (2006) Practical issues in usability measurement. Interactions 13(6): 42-43

3. Carver, C. S., & Scheier, M. F. (1998). *On the self-regulation of behavior*. New York: Cambridge University Press.
4. Burton, M and Walther, J (2001) The value of web log data in use-based design and testing. Journal of Computer-Mediated Communication, 6(3). jcmc.indiana.edu/vol6/issue3/burton.html
5. Cockton, G. (2004) From Quality in Use to Value in the World. *CHI 2004*, April 24–29, 2004, Vienna, Austria.
6. Cockton, G. (2008a) Putting Value into E-valu-ation. In: Maturing Usability. Quality in Software, Interaction and Value. Law, E. L., Hvannberg, E. T., Cockton, G. (eds). Springer.
7. Cockton (2008b) What Worth Measuring is. Proceedings of Meaningful Measures: Valid Useful User Experience Measurement (VUUM), Reykjavik, Iceland.
8. Groves, K (2007). The limitations of server log files for usability analysis. Boxes and Arrows. www.boxesandarrows.com/view/the-limitations-of
9. Harris, J. (2005) An Office User Interface Blog. http://blogs.msdn.com/jensenh/archive/2005/10/31/487247.aspx Retrieved January 2008.
10. Hassenzahl, M. (2002). The effect of perceived hedonic quality on product appealingness. International Journal of Human-Computer Interaction, 13, 479-497.
11. Hassenzahl, M. (2003) The thing and I: understanding the relationship between user and product. In Funology: From Usability to Enjoyment, M. Blythe, C. Overbeeke, A.F. Monk and P.C. Wright (Eds), pp. 31 – 42 (Dordrecht: Kluwer).
12. Hornbaek, K (2006). Current practices in measuring usability. Int. J. Human-Computer Studies 64 (2006) 79–102
13. ISO 9241-11 (1998) Ergonomic requirements for office work with visual display terminals (VDTs) Part 11: Guidance on Usability. ISO.
14. ISO FDIS 9241-171 (2008) Ergonomics of human-system interaction -- Part 171: Guidance on software accessibility. ISO.
15. ISO CD 9241-210 (2008) Ergonomics of human-system interaction -- Part 210: Human-centred design process for interactive systems. ISO.
16. ISO 13407 (1999) Human-centred design processes for interactive systems. ISO.
17. ISO TS 20282-2 Ease of operation of everyday products -- Part 2: Test method for walk-up-and-use products. ISO.
18. ISO/IEC 9126-1 (2001) Software engineering - Product quality - Part 1: Quality model. ISO.
19. ISO/IEC CD 25010.2 (2008) Software engineering – Software product Quality Requirements and Evaluation (SQuaRE) – Quality model
20. Ivory, M.Y., Hearst, M.A. (2001) State of the Art in Automating Usability Evaluation of User Interfaces. ACM Computing Surveys, 33,4 (December 2001) 1-47. Accessible at http://webtango.berkeley.edu/ papers/ue-survey/ue-survey.pdf
21. Jokela, T., Koivumaa, J., Pirkola, J., Salminen, P., Kantola , N. (2006) "Methods for quantitative usability requirements: a case study on the development of the user interface of a mobile phone", Personal and Ubiquitous Computing, 10, 345 – 355.Nielsen, J. (1993) Usability Engineering. Academic Press.
22. Norman, D. (2004) Emotional design: Why we love (or hate) everyday things (New York: Basic Books).
23. Väänänen-Vainio-Mattila, K., Roto, V., Hassenzahl, M. (2008) Towards Practical UX Evaluation Methods. Proceedings of Meaningful Measures: Valid Useful User Experience Measurement (VUUM), Reykjavik, Iceland.
24. Whiteside, J., Bennett, J., & Holtzblatt, K. (1988). Usability engineering: Our experience and evolution. In M. Helander (Ed.), Handbook of Human-Computer Interaction (1st Ed.) (pp. 791–817). North-Holland.

Towards Practical User Experience Evaluation Methods

Kaisa Väänänen-Vainio-Mattila
Tampere University of Technology
Human-Centered Technology
(IHTE)
Korkeakoulunkatu 6
33720 Tampere, Finland
kaisa.vaananen-vainio-mattila@tut.fi

Virpi Roto
Nokia Research Center
P.O.Box 407
00045 Nokia Group, Finland
virpi.roto@nokia.com

Marc Hassenzahl
University of Koblenz-Landau
Economic Psychology and
Human-Computer Interaction,
Campus Landau, Im Fort 7
76829 Landau, Germany
hassenzahl@uni-landau.de

ABSTRACT
In the last decade, User eXperience (UX) research in the academic community has produced a multitude of UX models and frameworks. These models address the key issues of UX: its subjective, highly situated and dynamic nature, as well as the pragmatic and hedonic factors leading to UX. At the same time, industry is adopting the UX term but the practices in the product development are still largely based on traditional usability methods. In this paper we discuss the need for pragmatic UX evaluation methods and how such methods can be used in product development in industry. We conclude that even though UX definition still needs work it seems that many of the methods from HCI and other disciplines can be adapted to the particular aspects of UX evaluation. The paper is partly based on the results of *UX evaluation methods in product development (UXEM)* workshop in CHI 2008.

Author Keywords
User experience, evaluation methods, product development.

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

INTRODUCTION

Companies in many industrial sectors have become aware that designing products and services is not enough, but designing experiences is the next level of competition [19, 22]. Product development is no longer only about implementing features and testing their usability, but about designing products that are enjoyable and support fundamental human needs and values. Thus, experience should be a key concern of product development.

VUUM2008, June 18, 2008, Reykjavik, Iceland.

There are many definitions for UX, but not an agreed one [16]. However, even the most diverse definitions of user experience all agree that it is more than just a product's usefulness and usability [2,6,17,18,23,26]. In addition, they stress the subjective nature of UX: UX is affected by the user's internal state, the context, and perceptions of the product [2, 6, 17].

However, definitions alone are not sufficient for the proper consideration of UX throughout product development. Product development in its current form needs tools from the very early concept to market introduction: UX must be assessable and manageable. An important element of this is a set of evaluation methods focused on UX.

Apparently, there is a gap between the research community's and the product developers' understanding of what UX is and how it should be evaluated (see Figure 1).

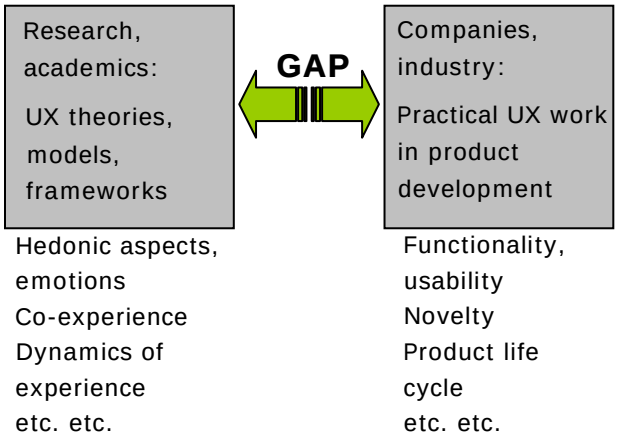

Figure 1. Currently the academic UX research and industrial UX development are focusing on different issues.

As an attempt to close the gap, we organised a workshop on UX evaluation methods for product development (UXEM) [25] in the context of the CHI 2008 conference on human factors in computing. The aim of the workshop was to

identify truly experiential evaluation methods (in the research sense) and discuss their applicability and practicability in engineering-driven product development. In addition, we hoped for lively and fruitful discussions between academics and practitioners about UX itself and evaluation methods. In this paper, we present the central findings of this workshop.

CURRENT STATE OF UX EVALUATION METHODS

Traditionally, technology-oriented companies have tested their products against technical and usability requirements. Experiential aspects were predominantly the focus of the marketing department, which tried to create a certain image of a product through advertising. For example, when Internet became an important channel in communicating the brand and image, technical and usability evaluations of Web sites needed to be integrated with more experiential goals [4,15]. Today, industry is in need of user experience evaluation methods for a wide variety of products and services.

User-centered development (UCD) is still the key to designing for good user experiences. We must understand users' needs and values first, before designing and evaluating solutions. Several methods exist for understanding users and generating ideas in the early phases of concept design, such as Probes [5] or Contextual Inquiry [3]. Fewer methods are available for concept *evaluation* that would assess the experiential aspects of the chosen concept.

EXPERIENTIAL EVALUATION METHODS

A number of evaluation methods were presented and discussed in the UXEM workshop. However, only a few were "experiential" in the sense of going beyond traditional usability methods by emphasizing the subjective, positive and dynamic nature of UX.

Isomursu's "experimental pilots" [11], for example, stress the importance of evaluating before (i.e., expectation), while (i.e., experience) and after product use (i.e., judgment). This acknowledges the subjective and changing, dynamic nature of UX: expectations influence experience, experience influences retrospective judgments and these judgments in turn set stage for further expectations and so forth. In addition, Isomursu points at the importance of creating an evaluation setting, which resembles an actual use setting. UX is highly situated; its assessment requires a strong focus on situational aspects. Roto and colleagues as well as Hoonhout [21,10] stress the importance of positive emotional responses to products and embrace the notion that task effectiveness and efficiency (i.e., usability) might be not the only source for positive emotions. Their focus is on early phases of development where idea creation and evaluation is closely linked and short-cycled.

Hole and Williams suggest "emotion sampling" as an evaluation method [9]. While using a product, people are repeatedly prompted to assess their current emotional state by going through a number of questions. This approach takes UX evaluation a step further, by focusing on the experience itself instead of the product. However, in the context of product development additional steps would have to be taken to establish a causal link between a positive experience and the product: how does the *product* affect the measured experience. Bear in mind, that product evaluation is not interested in experiences per se but in experiences *caused* by the product at hand.

Two further methods presented in the workshop (Repertory Grid, Multiple Sorting) [1,12] make use of Kelly's "personal construct psychology" [e.g., 13]. Basically, these are methods to capture the personal meaning of objects. They have a strong procedural structure, but are open to any sort of meaning, whether pragmatic or hedonic. Interestingly, the methods derived from Kelly's theory tend to provide both a tool for analysis *and* evaluation [7]. The results give an idea of the themes, topics, concerns people have with a particular group of products (i.e., content). At the same time, all positive and negative feelings (i.e., evaluations) towards topics and products become apparent.

Finally, Heimonen and colleagues [8] use "forced choice" to evaluate the "desirability" of a product. This method highlights another potential feature of UX, which may pose additional requirements for UX evaluation methods: There might be drivers of product appeal and choice, which are not obvious to the users themselves. Tractinsky and Zmiri [24], for example, found hedonic aspects (e.g., symbolism, beauty) to be predictive of product choice. When asked, however, participants gave predominantly pragmatic reasons for the choice. Note that the majority of the "experiential" methods discussed so far rely on people's self report. This might be misleading, given that experiential aspects are hard to justify or even to verbalize. In other words, choice might be driven by criteria not readily available to the people choosing. Forced choice might bring this out.

All in all, a number of interesting approaches to measure UX were suggested and discussed in the workshop. All of them addressed at least one key feature of UX, thereby demonstrating that "experiential" evaluation is possible. More work, however, has to be done to integrate methods to capture more aspects of UX simultaneously. In addition, methods need to be adapted to the requirements of evaluation in an industrial setting. So far, most suggested methods are still demanding in the sense of the skills and time required.

REQUIREMENTS FOR PRACTICAL UX EVALUATION METHODS

In industry, user experience evaluation is done in order to improve a product. Product development is often a hectic process and the resources for UX evaluation scarce. Evaluating early and often is recommended, as the earlier

the evaluations can be done, the easier it is to change the product to the right direction.

The early phases of product development are challenging for UX evaluation, since at that point, the material available about the concept may be hard to understand and assess for the participants [10, 21]. In the early phases, it is not possible to test the non-functional concept in the real context of use, although user experience is tied to the context [6]. We need good ideas for simulating real context in a lab [14]. Later on, when prototypes are stable enough to be handed for field study participants, UX evaluation becomes much easier. The most reliable UX evaluation data comes from people who have actually purchased and used a product on the market. This feedback helps improving the future versions of the product.

In summary, the UXEM workshop presentations and group works produced the following requirements for practical UX evaluation methods:

Valid, reliable, repeatable

- For managing UX also in a big company

Fast, lightweight, and cost-efficient

- For fast-pace iterative development

Low expertise level required

- For easy deployment (no extensive training needed)

Applicable for various types of products

- For comparisons and trend monitoring

Applicable for concept ideas, prototypes, and products

- For following how UX develops during the process

Suitable for different target user groups

- For a fair outcome

Suitable for different product lifecycle phases

- For improving e.g. taking into use, repurchasing UX

Producing comparable output (quantitative and qualitative)

For UX target setting and iterative improvement

Useful for the different in-house stakeholders

- As UX is multidisciplinary, many company departments are interested in UX evaluation results.

Clearly, it is not possible to have just one method that would fulfill all the requirements above. Some of the requirements may be contradictory, or even unrealistic. For example, a method which is very lightweight may not necessarily be totally reliable. Also, it might be challenging if not impossible to find a method which is suitable for different types of products, product development phases, and product lifecycle phases. We thus need to have a toolkit of experiential methods to be used for the different purposes.

In the UXEM workshop, we noticed that there is not always a clear line between the design and evaluation methods, since evaluating current solutions often gives ideas for new ones. On the other hand, companies do need evaluation methods that focus in producing UX scores or a list of pros and cons for a pool of concept ideas in an efficient way. After the product specification has been approved, the primary interest is to check that the user experience matches the original goal. In this phase, the methods applied are clearly about evaluation, not about creating new ideas.

DISCUSSION AND CONCLUSIONS

Obviously, applying and developing methods for UX evaluation requires an understanding of what UX actually is. This is still far from being settled. Although everybody in the workshop agreed that the UX perspective *adds* something to the traditional usability perspective, it was hard to even put a name to this added component: Is it "emotional", "experiential" or "hedonic"? The lack of a shared understanding on what UX means was identified as one of the major problems of UX evaluation in its current state. As long we do not agree or at least take a decision on what we are looking for, we cannot pose the right questions. Without an idea of the appropriate questions, selecting a method is futile. Nevertheless, once a decision is made ─ for example to take a look at the emotional consequences of product use ─ there seem to be a wealth of methods already in use within HCI or from other disciplines, which could be adapted to this particular aspect of evaluation.

Working with UX evaluation is a double task: We have to *understand* UX and make it *manageable* and *measurable*. Given the fruitful discussions in the workshop, a practice-driven development of the UX concept may be a valid road to a better understanding of UX. "UX is what we measure" might be an approach as long as there is no accepted definition of UX at hand. However, this approach requires some reflection on the evaluation needs and practices. By discussing the implicit notions embedded in the evaluation requirements and methods, we might be able to better articulate what UX actually should be. The UXEM workshop and this paper hopefully open up the discussion.

ACKNOWLEDGMENTS

We thank all participants of the UXEM workshop: Jim Hudson, Jon Innes, Nigel Bevan, Minna Isomursu, Pekka Ketola, Susan Huotari, Jettie Hoonhout, Audrius Jurgelionis, Sylvia Barnard, Eva Wischnewski, Cecilia Oyugi, Tomi Heimonen, Anne Aula, Linda Hole, Oliver Williams, Ali al-Azzawi, David Frohlich, Heather Vaughn, Hannu Koskela, Elaine M. Raybourn, Jean-Bernard Martens, and Evangelos Karapanos. We also thank the programme committee of UXEM: Anne Aula, Katja Battarbee, Michael "Mitch" Hatscher, Andreas Hauser, Jon Innes, Titti Kallio, Gitte Lindgaard, Kees Overbeeke, and Rainer Wessler.

REFERENCES

1. al-Azzawi, A., Frohlich, D. and Wilson, M. User Experience: A Multiple Sorting Method based on Personal Construct Theory, Proc. of UXEM, www.cs.tut.fi/ihte/CHI08_workshop/papers.shtml
2. Alben, L. (1996), Quality of Experience: Defining the Criteria for Effective Interaction Design. *Interactions*, 3, 3, pp. 11-15.
3. Beyer, H. & Holtzblatt, K. (1998). Contextual Design. Defining Customer-Centered Systems. San Francisco: Morgan Kaufmann.
4. Ellis, P., Ellis, S. (2001), Measuring User Experience. *New Architect* 6, 2 (2001), pp. 29-31.
5. Gaver, W. W., Boucher, A., Pennington, S., & Walker, B. (2004). Cultural probes and the value of uncertainty. *Interactions*.
6. Hassenzahl, M., Tractinsky, N. (2006), User Experience – a Research Agenda. *Behaviour and Information Technology*, Vol. 25, No. 2, March-April 2006, pp. 91-97.
7. Hassenzahl, M. & Wessler, R. (2000). Capturing design space from a user perspective: the Repertory Grid Technique revisited. *International Journal of Human-Computer Interaction, 12,* 441-459.
8. Heimonen, T., Aula, A., Hutchinson, H. and Granka, L. *Comparing the User Experience of Search User Interface Designs*, Proc. of UXEM, www.cs.tut.fi/ihte/CHI08_workshop/papers.shtml
9. Hole, L. and Williams, O. *Emotion Sampling and the Product Development Life Cycle*, Proc. of UXEM, www.cs.tut.fi/ihte/CHI08_workshop/papers.shtml
10. Hoonhout, J. *Let's start to Create a Fun Product: Where Is My Toolbox?*, Proc. of UXEM, www.cs.tut.fi/ihte/CHI08_workshop/papers.shtml
11. Isomursu, M. *User experience evaluation with experimental pilots*, Proc. of UXEM, www.cs.tut.fi/ihte/CHI08_workshop/papers.shtml
12. Karapanos, E. and Martens, J.-B. *The quantitative side of the Repertory Grid Technique: some concerns*, Proc. of UXEM, www.cs.tut.fi/ihte/CHI08_workshop/papers.shtml
13. Kelly, G. A. (1963). *A theory of personality. The psychology of personal constructs, paperback*. New York: Norton.
14. Kozlow, S., Rameckers, L., Schots, P. (2007). People Research for Experience Design. Philips white paper. http://philipsdesign.trimm.nl/People_Reseach_and_Experience_design.pdf
15. Kuniavsky, M. (2003), *Observing The User Experience – A Practitioner's Guide to User Research*. Morgan Kaufmann Publishers, Elsevier Science, USA
16. Law, E., Roto, V., Vermeeren, A., Kort, J., & Hassenzahl, M. (2008). Towards a Shared Definition for User Experience. Special Interest Group in CHI'08. Proc. Human Factors in Computing Systems 2008, pp. 2395-2398.
17. Mäkelä, A., Fulton Suri, J. (2001), Supporting Users' Creativity: Design to Induce Pleasurable Experiences. *Proceedings of the International Conference on Affective Human Factors Design*, pp. 387-394.
18. Nielsen-Norman group (online). User Experience – Our Definition. http://www.nngroup.com/about/userexperience.html
19. Nokia Corporation (2005), *Inspired Human Technology*. White paper available at http://www.nokia.com/NOKIA_COM_1/About_Nokia/Press/White_Papers/pdf_files/backgrounder_inspired_human_technology.pdf
20. Norman, D.A., Miller, J., and Henderson, A. (1995) What you see, some of what's in the future, and how we go about doing it: HI at Apple Computer. Proc. CHI 1995, ACMPress (1995), 155.
21. Roto, V., Ketola, P. and Huotari, S. *User Experience Evaluation in Nokia*, Proc. of UXEM, www.cs.tut.fi/ihte/CHI08_workshop/papers.shtml
22. Seidel, M., Loch, C., Chahil, S. (2005). Quo Vadis, Automotiven Industry? A Vision of Possible Industry Transformations. *European Management Journal,* Vol. 23, No. 4, pp. 439–449, 2005
23. Shedroff, N. An Evolving Glossary of Experience Design, online glossary at http://www.nathan.com/ed/glossary/
24. Tractinsky, N. & Zmiri, D. (2006). Exploring attributes of skins as potential antecedents of emotion in HCI. In P.Fishwick (Ed.), *Aesthetic computing* (pp. 405-421). Cambridge, MA: MIT Press.
25. Väänänen-Vainio-Mattila, K., Roto, V., & Hassenzahl, M. (2008). Now Lets Do It in Practice: User Experience Evaluation Methods for Product Development. Workshop in CHI'08. Proc. Human Factors in Computing Systems, pp. 3961-3964. http://www.cs.tut.fi/ihte/CHI08_workshop/index.shtml
26. UPA (Usability Professionals' Association): "Usability Body of Knowledge", http://www.usabilitybok.org/glossary (22.5.2008)

Exploring User Experience Measurement Needs

Pekka Ketola
Nokia
pekka.ketola@nokia.com

Virpi Roto
Nokia Research Center
virpi.roto@nokia.com

ABSTRACT
We conducted an empirical study on user experience (UX) measurement needs at different units and levels of Nokia product development and asked which kinds of UX measurements would be useful in different parts of the organization. In this report, we present the initial results of this study. We found that the needs for UX measures were not only about design details, but mostly about how the different touch points between user and company are experienced along the product experience lifecycle.

Author Keywords
User experience, Measurement

INTRODUCTION
In a big corporate like Nokia, measurements play an important role in all phases of product development as they enable systematic improvement of the products. Several company departments are interested in user experience measurement. To make the measurements relevant and useful, we first need to find out the *needs* of different stakeholders for UX measurements before starting to define the metrics.

In Human-Computer Interaction (HCI) field, measurements have traditionally been usability measures, such as efficiency, effectiveness, and satisfaction (ISO 13407); learnability, memorability, error prevention, and satisfaction (Nielsen 1993); effectiveness, learnability, flexibility, and attitude (Shackel 1991, 25); guessability, learnability, experienced user performance, system potential, and re-usability (Jordan 1998). Learnability is the common element that is included in all above measures of usability.

When usability evolved to user experience, the measurements broadened from pragmatic (easy and efficient) to experiential (delighting). Jordan upgraded his list to functionality, usability, pleasure, and pride (Jordan 2002). Norman set the goal in engaging users in visceral, behavioural, and reflective level (Norman 2003), and Nokia followed these lines with the Wow, Flow, Show model (Nokia 2005).

As UX highlights the emotional aspects, also emotion measurements have been investigated. Most emotion evaluations concentrate on identifying the emotion a user has *while* interacting with a product, and both objective and subjective methods are used to collect this information (e.g. Mandryk et al 2006; Desmet et al. 2001).

UX evaluation can take place also *after* interaction phase. For example, Hassenzahl (2003) has investigated the pragmatic and hedonic aspects of products from the perspective of product appraisal. This model helps to measure user experience in real life, preferably after long-term use.

Gartner (2007b) considers UX measurements from the perspective of return on investment: what is the monetary benefit of spending money on user experience improvement. They study the relations of brand experience, company experience, and the implications to related revenues and costs. User experience is claimed to be a subset of brand experience. According to Gartner research (2007a), the success of UX can be measured in hard metrics and as intangible benefits:

- Increased revenue: More orders per customer, More repeat engagements, More products per order.
- Reduced cost: Fewer support calls, Fewer returns due to mistake or misperceptions, More efficient server use
- Faster time to market due to accelerated development: Increased customer satisfaction, Improved brand image, Positive word of mouth.

THE EMPIRICAL STUDY
To our mind, people at different roles and levels of product development are the suitable population from whom to ask their view on user measurements and produced data. We selected to make a phenomenographic survey (Marton and Booth 1997) on the different measurement needs in their proper environment.

We use qualitative Email survey for data collection (Meho 2006). For practical reasons we invited 42 Nokia people from different corporate functions which we believe have interest in UX measurements, to act as subjects in this study. We sent our short question (below) to selected

VUUM2008, June 18, 2008, Reykjavik, Iceland.

specialists, senior specialists, managers, senior managers, directors and vice presidents). A total of 18 of them responded within one week

Which User Experience information (measurable data gained from our target users directly or indirectly),is useful for your organization? How?

The question is intentionally very open and can be interpreted in many ways. This way the study participants are not limited in describing their measurement needs, and can address any area they think is valid. Only one participant asked for further clarification for the question.

We analyzed the responses by using the content analysis. The free-text answers were first collected to an Excel document, matching the person, role and response. This grouping told us the key information needs for each discipline (Section "Measurement Needs for Different Groups") and the variations between roles within that discipline. Then, one researcher organized and grouped the answers using mind map technique (see http://en.wikipedia.org/wiki/Mind_map). The grouping was reviewed by the other researcher. This led to common grouping of topics across disciplines (Section "Common Needs for User Measurement"). We gave the draft of our report to the respondents to confirm that we interpreted and classified their responses correctly.

Measurement Needs for Different Groups

In this section we shortly summarize the key needs for each studied group. We will discuss only four groups that answered most actively.

Research (n=3)**.** This group represents people who work with research management or hands-on research, before development takes place. Measurement needs in this group are seen in two main groups:

- How users perceive and use new technologies?
- Which are the most important UX or usability problems in current products and services?

Development (n=4). This group represents people who manage and design concrete products and services, such as product manager or software designer. This group emphasizes the first use of products and services:

- Which functions are needed the most?
- What are the first impressions (overall experience, first use) and level of satisfaction?

Care (n=5). This group represents people who manage and provide numerous product supports and maintenance services in online forums and in local support centers. In most cases they have direct connection to customers. This group has very a rich set of measurement needs. Major point of interest is out of box readiness with products and services.

- How easy it is to start using product and services?
- What is customer experience in support activities?

Quality (n=6). This group consists of quality managers and specialists, working with concrete products or in company wide quality development activities. Respondents in quality are particularly interested in developing the quality measurement practises, and understanding the users' perceptions about both products and support services:

- Which metrics should be applied for experienced product quality?
- What is the perceived performance of products and services?

Common Needs for User Measurement

In this section we provide a consolidated grouping across all responses, based on a mind map categorization.

User experience lifecycle

Measurable information is needed not only when the user is using the product for its original purpose, but also when the user is planning to buy a new device, when the new device is being taken into use and when there is a shift from an old device to a new device.

What should be measured?	Examples of measures
Pre-purchase	The impact of expected UX on purchase decisions
First use	Success of taking the product into use
Product upgrade	Success in transferring content from old device to the new device

Table 1. Measurement areas in UX lifecycle

Retention

Retention is a concept and also measurement describing the loyalty of the customers. It is assumed that good user experience leads to high retention. Retention information would tell us how many customers continue with the brand, how many newcomers there are in customer base, and how many customer leave the brand. Among retention topics we can see non-UX information needs, such as the ownership of previous devices.

What should be measured?	Examples of measures
Expectations vs. reality	Has the device met your expectations?
Long term experience	Are you satisfied with the product quality (after 3 months of use)
Previous devices	Which device you had previously?
Engagement	Continuous excitement

Table 2. Measurement areas in retention

User groups and use of features

"What product features are used by different demographic groups" is the information that is mentioned by respondents from Research and Care organizations. The question has two key themes:

- What kind of differences there are in the use of product function across different user groups?
- What functions are most used (and underused)?

What should be measured?	Examples of measures
Use of device functions	What functions are used, how often, why, how, when, where?
Differences in user groups	How different user groups access features?
Reliability of product planning	Comparison of target users vs. actual buyers?

Table 3. Measurement in the use of functions

Where are the experience breakdowns?

Breakdowns and concrete problems seem to have link to almost all our respondents. A common challenge is to identify breakdowns and measure the improvements.

What should be measured?	Examples of measures
UX Obstacles	Why and when the user experiences frustration?
Malfunction	Amount of "reboots" and severe technical problems experienced.
Usability problems	Top 10 usability problems experienced by the customers.

Table 4. Measurement areas in breakdowns

User experience in customer care and maintenance

Organizations providing customer care and other support need to measure how users experience the given support, for example, was it helpful?

What should be measured?	Examples of measures
Customer experience in "touchpoints"	How does customer think & feel about the interaction in the touch points?
Accuracy of support information	Does inaccurate support information result in product returns? How?

Table 5. Measurement areas in customer care

Experiences with user interface localization

English is the primary language in technology communication, and often English language variants are provided first. There is a need to understand and measure how far this is enough, and what is the value of localized content in terms of user experience.

What should be measured?	Examples of measures
Effect of localization	How do users perceive content in their local language, what does it mean to them, how do they feel about it?

Table 6. Measurement areas in localization

Perceived performance

In product development it is essential to know the performance targets, and how those targets are met in completed products.

What should be measured?	Examples of measures
Latencies	Perceived latencies in key tasks.
Performance	Perceived UX on device performance
Perceived complexity	Actual and perceived complexity of task accomplishments.

Table 7. Measurement areas in device performance

Experiences with new technologies

There are several information needs mentioned by the people who represent research teams. Especially the reactions towards new proposed or realized solutions are interesting.

What should be measured?	Examples of measures
Change in user behaviour	How are usage patterns changing when new technologies are introduced
Innovation feedback	New user ideas & innovations triggered by new experiences

Table 8. Measurement areas with new technologies

DISCUSSION

Limitations

Primarily the findings should be used as a new data for further UX measurement development and research activities. Our findings are not complete nor universal since the study was conducted in only one firm and with a limited number of respondents.

Attention can be paid to the low response rate, but in the phenomenographic study the high response rate is not as important as the saturation level achieved. Alexandersson's survey (1994) on more than 500 phenomenographic studies concluded that the variation of a phenomenon reached saturation at around 20 research participants. Our number of informants (18) is close to that figure, and we can argue that the saturation took place in our study.

Practical Implications

Most of the UX measurement needs are familiar and already handled in existing practises. However, in our view this study provides new information revealing common cross-organizational needs for measurements. When new UX measurements are developed or existing measurements are improved, there should be sufficient cross-functional review to find out who else would benefit of the measures, and who else could be already measuring related topics or collecting similar data. The same finding can be extended to consider also cross-firm perspective, such as for developing UX measures together with business partners, 'third parties' and developers.

Implications to Research

To our mind the data from our survey is consistent with the evolution of measurements that are visible in previous research from a few different disciplines. As the discipline of user experience is now forming, it is beneficial for the field to be aware of the kinds of metrics needed in industry. It is also healthy to start UX metrics work from the needs of the audience that will use the results. This hopefully helps UX researchers to establish the boundaries for UX measurements and even for UX as a discipline.

Our research results are still tentative. The current data requires more thorough analysis and discussion that compares our findings with others. Our research will continue to look deeper at the responses and contextual factors (organizational environment), with the aim to develop and generalize useful model for further research and development.

REFERENCES

1. Alexandersson, M. (1994). *Metod och medvetande*, Acta Universitatis Gothoburgensis, Göteborg.
2. Desmet, P.M.A., Overbeeke, C.J., Tax, S.J.E.T. (2001). Designing products with added emotional value: development and application of an approach for research through design. *The Design Journal*, 4(1), 32-47.
3. Gartner. (2007a). Valdes R. and Gootzit D (eds.). U*sability Drives User Experience; User Experience Delivers Business Values*.
4. Gartner. (2007b). Valdes R. and Gootzit D (eds.). *A Value-Driven, User-Centered Design Process for Web Sites and Applications*.
5. Hassenzahl, M. (2003). The thing and I: understanding the relationship between user and product. In M.Blythe, C. Overbeeke, Monk, A. & Wright, P. (Eds.), *Funology: From Usability to Enjoyment* (pp. 31-42). Dordrecht: Kluwer.
6. ISO 13407:1999, *Human-Centred Design Proceses for Interactive Systems*. International Standardization Organization (ISO), Switzerland.
7. Jordan, P.W. (1998). *An Introduction to Usability*. Taylor & Francis.
8. Jordan, P.W. (2002). *Designing Pleasurable Products: An Introduction to the New Human Factor*s. Taylor & Francis.
9. Mandryk, R., Inkpen, K., Calvert, T.W. (2006). Using psychophysiological techniques to measure user experience with entertainment technologies. *Behaviour & Information Technology*, 25, 2, 141 – 158.
10. Marton. F. and Booth S. (1997). *Learning and awareness*. Lawrence Erlbaum, Mahvah N.J.
11. Meho L.I. (2006). E-Mail Interviewing in Qualitative Research: A Methodological Discussion. *Journal of the American Society For Information Science And Technology*. August 2006., Wiley Periodicals.
12. Nielsen J. (1993). *Usability engineering*. Academic Press. New York.
13. Nokia Corporation (2005). *Inspired Human Technology*. White paper available at http://www.nokia.com/NOKIA_COM_1/About_Nokia/Press/White_Papers/pdf_files/backgrounder_inspired_human_technology.pdf
14. Norman, D. (2004). *Emotional Design: Why We Love (or Hate) Everyday Things*. Basic Books.
15. Peppard, J. (2000). Customer relationship management (CRM) in financial services, *European Management Journal,* 18, 3, 312-327.
16. Shackel, B. (1991). Usability—context, framework, definition, design and evaluation. In *Human Factors For informatics Usability*, B. Shackel and S. J. Richardson, Eds. Cambridge University Press, New York, NY, 21-37.

Combining Quantitative and Qualitative Data for Measuring User Experience of an Educational Game

Carmelo Ardito[1], Paolo Buono[1], Maria F. Costabile[1], Antonella De Angeli[2], Rosa Lanzilotti[1]

[1]Dipartimento di Informatica
Università di Bari
70125 Bari, Italy
{ardito, buono, costabile, lanzilotti }@di.uniba.it

[2] Manchester Business School
The University of Manchester
Po BOX M15 6PB - UK
Antonella.de-angeli@manchester.ac.uk

ABSTRACT

Measuring the user experience (UX) with interactive systems is a complex task. Experiences are influenced not only by the characteristics of an interactive system, but also by the user's psychological state and the context within which the interaction occurs. Human-Computer Interaction researchers have emphasised the importance of quantitative measures for providing useful, valid and meaningful assessments. Until now, there is no consensus among researchers as to which specific techniques should be used for evaluating UX. This paper presents an evaluation methodology that combines different techniques to measure the UX. Some of this techniques provides quantitative data, others provides qualitative data. The method was applied in the evaluation of Explore!, a mobile learning system implementing a game to be played by middle school students during a visit to an archaeological park. Results demonstrate the importance of triangulating qualitative and quantitative evaluation data to provide a clearer assessment of the UX.

INTRODUCTION and motivation

Measuring the user experience (UX) with interactive systems is a complex task. Experiences are influenced not only by the characteristics of the interactive system (e.g. complexity, usability, functionality, etc.), but also by the user's internal states (such as predispositions, expectations, needs, motivation, mood, etc.), and by the context (or the environment) within which the interaction occurs (e.g. organisational/social setting, meaningfulness of the activity, voluntariness of use, etc.) [8]. Thus, the evaluation of UX has to consider not only functional aspects of the system (e.g., fast, easy, functional, error-free performance), but it has also to measure non-functional characteristics such as attractiveness (harmonious, clear), emotion (affectionate, lovable), engagement (fun, motivation and stimulation).

The evaluations of mobile systems open further challenges. User performance can be disrupted by external circumstances, such as noise, distractions, and competition for resources in multi-task settings, or other surrounding people. These factors are difficult to predict and can only be observed in the field. In fact, the comparison of laboratory and field evaluations reported in [7] demonstrated that users behave more negatively in the field than in the laboratory. Users also take longer time to perform certain tasks and also reported more negative feelings, such as dissatisfaction and difficult of use, to the use of the device in the field.

HCI (Human-Computer Interaction) researchers have long emphasised the importance of quantitative measures for providing useful, valid and meaningful evaluation. Until now, there is little consensus among researchers as to which specific techniques should be applied to understand the UX [4]. We think that different techniques should be used to measure different aspects of the UX. This approach has been followed in the evaluation of Explore!, an m-learning system implementing a game to be played by middle school students during a visit to an archaeological park. We have conducted a field study in which a range of different techniques were used for comparing the experience of playing the game with and without technological support. The evaluation study and its results are illustrated in details in [3]. We summarise here the main findings, discussing how quantitative and qualitative techniques should be integrated to provide a view of UX .

EXPLORE!

Explore! is an m-learning system that exploits imaging and multimedia capabilities of the latest generation mobile devices. Explore! implements an electronic game that support learning of ancient history, transforming a visit to archaeological parks into an engaging and culturally rich experience [1, 2]. The m-learning technique implemented in Explore! is inspired by the excursion-game, an educational technique that help students to learn history while they play in an archaeological park [5, 6].

VUUM2008, June 18, 2008, Reykjavik, Iceland.

In the study we have evaluated the "Gaius' Day in Egnathia" game, designed to be played in the archaeological park of Egnathia in Southern Italy. Gaius' Day is structured like a treasure hunt to be played by groups of 3-5 students, which have to discover meaningful places in the park following some indications. The game consists of three main phases: introduction, game, and debriefing. In the introduction phase, the game master gives a brief description of the park and explains the game. In the game phase, groups of students explore the park to discover some interesting places. This phase follows the metaphor of a treasure hunt, where students have to locate physical places by some cues provided to them in written messages. These locations have to be recorded on a map. Finally, in the debriefing phase, the knowledge which is implicitly learned during the game is reviewed and shared among students.

EVALUATION

The main objective of the evaluation was to compare the pupils' experience while playing Gaius' Day in its original paper-based version with the experience of the electronic version of the game. The results of this are reported in [3]. In this paper, we focus on describing the different techniques we applied in the study and discuss their relative advantages and disadvantages for measuring the user experiences of mobile learning.

The game type (paper-based vs. mobile) was manipulated between-subjects. Nineteen students, divided into 5 groups, played the paper-based version of the game; 23 students, divided into 6 groups, played the mobile version. The main difference between the conditions was the way in which individual missions were communicated to the students. In the mobile version of the game, the missions were communicated to students by text-messages and the location was recorded in the mobile phone by entering a code number identifying the place. The mobile also provide some contextual help, in the form of an 'Oracle', displaying a glossary of the terms directly related to the active mission.
In the paper-based condition, the missions were communicated to the participants in a letter format distributed at the beginning of the game. Participants had to record the right location on a map. A glossary containing the full list of terms was also delivered at the beginning of the study.

Instruments

In order to evaluate the overall user experience with the Gaius' Day game a wide range of different techniques were exploited, specifically: naturalistic observations, self-reports (questionnaires, structured interviews, and focus groups), post-experience elicitation techniques (drawings and essays), and multiple choice tests. A summary of the main factors addressed by our evaluation, along with the method and techniques applied is reported in Table 1. The use of these different techniques allowed us collect both qualitative data and quantitative data, that have been triangulated to provide a global view of the user experience.

Observations were based on event-sampling, an approach whereby the observers record all instances of a particular behaviour during a specified time period. Each group of children was shadowed by two independent observers, who had received in-depth training on data-collection. The events of interest referred to problem-solving strategies, social interaction processes (including collaboration and competition), and interaction with the artefacts (mobile, map, and glossary). These events were recorded in an observation grid organized on the basis of mission and time.

Behaviour	Naturalistic observations; questionnaire; focus group; essays and drawings.
Engagement	Open and closed questionnaires; naturalistic observations; focus group.
Learning	Observations during the debriefing session; questionnaire: learning self-assessment immediately after the debriefing; multiple-choice test administered in school on the day after the visit; essay writing in school.

Table 1. Instruments and techniques used in the evaluation

Two questionnaires were developed for this study based on the QSA, an Italian questionnaire measuring learning motivation, strategies and behaviour [9]. The first questionnaire was administered individually at the archaeological park immediately after the game phase. It included 20 Likert-scale items, in the form of short statements regarding their game experience from the viewpoint of the following factors: collaboration, competition, motivation, fun, and challenge. Responses were modulated on a five point scale, ranging from 1 (strongly agree), to 5 (strongly disagree). The second questionnaire was administered immediately after the debriefing to measure participants' evaluation on the discussion session and their opinions on how much they learnt during the game.

Learning was assessed in school on the day after the visit, via a multiple-choice test requiring the memory of facts, and knowledge application. The test was designed by researchers of Historia Ludens in collaboration with the school teachers.

Evaluation Study Results

Results addressing behaviour, engagement, and learning are reported in following subsections.

Behaviour

Efficiency and effectiveness during the game were analysed by traditional quantitative measures, namely time needed to complete the games and percentage of correct answers to

the missions. Both variables were significantly better in the paper-based condition of the game. Participants playing the game in the paper-based condition completed the challenge faster (mean = 29.5 minutes; std dev = 6.43) than those in the mobile condition (mean = 38.5 minutes; std dev = 7.66). The mobile condition was more prone to errors than the paper-based one, with a difference of some 20 percentage points.

The naturalistic observations performed during the game phase were instrumental in understanding the reasons for these differences. In fact, we observed that groups in the paper-based condition preferred to choose their missions according to their locations or contextual knowledge rather that following the order in which the missions were presented in the letter. Another parallel strategy, commonly adopted in the paper-based condition, was to read several items in the glossary at the same time, making it possible to compare the target details. Once again, this behaviour was not possible in the mobile condition, as the 'Oracle' displayed only the glossary entry directly related to the active mission.

Naturalistic observations were also used to understand problems with the game artefacts experienced during the game (i.e. cell phone, map, and glossary). In particular, we observed that the mobile groups had little problems in using the telephone, only in two cases at the start of the game did the technician need to intervene to explain how to use the phone. One group had difficulties in managing paper sheets, i.e. the wind complicated writing the answers. In both groups, students had some difficulties in reading the map, but the game master helped them in overcoming the problems.

Behavioural observations were also important in understanding social dynamics driving group behaviour. In particular, we looked at leadership, defined as the participants' willingness to take charge of the game, contributing ideas and suggestions and allocating tasks to the other members. This was done by analysis the notes collected in the field by a dedicated observer for each group and by video analysis. It was found that 50% of the participants who played the leader role in the mobile condition happened to be the one holding the cell phone, whereas no clear trends emerged in the paper-based condition.

Looking at participant's behaviour in the field it also appeared that the mobile groups were more competitive than the paper-based one. Usually, when they met they ignored each other and continued to carry out their own mission. They appeared very concentrated on their tasks and did not want to exchange any comments with their adversaries. The few questions they exchanged were aimed to get information that could be useful to them. Examples of such questions were: "*Have you found ****?*", "*Have you finished?*", "*What mission are you carrying out?*". In contrast, the paper-based groups were more talkative and often engaged in jokes and chit chat when they met. In general, however, winning appeared to be important to students, who often enquired about the other groups' performance during the game. We also witnessed a couple of attempts to cheat, where students tried to swap codes between different locations to make it impossible for others to win, or gave false answers to a direct enquiry.

Interestingly these differences did not emerge as significant in the analysis of self-reported items addressing sociability issues in the questionnaire.

Engagement

To analyse strengths and weakness of the game, we considered two open questions where participants reported the three best and the three worst features of the game. A total of 99 positive features were reported, and only 39 negative features. On the average participants in the mobile condition reported more positive features (mean per participant = 2.7) than the paper-based group (mean = 1.9). No differences in the number of negative features reported by the two groups emerged (mean =1).

Analysing the content of participants' self-reports it emerged a different trend in the two game conditions. The most frequently reported positive features in the mobile game addressed the artefacts used during the game, whereas in the paper-based condition participants referred most often to the archaeological park. A total of 11 out of 19 references to artefacts in the mobile condition directly addressed the cell phone or some interface features, such as the Oracle and the 3D reconstructions. Overall, the 3D reconstructions were given a score of 4.3 on a 5 point scale. One of the students commented "*The mobile and the game are a winning combination*".

The collaborative nature of the game was indicated as another winning factor by both groups. Children enjoyed playing together and demonstrated a good team spirit all over the game. The learning potential of the game was another positive factor in both conditions. Students, especially those in the mobile condition, appreciated the difficulty of the game: they enjoyed because "*It was challenging*", as reported by a participant in a focus group.

As regards negative features, the trend of results is more homogeneous between the experimental conditions, although the mobile groups was more likely to complain about the difficulty of the game and the paper-based group was more likely to complain about the duration of the game, normally considered to be too brief.

Learning

Learning was evaluated by a range of different techniques including self-evaluation and formal assessments performed

with multiple choices, essays writing and drawing. In the following some data are reported. On average the students were very positive about the educational impact of the systems and they all agreed they had learned something (mean = 4.1; std dev = 0.66). No significant differences emerged in the group comparison. Participants' opinions were confirmed by an objective test. A total of 36 tests were returned for analysis (21 from the mobile condition; 15 from the paper-based one). On average, students answered 9 out of 11 questions correctly (std dev = 1.65). No significant differences between the game conditions emerged. The distributions are strikingly similar (mobile mean=8.8 (SE=.38); paper-based mean=9 (SE=.40)).

Discussion

Summarizing, about the behaviour aspect, the evaluation study revealed that different problem solving strategies emerged. In the paper-based condition, students changed the mission order, either firstly performing those missions they perceived as easier or according to a personal strategy; while in the mobile condition students had to solve one mission after the other. This could be one of the reasons why, in the paper-based condition, students completed the challenge in less time and with less errors. From this result we have learnt that mobile games require more interaction freedom and context-dependent information to enhance the overall user experience.

The evaluation has also demonstrated that users enjoyed playing the game and, although we could not demonstrate the expected superiority of the mobile game condition by statistical comparison of questionnaire data, the introduction of the mobile appeared to be much appreciated as demonstrated by qualitative data. The use of the mobile was directly acknowledged as one of the best features of Explore!. We expect that, as we add to the interactive features of the mobile, we will also improve the user experience with Explore!.

Finally, no difference in learning between the two conditions was found. This can be explained by ceiling effect, as the participants' performance in both conditions was very high. We think that it is not a bad result because students are not distracted by the technology and also that e-learning is as equally valuable as traditional learning provided that appropriate techniques are used such as excursion-game, which is able to engage and stimulate students.

CONCLUSION

In this paper, we have illustrated an approach for the evaluation of m-learning. The study shows the importance of triangulation of different techniques and measures to capture all the different aspects that the user experience involves. Our understanding of the game was supported by a combination of qualitative data, collected through naturalistic observation, focus group, analysis of essays, and drawings, and quantitative data, collected through questionnaire, multiple-choice learning test, behavioural analysis.

The study results have highlighted that in some domains, such as the mobile one, collecting and analyzing qualitative data gives the possibility to discover UX aspects that should be neglected by only considering data that are quantitatively measurable. The outcome of the study suggests that in order to obtain *meaningful, useful and valid* results, qualitative and quantitative data should be combined. It is worth mentioning that qualitative data helped us explain some unexpected results obtained by quantitative data. For example, regarding engagement aspect, no difference emerged from quantitative data. Otherwise, from naturalistic observation, focus group, we can state that students enjoyed the use of cell phone and appreciated very much the 3D reconstructions both on phone during the game phase and during debriefing.

Quantitative data did not demonstrate the advantage of the electronic game. But, during the focus group, participants of the mobile condition referred that they were not distracted by the technology, while participants of the paper-based condition reported that paper annoyed them it was a windy day. Questionnaire revealed to be a useful source of information, but failed to discriminate between the experimental conditions. This may be due to a ceiling effect, as evaluations in both conditions were very high.

ACKNOWLEDGEMENTS

Partial support for this research was provided by Italian MIUR (grant "CHAT").

REFERENCES

1. Ardito C., Lanzilotti R. *Isn't this archaeological site exciting!": a mobile system enhancing school trips*. AVI 2008. Napoli, Italy, May 28-30, 2008. ACM Press (2008). In print.
2. Ardito C., Buono P., Costabile M., Lanzilotti R., and Pederson T. *Mobile games to foster the learning of history at archaeological sites*. VL HCC 2007. Coeur d'Alène, Idaho, September 23-27, 2007, 81-84.
3. Costabile M., De Angeli A., Lanzilotti R., Ardito C., Buono P., Pederson T. *Explore! Possibilities and challenges of mobile learning*. CHI 2008. Florence, Italy, April 5-10, 2008. ACM Press (2008), 145-154.
4. Bernhaupt, R., et al. (Eds). Proc. of workshop "*Methods for evaluating games-How to measure usability and user experience in games*", at ACE 2007.
5. Cecalupo, M., Chirantoni, E. *Una giornata di Gaio, in Clio al lavoro*. Technical report, Didattica della Storia, Università degli Studi di Bari, (1994), 46-57.
6. Ciancio, E., Iacobone C. *Nella città senza nome. Come esplorare l'area archeologica di Monte Sannace*, Laterza, Bari, 2000.
7. Duh, H., Tan, G., and Chen, V. *Usability evaluation for*

mobile device: a comparison of laboratory and field tests. Proc. MobileHCI'06, Helsinki, Finland, September 2006, ACM PRESS, New York, 2006, 181-186.

8. Hassenzahl, M., Tractinsky, N. User experience - a research agenda. *Behaviour & Information Technology*, 25, 2 (Mar-Apr 2006), 91-97.
9. Pellerey, M. *Questionario sulle Strategie d'Apprendimento* (QSA), LAS, Roma, 1996.

Is User Experience supported effectively in existing software development processes?

Mats Hellman
Product Planning User Experience
UIQ Technology
Soft Center VIII
SE 372 25 Ronneby, Sweden
Mats.hellman@uiq.com
+46708576497

Kari Rönkkö
School of Engineering
Blekinge Institute of Technology
Soft Center
SE 372 25 Ronneby, Sweden
Kari.ronkko@bth.se
+46733124892

ABSTRACT
Software process is one important element in the contextual foundation for meaningful and useful usability and UX measures. In this base usability has been around for a long time. Today the challenge in the mobile industry is User experience (UX), which is starting to affect software engineering processes. A common use or definition of the term UX is still not de facto defined. Industry and academy are both in agreement that UX definitely includes more than the previous usability definition. How do industry and manufacturers manage to successfully get a UX idea into and through the development cycle? That is, to develop and sell it in the market within the right timeframe and with the right content. This paper will discuss the above challenges based on our own industrial case, and make suggestions on how the case related development process needs to change in order to support the new quality aspects of UX.

Author Keywords
Software development processes, User Experience, Usability, Usability test, Management, Hedonic, Mobile, Product development, product validation, Concept validation

ACM Classification Keywords
H5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous. D.2.8 Metrics, D.2.9 Management.

INTRODUCTION
Mobile phones have reached a point beyond the level where technical hot news is not enough to satisfy the buyers, for today mobile devices also have to include the aspects of user experiences. Apple's iPhone is one indication of this change. Software engineering strives to make complex things manageable.

VUUM2008, June 18, 2008, Reykjavik, Iceland.

Engineering most often attempts to split the product complexity into smaller more manageable sub-functions. In the end all the sub-functions are put together and a product appears, hopefully as the designer or the idea maker intended. Deviations from the intended product idea are handled through iterated defect reporting and defect handling until the product is judged to have sufficient product quality. Hence, monitoring product quality is conducted by processes in which milestone criteria are measured mainly by different ways of controlling defect levels and defect status. So far this approach has been sufficient enough when striving to secure a product's quality from a task and goal perspective (classic usability view), but still no guarantee for enhancing the user experience (that increases the chances of product success). In the goal and task view three canonical usability metrics have dominated, i.e. effectiveness, efficiency and satisfaction. Where the latter, satisfaction, has been a term capturing the feeling experience on a very high level, i.e. without further dividing it into its diverse constituent elements. Today a new level of quality needs to be handled, i.e. user experience. Handling this quality forces us to divide satisfaction into soft values such as: fun, pleasure, pride, intimacy, joy, etc. [8, 4]

One problem that follows with this new approach towards quality is the risk of losing the holistic product view if we split it into smaller manageable sub-functions in the production process. In the quality of user experience apparently small changes made in different subparts can actually constitute a huge user experience change when put together in the final product. It is also difficult to predict the effects of such separately handled changes. Example of this could be that applications in the past have been more or less separate entities or islands in a mobile product. This has provided opportunities for application designers and engineers to apply their own solutions and create their own application specific components with "isolated" specific behavior to support a use case. Such isolated behavior can and will be a big threat to the total UX of a product. In this aspect designers and engineers need a generic framework of deliverable components to make sure that the total UX of

the product is consistent throughout the product development.

Pushing out ownership and responsibility to the separate parts is a common management strategy. Are organizational models that push ownership out to the leaves in organization really effective in the mobile industry with its many actors and stakeholders? Doesn't this model encourages handling risks via a focus on each constituent part rather than a holistic view on the end product or are there better and more efficient ways of making an idea appear in a product? Ways that could shorten the time to market, minimize the risk of fragmentation of the product, and in new effective ways help organizations to prioritize and secure a successful UX in a product. How can we maintain a holistic perspective despite multiple splits of functionality during development? For the goal and task related usability paradigm models simple division and delegation has been successful. In this era with a growing need for high level monitoring of UX in products we are still left with the goal and task oriented development models. Sorry to say but having good quality on each different part may not be sufficient and definitively not a guarantee for a good and successful end product. We need to find new ways to measure and monitor this combined quality aspect. To support UX efficiently a process with a clear product focus is needed in parallel with application development processes. Otherwise, because of the prevailing task and goal tradition, there is a risk that we talk about a holistic product view but in practice end up monitoring small identities. Still, we believe the engineering approach of separation is powerful and necessary in large projects. So - what are the possible approaches for ensuring an idea appears throughout the prevailing engineering approach of separating the development? The introduction of an overall design process is the solution we advocate, as a frame surrounding existing goal and task related development processes.

RESEARCH COOPERATION

The industrial partner is UIQ Technology AB, a growing company with currently around 330 employees, situated in the Soft Center Science Research Park, Ronneby, Sweden. The company was established in 1999, and has as one of its goals the creating of a world leading user interface for mobile phones. Their focus is "to pave the way for the successful creation of user-friendly, diverse and cost-efficient mobile phones" [12]. They develop and license an open software platform to leading mobile phone manufacturers and support licensees in the drive towards developing a mass market for open mobile phones. Their product, UIQ, is a media-rich, flexible and customizable software platform, pre-integrated and tested with Symbian OS, providing core technologies and services such as telephony and networking.

The research group U-ODD – Use-Oriented Design and Development [13], belongs to the School of Engineering at Blekinge Institute of Technology (BTH), and is a part of the research environment BESQ [1]. The research performed by U-ODD is directed towards use-orientation and the social element inherent in software development. Studies performed by U-ODD are influenced by the application of a social science qualitative research methodology, and the application of an end-user's perspective.

The process of cooperation is Action research according to the research- and development methodology called Cooperative Method Development (CMD), see [3] [11], chapter 8) for details.

HOLISTIC PRODUCT VIEW

There is a risk of losing the UX intent of a product if no support structure is in place. In order to keep the organization "mean and lean" and at the same time deliver UX focused products we need to secure the vision of a product throughout the development process. Today many companies have developed methods to validate concepts of the final product with end users. UIQ technology uses for instance their UTUM method. [12]. Unfortunately these kind of validation activities are too often handled by and within a UI Design/Interaction Design group and not as part of the overall design process, e.g. as ad hoc help in the design work at different stages.

Figure 1 visualizes the separateness product vision into many divided requirements and thereby risk not monitoring UX in a holistic way; it also represent today's goal and task oriented development models. The outcome/product includes the risk of becoming something that was not intended.

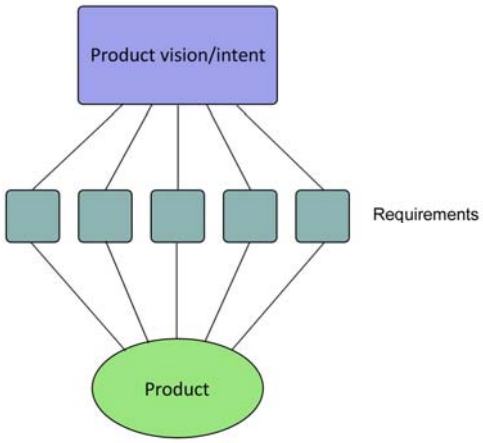

Figure 1.

How do we then keep the end user product in focus in a large product organization? Today most companies in the Mobile industry face the challenge of increasing demands on UX products. A lot of effort in improving the UX

capabilities in the software is undertaken but there are no new UX criteria included in the measurements of product quality even if most companies have both verbal and written UX statements and visions on their walls as lead goals for their businesses. A product quality definition is still based on different levels, measurements and predictions of defects as criteria, and seldom includes usability and/or UX quality criteria. This means there is no connection or possible way of measuring the "temperature" of UX in the product during the development between vision and final product. There is also a divergence between UX quality and existing product quality, meaning that we have processes and means (traditional Software Engineering) to monitor product quality by defects, but these are not a guarantee for achieving an envisioned high level of UX in the final product. The only way is to change the culture to become more UX and product focused and not as, in many cases today, focused on part delivery.

Therefore a more appropriate way to inject UX quality assurance into the development process would be by:

1. Gaining acceptance of a **vision** through user research with end users by means of methods like early prototype testing.

2. **Policing** the vision throughout the development process by internal review methods to secure UX product quality. UX quality criteria and milestones should be included in an overall design process influencing the development process. A new quality assurance role needs to be created for UX experts to act as guardians for the UX quality.

3. **Validating** the product and evaluating the result against the vision, again by formalizing existing methods like UTUM [1] in the development process. This is also visualized in figure 2:

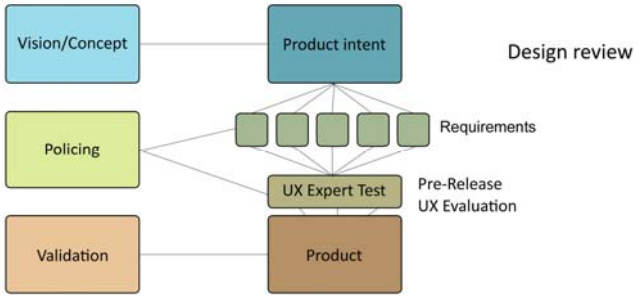

Figure 2.

An organizational set up like the one described in Figure 2 would be a better guarantee that the product vision and intent is what will be delivered in the end compared with an organizational set up as described in figure 1. Meaning that the whole of the organization needs to understand and prioritize the end result. The way to secure product quality and to include UX into the product quality aspect has to be to introduce "UX guards" in all levels of development. Their role would need to be to police the fulfillment of the UX quality criteria in the process defined and decided checkpoints. These checkpoints could e.g. be expert reviews of requirements and expert UX reviewers to get the authority to set a pass/not pass stamp on the intended delivery. This needs to be agreed and formalized into the development process.

UX DEFINITION AND MEASUREMENT

Beside the introduction of an overall design process monitoring the UX visions we also need to develop ways to both define and measure it. Today we use the UTUM method to measure usability according to efficiency, effectiveness and satisfaction and visualize the result for management as shown in figure 3.

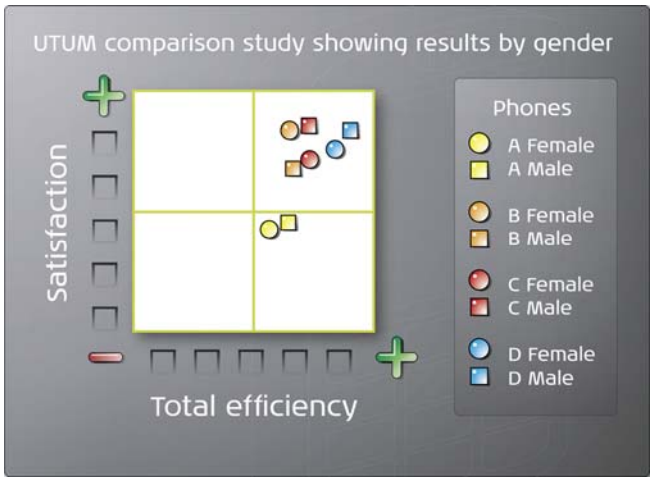

Figure 3.

P. Jordan [6] writes about Functionality, usability and pleasure as the essential ingredients for a successful product or service to provide a consumer experience, see figure 5. (See also [7] for discussion of contextual factors) Jordan explains the three levels of consumer needs that have to be supported as:

Functionality: Meaning that a product without the right functionality is useless and causes dissatisfaction. A product has to fulfill the needs of a user.

Usability: When a product has the right functionality the product has to be easy to use. A user doesn't just expect the right functionality they expect ease in use.

Pleasure: When the user is used to usable products they want something more. Meaning, when functionality and usability have become levels of plain "product hygiene" they want products that just not bring functional benefits but also give the user emotional benefits.

We think Jordan's view on UX is a good starting point and is in line with our thoughts and present thinking. For us the Hedonic side of an experience is much about the users

experience and the Pragmatic side is about the user's expectations.
Meaning that the Hedonic aspects of UX is more about how the product creates pleasurable and exciting experiences that delights the user in sometimes unexpected but hopefully positive ways compared with
the Pragmatic aspect of UX that is about how well/bad a product lives up to the users expectation of that product's in functionality, ease-of-use etc.
This is also applicable to our thinking in this paper about the need to increase the product quality aspect. By developing methods to understand users in all three areas, and by making it possible to also measure pleasurable aspects of a product, it would be possible to measure UX product quality.

Today we measure in and cover the areas of "Usability" and "Functionality" but just partly in the area of "Pleasure" by doing traditional ISO standard measurements as Efficiency, effectiveness and user satisfaction. But we propose a shift in to include aspects of "Hedonic", "Pragmatic" and "Business cost" (Development cost, infrastructure, technology maturity etc.) in measuring UX. It is also vital to understand the context in which the product will be used, as visualized in figure 4.

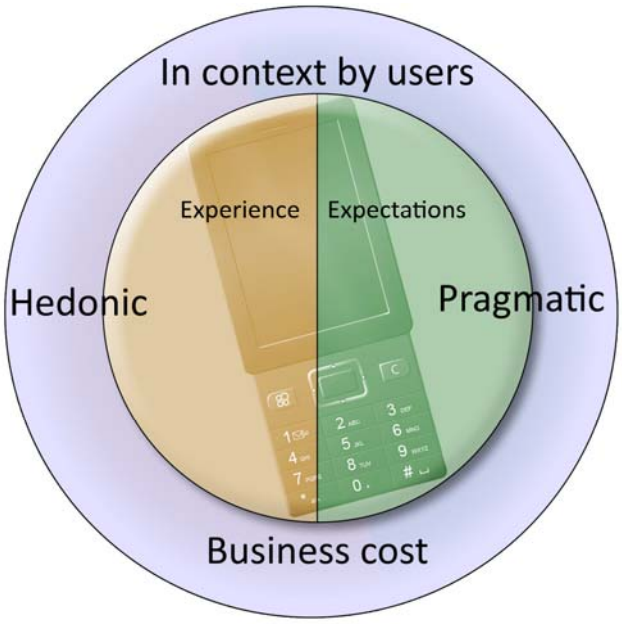

Figure 4

EXTENDING UTUM WITH UX QUALITY

UTUM is a usability test package for mass market mobile devices, and is a tool for quality assurance, measuring usability empirically on the basis of metrics for satisfaction, efficiency and effectiveness, complemented by a test leader's observations.

The current test package is the result of several years of joint research cooperation and experimentation. During the period 2001 to 2004 an attempt was made, initiated by Kari Rönkkö from the research group U-ODD from Blekinge Institute of Technology (BTH), to use Personas in UIQ's development processes. The intention was to find something that could bridge the gap between designers and developers and others within the company, and find a way of mediating between many different groups within the company. The Personas attempt was finally abandoned in 2004, as it was found not to be a suitable method within the company, for reasons which can be found in [10]. In 2001, Symbian company goals included the study of metrics for aspects of the system development process. An evaluation tool was developed by the, at that point of time, head of the interaction design group at Symbian, Patrick W Jordan, together with Mats Hellman from the Product Planning User Experience team at UIQ Technology, and Kari Rönkkö from U-ODD/BTH. The first part of the tool consisted of six use cases to be performed on a high fidelity mock-up [9] within decided time frames. The second part of the tool was the System Usability Scale (SUS) [2].

The first testing took place during the last quarter of 2001. The goal was to see a clear improvement in product usability, and the tests were repeated three times at important junctures in the development process. The results of the testing process were seen as rather predictable, and did not at this time measurably contribute to the development process, but showed that the test method could lead to a value for usability. During the period 2004 to 2005, a student project was performed, supported by cooperation with representatives from U-ODD, which studied how UIQ Technology measured usability. The report from the study pointed out that the original method needed improvements and that the process should contain some form of user investigation, a way of prioritizing use cases, and that it should be possible to include the test leader's observations in the method. The test method was refined further during 2005, leading to the development of UTUM v 1.0 which consisted of three steps. First, a questionnaire was used to prioritize use cases and to collect data for analysis and statistical purposes. The use cases could either be decided by the questionnaire about the user's usage of a device (in this choice Jordan's level of functionality is visible) or in advance decided by the company if specific areas were to be evaluated. After each use case the user performs a small satisfaction evaluation questionnaire explaining how that specific use case supported their intentions and expectations. Each use case is carefully monitored videotaped if found necessary and timed by the test expert. The second step was a performance metric, based on completion of specified use cases, resulting in a value between 0 and 1. The third was an attitudinal metric based on the SUS, also resulting in a value between 0 and 1. These values were used as parameters in order to calculate a Total Usability Metric with a value between 0 and 100. Besides these summative

usability results the test leader also through his/her observation during the test can directly feed back formative usability results to designers in the teams within the organization, giving them user feedback to consider in their improvement and redesign work.

It is also the test expert observation that is decisive in what the trade offs are in the test. One example could be a use case that takes too long to perform compared to the norm, which still is rated as high usability on the satisfaction scale, by the user. It is then up to the test expert to explain and describe that trade off.

In 2005, the work with UTUM gained impetus. A usability engineer was engaged in developing the test package, and a new PhD student under Rönkkö's supervision from the research group U-ODD became engaged in the process and began researching the field of usability, observing the testing process, and interviewing staff at UIQ. In an iterative process in 2005 and 2006, the method was further refined and developed into UTUM 2.0. UTUM 2.0 is "a method to generate a scale of measurement of the usability of our products on a general level, as well as on a functional level." [12]. The test package was presented for the industry at the Symbian Smartphone show in London, October 2006, and the philosophy behind it is presented in full detail on UIQ's website. [12] Since UTUM is available via the Internet, we do not provide detailed instructions of the testing procedures in this paper.

Increasing the UTUM test package to a more in depth focus on UX and measurements of pleasurable and "holistic" aspects through for example, questionnaires would help the organization to focus on the UX aspects within the software development process. The visual presentation for management for the achieved UX in a product at a specific time could then be presented as a radar diagram containing the current status in each area of UX. (See Figure 5)

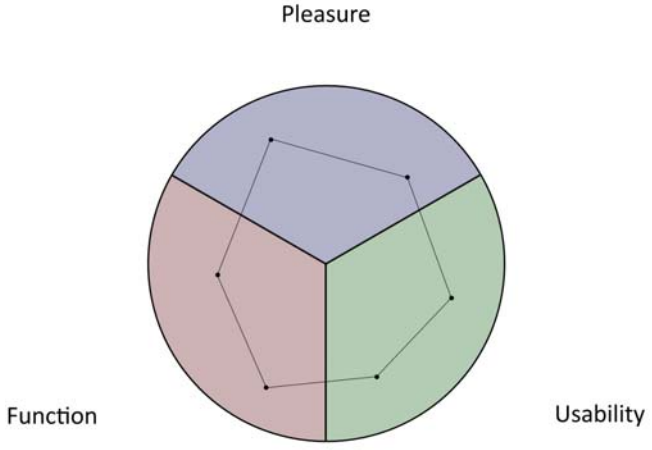

Figure 5

This could be a way to include UX quality measurements into the software development process as a quality aspect, to act upon the result in a similar way as we do with defects. This would also be a possible way to undertake regular measurements along with existing processes and using existing forums for displaying and acting on the received result.

NEW PROCESSES NEEDED

Today many organizations, including UIQ Technology AB, have decided to work in multidisciplinary teams to secure quality and focus on deliveries. It is our belief that this is necessary, but there is a big risk that an organization that focuses on securing component quality loses sight of the actual holistic product intent, meaning that the delivered product doesn't match the actual product intent. In order to secure the vision of product intent in complex and multi requirement projects the organization needs to acknowledge the need for policing. Not just defect levels, but also and maybe even more important, the holistic product intent throughout the development cycle and in all different teams participating in the development process.

This is needed to secure an efficient and effective way of working towards a successful product. Today it becomes more and more important to deliver UX focus products faster and faster, whereby it also becomes vital for an organization to decrease development times.

"Everything about mobile phone design and production has to be quick, so it's months from when there is an idea for a phone to the roll out on the market," said James Marshall, Sony Ericsson's head of product marketing, who is in Las Vegas this week for the trade fair. "The market moves very quickly, so you have to minimize development times."[5]

Our suggestion is that companies organize in such a way so that UX requirements developed by end user understanding and user knowledge are monitored throughout the development cycle. Companies need to develop a better holistic understanding of their product, and the attempt that product should support and aimed for, even if they optimize their efforts in multi-disciplinary teams responsible only for specific components in a system. This could be done by having UX guarding functionality in leading positions in the development process. People that monitor the holistic view of the product and who have the mandate take necessary actions whenever it is needed to secure the product intent.

CONCLUSION AND SUMMARY

A new approach to this problem that focuses on the organization per se, as well as ways of working, is needed. By prioritizing UX in products a new perspective needs to be adopted if the intended result should appear in products in the market and appreciated by users. On a high level there needs to be a cultural shift into a more UX oriented and UX driven mentality within the whole product development organization. On an organizational level UX

quality assurance needs to be established by recognizing and given authority to UX expertise that can secure the total UX product quality in all levels of development. In this paper we want to throw light upon the fact that if mobile manufactures want to be competitive and able to deliver high quality UX products fast and with regularity and predictability the need for UX "guardians" are necessary, see Figure 6.

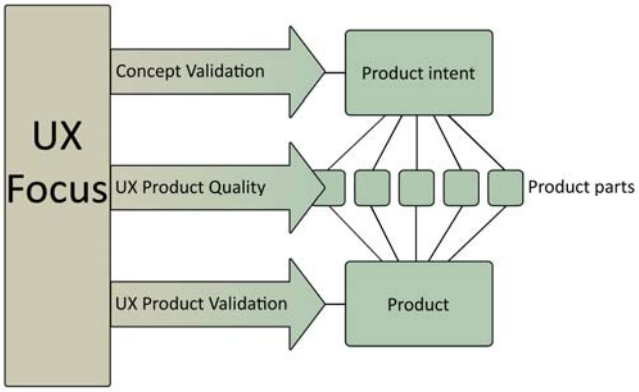

Figure 6

We also point out and argue the need for both cultural as well as organizational changes. Today we monitor and define product quality by measuring defects levels in different ways. This will still be needed but must be complemented by UX quality measurements. The product quality definition needs to be increased and widened to include measurements from the UX area and these new quality criteria need to be accounted for with the same priority as other quality criteria. More organizational effort should be spent on developing Metrics and KPI's for monitoring and securing UX product quality.

ACKNOWLEDGMENTS

We wish to thank Gary Denman from UIQ Technology AB for providing valuable support and insights. This work was partly funded by Vinnova (a Swedish Governmental Agency for Innovation Systems) under a research grant for the project WeBIS [14]; and the Knowledge Foundation in Sweden (work to make Sweden more competitive) under a research grant for the software development project "Blekinge – Engineering Software Qualities"[1].

REFERENCES

1. BESQ. Blekinge - Engineering Software Qualities, http://www.bth.se/besq , 2008-05-22.
2. Brooke, J., "System Usability Scale (SUS) - A quick and dirty usability scale, http://www.usabilitynet.org/trump/documents/Suschapt.doc , 2008-05-22
3. Dittrich, Y., Rönkkö, K., Erickson, J., Hansson, C. and Lindeberg, O. Co-operative Method Development: Combining qualitative empirical research with method, technique and process improvement, *in the Journal of Empirical Software Engineering*, Springer Netherlands, (online 18 Dec 2007) 1382-3256 (Print) 1573-7616 (Online), 2007.
4. Hassenzahl, M. and Tractinsky, N. User experience – a research agenda, in Behaviour & Information Technology, Vol. 25, No. 2, March-April, pp. 91 – 97, 2006
5. International Herald Tribune/Technology and Media. *Iphone pushes mobile makers to think simpler*. By Eric Sylvers. http://www.iht.com/articles/2008/01/09/technology/wireless10.php, 2008-05-22
6. Jordan, Patrick W., 2008. The four Pleasures: Understanding User's holistically. In proceedings of applied ergonomics International, (AHFE International), 2008.
7. Karapanos, E., Hassenzahl, M., and Martens, J. User experience over time. In *CHI '08 Extended Abstracts on Human Factors in Computing Systems* (Florence, Italy, April 05 - 10, 2008). CHI '08. ACM, New York, NY, 3561-356, 2008.
8. Law, E., Hvannberg, E. and Hassenzahl, M. , User Experience: Towards a Unified View, in proceedings of the NordiCHI 2006 Workshop, Oslo, Norway, 14 Oct. 2006.
9. Rettig, M., Prototyping for tiny fingers," Communications of the ACM, vol. 37, pp. 21 - 27, 1994.
10. Rönkkö, K., B. Kilander, M. Hellman, and Y. Dittrich, "Personas is Not Applicable: Local Remedies Interpreted in a Wider Context," presented at Participatory Design Conference PDC '04, Toronto, Canada, 2004.
11. Rönkkö, K. Making Methods Work in Software Engineering: Method Deployment as a Social achievement *School of Engineering*, Blekinge Institute of Technology, Ronneby, 2005.
12. UIQ Technology. UIQ Technology Usability Metrics, UIQ Technology, http://uiq.com/utum.html. 2008-05-22
13. U-ODD. Use-Oriented Design and Development, http://www.bth.se/tek/U-ODD, 2008-05-22
14. WeBiS, http://www.webis.se

On Measuring Usability of Mobile Applications

Nikolaos Avouris, Georgios Fiotakis and Dimitrios Raptis
University of Patras, Human-Computer Interaction Group
GR-26500 Rio, Patras, Greece
avouris@upatras.gr, fiotakis@ece.upatras.gr, draptis@ece.upatras.gr

ABSTRACT
In this paper we discuss challenges of usability evaluation of mobile applications. We outline some key aspects of mobile applications and the special characteristics of their usability evaluation that have recently lead to the laboratory vs field discussion. Then we review the current trend and practices. We provide an example of a usability evaluation study from our own background. We conclude with discussion of some open issues of usability evaluation of mobile applications.

Author Keywords
Usability evaluation measures, mobile applications, field evaluation studies, user studies.

ACM Classification Keywords
H5.2. Information interfaces and presentation (e.g., HCI): User Interfaces – Evaluation/Methodology, H5.4 Information interfaces and presentation (e.g., HCI): Mobile Applications, Usability Evaluation Methods, Usability Measures.

INTRODUCTION

Usability evaluation of mobile applications is an active area of research and practice in continuous evolution, due to the high demands and the evolving context in which mobile applications are developed and used. A fundamental concern of researchers and practitioners of this area is the use of adequate usability methods and measures. This has taken recently the form of a debate between the advocates of field versus laboratory study approaches, e.g. [18] and [19], that relates to qualitative vs quantitative measures. This debate seems to have its roots in the historical distinction between subjective and objective knowledge and corresponding epistemological implications, taking the form of a discussion on the merits of the scientific versus the design research method [4].

VUUM2008, June 18, 2008, Reykjavik, Iceland.

As Annett [2] discusses in the debate paper of a special issue on subjective versus objective method in ergonomic practice: "*..all knowledge is based on subjective experience. What really matters in establishing scientific truth is the method by which independent observers agree on the meaning of their individual observations. Complex features of the external world may be judged subjectively by experts, but the reliability of these judgments depends heavily on the use of agreed criteria … that provide the best assurance of inter-subjectivity*".

These agreed criteria however need to be based on deep understanding of the characteristics of mobile applications. So, to start with, the evaluation of mobile applications necessitates examination of the factors that affect user experience and interaction with such applications. It is therefore understandable that during this exploratory phase existing evaluation techniques involving established measures and qualitative approaches are mostly to be used. So, in a survey of usability attributes in mobile applications by Zhang and Adipat [31] nine attributes were identified that are most often evaluated: learnability, efficiency, memorability, user errors, user satisfaction, effectiveness, simplicity, comprehensibility and learning performance. All nine of them are well defined and extensively used measures of usability in more traditional desktop applications. Yet it is apparent that mobile applications introduce new aspects that need to be considered. We cannot limit the evaluation only to the device (typical scenario in desktop applications) but we must extend it, including aspects of context, which often bears dynamic and complex characteristics. There is the possibility that a single device is used in more than a single context, in different situations, serving different goals and tasks of a single or a group of users. Also, group interactions, a common characteristic in mobile settings, influence the interaction flow and increases the complexity of the required analysis as well as the necessity of analysis of complex observational data. The process of selecting appropriate usability attributes to evaluate a mobile application depends on the nature of the mobile application and the objectives of the study. So far, a variety of specific measures (e.g., task execution time, speed, number of button clicks, group interactions, support sought, etc.) have been proposed to be used for evaluation of different usability attributes of specific mobile applications, as discussed in the survey section of the paper. Is this due to the limited way of measuring user behavior and user

experience? Wilson and Nicholls [29] point out in discussing performance during the evaluation of virtual environments, "There are only a limited number of ways in which we can assess people's performance: we can measure the outcome of what they have done, we can observe them doing it, we can measure the effects on them of doing it or we can ask them about either the behavior or its consequences." [29]. This leads to Baber [3] suggesting that perhaps what is required is not so much a set of new measures, as an adaptation of existing approaches that pay particular attention to the relatively novel aspects of the environment and activity that pertain to mobile devices. Along these lines, a new breed of techniques for usability evaluation has been proposed [17], [18], [9], [14]. In a survey of research methods for mobile environments [13] a number of new data sources have been identified, to be considered in design and evaluation of mobile applications. These new data sources, may lead to new measures of usability and user experience. *Mediated* data collection includes a range of approaches for collecting data remotely by relying on the participant or mobile technologies themselves. *Simulations* enable knowledge about physical movement, device input and the ergonomics of using a device while mobile. *Enactments* enable researchers to know more about why we carry these mobile devices with us and what these devices give us the potential to do.

On the methodological aspect of usability evaluation of mobile applications, influencing the usability measures, there is a growing interest in the development of scenarios, personas or applying performance techniques [14]. Along these lines, de Sa et al. [8] suggest a set of guidelines for generation of scenarios in usability evaluation of mobile applications. They claim that specific details of the design should be the focus of evaluation studies to be conducted in the field or the lab. While no specific metrics are suggested, a framework for definition of usability evaluation scenarios is introduced that covers the following aspects: (1) *Locations and settings*: lighting, noise, weather, obstacles, social environment. (2) *Movement and posture*: variations for sitting, standing and walking (3) *Workloads, distractions and activities*: Critical activities, settings or domains requiring different degrees of attention, cognitive distractions (e.g., phone ringing, etc), to study cognitive recovery, physical distractions (4) *Devices and usages*: Single vs dual handed interaction, and stylus/finger/keyboard/numeric pad, different devices (e.g., PDAs, smart phones, etc). (5) *Users and Personas*: movement and visual impairment, Heterogeneity – age (small/large fingers), dexterity, etc. For each of these aspects a set of variables has been suggested to be used by designers while arranging their scenarios and preparing the usability evaluation study.

Various aspects related with mobility need to be measured. For instance the effect of mobility on the subjects of field experiments was demonstrated by Kjeldskov and Stage [18], who found that having participants report usability problems while sitting down in the laboratory led to more usability problems being reported than when the participants performed the evaluation while walking. They suggested that this result might have arisen from different demands on *attention* – in the seated condition, there was little distraction from the product and so participants were able to devote most of their attention to it, but in the walking conditions attention needed to be divided between the device and the task of walking.

Tools for Mobile Usability Evaluation

A number of tools have been proposed to support this diversity of approaches. Some tools gather usage data remotely through active (e.g., Experience Sampling Method - ESM, Diary studies) and passive modes (e.g., logging), enabling field evaluation on real settings. The Momento [7] the MyExperience [10] tools are two examples of such systems. Both systems support remote data gathering. The first relies on text and media messaging to send data to a remote evaluator. The second also incorporates a logging mechanism that stores usage data for synchronization. An innovative approach is suggested by Paterno et al. [21]) which combines a model based representation of the activity with remotely logged data of mobile application user activity. An alternative approach is proposed by Yang et al. [30] who have applied UVMODE a mixed reality based usability evaluation system for mobile information device evaluation. The system supports usability evaluation by replacing real products with virtual models. With this system, users can change the design of a virtual product, and investigate how it affects its usability. While users can review and test the virtual product by manipulating it, the system also provides evaluation tools for measuring objective usability measures, including estimated design quality and users' hand load. The metrics supported by this approach include direct measure of physical loads given to the user's hand when user grabs and manipulates a virtual product. A custom built glove interface, equipped with pressure and EMG (electromyography) sensors, is used for measuring the hand load. The collected information is visualized in real-time, so that usability experts could investigate and analyze the hand load while the user manipulates a virtual product.

Given the theoretical and methodological aspects of the mobile usability evaluation problem, it is worth examining the current trends in mobile usability evaluation practice; this is the subject of the next section.

CURRENT PRACTICE: USABILITY EVALUATION OF MOBILE APPLICATIONS

In this section we outline typical examples of current mobile usability evaluation practice. First a survey of reported evaluation studies that involved mobile applications has been performed, based on the ACM CHI 2008 Conference. Five cases that were found are briefly

discussed here in terms of context of the study and methodological approach, see Table 1.

Jokela et al. [15] described the design of the Mobile Multimedia Presentation Editor, a mobile application that makes it possible to author multimedia presentations on mobile devices. To validate and evaluate the application design, they have conducted a series of laboratory evaluations, that involved 24 participants, and a field trial with 15 participants. No quantitative measures are reported from these studies, except of descriptive statistics of participants' behaviour.

Guo et al. [12] introduced a new device (based on the Nintendo Wiimote and Nunchuk) for human-robot interaction. To evaluate this technique, they have conducted a lab based comparative user study, with 20 participants. This study compared the Wiimote/Nunchuk interface with a traditional input device – keypad in terms of *speed* and *accuracy*. Two tasks were employed: the posture task utilized a direct mapping between the tangible interface and the robot, and the navigation task that utilized a less direct, more abstract mapping. A follow-up *questionnaire* recorded the participants' opinion about the preferred technique for controlling the robot in these tasks.

Source	*System*	*Number of Participants*	*Evaluation Method*	*Metrics Used*
Jokela et al. (2008)	Mobile Multimedia Presentation Editor	24 (lab) 15 (field)	Laboratory evaluation and Field study	Qualitative measures of user behaviour.
Guo et al. (2008)	Nintendo Wiimote and Nunchuk – based controller of a robot	20	Lab based comparative user study	Speed and accuracy in both tasks and user preference through questionnaire
Riegelsberger et al. (2008)	Use of Google Maps in Mobile Devices	24	Field study in four different locations.	Qualitative measures and usability problems found using group briefing sessions, recorded usage, multiple telephone interviews and debriefs in a lab setting
Sanchez et al. (2008)	*AudioNature*, A pocketPC device for science learning for the blind	10	Case study in lab involving typical users	Qualitative measures of effectiveness and performance through pre and post tests and questionnaires
Bellotti et al. (2008)	Leisure guide *Magitti*	11	Field study over a period of several days	Qualitative measures of user experience recorded through questionnaires

Table 1. CHI 2008 usability evaluation studies of mobile applications

Riegelsberger et al. [22] discuss a 2-week field trial of Google Maps for Mobile phones, with 24 participants, that was conducted in various parts of the world. The field trial combined many methods: group briefing sessions, recorded usage, multiple telephone interviews for additional context around recorded use, and 1:1 debriefs in a lab setting with the development team observing. No metrics were reported. As a result over 100 usability problems were found. Insights were gained along several dimensions: user experience at different levels of product familiarity (e.g. from download/install to habitual use) and identified hurdles to user experience arising from the mobile ecosystem (e.g. carrier and handset platforms).

Sanchez et al. [23], evaluated the usability of AudioNature, an audio based interface implemented for pocketPC devices to support science learning of users with visual impairments. The usability and the cognitive impact of the device were evaluated. The usability evaluation methodology used was that of a case study involving 10 blind participants in a lab setting. An end-user usability test was conducted that contained 24 statements in a Likert scale (12 statements regarding interaction with AudioNature and 12 statements related to the device used). In addition in order to evaluate the impact of AudioNature on learning, preliminary pre-tests and post-tests were administered. Cognitive testing consisted of different questionnaires to evaluate the learning of biology concepts, the behaviours and skills of visually impaired users and their performance with the cognitive tasks.

Bellotti et al. [5] presented an evaluation study of a context-aware mobile recommender system, the leisure guide Magitti. A field evaluation of this mobile application was conducted, in which 11 volunteers participated. The main features of Magitti were assessed in this study. The field study involved 60 visits (places to eat, to buy, and to do) in the Palo Alto area. After each outing, participants filled out a questionnaire about their activities. In addition, all participants actions with the device were logged and map traces of outings were collected. An experimenter accompanied each participant on one of their outings to observe their use of the system. Finally, participants were interviewed after they completed all their outings. The

reported metrics were related to the user view on these characteristics, while no quantitative usability measures were applied.

Quantitative Measures of Mobile Usability

While the CHI 2008 conference papers, typical of the current practice on mobile applications usability evaluation, applied mostly qualitative measures in field studies, there have been cases in which mobile evaluation was based on specific quantitative usability measures. Two typical such examples are the evaluation study of Goodman et al. [11] and the study reported in [28].

In the first case, [11] a number of measures of usability are suggested to be used in location sensitive mobile applications that are, like mobile guides, to be tested in controlled field conditions (a setting termed Field Experiments). The following measures and methods are suggested: *Performance*: measured through timings and number of errors, *identification of particularly hard points or tasks*: by number of errors or by success in completing the tasks or answering questions correctly, *Perceived workload*: User satisfaction Through NASA TLX scales and other questionnaires and interviews, *Route taken and distance traveled* measured using a pedometer, GPS or other location-sensing system, or by experimenter observation, *Percentage preferred walking speed (PPWS)* Performance By dividing distance traveled by time to obtain walking speed, *Comfort: User satisfaction*. *Device acceptability* using the Comfort Rating Scale and other questionnaires and interviews, *User satisfaction and preferences* and *Experimenter observations*. In effect, in this scheme, the authors suggest in addition to established measures, to take in consideration issues like walking speed and route taken, while they also suggest that existing measures like the Task Load Index and Comfort Rating Scale to be adapted for mobile use. For instance the NASA Task Load Index (TLX) has been extended for mobile applications, by addition of the *Distraction* scale.

In the second case, Tullis and Albert [28] report a case study of usability evaluation of a mobile music and video device, conducted by practitioners S. Weiss and C. Whitby (pp. 263-271) in lab conditions. In this study a number of alternative devices where used by participants for executing a number of tasks relating to purchase and playback of music and video, using three different mobile devices. The measures used included: the time to complete the task, the success or failure of each task, the number of attempts, perception metrics, like feeling about the handsets before and after one hour's use (affinity), perceived cost, perceived weight of handset, ease of use, perceived time to complete the tasks and satisfaction. In the summary findings they included in addition to qualitative findings, the summative usability metric (SUM), based on work by Sauro and Kindlund [24] This is a quantitative model for combining usability metrics into a single score. The focus of SUM is on task completion, task time, error counts per task and post-task satisfaction rating, each of these four metrics, once standardized, contributing equally to the calculation of the SUM score. While this was a large scale usability evaluation study, in which both usability quantitative and qualitative measures were used, the produced report, seemed to include both usability findings and comparative quantitative measures that were effective in communicating the quality in use of the evaluated devices. One concern related with this study, is the fact that there seem to be lack of consideration on issues of mobility and the effect of the mobility dimension on the user experience.

A similar case is the study of Jokela et al. [16], who proposed two composite attributes the *Relative average efficiency* (*RAE*) and the *Relative overall usability* (*ROU*) which however do not take in consideration the specific characteristics of mobile devices. Finally an alternative approach, focusing on social interaction, is the evaluation study reported by Szymanski et al. [26], which analyses user activity while touring a historic house with the Sotto Voce mobile guide.

IN SEARCH OF A USABILITY EVALUATION METHOD

In the rest of the paper, we describe our own experience with conducting usability evaluation studies of mobile applications over the last years and conclude with an outline of a methodological proposal for measuring usability in such contexts. An example of a typical usability evaluation study of a mobile application from our own experience is a collaborative game supported by PDAs in a cultural-historical museum [27], [25], an earlier version of which was discussed in [6].

One of the characteristics of these mobile applications was that they were context sensitive, allowed interaction with objects in the environment (e.g. scan RFID tags), in order to harvest information or make gestures for interacting with the application, while there was a strong social aspect as users acted mostly in groups. Evaluation studies were conducted in various phases of design of these applications. Earlier in design time, these were laboratory based. One limitation of the lab evaluation approach was that it was particularly difficult to reconstruct the context of use, given the above characteristics (social aspect, interaction with objects of the environment etc.). One approach used was to create scenarios and run them in a simulated environment. Sketches of the scenarios that included incidents of interaction were built at these early stages (see Figure 1)

Figure 1. Inheritance museum game sketch

Then these scenarios were transformed in more detailed models of the activity, see figure 2 for a task model of the Inheritance Museum Game, using CTT [20]. These models were then used for formative evaluation of the design in the lab, using low fidelity prototypes. No quantitative measurements were made during this phase, while interaction was fragmented and focused in specific aspects of the interaction and tasks, like navigation, scan of exhibits, exchange of harvested information between group members, etc.

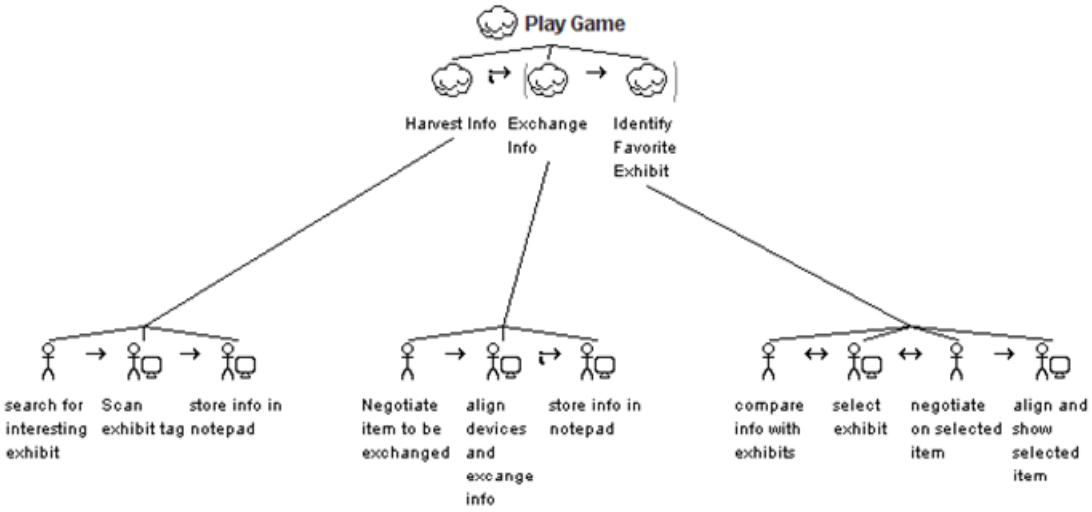

Figure 2. Task model of the inheritance museum game using CTT notation [20].

High Fidelity Prototypes Usability Evaluation

As soon as high fidelity prototypes of the application were made available, more extensive and systematic studies were conducted in the lab, in terms of subjects and tasks. In this case the task execution was recorded using video and audio recording equipment. In the final phase, a study was conducted in the field, following a micro-ethnographic approach, which involves, typical users, engaged with the application for a limited period of time, following a given scenario, without any intervention of the evaluators.

In order not to miss important contextual information, multiple video cameras were used in this study. Two of them were steadily placed in positions overlooking the museum halls, while the third one was handled by an operator who tenderly followed the users from a convenient distance. One member of each group wore a small audio recorder in order to capture the dialogues between them, while interacting with the application and the environment. Furthermore, snapshots of the PDA screens were captured during the collaborative activity at a constant rate and stored in PDA's memory. After the completion of the study the guide, who was member of the evaluator team, interviewed the users, asking them to provide their opinion and experiences from the activity in the museum, while a week later they were asked to report their experience.

In order to analyze all the collected data, we used ActivityLens (Figure 4) that has already been effectively used in similar studies [6, 25].

Three usability experts, with different level of experience, analyzed the collected data, in order to increase the reliability of the findings. Initially, we created a new ActivityLens Study including 4 projects (each project concerns the observations of a team). We studied then the integrated multimedia files and annotated the most interesting interaction episodes. It should be clarified that we wished to evaluate the performance of each team and not individual team members. The performed analysis through ActivityLens revealed several problems related to user interaction with the device and the overall setting, given the social and physical context of use. In figure 3 the dimensions of analysis used are outlined, these include in addition to typical user device interaction, user-user interaction, user –setting interaction, while observed phenomena were related to other aspects, like interaction of the mobile devices with the infrastructure and other applications.

Various usability problems observed were unexpected and need further investigation, for instance the use of scrollbar in the textual description of the exhibits. The typical users were not familiar with the procedure of scrolling on a PDA using a stylus, a task necessitating both hands, while various cases of split attentional resources were observed.

Quantitative measures of usability beyond the standard performance and user satisfaction measures need however yet to be defined. Inspired by the literature, discussed also in the survey part of this paper, we currently work towards defining a number of combined measures related to the dimensions outlined in Figure 3.

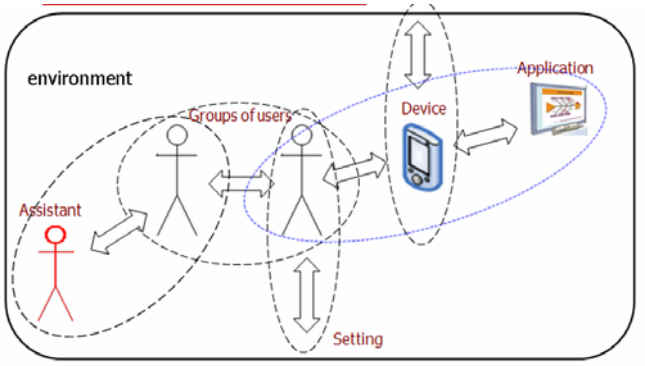

Figure 3. Dimensions of analysis of observed behaviour

In addition, and using the ActivityLens tool that integrates a model-based view with observational data, we plan to relate usability problems found to the task structure, the dimensions of interaction and a classification of usability problems scheme (e.g. the user action framework [1]), extending it in order to accommodate the characteristics of mobile applications.

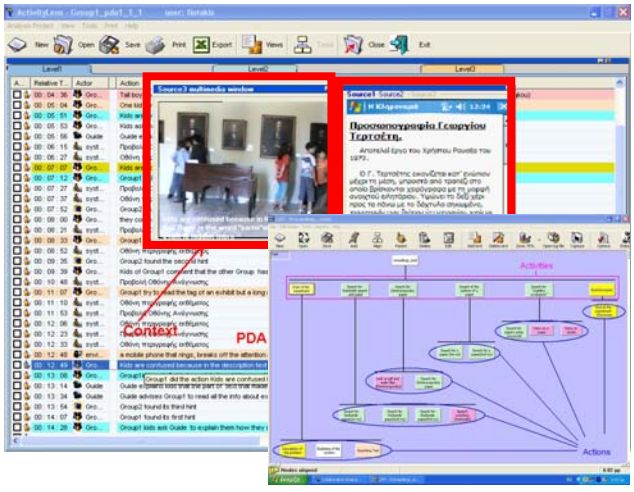

Figure 4. The ActivityLens analysis environment, annotation of video recording and mapping to task model.

CONCLUSION

In this paper, we discussed issues related with measuring usability of mobile applications. It has been argued that the nature of these applications necessitate new methods and tools for measuring their usability. Issues like study of attentional resources and mobility need to be included in the usability measures. Survey of current practice demonstrated that so far most usability evaluation studies are of qualitative nature, while quantitative measure of usability used are mostly related to well established general purpose attributes of usability, like effectiveness, efficiency and user satisfaction. However as the mobile applications field matures and the usability evaluation methods become more reliable and theoretically understood, the specific characteristics and requirements of mobile applications, that become pervasive in modern societies, we feel that there will be accordingly adapted.

REFERENCES

1. Andre, T.S., Hartson, H.R., Belz, S.M. and McCreary, F.A., The user action framework: a reliable foundation for usability engineering support tools. Int. J. Human-Computer Studies 54, (2001), pp. 107-136.
2. Annett, J., 2002, Subjective rating scales: science or art? Ergonomics, 45 (14), pp. 966 – 987, 2002
3. Baber C., Evaluating Mobile-Human Interaction, in J. Lumsden, "Handbook of Research on User Interface design and Evaluation of Mobile Technology, 2007, Hershey, PA, IGI Global
4. Bartneck, C., What is good? A Comparison between the Quality Criteria Used in Design and Science, CHI 2008, pp. 2485-2492, 2008 • Florence, Italy
5. Bellotti V., Begole B., Chi E. H., Ducheneaut N., Fang J., Isaacs E., King T., Newman M. W., Partridge K., Price B., Rasmussen P., Roberts M., Schiano D. J., Walendowski A., Activity-Based Serendipitous Recommendations with the Magitti Mobile Leisure Guide. CHI 2008, pp. 1157-1166, 2008, Florence, Italy
6. Cabrera, J. S., Frutos, H. M., Stoica, A. G., Avouris, N., Dimitriadis, Y., Fiotakis, G., and Liveri, K. D. 2005. Mystery in the museum: collaborative learning activities using handheld devices. In *Proceedings of the 7th* MobileHCI '05, (Salzburg, Austria, September 19 - 22, 2005). vol. 111. ACM Press, New York, NY, 315-318.
7. Carter S., Mankoff J., Heer J., Momento: Support for Situated Ubicomp Experimentation. CHI'07, pp. 125-134, ACM
8. de Sá M. , Carriço L., Duarte L.. A Framework for Mobile Evaluation. CHI 2008, Florence, Italy, pp. 2673-2678, ACM Press, April, 2008
9. de Sa M., Carrico L., Defining Scenarios for Mobile Design and Evaluation, CHI 2008, pp. 2847-2852, 2008 • Florence, Italy

10. Froehlich J., Chen M.Y., Consolvo S.,Harrison B., Landay J.A., MyExperience: A System for In Situ Tracing and Capturing of User Feedback on Mobile Phones. MobiSys'07, pp. 57-70, ACM.
11. Goodman J., Brewster S., and Gray P., Using Field Experiments to Evaluate Mobile Guides, in Schmidt-Belz, B. and Cheverst, K. (ed.), Proc. Workshop "HCI in Mobile Guides", *Mobile HCI 2004*, Glasgow, UK.
12. Guo C., Sharlin E., Exploring the use of tangible user interfaces for human-robot interaction: a comparative study, CHI 2008, pp. 121-130, Florence, Italy
13. Hagen P., Robertson T., Kan M., Sadler K., Emerging research methods for understanding mobile technology use, Proc. OZCHI '05, Canberra, Australia, 2005
14. Jacucci G., Interaction as Performance, PhD Thesis, University of Oulu, Oulu University Press, 2004
15. Jokela T., Lehikoinen J.T., Korhonen, H., Mobile Multimedia Presentation Editor: Enabling Creation of Audio-Visual Stories on Mobile Devices, CHI 2008, pp. 63-72, Florence, Italy
16. Jokela, T., J. Koivumaa, J. Pirkola, P. Salminen and N. Kantola (2006). "Quantitative Usability Requirements in the Development of the User Interface of a Mobile Phone. A Case Study." Personal and Ubiquitous Computing 10(6): 345-355.
17. Kjeldskov J. & Graham C. (2003) A Review of Mobile HCI Research Methods. 5th Int. Mobile HCI 2003, Udine, Italy, Springer-Verlag, LNCS
18. Kjeldskov J., & Stage J., New Techniques for Usability Evaluation of Mobile Systems, International Journal of Human-Computer Studies, 60, (2004), pp. 599-620.
19. Nielsen C.M., Overgaard M., Pedersen M.P., Stage J., Stenild S., It's worth the hassle!: the added value of evaluating the usability of mobile systems in the field, Proc. 4th Nordic CHI, Oslo, 2006, pp.272 - 280.
20. Paternò F., Model-Based Design and Evaluation of Interactive Applications, Springer, 2000.
21. Paternò F., Russino A., Santoro C., Remote Evaluation of Mobile Applications, TAMODIA 2007, LNCS 4849, pp. 155 – 169, 2007
22. Riegelsberger J., Nakhimovsky Y., Seeing the Bigger Picture: A Multi- Method Field Trial of Google Maps for Mobile, CHI 2008, pp. 2221-2228, Florence, Italy
23. Sánchez J.,Flores H., Sáenz M., Mobile Science Learning for the Blind, CHI 2008, pp. 3201-3206, 2008 • Florence, Italy
24. Sauro J., Kindlund E., A method to standardize usability metrics into a single score, Proc. CHI 2005, pp. 401-409.
25. Stoica A., Fiotakis G., Raptis D., Papadimitriou I., Komis V., Avouris N., Field evaluation of collaborative mobile applications, chapter LVIX in J. Lumsden (ed.), "Handbook of Research on User Interface Design and Evaluation for Mobile Technology", 2007, pp. 994-1011, Hershey, PA, IGI Global
26. Szymanski, M. H., Aoki, P. M., Grinter, R. E., Hurst, A., Thornton, J. D., and Woodruff, A. 2008. Sotto Voce: Facilitating Social Learning in a Historic House. *Comput. Supported Coop. Work* 17, 1 (Feb. 2008), 5-34.
27. Tselios N., Papadimitriou I., Raptis D., Yiannoutsou N., Komis V., Avouris N. (2007), Designing for Mobile Learning in Museums, in J. Lumsden (ed.), "Handbook of Research on User Interface Design and Evaluation for Mobile Technology", Hershey, PA, IGI Global
28. Tullis T., Albert B., Measuring the User Experience, Morgan Kaufmann, 2008.
29. Wilson J.R., Nichols S.C., 2002, Measurement in virtual environments: another dimension to the objectivity/ subjectivity debate, Ergonomics, 45 (14), pp. 1031 – 1036
30. Yang U., Jo D., Son, W., UVMODE: Usability Verification Mixed Reality System for Mobile Devices, CHI 2008, pp. 3573-3578, Florence, Italy
31. Zhang D. and Adipat B., Challenges, Methodologies, and Issues in the Usability Testing of Mobile Applications, Int. Journal of Human-Computer Interaction, 18, 3, (2005), 293-308.

User Experience in the Systems Usability Approach

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

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

ABSTRACT
In this paper we propose how and why the concept of user experience (UX) could be used in the analysis of user interfaces of complex systems. Complex systems often relate to work and professional usage of tools. In this context UX is typically not considered a meaningful attribute. As opposed to that, we feel that experience is especially important in work-related contexts as it has a link to development of good work practice, job satisfaction, and motivation. Thus, in this paper we describe how we think that UX should and could be made operational in the analysis of ICT tools used in various kinds of work.

Author Keywords
User experience, activity theory, practice, systems usability.

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

INTRODUCTION

In this paper we describe how the concept of user experience UX could be, and why it should be considered in the analysis of complex system user interfaces. With complex systems we refer to ICT (information and communication technology) based tools that are used in work that can be characterised as outcome critical from the point of view of society. Such works include for example control of power plants (including nuclear), control of other industrial plants (manufacturing and process industry), manoeuvring of large ships, command and control of emergency services, command and control of traffic systems etc. Also smaller scale systems such as patient monitoring can be considered to belong to the class of complex systems. What is in common with all of the above mentioned systems is that the work conducted with them is demanding and requires thorough training.

VUUM2008, June 18, 2008, Reykjavik, Iceland.

The systems can also be characterised complex socio-technical systems [1] because they comprise both social and technical components which have to work together in a uniform manner to produce the desired outcome.

The traditional approach to the analyses of the quality and functioning of complex socio-technical systems has been *safety*. From a human factors point of view this line of thought means the effort of minimising the possibility of human errors.

Recently, the focus of the human factors research in safety critical complex systems has been extended to cover also the possible positive effects of the human components of the systems for example in tackling unexpected situations. But even a more systemic approach has been proposed which emphasises the transaction of the technical and human elements, and their mutual development [2]. A new notion of "joint cognitive systems" (JCS) proposed recently by Hollnagel and Woods represents the system-oriented point of view [3]. JCS presents the idea that from a functional point of view, human and technology constitute a unified system. While in the traditional Human-Computer Interaction (HCI) or Cognitive System Engineering (CSE) approaches a structural perspective was dominant and the information flow between two separate elements, i.e. human and technology was in focus, in the JCS approach the intelligent features and patterns of functioning of the unit should be the focus of design. From this perspective, rather than safety alone, an overall adaptive functioning of the system becomes a relevant success criterion. For example features that are typically considered within *usability* research may be linked to more global risk and reliability features.

Our approach proposed recently under the title "systems usability" [4] is another example of a system-oriented approach. We have proposed how to tackle safety and usability issues in an integrated way in empirical analysis and practical design connections. We have also demonstrated that exploiting the concept of activity of the cultural historical activity theory (CHAT) opens a possibility to consider system functionalities more comprehensively than just dealing with the user interface [see also 5]. In our work we have become aware that *experience of usage* of complex tools is an intrinsic element in understanding capabilities and willingness of users to adopt new tools and technologies. We see however, that research and innovation is needed to develop an approach that in an appropriate way comprehends user experience of complex professional technologies. The aim of this paper is an

outline on which bases and how we would like to proceed in this work.

EXPERIENCE IN THE DYNAMICS OF ACTIVITY

Considering user experience is not a well developed research area with CHAT. Kaptelinin and Nardi (2006) in their recent account on "activity theory and interaction" touch the topic only briefly. They see that CHAT has potential to consider the role of emotions in the design and use of technologies but that challenging theoretical and conceptual work has to be accomplished.

Developing Mediations in Action

We take one of the basic ideas of CHAT that all human action is mediated by tools and signs [6] as our starting point. CHAT has traditionally focused particularly in the role of material tools [7]. Symbolic mediation has been the focus of semiotics. Within it, the pragmatist semiotics founded by Charles Sanders Peirce [8], has close resemblance to CHAT, and great relevance to understanding mediated action of the modern age. In CHAT it is also maintained that development of human activity becomes evident in human beings becoming capable of making environmental features as signs of new possibilities of action. People experience new mediations as emotionally meaningful events, "decisive events" [9]. These events provide leverage for activity and create new motives. Drawing on Ch. S. Peirce, Koski-Jännes notes further that the strength of mediation is then greater when all the different sign aspects - iconic showing resemblance, indexical pointing to or identifying specifics, and symbolic creating generalized connections – are all present. A symbolic connection has emerged when a sign awakens awareness of a more comprehensive context, significance and worth of the event by the actor.

A further basic notion of CHAT that we consider relevant is, that creating new meaningful and emotionally experienced mediations is a process that takes place in a "potential space" in which (a child and adult) the user and the designer jointly experience and develop the mediation function. In this space the tool or sign does not yet have established the mediating function but receives attention and interest as it is, as is typical for play or art. We have borrowed these ideas from Leiman [10], but they are close to those of the so-called instrument genesis theory, too [11]. The "potential space" is clearly an intercommunicational space.

The principles are the basics of understanding the development of mediated action. Hence, they should be relevant in understanding how practices and tools should be developed and how to proceed in evaluating their maturity. At the present state, we observe tools and tool usage from two distinct perspectives. First we pay attention to the function of tools in activity, and, second, we consider how people experience technology in usage. Both aspects are incorporated in the analysis frame we have labelled "contextual assessment of systems usability".

Functional Approach to Analysis of Use Technology

In a technology development oriented analysis of an activity system the tools that are used in the activity are the key element of analysis. According to the theory the tools mediate the role between the actor and the object of activity; it is with tools that the actor can affect his/her environment. The tool's effect on the environment is called its *instrumental function*. Our interpretation is that the instrumental function refers to the same phenomena as "effectiveness and efficiency" in the prevailing ISO definition [12] of usability. Exploiting activity theory, the tools in an activity system have also further functions from instrumental. The second function of a tool is its *psychological function.* This refers to the tool's ability to shape the actor's understanding of the world. Use of a tool gives the actor concepts and schemas concerning the world, which enable some cognitive processes such as reflection of own behaviour and learning. Recently an addition to the functions of tools has been proposed by a prominent advocate of media theory Georg Rückriem [13]. Thus the third function of a tool is its *communicative function*. This means that tools act as mediators that shape our shared understanding and awareness of the surrounding world. Tools are a manifestation of our culture and communicate meanings among human beings.

Our claim is that good tools must satisfy all of the three above mentioned functions before they can be considered appropriate for a particular activity.

Phenomenological Approach to Analysis of Use of Technology

The phenomenological tradition in philosophy and social sciences has traditionally considered interesting and worth while to understand how people experience world, and also technology. Drawing on this tradition John Ihde has proposed his idea of "phenomenology of technics" [14]. Ihde identifies three forms of experiencing human-technology-environment relationships. These are the *embodiment, hermeneutic* and the *alterity* relationships. In the first variant, technology, when functioning well, is experienced invisible and as merging with the human body. In the second technology is experienced as part of the environment and interpreted as signs denoting environmental possibilities. In the third variant the focus of experience is on technology as another agent like the human him/herself. This approach does not assume that people necessarily conceive technology as mediator or in a particular function. It is interested in the user's experience of *coordination with technology*. This is, of course, a very relevant aspect when considering usability of technology from UX point of view. We see, that understanding experience of this coordination provides added value to our attempt to understand users' experience of the functions of tools and technologies.

Systemic Approach to Analysis of Use of Technology

We have earlier introduced the concept of *systems usability* which provides a systemic approach to understanding use of technology. In the approach we have conceptualised what constitutes quality of tools in an activity system. We see that the quality of tools becomes visible in the *practice of usage*. Good tools promote good work practices! Further, the good practice, and its measures can be made operational by analysing the domain and the result critical functions of the particular activity.

The systems usability approach presumes the following principles for the analysis of use of technology.

1. The analysis should be contextual. The quality requirements for tools are determined by the objectives and purposes of the domain. The domain knowledge is elicited by modelling the critical functions and resources of the domain
2. The analysis should be holistic. The users, tools, and the environment constitute one system the functioning of which is the focus of analysis. We do not only analyse how the users affect the environment with the technical system, but also vice versa: how the environment and the system have an effect on user.

We have used the systemic approach to the analysis of use of technology in the development of evaluation framework: Contextual Assessment of Systems Usability (CASU). In evaluating technology with CASU we make a difference between *the outcome* of action and *the mode* of action. This means that although a task may be completed (the outcome is the desired one) there might still be problems in the *way of* reaching the outcome. For example the practice might be such that it somehow expends resources excessively. Thus the result cannot be considered as good as possible even though the objective has been accomplished. The two aspects of activity is included in the Core-task analysis approach, in which actual activity is distinguished from what it is a sign of, or what internal logic it refers to. The latter is the semiotic point of view which is represented by the concept of mode of activity.

The third dimension (in addition to the outcome and mode) in the analysis of technology use is the experience of use. Thus in this type of analysis of technology use we combine the functional and the phenomenological approaches.

EVALUATION OF EXPERIENCE

Evaluating experience is not easy. Especially so-called objective evaluation of experience is at least extremely difficult, but it maybe also unnecessary as the phenomenon itself (the experience) is not objective by nature. We have connected the concept of experience of technology use to the functions that the tool has in an activity system (Table 1). With this we are able to infer which experiences have meaning in the particular context in which the technology is used.

focus of analysis tool functions	outcome of action	mode of action	experience
instrumental	task achievement, time, errors, objective reference	meaningful routines	experience of appropriate functioning
psychological	cognitive measures, SA, mental models	coordination, control of own activity	experience of fit for human use, experience of own competence, experience of trust in technology
communicative	amount and content of communications	meaning of actions and communications	experience of joint culture, tool as a sign of shared culture, experience of fit for own style

Table 5. Classes of measures used in CASU framework.

Measures of Experience

The measures of experience are connected to the functions of the tools in the activity system. Above (Table 1) are stated also the classes of measures that are used in the functional evaluation of the tool. But only the measures related to experience are described in more detail.

When the instrumental function is considered, the experience that we are trying to unveil is the experience of the appropriate, and probably embodied, functioning of the tool. This is the feeling that the tool functions as expected and that it is possible to use the tool for the purpose that it was designed for.

In the psychological function the user experience is related to the tool's fit for human use. For example the feeling that the tool functions as expected and that the user is able to use it are characterisations of experience in psychological function. If the experience is positive the users feel that they are competent users and they are able to formulate appropriate trust in technology (not over or under trust).

The communicative function of technology is experienced in the community. If an individual user has an experience of belonging to a community (of practice) that shares the same objective and tools and also rules and norms, which all enable him to interpret the meaning of his fellow actors' behaviours and tool usages. These are embodiments of positive user

experience concerning the communicative function of a tool. Further, the experience of the communicative function assumes that users' are able to use the technology with their own style the fits their own identity and how they want others to perceive them.

Data Concerning Experience

In order to evaluate or measure the experience we need data concerning it. In our evaluations we have gathered data about the experiences with mainly two methods: observations and interviews.

In the careful analysis of observation data it is possible to recognise experience related features in the users' behaviour. For example communications and statements about own actions give information about experiences. Very careful analysis of the usage is needed in order to understand the behavioural markers that are related to experiences.

The other method to acquire data about experiences of use is the interview method. We have used interface interviews that are conducted after the use of the tool. In the interviews we specifically ask the user to comment how in his/her view the system works. Also a stimulated interview method in which the user comments his or her own use of technology is suitable in eliciting data concerning UX.

CONCLUSIONS

In the development of complex system interfaces it is especially important to understand how the potential users experience the new technology. The experience of work tools is an evolving phenomenon that might take different forms during the technology life cycle. There is a difference in comparison to the development of different consumer products in which the experience is often connected to the decision to buy the product. With work related tools the user is not necessarily the same person who makes the decision to buy the particular tool. This means that the role of experience is also different. The need for positive user experiences is rooted in the need for the users to have high motivation in their work and also to carry out their work with appropriate work practices. Experience is related to both these concepts. If the tools in the work are experienced somehow negatively this has an effect on the whole activity system. An contrary, positive experience and activeness is raised if tools can be seen to support professionally and personally worthy development in the activity.

We see that development of new mediations that technologies enable takes place in the "potential space" in which the role and features of technology have not yet been crystallised. In this space a communication and mutual understanding between technology developers and users is a natural and indispensable element. UX is part of this dialogue.

REFERENCES

1. VICENTE, K.J., *Cognitive Work Analysis. Toward a Safe, Productive, and Healthy Computer-Based Work*. 1999, Mahwah, NJ: Lawrence Erlbaum Publishers,(1999).
2. FLACH, J., et al., *Global perspectives on the ecology of human-machine systems*. 1995, Hillsdale, New Jersey: Lawrence Erlbaum Associates,(1995).
3. WOODS, D. and E. HOLLNAGEL, *Joint cognitive systems - patterns in cognitive systems engineering*. 2006, Boca Raton: Taylor & Francis,(2006).
4. SAVIOJA, P. and L. NORROS, *Systems usability - promoting core-task oriented work practices*, in *Maturing Usability: Quality in Software, Interaction and Value*, E. Law, et al., Editors. 2008, Springer: London. p. 123-143.
5. KAPTELININ, V. and B.A. NARDI, *Acting with technology. Activity theory and interaction design*. 2006, Cambridge, Massachusetts: The MIT Press. 333,(2006).
6. VYGOTSKY, L.S., *Mind in Society. The Development of Higher Psychological Processes*. 1978, Cambridge, Mass.: Harvard University Press,(1978).
7. LEONT'EV, A.N., *Activity, Consciousness, and Personality*. 1978, Englewood Cliffs: Prentice Hall,(1978).
8. PEIRCE, C.S., *Pragmatism*, in *The essential Peirce. Selected philosophical writings*, T.P.e. project, Editor. 1998, Indiana University Press: Bloomington and Indianapolis. p. 398-433.
9. KOSKI-JÄNNES, A., *From addiction to self governance*, in *Perspectives on activity theory*, Y. Engeström, R. Miettinen, and R.-L. Punamäki, Editors. 1999, Cambridge University Press: Cambridge, UK.
10. LEIMAN, M., *The concept of sign in the work of Vygotsky, Winnicot, and Bakhtin: Further integration of object relations theory and activity theory*, in *Perspectives on activity theory*, R. Engeström, R. Miettinen, and R. Punamäki, Editors. 1999, Cambridge University Press: Cambridge, Massachusetts. p. 419-434.
11. RABARDEL, P. and P. BEGUIN, Instrument mediated activity: from subject development to anthropometric design. *Theoretical Issues in Ergonomics Science*. **6**(5): p. 429-461(2005).
12. ISO, ISO 9241-11. Ergonomic requirements for office work with visual display terminals (VTDs) - Part 11. Guidance on usability. (1998).
13. RÜCKRIEM, G.,Tool or medium? The Meaning of Information and Telecommunication technology to Human Practice. A quest for Systemic Understanding of Activity Theory. University of Helsinki, Center for Activity Theory and Developmental Work Research: http://www.iscar.org/fi/.
14. IHDE, D., *Technology and the lifeworld. From garden to earth*. 1990, Bloomington, Indianapolis: Indiana University Press,(1990).

Developing the Scale Adoption Framework for Evaluation (SAFE)

William Green, Greg Dunn, and Jettie Hoonhout
Philips Research
High Tech Campus 34
5656 AE, Eindhoven, NL
{first name.last name}@philips.com

ABSTRACT
A growing number of psychometric scales have been developed to measure a plethora of constructs within the umbrella term, *User Experience* (UX). Unfortunately, selecting an appropriate scale for UX can be difficult for the usability practitioner. Based on psychometric scale development literature, the Scale Adoption Framework for Evaluation (SAFE) has been designed to support practitioners in selecting appropriate scales. The SAFE is presented in this paper, with an example of how it could be used. However, utilizing the SAFE has emphasised the difficulty in selecting suitable measures, revealing that the SAFE will not solve this issue in isolation. Two main challenges are presented for the VUUM workshop: more explicit and transparent descriptions of psychometric scale development, and a need for more sensitive psychometric scales. Future work should build on these challenges by consulting usability practitioners directly.

Author Keywords
Psychometric scale development, evaluation, UCD.

ACM Classification Keywords
H.1.2.,User/Machine System: Design, Experimentation, Human Factors, Measurement.

INTRODUCTION
This paper is concerned with psychometric scale development and its adoption by practitioners within the User-Centred Design (UCD) process. There are many illustrative texts and practical examples of psychometric scale development in psychology literature, e.g., [4, 21, and 26]. Interpreting this literature can be difficult however; the majority of psychometric literature is written for an assumed audience that will have either psychology or statistics expertise. In this paper, the Scale Adoption Framework for Evaluation (SAFE) is presented and has been developed in an effort to support practitioners before selecting psychometric scales. A psychometric scale is a form of measurement used to assess abstract qualia subjectively experienced by an individual (e.g., individuals' personality, their emotional states, and interpersonal trust). From this point forward, 'psychometric scale,' will be referred to simply as, 'scale.' The SAFE is intended to help inform practitioners, who may not have a psychology or psychometric background, about the essential properties of valid and reliable measures. To achieve this, the SAFE illustrates the most important elements that a practitioner should look for when adopting a measure, to ensure that it is sufficiently robust and accurately measuring the construct that the practitioner needs, and it claims to measure.

Before presenting the SAFE, the UCD process is briefly reviewed emphasising how psychometric measurement of users' subjective experience supports the evaluation of products. The reasoning behind why it was believed that the SAFE would be helpful for practitioners is then presented. In the next section, we discuss development and evaluation of psychometric measures, specifically focusing on what is particularly *pertinent to practitioners* when selecting a measure. This led to the development of the SAFE. An example of how to utilise the SAFE is illustrated. Initial use of the SAFE to select a measure of emotion indicated several challenges, which has led to conclusions and recommendations for future research.

Measuring Usability and User Experience for UCD
User-Centred Design (UCD) is a design philosophy that focuses on the needs and interests of users and emphasises making products usable and understandable [20]. There have been many descriptions of UCD in the literature that vary in the nomenclature used. Nonetheless, [5] identified three principles for user-centred system design, which are still prevalent in the majority of the UCD descriptions, these are: 1) a user focus from the beginning and throughout the process, 2) measurement of system usage, and 3) iterative design. As [3] noted, the two editions of the Handbook for HCI, e.g., [7], also emphasise the three UCD principles proposed by [5]. Despite recent suggestion that the emphasis on evaluation in [5] and [7] is dated [3] and

VUUM2008, June 18, 2008, Reykjavik, Iceland.

not always appropriate [6], evaluation is undeniably an important aspect in UCD, often achieved via measurement.

Measurement of system usage by practitioners involved in UCD is often concerned with usability evaluation, i.e., are the products easy and comfortable to use, safe, effective, efficient, and easy to learn how to use? These usability evaluations have been done through objective (e.g., time or physiology) and subjective (e.g., perceptions, attitudes, and other scales of psychological constructs) measurement and are typically based on at least one of three dimensions outlined by ISO 9241-11 [13], which are: *Efficiency*, *Effectiveness*, and *Satisfaction*.

Subjective evaluation is often done via questionnaires, described as both the most versatile, but also most often misused research tool for HCI [19]. Despite [19]'s caveat and support for questionnaire design, development, and appropriate use, it appears that the questionnaire remains misused almost twenty years later. For example, using [13]'s dimensions, [12] demonstrated that while the measurement of both *Efficiency* and *Effectiveness* is relatively straightforward and measured in similar fashion among UCD practitioners, the measurement of users' *Satisfaction* is diverse. This diversity is illustrated by findings indicating that only 11% of satisfaction scales used by practitioners are established (i.e., valid and reliable) [12]. Furthermore, [12] reported that the 89% of "homegrown" measures vary greatly in terms of reliability, and so, recommend that practitioners should use established scales. This is not a new recommendation. For example, [15] emphasised that, "questionnaires should be elegant in terms of their reliability, validity and human factors appeal" (pg. 210). It is in this vein that [11] proposed a model to support practitioners when selecting the best approach for usability evaluation. His model outlines six approaches for evaluating usability, both objectively and subjectively for each of the three usability dimensions noted within the ISO standard above. Extending from [11]'s model, this paper focuses on subjective approaches to usability evaluation.

Obviously, performance criteria related to *Efficiency* and *Effectiveness* are important for consumer products, especially in the case of safety, comfort, and learnability. Particularly for consumer products, however, it has been increasingly accepted that other requirements related to *Satisfaction* should also be considered. Using a product should be enjoyable, engaging, and appealing [e.g., 1, 8]. These requirements have often been discussed as part of the *User Experience* (UX). UX has been particularly important given the migration from the workplace, ubiquity and growth of technology in the home, and the emergence of 'intelligent' and perhaps more complex products. Nevertheless, like usability satisfaction, UX has numerous subjective meanings that are dependent not only on the individual, but also the context of interaction. Given the growing importance of UX evaluation, [12] findings regarding the low percentage of established satisfaction scales being used by practitioners becomes even more poignant. If practitioners are creating or implementing scales that do not provide valid or reliable results, this could seriously compromise evaluations of constructs critical to the success of a product (e.g., trust measures for users' perceived trustworthiness of an e-commerce website). This paper supports [12]'s position recommending that practitioners should use established scales, but greater emphasis is placed on supporting practitioners when selecting existing scales. To do this, steps are taken to further understand the benefits of using scales from a psychometric perspective, and their contribution to the UX evaluation process.

Using Scales in User Experience

Psychometric scales have often been employed for evaluation when measurements are required across a number of studies to determine some psychological construct (e.g., emotion, trust). This method enables users to subjectively quantify their experiences, which map onto a construct. These findings then indicate to practitioners how a product could be improved. This use of scales has provided a number of advantages: 1) it enables testing of large quantities of participants over short periods of time, at relatively low cost, which is practical since time and financial resources to conduct user-studies have usually been limited, 2) it is an easy to apply technique (although scale development is more time-consuming and complex), and 3) it has usually been non-intrusive for participants.

To support design decisions regarding the dimensions of UX through evaluation, scales must first be developed based on a UX construct definition. For example, emotion and affect are ambiguous terms but are often postulated as integral to designing UX, e.g., [14]. This ambiguity stems from the literature itself. [22] differentiate between the terminology 'affect', 'emotions', 'moods dispositions' and 'preferences' (pg. 124). They note that difficulty in answering the question, "what is an emotion?" is related to the interchangeable use of the terms 'emotion', 'emotional', and 'affect.' An example of this can be seen in [23] in an attempt to clarify this ambiguity in their use of terminology, "emotional and affective will be used interchangeably as adjectives describing either physical or cognitive components of emotion, although 'affective' will sometimes be used in a broader sense than 'emotional" (pg. 24). This ambiguity stresses the importance of evaluating UX dimensions with scales based on specific constructs.

One possible reason for the apparent lack of psychometric scale usage for UX could be due to practitioners simply not being familiar with the exhaustive process needed to create a new scale. Choosing whether a scale is appropriate for a particular product is difficult given that part of this decision involves understanding scale construction, which is perhaps limited to those with psychology or statistics expertise. It is

anticipated that new psychometric scales will be developed to evaluate UX [e.g., 10]. Given the high percentage of homegrown measures used to measure satisfaction and the similarities between UX and satisfaction (both are subjective and abstract constructs), it is possible that the nature of UX will inadvertently encourage its practitioners to develop their own scales without properly recognising issues such as scale validity and reliability. If this were to occur, then it would only serve to propagate questions concerning the validity and reliability of UX evaluation. To prevent this possibility, this paper presents the Scale Adoption Framework for Evaluation (SAFE), which is intended to encourage and support practitioners to select established scales for UX, rather than utilising homegrown ones. But if a homegrown one must be used, then the SAFE also provides a starting point for practitioners to consider what is vital when attempting to construct a valid and reliable scale.

THE SAFE DEVELOPMENT APPROACH

Based on several methodological and statistical texts that discuss (aspects of) scale development [e.g., 2, 21, 26], the SAFE was created to portray the key elements of scale development that are, debatably, essential for scale construction. As shown in Figure 1, the SAFE is composed of three elements. Within each element, on the left, a description of the aim that developers have when considering the scale in scale construction is provided. Additionally, for each of these elements three key aspects are emphasized for practitioner's to consider when deciding on a scale for their usability or UX evaluation. The pertinent aspects of these elements are as follows:

1. Construct definition – confirming that the scale construct (abstract qualia to be measured) is suitable.

1a) Theory – is concerned with whether scale developers have based their proposed construct on a strong theoretical foundation following a literature review, experiments, or consultation with domain experts. If this is not done, then there are fundamental questions regarding what the scale is measuring, despite possibly being valid and reliable (see below).

1b) Clarity – recommends that the theory behind the construct is clear, coherent, and straight to the point. If this is not the case, it becomes more difficult for the practitioner to decide whether the theory adequately describes the construct.

1c) Discriminating – acknowledges that it is important for scale developers to also consider what the construct does not include at a theoretical level. For the practitioner, this will simply mean that it will make it easier to see what the scale does and does not measure.

2) Scale validity – the extent to which a tool or method measures what it is intended to measure.

2a) Construct validity – it is important that practitioners see evidence from developers that the constructed scale shows some sort of relation (via prediction or correlation) with existing measures/behaviours connected to the construct. For further information see [2] and [26].

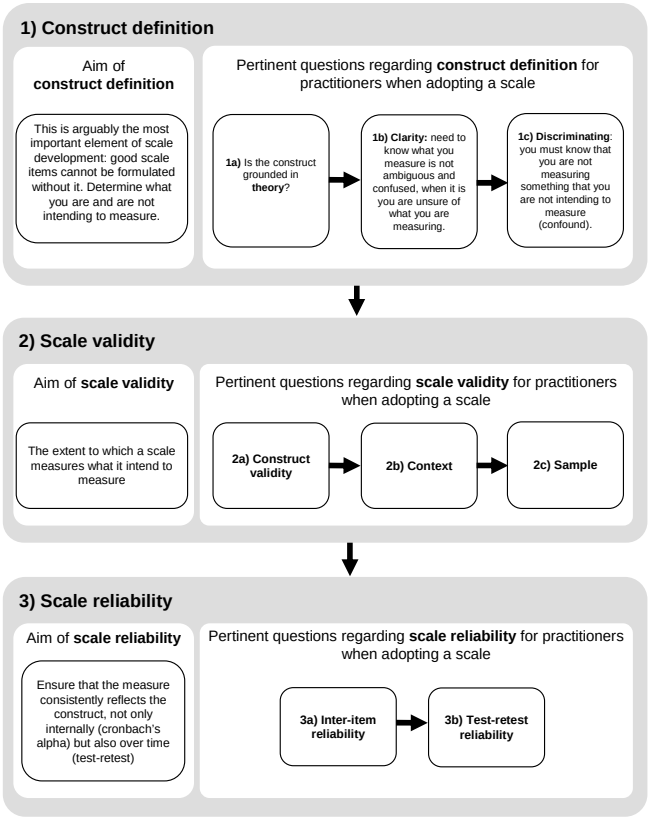

Figure 1: Scale Adoption Framework for Evaluation (SAFE)

2b) Context – practitioners must be aware of the context of use that the developers first envisaged when they constructed their scale. If practitioners use a scale in a context that was not originally envisaged by developers, then the validity (and reliability) of the scale could be compromised.

2c) Sample – similar to Context, this aspect asks practitioners to be aware of the sample (i.e., users) that developers used when constructing the scale. If, in the opinion of the practitioner, this is distinctly different from the sample that s/he is intending to recruit, then this could compromise scale validity and reliability.

3) Scale reliability – the extent the tool provides stable results across repeated administrations.

3a) Inter-item reliability – Measured via Cronbach's alpha or split-half reliability, it determines what extent scale items measure the same construct that it is said to be measuring accurately. Generally, Cronbach's alpha of .7 is a recommended minimum, though higher

Cronbach's alpha are said to be necessary depending on the construct that is said to be measured.

3b) Test-retest reliability – Much like Inter-item reliability, this aspect tests how reliable the scale is when individuals are retested with the same scale. The level of test-retest reliability depends on what is said to be measured, but again, a good rule of thumb is .7.

Example of the SAFE usage

To initially ascertain whether or not the SAFE would support usability practitioners, the authors of this paper utilised the SAFE to review four established psychometric scales (Table 1). For this, the original publications of these psychometric scales were obtained. The scales have all been referenced in relation to their use in usability evaluations and are purposively concerned with disparate constructs that are all related to usability, satisfaction, and UX. In Table 1 the four psychometric scales and their adherence to the attributes of SAFE are tabulated.

As seen in Table 1, the four scales do not adhere to all of the SAFE attributes. Nonetheless, all of these scales are established and have been used to measure their respective constructs in publications. It is noted then, that it is not always possible or necessary to create scales that fully meet the SAFE requirements discussed above. This decision is down to the practitioner and whether or not the scale can be accepted. To determine how the use of SAFE can support practitioners in selecting appropriate scales, several case studies conducted in an industrial setting are now discussed.

Determining the Affective Response of Participants: Feedback from Case Studies

In many consumer application studies, a key question often addressed is, what is the affective response of participants when using the product being tested? In several projects, ranging from studies investigating users' affective response to environments that present coloured lighting effects [9, 16], to game applications [28], different measures to obtain participants' affective state were used: Self Assessment Manikin [17], Activation-Deactivation (AD-ACL, [27]), pleasure and arousal as evoked by environmental factors [24], and intrinsic motivation inventory (IMI [29]). These instruments are all claimed to be established measures. It is important to realise though that these instruments, as well as other affect scales that are often used in the context of UX studies, were never developed with possible application of usability testing for consumer applications in mind. These scales originate, for example, from research in the field of organizational psychology (e.g., IMI), or clinical psychology (e.g., AD-ACL).

Unfortunately, alternative measures do not seem to be readily available. Other means such as psycho-physiological measures have many drawbacks, especially in the context of testing consumer applications. So, it seems inevitable to adopt these scales, such as the SAM, which is being used extensively in usability testing. Without going into a detailed description of the above mentioned studies, this section provides a number of discussion points with respect to the use of these scales, and some lessons learned.

Attribute of SAFE		Scale and adherence to SAFE			
		Pleasure, arousal & dominance [17]	**Interpersonal trust** [29]	**Aesthetics** [18]	**Usability satisfaction** [19]
1) Construct	a) Theory	Yes	Yes	Yes	Yes
	b) Clarity	Yes: multi-dimensions	Yes: uni-dimentional	Yes: bi-dimensional	Yes: multi-dimensions
	c) Discriminat'	Yes	Yes	Yes	Yes and No
2) Validity	a) Construct validity	Yes	Yes	Yes	Yes
	b) Context	Yes	Yes	Yes	Designed for scenario based evaluation
	c) Sample	N=78: 45 Male	N=222	N=384: 211 Male	N=377
3) Reliability	a) Inter-item reliability	No	Split-half reliability: 0.92	Between 0.60 & 0.78	Exceeding 0.89
	b) Test-retest	No	No	Yes	No

Table 1: Example use of the SAFE[2]

- A general issue with several of the scales used is that quite often the scores obtained for different conditions do not show any significant differences. Obviously, a range of causes might have contributed; the conditions not being distinct enough, too many variables in play, selection and number of participants, or the scales not being sensitive enough. The first two possibilities are related to the fact that the studies in question are conducted in an ecologically valid setting, resulting in conditions that are not too extreme, and relatively rich in features. After all, in an industry context, testing applications in a realistic setting is very important.
- One could also wonder if these measures are effective in typical usability test conditions that often involve a relatively limited number of participants.
- The SAM consists of three subscales, pleasure, arousal, and dominance. The last subscale, however, is quite often not well understood by participants, even when provided with the official instructions as recommended [17]. A growing number of researchers have decided to simply no longer include the dominance subscale in their tests. This raises questions regarding the underlying construct of the scale, or at least questions with regard to applicability of part of the construct in a usability test.
- Valence measures (i.e., pleasure scales) quite often show ceiling effects; obtained scores tend towards the positive end of the scale. In many cases, such strong positive scores are already obtained in connection with a baseline condition. That is, even before the participants experienced the device or condition under investigation,

[2] Table will be illustrated further in the VUUM workshop.

they are in a positive state, and this continues for the test, unless something dramatic happens to impact mood.

- Finally, self-report measures are associated with a number of well-known issues:
  - Participants' answers might be biased or guided by what the participants think is the "right" answer, or by socially desirable answers. Furthermore, the discussed self-reports are retrospective and, thus, potentially subject to distortions.
  - In the case of SAM, such self-report issues might be strengthened because it is a measure that involves only three subscales, making it easy to remember what one filled in before the test condition, for example.

DISCUSSION

The aim of the SAFE is to support the selection of psychometric scales for usability and UX evaluations. The development of the SAFE has built on previous tools to support practitioners in selecting and adopting psychometric scales. For example, [11]'s proposed model that outlined six approaches to measuring usability. The SAFE has been developed to support the selection of psychometric scales for (subjective) usability satisfaction and UX, supporting [12]'s and [15]'s position recommending that practitioners should use established scales. The example use of the SAFE in this paper considered existing scales related to the broad notion of (subjective) usability satisfaction and UX. A number of discussion points emerged.

First, it became apparent that gathering the information to complete the SAFE was not as easy as initially anticipated. The main reason for this was that scales are developed in many different ways, (e.g., using different formats and statistical tests). This diversity is also reflected in the language used to describe scale development. Translating the original publications of the scales to complete the SAFE required an understanding of the development of psychometric scales. It is not anticipated that all practitioners wishing to conduct an evaluation will have this required expertise or time. It is proposed that developers are more transparent in their scale description when publishing their developed scales; [19] is a good example to follow.

Secondly, the majority of usability scales were developed in the 1990s, between 10 and 20 years ago. There are potential issues surrounding measures of this age, which practitioners should consider. For example, the majority of usability scales were developed for workplace systems where tasks are structured and efficiency is of primary importance. When choosing measures to evaluate systems for the home, these factors may not be as prominent in design.

In addition, the relative age of these scales may affect validity, considering the type of systems that would have been used when these scales were developed. Therefore, the practitioner should consider that the original context of scale development may have an effect on the reliability and validity versus today. Thus it is suggested that practitioners should at least acknowledge this and, if there is no time to develop another scale, be aware of the potential problems that could arise when reviewing findings.

Thirdly, to the knowledge of this paper's authors, there is no authority textbook to support the selection of psychometric measures specifically for UCD and UX. Given the abstract constructs that practitioners will be required to measure in UX, a textbook to support and explain to practitioners the advantages of using psychometric measures, and how they are developed, would perhaps increase the validity and reliability of UX scales utilised by practitioners, regardless of whether or not these scales are homegrown. It should, however, be noted that there are some good texts related to the development of scales for the HCI domain. For example, [15] describe a 'basic lifecycle of a questionnaire' and call for a 'battery' of scales to be developed for HCI. They describe the need to achieve validity and reliability, but not how this is done. Another important focus is the importance of language and choice of scale, amongst others. The SAFE builds on this work by supporting the choice of appropriate scales.

It seems that the call for a battery of psychometric measures [15] has been met and the HCI community is now entering a second generation of evaluation methods. Whilst the community has a battery of scales to support evaluation and design, it is clear that the context for usage is now being stretched. It is hoped that the SAFE will iterate to practitioners that just utilising a psychometric scale does not determine if constructs being measured are valid in every situation, and that using a psychometric scale does not guarantee significant findings. In other words, a psychometric scale can be the wrong measure for evaluating a system. Using the SAFE should encourage practitioners to question appropriate use of scales. Nonetheless, the example use of the SAFE to select a psychometric scale in the UX example SAM [17], among others, illustrates the need for new scale development.

Please note that this paper does not discuss scale norms. Psychometric scale norms are not reported in many UX studies. Norms would provide the researcher with the ability to compare findings with other studies (assuming the studies were sufficiently controlled) in the same way that personality and intelligence of individuals can be compared.

CONCLUSIONS

The preliminary development and example use of the SAFE, a tool to support practitioners in selecting appropriate psychometric scales, was illustrated in this paper. Two conclusions have emerged, these are:

- It can be daunting and difficult to gauge whether or not a scale is based on a robust construct, and if the scale is

valid and reliable. This should be clearer in future publications of scales.

- New scales are required to measure new constructs and overcome ceiling effects due to the new emphasis on augmenting existing and enjoyable experiences.

These challenges are pertinent to the VUUM workshop, and can be elaborated on for presentation and discussion.

The SAFE has been developed to support practitioners in adopting established psychometric measures. Nonetheless, the SAFE model still has to be evaluated by practitioners. We intend to build on the SAFE and develop it into a useful and usable design tool, via evaluation with practitioners.

ACKNOWLEDGMENTS

We thank Jan Engel, Nele Van den Ende, and Janneke Verhaegh for lending ideas and comments in support of this research. We thank Marie Curie for funding William Green's research (SIFT project: FP6-14360) and Greg Dunn's research (CONTACT project: FP6-008201).

REFERENCES

1. Blythe, M. A., Overbeeke, K., Monk, A. F., & Wright P. C. (Eds.). Funology: From usability to enjoyment. Dordrecht, The Netherlands: Kluwer Academic, 2003.
2. Clark, L.A., Watson, D. Constructing validity: basic issues in objective scale development. *Psychological assessment*, 7, 3 (1995), 309-319.
3. Cockton, G. Revisiting usability's three key principles. In *Proc. CHI 2008:alt.CHI.* Florence, Italy, ACM, (2008), 2473-2484.
4. Cohen R. J. Swerdlik, M. E., and Philips, S. M. *Psychological testing and assessment. An introduction to tests and measurement.* Mayfield, 1996.
5. Gould, J. D. and Lewis, C. Designing for usability: key principles and what designers think. *Communications of the ACM* 28, 3 (March 1985), 300-311.
6. Greenberg, S. and Buxton, B. Usability evaluation considered harmful (some of the time). *Proc. CHI 2008*. Florence, Italy, ACM, (2008), 111-120.
7. Helander, M.G. Landauer, T.K. and Prabhu, P. (eds.). Handbook of Human-Computer Interaction, 2nd Edition. Amsterdam, North-Holland, 1997.
8. Helander, M.G. Khalid, H.K., and Tham, (Eds.). *Proc of the Int. Conf. on Affective Human Factors Design*. London: Asean Academic Press, 2001.
9. Hoonhout, H.C.M., Human Factors in Matrix LED illumination. In: Aarts, E., Diederiks, E. (eds.). *Ambient Lifestyle*. Amsterdam: BIS Publishers, 2006.
10. Hoonhout, H.C.M. How was the experience for you just now? Inquiring about people's affective product judgments. In: Westerink, J., Ouwerkerk, M., Overbeek, T., Pasveer, F. & de Ruyter, B. (Eds.). *Probing Experiences,* Springer, Dordrecht, NL, 2008.
11. Hornbæk, K. Current practice in measuring usability: Challenges to usability studies and research. *International Journal of Human-Computer Studies*, *64*, *2* (2006), 79-102.
12. Hornbæk, K. and Law, E.L. Meta-analysis of correlations among usability measures. In *Proc. CHI 2007*, ACM Press (2007), 617-626.
13. ISO 9241. Ergonomic requirements for office work with visual display terminals (VDTs) – Part 11: Guidance on usability. *ISO 9241-11*, Geneva, CH, 1998.
14. Isomursu, M., Tähti, M., Väinämö, S., and Kuutti, K. 2007. Experimental evaluation of five methods for collecting emotions in field settings with mobile applications. *International Journal of Human-Computer Studies.* 65, 4 (2007).
15. Kirakowski, J, and Corbett, M, *Effective Methodology for the Study of HCI.* Amsterdam: North-Holland, 1990.
16. Kohne, R. Subjective, behavioral and psychophysiological responses to lighting induced emotional shopping. Unpub master thesis. Utrecht / Eindhoven: Universiteit Utrecht / Philips Research, 2008.
17. Lang, P. J. Behavioral treatment and bio-behavioral assessment: computer applications. In J. B. Sidowski, J. H. Johnson, & T. A. Williams (Eds.), *Technology in mental health care delivery systems*. Norwood, NJ: Ablex, (1980), 119-137.
18. Lavie, T., and Tractinsky, N. Assessing dimensions of perceived visual aesthetics of web sites. *International Journal of Human-Computer Studies*, 60, 3 (2004), 269-298.
19. Lewis, J. IBM Computer Usability Satisfaction Questionnaires: Psychometric Evaluation and Instructions for Use. *International Journal of Human-Computer Interaction,* 7, 1 (1995), 57-78.
20. Norman, D.A. *The psychology of everyday things*. New York: Basic Books, 1988.
21. Nunnaly, J. C. *Psychometric theory.* New York: McGraw-Hill, 1978.
22. Oatley, K, & Jenkins, J. M. *Understanding Emotions.* Cambridge MA: Blackwell Publishers, 2006.
23. Pickard, R. *Affective Computing.* Cambridge, MA: MIT Press 1998.
24. Russell, J.A., and Pratt, G. A description of the affective quality attributed to environments. *Journal of Personality and Social Psychology.* 38 (1981), 311-322.

25. Ryan, R.M. Intrinsic Motivation Inventory: http://www.psych.rochester.edu/SDT/measures/intrins.html. Accessed on April 13 2008.

26. Spector, P.E. *Summated rating scale construction. An Introduction.* Newbury Park: Sage, 1992.

27. Thayer, R.E. Measurement of activation through self-report. Psychological reports, 20 (1967), 663-678.

28. Verhaegh, J., Fontijn, W., and Hoonhout, J. TagTiles: optimal challenge in educational electronics. Proc. of the Tangible and Embedded Interaction conference, Baton Rouge, LA, ACM, 2007.

29. Wheeless, L. and Grotz, J. The Measurement of Trust and Its Relationship to Self-Disclosure. *Human Communication Research*, 3, 3 (1977), 250–257.

A Two-Level Approach for Determining Measurable Usability Targets

Timo Jokela
Joticon, P.O. Box 42
90101 Oulu, Finland
timo.jokela@joticon.fi

ABSTRACT

Usability measures are required for defining clear, unambiguous usability targets for a system or product to be developed. A two-level approach for defining usability targets is proposed. Strategic usability measures are used to define the usability targets at business level. Operational usability measures are used to define the usability targets at the level of user performance. Different usability measures are required for each of the two levels.

INTRODUCTION AND RATIONALE

It is generally agreed as a good project management practice to define clear, measurable targets for product[3] quality characteristics. Clear quality targets provide a clear direction of work and acceptance criteria for the product under development.

In practice, however, usability targets are quite seldom among the measurable goals in system or product development projects. One of the consequences of not having usability goals is that other project objectives dominate and usability is considered only as a secondary objective of a project. The result is a product with usability problems.

Thereby, usability practitioners should aim for defining measurable usability targets and thereby also make their work and role more important in the projects.

The relevance of measuring usability is generally accepted in literature, for example in (Good, Spine et al. 1986) (Wixon and Wilson 1997) (Gulliksen, Göransson et al. 2005) (Jokela, Koivumaa et al. 2006) (Whiteside, Bennett et al. 1988), (Hix and Hartson 1993), (Gould and Lewis 1985)m (Nielsen 1993) (Mayhew 1999) (Dumas and Redish 1993) (Bevan and Macleod 1994) (Macleod, Bowden et al. 1997) (Maguire 1998) (Bevan, Claridge et al. 2002) (ANSI 2001) (Butler 1985) (Karat 1997).

In this paper, we propose a two-level approach for measurable usability targets. *Strategic usability measures* are used to define the usability targets at business level. *Operational usability measures* are used to define the usability targets at the level of user performance. Different usability measures are required for each of the two levels.

The approach presented in this paper is not fully evaluated. Examples of partial application of the approach are provided.

ON USABILITY TARGETS AND MEASURES

Usability is not a generic "on/off" characteristic - one basically cannot state that a product is usable or not usable. Instead, usability is a continuous variable. Thereby, when a product is developed, it is not viable to aim at "good usability" at a generic level. Instead, one should define the level of usability that one is aiming at. In other words, clear measurable usability targets are required.

We basically have two questions when defining usability targets:

1. What are the appropriate usability measures to be used in the development project?
2. What are the appropriate usability targets, i.e. the target values for the measures?

But the matter, however, is more complex. The challenging question is: what is the *process* of finding answers to the questions above? In other words:

1. How does one determine the appropriate usability measures?
2. How does one set the appropriate usability targets (target values for the measures)?

While usability is not an on/off thing, neither is the process of developing usability: one can carry out some basic low-cost usability activities, or one can use a lot of resources for the usability activities. If a project team aims to achieve

VUUM2008, June 18, 2008, Reykjavik, Iceland.

[3] or system

ambitious usability targets, they obviously need more resources for usability activities. If the usability targets are modest, less usability work is probably adequate.

Therefore, setting usability targets is necessarily a question of how much usability resources to assign for a development project. And thereby *setting usability targets becomes a business issue*: how many resources and how much money to assign on usability activities.

The conclusion is that the basis defining usability targets - and usability measures - is the business context of the product under development. One should analyse and understand the business context, in order to be able to set the appropriate usability targets.

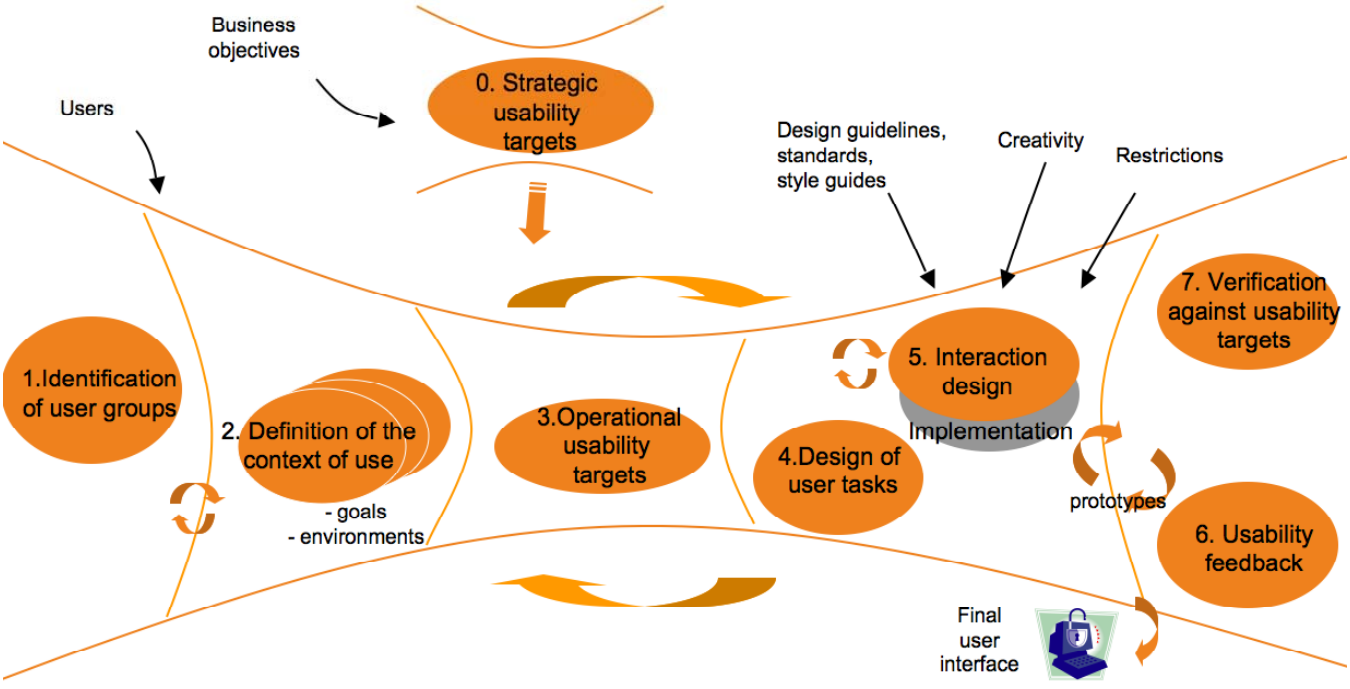

Figure 1. The JFunnel usability life-cycle model

THE TWO SETS OF USABILITY TARGETS

We use the JFunnel[4] model (illustrated in Figure 1) as a reference of usability life cycle.

There are two activities related to the determination of usability targets, and thereby to the determination of appropriate usability measures:

- (0) Strategic usability targets
- (3) Operational usability targets

The strategic usability targets define usability goals at the business level. The operational usability targets define the usability goals the product level. The achievement of the usability targets is later checked in the usability verification activity (7).

[4] JFunnel is a process model used by Joticon. Its earlier version is the KESSU model of the author, reported e.g. in Jokela, T. (2007).

Strategic Usability Targets

Strategic usability targets are used to define the impact of usability at business level: what is the economical and competitive advantage of usability. Strategic usability targets can be defined in terms of attributes known as 'business benefits of usability" (Bias and Mayhew 1994) , such as:

- Reduced training time
- Better system acceptance
- Savings in support costs
- No need of a user manual, or a smaller one
- Improved efficiency of users' work
- Better user satisfaction
- Savings in development costs
- Etc.

A company may choose different levels of ambition in strategic usability objectives. High-level ambition leads to a need of more usability resources; lower level strategic usability objectives can be reached with fewer resources. For example, a system that is learnable without training is obviously more challenging to develop than one without such a target.

Strategic usability targets should be discussed and decided with business management.

Operational Usability Targets

Strategic usability targets form the basis for operational usability targets.

While strategic usability targets are the ultimate goals for usability, they are too high-level issues to concretely guide the development process. Operational usability targets are more concrete, and they provide guidance for the design.

The main reference of defining usability measures is probably the definition of usability from ISO 9241-11: "the extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use" (ISO/IEC 1998). In brief, the definition means that usability requirements are based on measures of users performing tasks with the product to be developed.

- An example of an effectiveness measure is the percentage of users who can successfully complete a task.
- Efficiency can be measured by the mean time needed to successfully complete a task.

The operational usability targets are derived from the strategic ones. For example:

- If a strategic target is set to be that the product under development should be learnable without training, the operational usability targets define what exactly the users should learn.
- If a strategic target is to improve the efficiency of users' work, operational usability targets define what exactly are those user tasks that should be more efficient.

EXAMPLES

The approach is partially applied in a couple of case studies.

A Health Care System

The user analysis of a healthcare system revealed that doctors do not attend training sessions of new information systems. Thereby, it was seen a business benefit if the new system would be learnable without doctors attending the training sessions. Through this, one would achieve good acceptance of the system by doctors. In other words, we had an obvious strategic usability target: the system should be learnable without training.

The operational level usability targets would then define what exactly are those things that should be "learnable without training". In practice, this would mean the definition of those tasks that the user should be able to do without training.

At the context of this project, usability work has not (yet) continued beyond the level of defining strategic usability targets.

A Position Tracking Product

This is a type of product the use of which requires special expertise by the user. In other words, it is not feasible that a person without any background knowledge would be able to use the product just by intuition.

When defining the business objectives, it was therefore decided that the users would be categorised into two main groups: those that have previously used a similar product, and those that are novices. It was decided as a design target that those users with earlier experience should be able to use the product intuitively, without training or user manual. This is an obvious strategic usability target: the product should be intuitively learnable by non-novice users.

During the next step, the user tasks were identified, from the introduction of the product to the daily use of the product. Operational targets were that the users should be able to carry out these tasks without training.

For the other user group – novices – the objective was set that a user should be able to take the product into use without assistance of sales person. (the detailed discussion of this is not in the scope of this paper.

This case study is at the phase where the user interface has been designed at the 'paper prototype' level and software development is in progress.

DISCUSSION

In this paper, a two-level approach for defining usability targets is proposed. The approach is meant to "identify practical strategies for selecting appropriate usability measures and instruments that meet contextual requirements, including commercial contexts" (in the Call for papers of VUUM).

The specific feature of the proposed approach is that usability targets should be business driven. In some business contexts ambitious usability targets are appropriate, while in other contexts more modest usability targets may be justified.

Two preliminary case studies are presented. In both of these cases, a strategic business benefit (among others) was identified as "learnability without training". The strategic usability goals would be most likely different in other cases, depending on the business case and application domain.

REFERENCES

1. ANSI (2001). Common Industry Format for Usability Test Reports. NCITS 354-2001.
2. Bevan, N., N. Claridge, et al. (2002). Guide to specifying and evaluating usability as part of a contract, version1.0. PRUE project. London, Serco Usability Services: 47.
3. Bevan, N. and M. Macleod (1994). "Usability measurement in context." Behaviour and Information Technology 13(1&2): 132-145.
4. Bias, R. and D. Mayhew, Eds. (1994). Cost-Justifying Usability. San Diego, California, Academic Press.
5. Brooke, J. (1986). SUS - A "quick and dirty" usability scale, Digital Equipment Co. Ltd.
6. Butler, K. A. (1985). Connecting theory and practice: a case study of achieving usability goals. SIGCHI 1985, San Francisco, California, United States, ACM Press New York, NY, USA.
7. Card, S. K., T. P. Moran, et al. (1983). The Psychology of Human-Computer Interaction. Hillsdale, Lawrence Erlbaum Associates.
8. Chin, J. P., V. A. Diehl, et al. (1988). Development of an instrument measuring user satisfaction of the human-computer interface. Proceedings of SIGCHI '88, New York.
9. Dumas, J. S. and J. C. Redish (1993). A Practical Guide to Usability Testing. Norwood, Ablex Publishing Corporation.
10. Good, M., T. M. Spine, et al. (1986). User-derived impact analysis as a tool for usability engineering. Conference Proceedings on Human Factors in Computing Systems, Boston.
11. Gould, J. D. and C. Lewis (1985). "Designing for Usability: Key Principles and What Designers Think." Communications of the ACM 28(3): 300-311.
12. Gulliksen, J., B. Göransson, et al. (2005). Key Principles of User-Centred Systems Design. Human-Centered Software Engineering: Bridging HCI, Usability and Software Engineering. M. Desmarais, J. Gulliksen and A. Seffah.
13. Hix, D. and H. R. Hartson (1993). Developing User Interfaces: Ensuring Usability Through Product & Process. New York, John Wiley & Sons.
14. ISO/IEC (1998). 9241-11 Ergonomic requirements for office work with visual display terminals (VDT)s - Part 11 Guidance on usability. ISO/IEC 9241-11: 1998 (E).
15. John, B. E. (1995). Why GOMS? Interactions. 2: 80-89.
16. Jokela, T. (2005). Guiding designers to the world of usability: Determining usability requirements through teamwork. Human-Centered Software Engineering. A. Seffah, J. Gulliksen and M. Desmarais, Kluwer HCI series.
17. Jokela, T. (2007). Characterizations, Requirements and Activities of User-Centered Design - the KESSU 2.2 Model. Maturing Usability. Quality in Software, Interaction and Value. E. Law, E. Hvannberg and C. Cockton. London, Springer-Verlag: 168-196.
18. Jokela, T., J. Koivumaa, et al. (2006). "Quantitative Usability Requirements in the Development of the User Interface of a Mobile Phone. A Case Study." Personal and Ubiquitous Computing 10(6): 345-355.
19. Karat, J. (1997). User-Centered Software Evaluation Methodologies. Handbook of Human-Computer Interaction. M. G. Helander, T. K. Landauer and P. V. Prabhu. Amsterdam, Elsevier Science.
20. Kirakowski, J. and M. Corbett (1993). "SUMI: The software usability measurement inventory." British Journal of Educational Technology 24(3): 210-212.
21. Macleod, M., R. Bowden, et al. (1997). "The MUSiC Performance Measurement Method." Behaviour & Information Technology 16(4 & 5): 279 - 293.
22. Maguire, M. (1998). RESPECT User-centred Requirements Handbook. Version 3.3, HUSAT Research Institute (now the Ergonomics and Saftety Research Institute, ESRI), Loughborough University, UK.
23. Mayhew, D. J. (1999). The Usability Engineering Lifecycle. San Fancisco, Morgan Kaufman.
24. Nielsen, J. (1993). Usability Engineering. San Diego, Academic Press, Inc.
25. Whiteside, J., J. Bennett, et al. (1988). Usability Engineering: Our Experience and Evolution. Handbook of human-computer interaction. M. Helander. Amsterdam, North-Holland: 791-817.
26. Wixon, D. and C. Wilson (1997). The Usability Engineering Framework for Product Design and Evaluation. Handbook of Human-Computer Interaction. M. Helander, T. Landauer and P. Prabhu. Amsterdam, Elsevier Science B.V: 653-688.

What Worth Measuring Is

Gilbert Cockton
School of Computing & Technology, Sir Tom Cowie Campus, University of Sunderland,
St. Peter's Way, Sunderland SR6 0DD, UK.
gilbert.cockton@ sunderland.ac.uk

ABSTRACT
Worth-Centred Development uses worth maps to show expected associations between designs, user experiences and worthwhile outcomes. Worth map elements focus support for Element Measurement Strategies and their selection of measures, targets and instruments. An example is presented for a van hire web site to relate how and why user experience is measured in a worth-centred context.

Author Keywords
Evaluation, Element Measurement Strategy (EMS), User Experience, Value, Worth Centred Development (WCD).

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

INTRODUCTION
The title of this paper is not in the style of Joda from Star Wars, but instead promises a worth-centred answer to "What is Worth Measuring". Interaction design has inherited a strong evaluation tradition from HCI, which largely reflects values from experimental psychology, bringing measures that will not be appropriate for all evaluations [4]. Rather than choose measures that matter to what a design is trying to achieve, usability evaluations re-use procedures and measures from cognitive psychology. In early usability work, psychologists applied measures that they were familiar with, using instruments and procedures refined for experimental validity [15].

There is a strong risk that affective psychology in user experience (UX) research will again privilege disciplinary values from psychology over professional practice values of Interaction Design. The primary issue for choosing any measure or instrument is relevance to design purpose. Measures should reveal what we *really need to know about a design*, and not simply what psychology can 'best' tell us.

VUUM2008, June 18, 2008, Reykjavik, Iceland.

There is thus a tension in the title of this workshop. *Valid Useful User Experience Measurement* (VUUM) may only be possible *on balance*, that is, we may have to trade-off validity, when judged by psychometric standards, in return for relevance, that is, what makes measures matter. The tension goes once we ignore psychometric disciplinary values, and see validity as a question of relevance rather than repeatability or objectivity. Even so, prejudice either way in relation to psychometric standards would be wrong: we want neither a 'tail' that 'wags the dog', nor a 'baby' that is 'thrown out with the bathwater'. Psychometric approaches from affective psychology should be allowed to prove their worth, but we should not assume this. There is enough evidence from usability engineering that cognitive performance such as time on task, error rate, task completion and task conformance are not always the appropriate measures. Too often, usability engineers were like 'drunks under the lamppost', looking for dropped keys where the light is best, and not where they actually lost them. A starting place for relevant measures, to continue the analogy, is where the keys were dropped, and not where the light is best: we should ground all evaluation purpose in design purpose, not in favoured measures.

We need to clearly express what an interaction design seeks to achieve, and from this derive and/or anchor evaluation strategies. This workshop paper motivates and outlines an approach to grounding *Element Measurement Strategies* (EMS) in worth maps, the anchor approach in Worth-Centred Development (WCD). Next, the logic of WCD is briefly motivated, and the evolution of worth maps is summarised. Next, the anchoring of an EMS is presented. The final section presents a van hire example.

WORTH AND EVALUATION
Worth is a *balance* of value over costs. A worthwhile design delivers sufficient value to its intended beneficiaries to outweigh costs of ownership and usage. The benefits of the *ends* (design purpose) outweigh costs arising from the *means* (design features and composition). Although originally used as a *near synonym* for value, to avoid confusing value and values [3], worth actually subsumes value as one half of its cost-benefit balance. *Worth-Centred Development* (WCD) thus needs a *survey of costs*, including *adverse* outcomes (as opposed to worthwhile ones). Designing becomes more complicated, requiring constant

consideration of purchase, ownership and usage costs in context, and reflecting on the extent to which achievable value could sufficiently offset such costs.

In [3], *worth* was introduced in the sense of "such as to justify or repay; deserving; bringing compensation for". Elaborating on this, 'worth' deserves (or compensates) whatever is invested in it, whether this is money (repay), or time, energy or commitment (justify). In short, worth is a motivator. Thus differing *motivations* of users, sponsors and other stakeholders must be understood, as must factors that will de-motivate them, perhaps so much that (perceived) costs outweigh (perceived) benefits, when a design will almost certainly fail in use and/or the market.

WCD broadens designing's view on human motivation, realising that worth is the outcome of complex interactions of Herzberg's *motivators* and *hygiene factors* [11]. The former are satisfiers and the latter are dissatisfiers. For example, at work, salary is a hygiene factor, whereas advancement is a satisfier. Cognitive and affective HCI have largely focused on hygiene factors. Motivational HCI shifts the focus to Herzberg's factors, and does not automatically associate positive impact with favourable hedonic factors, which may be short-lived and transient, with no long term impact. This is immediately relevant to choosing UX measures. Contrary to much current thought in HCI, there may often be no need, in summative evaluation at least, to measure hedonic factors. If value can be demonstrated to be achieved in a worthwhile manner, i.e., with benefits outweighing costs, then one could assume that all contributory hedonic factors are *good enough*, and thus detailed measures of specific emotions would add nothing to a summative evaluation. However, for formative evaluations where adverse hygiene factors do appear to increase costs, degrading or destroying intended worth, then measurement of hedonic factors could be essential.

Worth thus turned out to be not only less ambiguous than value(s), but it also broadened design spaces to cover interacting benefits/value(s) and costs (price, usage and learning effort, synergies with existing product-service ecosystems). However, moving from *models of cost-benefit balances* into evaluation criteria needs to be structured, as WCD as presented in [3] is less concrete and focused than Value-Centre Design (VCD, [2]). An approach was needed to bridge between understandings of worth (in terms of costs and benefits) and worth-centred evaluation criteria. *Worth Sketches* and *Worth Maps* have provided this bridge.

Worth Sketches and Worth Maps

Korean research extended the VCD framework with approaches from consumer psychology. Lee, Choi and Kim [12] took approaches from advertising and marketing and applied them to an existing design to explore the feasibility of grounding VCD in laddering and *hierarchical value models* (HVMs).

Laddering has its origins in construct psychology's repertory grids, developed as a clinical instrument to elicit people's understandings of their relationships. Having identified the significant people in their life, a respondent would be asked to contrast them, in the process exposing constructs that they used to describe personality and social relations. Laddering takes elicitation further by asking what each construct means to the respondent, and applies this questioning recursively to each answer. The result is a rich structure of what matters to someone, with personal constructs related via progressive questioning to an individual's fundamental values. Approaches have been transferred from therapeutic settings to marketing [10], where ladders become *means-end chains*, exposing perceived product benefits, especially ones relating to personal values. Such elicited benefits form a basis for advertising messages or market positioning.

Lee and colleagues' published work is in Korean (*Journal of the HCI Society of Korea* 2(1), 13-24) but there is a draft manuscript in English [12]. Rather than apply laddering to existing designs (as in [10]), laddering concepts can be adapted into a design management approach. [5] and [6] present WCD's early keystone approach, *Worth/Aversion Maps* (W/AMs).

W/AMs retained much of the structure of HVMs (formed by merging ladders at points of intersection). Specifically, they incorporated design elements as *concrete* and *abstract product attributes* and human elements as *functional* and *psychosocial* usage *consequences*. These initial categories were quickly extended by a few example W/AMs in [5,6] to include physiological, environmental and economic consequences. Use of Rokeach's *Instrumental* and *Terminal Values* [13] in original HVMs was replaced by a single W/AM element type (worthwhile outcomes).

In W/AMs, positive *means-ends chains* thus began with concrete product attributes and then could connect with worthwhile outcomes via qualities and a range of usage consequences. Negative means-end chains linked downover from concrete product attributes to adverse outcomes via defects and adverse usage consequences.

Once people other than W/AMs' inventor began to use them, it became clear that modifications were needed. During a WCD workshop for the VALU project (webhotel.tut.fi/projects/uservalues), a need became clear to split concrete product attributes into *materials* (basic unmodified technical system components) and *features* (technically coherent parameterisations, configurations or compositions of materials). When applying W/AMs at Microsoft Research, it became clear that the extensive usage consequences were unwieldly. Information hiding techniques were thus applied to replace all basic consequence types by a single human value element type of *user experiences*. A comparison of Figure 1 below and

[5,6] will highlight significant changes from W/AMs to Worth Maps. There is now no named distinction between Worth Maps and W/AMs, since worth by definition considers cost issues associated with product defects, adverse experiences and outcomes.

EVALUATION MEASUREMENT STRATEGIES

Worth Maps are network structures, drawn as box and arrow diagrams, with elements as nodes and associations as arcs (see Figure 1 later). Paths up from feature elements to worthwhile outcomes constitute positive *means-end chains*, where associations between elements indicate a hoped for causal relationship. These causal associations hopefully combine to deliver intended value in usage. Paths down from features to adverse outcomes constitute negative means-end chains. For a worthwhile interaction design, positive chains will outweigh negative ones in usage.

Worth Map elements focus *Element Measurement Strategies* (EMSs), which can begin at the *worth sketching* stage [8], before associations between elements. There are three types of *positive element* (worthwhile outcomes, worthwhile user experiences, qualities), three of *negative elements* (adverse outcomes, adverse user experiences, defects), and two of *neutral element* (features, materials). Elements also divide into *human value elements* (outcomes and experiences) and *design elements* (the rest).

Initial worth sketch elements can be inspired, grounded, derived or otherwise created from a broad range of *sensitivities* [8,9]. A project team's *receptiveness* to facts, theories, trends, ideas and past designs spreads the breadth of these sensitivities. Some directly address human value(s) independently of specific technology usage. Such human sensitivities form a basis for adding human value elements to a worth sketch (i.e., a worth map without associations [8,9]). Other sensitivities concern creative craft possibilities and technological opportunities. These form a basis for adding design elements to a worth sketch.

An EMS decides on measures, instruments and targets for worth-centred elements. It can start from a worth sketch, but is better supported once elements are associated. Furthermore, WCD *survey data* [5,6,9] that makes a worth map more receptive and expressive, and provide a rich context for the simple box label of a worth element. The intended value statements of [2] are replaced by the balance of worth indicated by human value elements.

Available information guides EMS selection of measures, instruments and targets. An advantage of WCD is that one can identify opportunities for *direct worth instrumentation* (DWI, [6]) (self-instrumentation in [5]). Here system features are added to capture evaluation data above low level system logging, substantially automating direct collection of relevant measures. For example, if the intended worth for a university website includes increased student recruitment, then design could include a sales pipeline that tracks potential students from initial interest through downloads and registrations of interest, to interaction with admissions. If, as preferred in VCD, evaluation planning is completed up to initial outlines of procedures (e.g., test plans) before detailed design completes, then this lets sales pipeline instrumentation features be added to the design (e.g., registration forms, downloadable content for registered prospects, information and advice service, active management of relationship with potential students through email and/or personalised web pages, integration with events and application process). These example features can measure engagement of potential students, tracking them through different stages of interest (i.e., registration, information downloads, use of information and advice service, use of personal pages, attendance at events, applying for entry). WCD's close synergies between design and evaluation would actually shape the design here. DWI, like much measurement, changes the behaviour of what is measured, in this case through design features that not only measure engagement, but form a sales pipeline structure that actively increases engagement. With DWI, the relative effectiveness of different features can be assessed, based on targets set for maximum drop outs at each stage of the sales pipeline. Ineffective features can be dropped and perhaps replaced, and poorly performing features can be improved.

Evaluation should never be separated from design. There will be no *interplay* between them if they are co-ordinated from the outset. Interplay occurs between loosely coupled and potentially independent phenomena. In HCI, design and evaluation have been so separated that their main relationship is interplay, but this is a fault line from HCI's forced marriage between computing and psychology. Aligning evaluation purpose with design purpose replaces fault lines with synergies beyond unpredictable interplay. In particular, DWI goes beyond measuring design effectiveness to actually improving it (by adding features that combine instrumentation and user support).

Evaluation in WCD

Worth sketches and maps identify the main summative evaluation criteria for designs as worthwhile and adverse outcomes. Evaluation criteria selection is an automatic by-product of worth sketching. Worth maps refine sketched worthwhile and adverse outcomes as associations are added, since this often results in reflection on elements and consequent revisions to them. For formative evaluation, all other elements may be measured to provide diagnostic information for iteration. User experiences in worth maps are treated as means to the ends of worthwhile outcomes, but the quality of the UX will impact on the worth achieved by a design. Thus formative measures of UX can become summative when costs of interaction reduce achieved worth. However, the primary evaluation focus remains worthwhile outcomes, with UX only considered in so far as

it adds or detracts from the balance of worth. Appropriate methods are indicated by the *epoch* within which evaluation criteria form. Laboratory testing will thus be inadequate to measure value that takes days or weeks to be realised.

AN EXAMPLE: A VAN HIRE WEB SITE

Figure 1 below shows a *worth map* for a hypothetical van hire site, based on previous commercial UX work at Sunderland. The map is populated by design elements (materials, features, qualities) and value elements (user experiences, outcomes). Element colours are: yellow for *worthwhile outcomes*, pink for *user experiences*, light blue for *qualities*, grey for *features*, white for *materials*, and red edged for *adverse outcomes*. Other negative elements (defects, adverse experiences) are omitted for simplicity.

Any focus on UX measures would relate to the three user experience elements. Figure 1 shows a divide between the experience of forming a *good plan* (at point of order) and the experience of being *in control* (seeded by a good plan but unfolding during actual van use). This corresponds to a hiatus in user activity, which lets user testing in the hiring context (e.g., home, work) formatively explore confidence in quality of van hire plans up to confirmation of booking. Such formative evaluation should not be confused with summative evaluation, which should be worth-centred. In WCD, measuring achieved worth takes precedence over measuring UX.

The main requirements for summative evaluation are to measure achievement of worthwhile outcomes (i.e., worthwhile economic transaction, pleasant sequel, successful gift/transfer/disposal and/or nicer home), plus the extent of the adverse outcomes (costs out of control, arrive late, load won't fit, can't collect van/ find depot). The diamond ended arcs in Figure 1 are *aversion blocks*, which express claims that associated design elements will prevent adverse outcomes. Measures can test these claims. Most need instrumentation at the van hire depot, with data provided directly by customers/and or depot staff. Others require instrumentation of customers soon after the end of a hire, again either with data provided directly at the depot by customers and/or staff, or via customer feedback web pages, phone or email surveys. The 'moments of truth' of all the outcome measures are such that none can be measured through web-site user testing, and most cannot be fully measured until a customer has completed removal of goods, or abandoned delivery for some reason. Referring back to this paper's title, this is what "worth measuring" is.

Worth measuring measures what matters, and what matters is worth. In contrast, [1] associates the van hire W/AM from [5] with a user dissatisfaction and two difficulties:

- Not offering exactly preferred type of van
- Mistakenly booking for wrong dates or wrong type of van
- Booking process taking longer than competitor systems

The first and third have no impact relative to the worth map in Figure 1. If saved time relative to any competitor matters

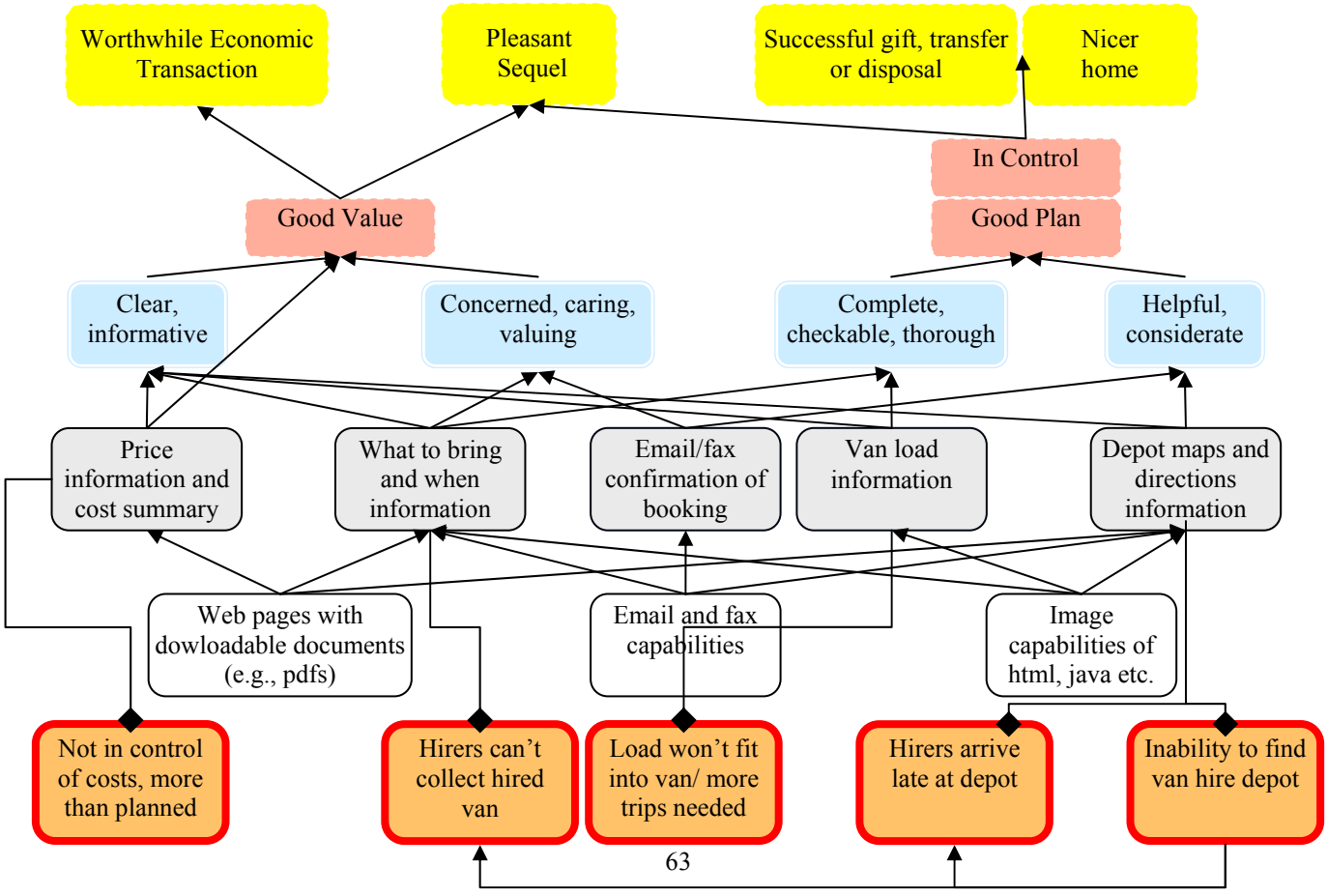

Figure 1. Worth Map for Hypothetical Van Hire Site.

(regardless of overall achievable worth), then it must be added to the worth map as a worthwhile outcome. Similarly, if a specific type of van is required, in terms of the example worth map, its unavailability must impact on load handling, and would be covered by the failure to achieve a *good plan*. Just liking Citroen Berlingos isn't an issue though. Clearly, date or van selection errors would lead to users experiencing a bad actual plan, and by implication a failure of the qualities in Figure 1. In short, these are at best low level formative evaluation concerns that could be explored following failures in worthwhile outcomes. If not, then the worth map is wrong. Such a conclusion may often be reached during EMS formation, where designs can be fixed before they even form in detail.

[1] endorses advice in [15] to base targets on an existing system or previous version, strongly competitive systems, or performing tasks with no computer. However [15] advised that such (revisable) targets be chosen by the development team, engaging engineering staff in user-centred design but giving them control of development risks [6]. Importantly (but overlooked in [1]) over half of [15] addresses why they *abandoned* this approach [6].

For the VUUM focus, we must state when UX measures become important. In WCD, they are secondary to worth measures. Note that a *pleasant sequel*, that is good times after hiring a van, is a worthwhile outcome that may follow from the UXs of receiving *good value* and *feeling in control*. The coalescing of these experiences minimizes potential attrition from van hire, which may otherwise lead to disappointment, frustration and/or fatigue. It is important to include worth elements that capture the impact of interactions on user resources for subsequent activities.

Where UX measures do not concern worthwhile outcomes, their role is largely formative. UX measures tend to be associated with emotion (e.g., joy) or meaning (e.g., fun, trust). A now common view in HCI is that these can and should be measured in isolation. This however ignores the reality that UXs are holistic, taking their coherence from the meanings that coalesce during them. Feelings, beliefs, system usage, system response and actions in the world are almost inseparable in the unfolding of coherent experiences. Measuring emotions has to be related to their role. At each point in a UX, emotions increase or reduce a user's confidence and comfort with the progress of interaction. Each strong emotion both evaluates the immediate past and judges the immediate future. Emotions indicate comfort with the experience so far, and expectations for the remaining experience. They are bound up in interpretation and anticipation *of something*, from which they are inseparable. Emotions must thus be measured in the context of UXs with the aim of understanding their impact within the dynamics of UX. Once again, formal measures should not become relevant until summative evaluation of achieved worth has indicated degradation or destruction as a result of poor design. In short, you don't measure UX until you know you have specific problems with worth. When you do, measures become valid and useful in so far as they can locate causes of degraded or destroyed worth in poor UXs that fail to coalesce into their intended positive meanings (or, for adverse experiences, achieve an anticipated but adverse meaning that must be avoided).

In WCD, UXs are named to reflect their expected meanings. When summative evaluation reveals a shortfall in achieved worth, formative evaluation needs to explore why. Working into the centre of a worth map from outcomes, the next elements to look at are worthwhile and adverse UXs. UX Frames (UEFs) are a WCD notation that supports a detailed focus on the unfolding of UX. They have a flexible table format. The example schematic UEF in Figure 2 includes feelings, beliefs, system usage, system response and actions in the world but other UX aspects

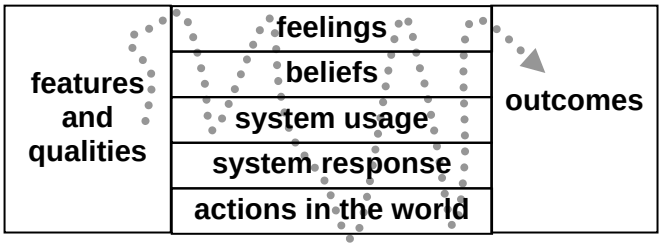

Figure 2. Schematic User Experience Frame (UEF)

could be added, and existing ones removed. The contents of a UEF depend on the UX being represented.

The schematic rotates a UEF, which is normally a table of columns below a header of outcomes and a footer of features and qualities. Header and footer items correspond to associated worth map elements for the UX. Column items are time ordered, with the most recent at the top. In the schematic, time runs left to right. The dotted arrow represents an abstract scenario, indicating how specific features and/or qualities give rise to feelings, then beliefs, that lead on to usage, further feelings and beliefs, further system usage and response an action in the world and then one last feeling that closes the meaning of the user experience and gives rise to associated outcomes.

There is not space to provide a detailed UEF example. The purpose of Figure 2 is to illustrate the context within which emotions and meanings form. It is this context that identifies relevant emotions and user interpretations that *could* be measured in formative evaluation. A range of instruments could be designed to collect specific measures, including semantic differentials, questionnaires, playback (think aloud in response to video replay), facial expression monitoring, and video analysis. Similarly, user perceptions of qualities in Figure 1 can be measured using some of these instruments, especially semantic differentials.

CONCLUSION

The purpose of this paper is to show how and when relevant measures can be identified. Only once success criteria and related design and interaction elements are identified in context can selection of measures, targets and instruments proceed. Evaluators have to be instrument makers, but this should not be their key skill. Instead they should work closely with design and related roles in a project team to isolate the critical success and failure factors for a proposed system. Sensitivities, worth sketching and worth mapping are WCD approaches that structure a process that identifies critical factors as worthwhile and adverse outcomes. These become the major focus for summative evaluation in WCD.

Much UX research in HCI is currently approaching evaluation from another position, which assumes the value of measuring emotions in isolation, without regard to their actual role in UX. This approach also looks for affective psychology for measures and instruments. As with usability in 1980s HCI, this position starts with a set of disciplinary assumptions that do not hold for designing as a creative, holistic, judgemental activity. We do not design for specific emotions without regard to context. Instead, we design for positive experiences of which emotions are only one aspect. Meanings and outcomes bring emotions with them, and not vice-versa. We must thus measure emotions only where needed, and mostly in formative evaluation.

We must trade proven validity of generic psychometric tools for relevance of evaluation measures in design contexts. The purpose of evaluation is to support design, not psychology. This paper has focused on WCD approaches to highlight what matters in evaluation, as opposed to what can be easily or reliably measured. Where the two conflict, design's values take precedence over science's. Science can support design, but not direct it. Designing is a creative activity involving much judgement and interpretation, and this must extend to evaluation too. If HCI evaluation remains a (pseudo-) scientific silo alongside interaction design, its influence and effectiveness will stagnate. We have an opportunity in early UX research to avoid the mistakes of usability evaluation, which uncritically applied measures, instruments and procedures from cognitive psychology. It would be wrong for affective psychology to take UX evaluation into the same blind alley of evaluation methods with little downstream utility.

The D in WCD stands for *development*, not design. WCD takes a holistic view of development that allocates clear synergistic roles to design and evaluation, unified around the common purpose of achieving intended worth. This holistic context, supported by representations such as UEFs and worth maps, lets us identify key criteria for summative evaluation and potential diagnostic measures for formative evaluation. UX measures largely relate to the latter. If intended worth is demonstrably achievable through interaction with a design, there is little need for detailed measurement of UX. The latter is generally a *means to an end*, not an end in itself. By placing UX in a worth-centred context, WCD guards against inappropriate use of measures and instruments from affective psychology. The key requirement is to understand evaluation in a development context, and not as a scientific enterprise.

ACKNOWLEDGMENTS

The WCD framework was developed with support from a NESTA Fellowship (www.nesta.org). User Experience Frames (UEFs) were introduced to replace consequence elements of Hierarchical Value Models (HVMs) at Microsoft Research Cambridge when I was a visiting researcher. Alan Woolrych directed the commercial UX testing of van hire websites on which the example is based. My initial interest in adapting HVMs for WCD was inspired by research by Lee and colleagues [12].

REFERENCES

1. Bevan, N. (2008) "UX, Usability and ISO Standards," in *CHI 2008 Workshop on User Experience Evaluation Methods in Product Development*, last accessed 25/4/08 as www.cs.tut.fi/ihte/CHI08_workshop/papers/Bevan_UXEM_CHI08_06April08.pdf
2. Cockton, G. (2005) "A Development Framework for Value-Centred Design," in *CHI 2005 Extended Abstracts*, ed. C. Gale, ACM, 1292-95,
3. Cockton, G. (2006) "Designing Worth is Worth Designing," in *Proceedings of NordiCHI 2006*, eds. A.I. Mørch, *et al.*, ACM. 165-174.
4. Cockton, G. (2007) "Make Evaluation Poverty History", *alt.chi 2007*, last access 19/4/08 at www.viktoria.se/altchi/submissions/submission_gilbert_0.pd
5. Cockton, G. (2008) "Putting Value into E-valu-ation," in *Maturing Usability: Quality in Software, Interaction and Value*, eds. E. Law et al., 287-317, Springer.
6. Cockton, G. (2008) "Some Experience! Some Evolution," in *HCI Remixed*, eds. T. Erickson and D.W. MacDonald, MIT Press, 215-219, 2008
7. Cockton, G. (2008) "Revisiting Usability's Three Key Principles", in *CHI Extended Abstracts*, 2473-2484.
8. Cockton, G. (2008) "Sketch Worth, Catch Dreams, Be Fruity, in *CHI 2008 Extended Abstracts*, 2579-2582.
9. Cockton, G., (2008) "Designing Worth: Connecting Preferred Means with Probable Ends," to appear in *interactions*, 15(4)
10. Heitmann, M., Prykop, C. and Aschmoneit, P. (2004): "Using Means-End Chains to Build Mobile Brand Communities," in *HICSS 2004*, last accessed 29/4/08 at http://www.hsw-basel.ch/iwi/publications.nsf/id/289
11. Herzberg, F. (1966) *Work and the Nature of Man*, Ty Crowell Co; Reissue edition.

12. Lee, I, Choi, B. and Kim, J. (2006) *Visual Probing of User Experience Structure: Focusing on Mobile Data Service Users*, HCI Lab, Yonsei University, Seoul.
13. Rokeach, M. (1973) *The Nature of Human Values,* Free Press.
14. Rosenbaum, S. (2007) "The Future of Usability Evaluation: Increasing Impact on Value," in *Maturing Usability: Quality in Software, Interaction and Value*, eds. E. Law, E. Hvannberg and G. Cockton, Springer.
15. Whiteside, J., Bennett, J., and Holtzblatt, K. (1988) "Usability engineering: Our experience and evolution," in *Handbook of Human-Computer Interaction*, 1st Edition, ed. M. Helander., North-Holland, 791-817,

Is what you see what you get? Children, Technology and the Fun Toolkit

Janet C Read
ChiCI group
University of Central Lancashire
Preston, UK
+44(0) 1772 893285
jcread@uclan.ac.uk

ABSTRACT
There are several metrics used to gather experience data that rely on the user opinions. A common approach is to use surveys. This paper describes a small empirical study that was designed to test the validity, and shed light on, opinion data from children that was gathered using the Smileyometer from the Fun Toolkit.

This study looked at the ratings that children gave to three different interactive technology installations, and compared the ratings that they gave before use (expectations) with ratings they gave after use (experience). Each child rated at least two of the installations and most rated all three

Different ratings were given for the different installations; this suggests that children can, and do, discriminate between different experiences. In addition, three quarters of the children, having encountered the product, changed their minds on at least one occasion and gave a different rating after use than they had before. Finally, the results showed that in most cases, children expected a lot from the technologies and their after use rating confirmed that this was what they had got.

KEYWORDS
Fun Toolkit, User Experience, Children, Smileyometer, Errors

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

VUUM2008, June 18, 2008, Reykjavik, Iceland.

INTRODUCTION

Asking children about technologies is a commonly used method for evaluation, especially where opinions or some measure of experience is required. Determining what children think about technologies can give insights into software appeal and can provide a designer or developer with feedback as well as with a very useful endorsement of a product. Many academic papers, especially design papers, use the opinions of end users as an indication of value or worth.

Collecting children's opinions is not always straightforward and there are several known hurdles that need to be overcome if the opinions that are gathered are to have real value. Where opinions are elicited by questioning, an area of concern is the degree to which the child understands the question being asked. Answering questions in complex as not only does the child need to understand and interpret the question being asked, retrieving relevant information from memory, and integrating this information into a summarised judgment, in most cases the child then has to code the answer in some way, typically using rating scales. .Researchers often discuss the importance of the question-answer process in determining the reliability of responses provided by children in surveys [1].

Factors that impact on question answering include developmental effects (the child may be too young to understand); language (the child may not have the words needed), reading age (the child might be unable to read the question), and motor abilities (the child may be unable to write neatly enough), as well as temperamental effects including confidence, self-belief and the desire to please. Another factor, especially evident in surveys where respondents are being asked to pass attitudinal judgments [2], is the degree to which the respondent satisfices or optimises his or her answers.

For validity in a survey, optimising is the preferred process. Optimising occurs when a survey respondent goes thoughtfully and carefully through all three or four stages of the question and answer sequence. Thus, the respondent will ensure that he or she understands what the question is really about, will take care to retrieve an appropriate answer,

will think about how the response should be articulated in order to make it clear what he or she is thinking, and, where there is a rating to apply, will carefully consider where on the rating scale each answer should be placed.

Satisficing is a midway approach (less good than optimising but better than guessing!) and occurs when a respondent gives more or less superficial responses that generally appear reasonable or acceptable, but without having so carefully gone through all the steps involved in the question-answer process. Satisficing is not to do with guessing answers or choosing random responses, it is to do with picking a suitable answer. The degree or level of satisficing is known to be related to the motivation of the respondent, the difficulties of the task, and the cognitive abilities of the respondent [3].

Especially because of motivation and cognitive maturity, it is the case that children are prone to satisficing - they find survey participation difficult and often lack the abilities to articulate answers or to fully understand the questions.

As well as there being concerns with children's responses to questions given at a moment in time, there are also concerns about the stability of these responses over time, both of these concerns affect the reliability of self-reported data from children. It is generally considered that all that can be elicited from a survey is a general feel for a product or a concept with a particular group of children at a particular time, e.g. [4]. This is not especially good news for researchers hoping to evaluate user experience with children, but there is some comfort in that, in HCI work, the reliability of responses is generally not critical (as could be the case where a child is being interviewed as part of a criminal investigation).

Given the general difficulties of surveying children, there are several useful approaches that can be taken to make the process more valuable and at least satisfactory for all the parties. One of these is to use specially designed tools that fit the task of evaluating user experience for children [5], [6]. These tools may be easy to use but because they are relatively novel, they may be unreliable and may be used inappropriately. The remainder of this paper describes a suite of tools, the Fun Toolkit [7], and then looks at one tool in particular, the Smileyometer, before describing a study that tested the reliability of the Smileyometer and the use of it with children.

THE FUN TOOLKIT

The fun toolkit is a selection of tools that have been designed to gather opinions from children about interactive technology. They use only essential language, lend themselves well to the use of pictures and gluing and sticking for input, they are fun and attractive, and they reduce some of the effects of satisficing.

The Three Main Tools

The Fun Toolkit was originally proposed to have four tools. Over time, this original fun toolkit, has been tested and evaluated with the result that it is now considered to be sufficient to have three main tools for use. These are the tools described here.

The Smileyometer (shown in Figure 1) is a discrete visual analogue scale (VAS). Research has shown that children aged around seven and over can use these VAS effectively and so, for the Fun Toolkit, this was an obvious choice of tool to consider [8]. In validations of the Smileyometer, studies by the authors and others have shown VAS to be useful for considerably younger children than was previously reported, but it should also be noted that when these scales are used to elicit opinions about software or hardware products, younger children are inclined to almost always indicate the highest score on the scale [9].

During use children are asked to tick one face. This makes it a very easy tool for the children, but it also includes textual information to improve the validity. The Smileyometer can be easily coded; it is common to apportion scores of 1 to 5 for the different faces; if used in this way.

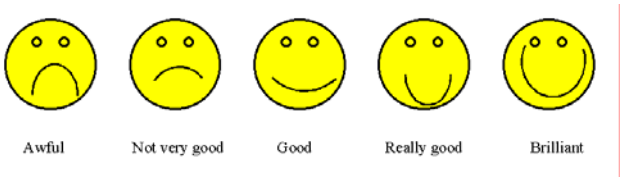

Figure 1. The Smileyometer

The Smileyometer has been used in many studies to rate experienced but also anticipated fun; in the latter case the child is asked, before an experience, to tick a face according to 'How much fun will it be?'

In many evaluation studies, the desire is not to carry out an isolated evaluation but to rank a series of connected or competing activities or technologies. These comparative evaluations can help the evaluator determine which may be more appealing or which may be least fun. Repeated instances of the Smileyometer can be used but the Fun-Sorter, a variation on a repertory grid test [10], can be useful in this context. The Fun-Sorter (shown in Figure 2) has one or more constructs and a divided line (or table) that has as many spaces in it as there are activities to be compared. The children either write the activities in the spaces, or for younger children, picture cards can be made and placed on the empty grid.

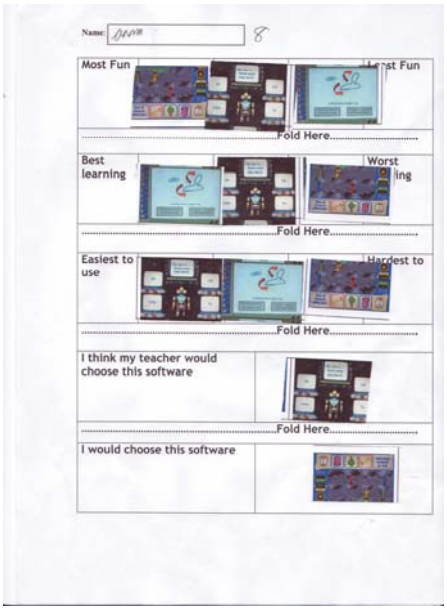

Figure 2. A completed Fun-Sorter

The attractiveness of the Fun-Sorter, over repeated use of the Smileyometer or the Fun-Sorter is that it forces the children to rank the items or activities in order. Where the Smileyometer has been used for comparative studies, the tendency of children to give maximum ratings to products inevitably makes it difficult to see which item, from a set of four, is the least fun.

The 'Again–Again' tool (see Figure 3) can also be used to compare activities / events. This table lists some activities or instances on the left hand side, and has three columns headed Yes, Maybe, and No.

Would you like to do it Again?

	Yes	Maybe	No
Visit U Boat	✓		
Puppet show		✓	

Figure 3. Part of an Again–Again table

The child ticks either yes, maybe or no for each activity, having in each case considered the question 'Would you like to do this again?' The background to this is that, for most children, a fun activity would be wanted to be repeated!

Tests of the Validity of the Fun Toolkit Instruments

The Fun toolkit has been around for quite some time now and has been evaluated in several studies. The major findings to date have been that;

- Children almost always pick 5 for everything, this is more the case the younger the children (REF Removed) – Either, everything children see is great, or, children see everything as great!
- Children tend to report their experience as being the same as their expectation (REF Removed) – This suggests that what is seen at the beginning is very important
- The Smileyometer is measuring Fun rather than Usability (REF Removed)
- The tools, when used together, give similar results (REF Removed) – This suggests that the tools are fundamentally not flawed and that children know what they are doing when they use them.

Of these findings, there are several research questions that merit further study. In the work that reported that children 'got what they expected', the technologies being evaluated were all attractive and were quite 'even' in their attractiveness. Whether this result would hold with less attractive technology was worth investigating as it was hoped, but not known, that children would be able to discriminate more than that research study suggested!

THE STUDY

In this study, a group of children were given the Smileyometer to rate three different technology experiences. There were three hypotheses:

1. Children would give different overall experienced (after fun) ratings according to the predicted experiences of the three technology installations
2. Children would generally rate the technologies highly, with 5 being a common score
3. Some of the children would demonstrate discriminatory behaviour according to their expected and experienced fun that would align with the predictions of the evaluators.

Method

The study took place at a Mess day at the University of Central Lancashire. 24 children aged 6 and 7 (9 girls, the rest boys) came to the lab and each child took part in 10 minute activities on the three technologies in question. The technologies were intended to offer different user experiences.

- Interface A was a tangible block game that was entirely novel to the children, it looked fun, was very easy to play, and it was expected that all the children would enjoy playing it.

- Interface B was a PDA application that was probably new to most of the children, it looked interesting, it was not especially easy to use (due to the size of the stylus), and the children were expected to find it rather dull as they were carrying out a writing activity.
- Interface C was a laptop PC with paint installed on it. This was not especially novel but was expected to be greeted with enthusiasm by the children. However, there was a catch as the interface was relatively difficult to work due to there being no mouse attached to the computer and so children had to draw using the touch pad. Thus, it was expected that most children would find this hard.

Children moved around the activities in a predetermined order. Each child moved from A to B to C but this was in a cyclic order and so some began with C (then moved on to A and B), others started with B (and moved to C then A). Some children only visited two applications as there were other activities happening in the room that they were involved in. The children came to each application in groups of four and each child participated.

On arriving at an activity, the activity was briefly described, the children were told what they were going to do and shown the technology (but it was not demonstrated), then the child completed a 'Before Smileyometer' to indicate how much fun they thought it was going to be. Having completed the activity, the child then completed an 'After Smileyometer' (the before version was hidden from view at this point) to indicate how much fun the activity had been.

The scores from the children were coded as follows: Brilliant = 5, Really good = 4, Good = 3, Not very good = 2, Awful = 1.

At the end of the activity the children were thanked for their assistance and given certificates to take home.

Results

In total, 59 pairs of ratings were gathered. 16 of these related to Interface A, 21 related to Interface B, and 22 related to Interface C. The average before and after ratings for the three interfaces are shown in Table 1:

	Interface A	**Interface B**	**Interface C**
Before	4.75	3.71	4.36
After	4.75	3.52	3.64
Number rating	16	21	22

Table 6. Average ratings, before and after

For Interface A, all but two of the children gave a 5 rating for expected (before) fun and all but two children (a different two), gave a 5 for experienced (after) fun.

For Interface B, 10 children gave it 5 before playing and 9 gave it 9 after playing with the other children rating it from 1 – 4 in both cases.

Interface C was rated at 5 by 14 children before it was played and by 9 after playing. The other children rated it at 3 or 4 before playing and between 1 and 4 after playing.

Across the interfaces there were some trends in respect of individual children, six of the twenty four, in every instance gave the same ratings before they used the technology as they did after use. Another six never changed their ratings by more than one from before use to after use, whereas half the children, on at least one occasion made a shift of two or more. Overall, the score of 5 before and 5 after was the most popular choice. Table 2 shows the frequency and spread of the different scores, the row indicates the after score, the column the before score.

25	3	3	1		**5 after**
3	3	3	1		**4 after**
5	2	1	1		**3 after**
5					**2 after**
1		1		1	**1 after**
5 before	**4 before**	**3 before**	**2 before**	**1 before**	

Table 7. The 59 results

This table shows that out of the 59 pairs, 25 were rated (5,5) – it also shows that only one interface was rated as 'awful' before use, and only 3 rated as awful after use.

Discussion

The results found in this study confirmed the three hypotheses whilst also indicating some interesting areas for further study.

Looking at the average ratings in Table 1, it appears that children rated fun differently in each interface with there being a large difference between interface A and the other two. Given that interface A was designed to be a game, and the aim of the interface was to have fun, this result seems to indicate that children are able to discriminate for fun even when using such a raw tool as the Smileyometer. This was not previously reported as in earlier evaluations of the Fun Toolkit; the technologies being compared were very similar.

Table 2, once again shows (as demonstrated in studies in [7] that on the whole, children will rate things with a maximum (5) score. 46 of the 59 ratings given overall included at least one five showing that children generally expect and experience high levels of fun, with their expectations (39 with a 5), in this instance, appearing slightly higher than their experienced fun (32 with a 5), and evidenced by lower average ratings as seen in Table 1.

On interface C, the paint application, at least half the children experienced significant difficulties with the touch pad but the interface still got an average score that would indicate it was better than good and almost 'really good' and this confirms earlier results that suggest that children 'put up' with poor usability if what they experience is fun, and that fun, therefore, is more important than usability (a finding noted in [11]).

Considering how the ratings changed from before to after it is interesting to note that interface C, which looked easy and familiar had a high rating for expected fun which fell once it had been played. In this regard the children's scoring matched the predictions of the evaluators.

CONCLUSION

This work shows some important trends for measuring the experience of children with technology. It also adds to the already published work on the use of the Fun Toolkit and, specifically, the Smileyometer. The relationship between expected and experienced fun has been shown to vary according to the technology being evaluated.

Where good, fun technology is being evaluated, the children will, by and large, rate it very highly and the expected and experienced fun will hardly vary. Where the technology is familiar but difficult to use, expectations will be quite high but children will adjust their ratings based on the experience with the product. Technology that appears uninteresting will be rated lower before use than technology that appears interesting.

Further work will explore the use of the other Fun Toolkit metrics with 'less attractive' and 'less functional' technologies and will test the ideas developed in this small study with a larger cohort of children and across a more diverse set of technologies.

ACKNOWLEDGEMENTS

Thanks to the researchers of the ChiCI group at UCLan and the children of the local primary school for their assistance in this study.

REFERENCES

1. Borgers, N., J. Hox, and D. Sikkel, *Response Effects in Surveys on Children and Adolescents: The Effect of Number of Response Options, Negative Wording, and Neutral Mid-Point.* Quality and Quantity, 2004. 38(1): p. 17 - 33.
2. Krosnick, J.A., *Response Strategies for coping with the Cognitive demands of attitude measures in surveys.* Applied Cognitive Psychology, 1991. 5: p. 213 - 236.
3. Borgers, N. and J. Hox, *Item Non response in Questionnaire Research with Children.* Journal of Official Statistics, 2001. 17(2): p. 321 - 335.
4. Vaillancourt, P.M., *Stability of children's survey responses.* Public opinion quarterly, 1973. 37: p. 373 - 387.
5. Airey, S., et al. *Rating Children's Enjoyment of Toys, Games and Media*. in *3rd World Congress of International Toy Research on Toys, Games and Media*. 2002. London.
6. Hanna, L., K. Risden, and K. Alexander, J, *Guidelines for usability testing with children.* Interactions, 1997. 1997(5): p. 9-14.
7. Read, J.C. and S.J. MacFarlane. *Using the Fun Toolkit and Other Survey Methods to Gather Opinions in Child Computer Interaction*. in *Interaction Design and Children, IDC2006*. 2006. Tampere, Finland: ACM Press.
8. Shields, B.J., et al., *Predictors of a child's ability to use a visual analogue scale.* Child: Care, Health and Development, 2003. 29(4): p. 281 - 290.
9. Read, J.C., et al. *An Investigation of Participatory Design with Children - Informant, Balanced and Facilitated Design*. in *Interaction Design and Children*. 2002. Eindhoven: Shaker Publishing.
10. Fransella, F. and D. Bannister, *A manual for repertory grid technique*. 1977, London: Academic Press.
11. Read, J.C., *Validating the Fun Toolkit: an instrument for measuring children's opinions of technology* Cognition Technology and Work, 2008. 10(2): p. 119 - 128.

Comparing UX Measurements, a case study

Arnold P.O.S. Vermeeren
TU Delft, Industrial Design Engineering
Landbergstraat 15,
2628 CE Delft, The Netherlands
E: a.p.o.s.vermeeren@tudelft.nl,T: +31(0)152784218

Anita H.M. Cremers
TNO Human Factors
Kampweg 5, P.O. Box 23
3769 ZG Soesterberg, The Netherlands
E: anita.cremers@tno.nl, T: +31(0)346356310

Joke Kort
TNO Information and Communication Technology
Eemsgolaan 3, P.O. Box 1416
9701 BK Groningen, The Netherlands
E: joke.kort@tno.nl, T: +31(0)505857751

Jenneke Fokker
TU Delft, Industrial Design Engineering
Landbergstraat 15
2628 CE Delft, The Netherlands
E: j.e.fokker@tudelft.nl, T: +31(0)152789677

ABSTRACT
In this paper we present our preliminary findings of an informal comparison of different types of User eXperience (UX) measurements and methods during the field trial of a peer-to-peer file sharing application called Tribler.

Author Keywords
User experience (UX), usability, framework, measurement, methods, requirements, peer-to-peer file sharing.

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

INTRODUCTION
Within the TUMCAT[5] project, a generic, in-situ Testbed for UX measurements of Mobile, Context-Aware applicaTions is being developed. An informal comparison was made of TUMCAT findings from a first Tribler field trial, a laboratory usability test and an internet survey [2, 3]. In this paper we report some preliminary findings of a second comparative study. In this second Tribler field trial three studies (a longitudinal study with TUMCAT tools, an expert review and a laboratory test) are cross-validated on their ability to measure different User eXperience (UX) aspects. The measurements within the three studies were set up to answer UX research questions based on the UX framework as described in [3]. Below, the UX framework and different TUMCAT measurement tools are explained in more detail, followed by a description of the field trial, the results and some conclusions and future work.

VUUM2008, June 18, 2008, Reykjavik, Iceland.

UX FRAMEWORK
Shifts in focus from functional aspects of Information and Communication Technology (ICT) towards ICT as an integrated part of a user's everyday life create a situation in which the success of a product in terms of value for the user can no longer be fully understood based on mere usability (e.g. efficiency, effectiveness, satisfaction, learnability, etc.). It becomes more apparent that the nature of user activity is opportunistic and situated, based on the relation between individual predispositions (personality, norms/values, emotions/moods, goals and preferences, earlier experiences/knowledge, etc), product interaction and context (social, physical as well as virtual) [4].Furthermore ICT increasingly influences our everyday experiences and emotions, which is often referred to as the UX [4, 5]. In [3] we presented a tentative framework for UX addressing the changes mentioned above. This framework draws heavily on research by Wright & McCarthy [4] and by Desmet & Hekkert [6], as well as on our own previous research [1]. The framework is meant to provide an overview of all possible aspects of UX. Further, it should help to identify (1) possible important product (interaction) aspects, such as design and context features that may play a role in the UX, and (2) suitable evaluation methods and instruments (see Section "From UX Framework to Measurements") for measuring resulting emotions and experiences, at appropriate moments before, during and/or after product interaction.

Design elements (outer circle of Figure 1) are the product features a designer can manipulate, such as form, colour (providing aesthetics), interaction flow or choice of

[5] For more information on TUMCAT see: http://www.freeband.nl/project.cfm?id=1126&language=en

specific application features/functionality (providing compositional experiences and meaning).

Aesthetic design aspects (middle circle) relate to a product's capacity to delight one or more of our sensory modalities [6]. These aspects are closely related to design elements such as look, feel, sound, colour, form and their specific composition. Aesthetic aspects hardly involve cognitive processing and lead to emotions (inner circle) such as thrill, fear, excitement, unease, awkwardness, the perception of speed, time and its boundaries [4, 6].

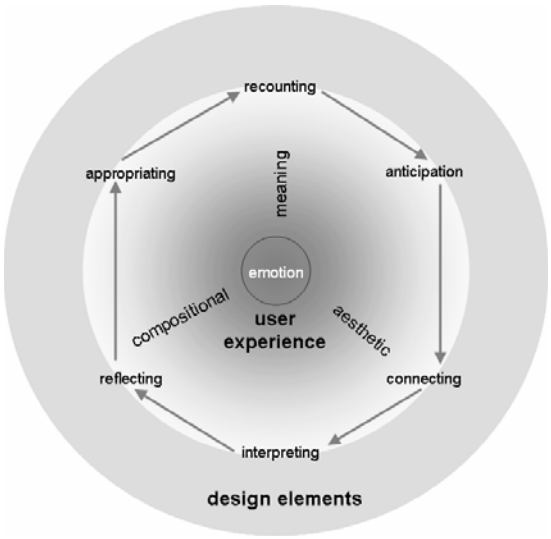

Figure 2. UX framework

Compositional design aspects (middle circle) are closely related to usability, pragmatic and behavioural characteristics of an interactive product. They encompass aspects directly related to interaction with the product, such as the design of interactive elements, action possibilities, narrative structure [4], intended use and function [6]. Compositional aspects result in emotions (inner circle), such as (mis)understanding of a product's workings, (un)predictability of a product's behaviour or outcomes, and feelings of (not) making progress. Aspects related to meaning are those through which a designer tries to realize the users' higher order goals, by making users attribute *meaning* (middle circle) to the product. Through sense-making processes people are able to recognize metaphors, assign personality or other expressive characteristics, and assess the personal or symbolic significance of a product [6]. Attributed meaning can result in emotions (inner circle) such as anger, joy, satisfaction, fulfilment, fun, bliss, closeness to one's own identity or image, inspiration and regret [3, 5].

According to Wright & McCarthy [4], experiences come to life via a process of sense-making. Sense-making refers to processes such as: *Anticipation*, expectations about our experience with a product; *Connecting*, our first experience (sense of speed, thrill, openness, etc.) with the product and its interaction without giving meaning to it; *Interpreting,* interpretation of our interaction ((non-) instrumental, (non-) physical [6]) with the product, relating it to our goals, desires, hopes, fears and previous experiences, leading to experiences such as anxiety, unease, desire, willingness to continue); *Reflecting,* making judgements about our experiences, evaluating them and comparing them with other experiences, resulting in satisfaction, excitement or boredom, a sense of achievement, etc; *Appropriating,* trying to identify ourselves with the experience or changing our sense of self as a consequence of the experience; *Recounting,* reliving experiences by recounting them to ourselves and others, finding new possibilities and meaning in them. In some sense-making processes such as *connecting*, cognitive processing is hardly involved and experiences are a direct result of the perception or sensation of the aesthetics of the product [7]. In other sense-making processes cognition plays an important role in making sense of the product's workings (compositional structure) and its attributed meaning (e.g. in anticipation, interpreting, reflecting, appropriating and recounting). The sense-making processes described here can all happen in parallel or successively [4].

FROM UX FRAMEWORK TO MEASUREMENTS

In earlier work we identified requirements for new UX measurement tools; such tools should [1]:

- Measure *qualitative* as well as *quantitative* and *subjective* as well as *objective* data related to the UX, as unobtrusively as possible. For example, they should capture users' behaviour (e.g. application feature usage) and users' opinions and emotions about the application's features and relate both kinds of data to understand the results in the context in which the measurement was taken.
- Support *long term studies* and *timed* or *continuous measurements* in which different measurement qualities (as mentioned above) are combined to create a picture of the overall UX or the dynamics therein (e.g. changes over time, at a specific moment in time, as a mean over time).
- Enable researchers to perform *situated measurements* to approach or realize UX measurements that reflect users' experiences in their every day lives.

The above requirements were used to develop four different kinds of TUMCAT measurement tools [1]:

- Logging tools: automatic capture of user's behaviour or product usage (e.g., mouse-clicks, keystrokes, application feature usage).
- Sensing tools: automatic capture of a user's context (e.g. physical, social and/or virtual such as physical location, buddy lists and buddies online, active application window, system/application status).

- Experience sampling[6] tools: automatically generated self report requests sent once or multiple times to the user, triggered by user activity (logged data), specific context aspects (sensed data) or a predefined time schedule.
- User generated content tools: provided to the user to give feedback about any desired topic when a user feels like it (e.g. feedback button linked to a feedback form, discussion forum related to the product).

CASE STUDY: TRIBLER

Peer-to-peer (P2P) networks are networked computer devices (nodes, terminals) that permit other computer devices to utilize locally available storage space, communication bandwidth, processing capacity, and sometimes even hardware components. P2P technology has brought clear advantages over client-server architectures [8]. Yet, the success of any P2P system fully depends on the level of cooperation among users. Technical enforcement of this cooperation is limited. Therefore, within the Tribler project an alternative approach is chosen: making use of knowledge from (social) psychology on altruistic behaviour for developing cooperation inducing features [8].

What is Tribler?

Tribler is a peer-to-peer television (P2P-TV) system for downloading, video-on-demand and live streaming of television content [2]. The system not only gives users access to all discovered content and other users in the network, but also provides the means to browse personalized content with a distributed recommendation engine and an advanced social network that each user creates implicitly and explicitly. An advantage of having trustworthy friends in Tribler is that they can speed up the downloading process by donating their own idle bandwidth. Figure 2 shows the main screen of Tribler version 4.0.

Measuring Experiences with Using Tribler

Measurements were taken by conducting a longitudinal field study using the TUMCAT measurement tools, as well as by conducting a laboratory study in which test participants were asked to perform some tasks with Tribler and by asking international experts on usability and user experience from the COST294 network to provide their expectations on the users' experiences with Tribler.

Longitudinal field study

In the field trial 39 users were asked to download Tribler and use it at home for five weeks in any way they wanted. Together with their download of the Tribler software they also received a customized version of the uLog[7] software that *logged* all activities users performed with Tribler. An other tailor-made software package sent the loggings to a central server that the researchers used for data gathering. This software also *sensed* application and user related data that Tribler stores on the user's computer (e.g., a list of the user's friends on Tribler and the library of downloaded files). To be able to gather subjective data from the test participants, the server also sent out *Experience Sampling* questions to users which opened in a browser window. These questions were automatically triggered based on specified logged activities or combinations of activities.

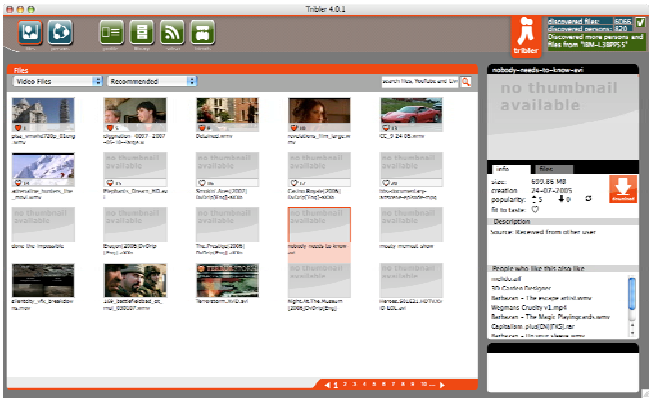

Figure 3. Tribler 4.0 main screen.

Responses to these questions were automatically sent back to the server. In addition, test participants had a possibility to provide *user generated content* to the research team by clicking a feedback icon in the system tray of their computer and then entering text in a text window that would pop up. By clicking the submit button this feedback was sent to the server and was merged with the logged data file. Test participants were gathered through personal contacts, as well as through flyers and posters at institutes of higher education as well as technical institutes of applied science. They were given a € 25 reward after the five weeks had passed. In the fourth week of the study they were asked to perform some specified tasks and to fill in a questionnaire. Tasks and questionnaire were similar to those of the laboratory study

Laboratory study

In the laboratory study ten test participants were asked to download the Tribler software, explore its use and then perform some tasks with it, in the presence of a facilitator. Sessions lasted about one hour to one hour and a half. Example tasks were: *accepting an invitation to become*

[6] Experience sampling is a set of empirical methods that are designed to repeatedly request people to document and report their thoughts, feelings, and actions outside the laboratory and within the context of everyday life.

[7] For more information on the uLog software see: http://www.noldus.com/site/doc200603005

friends with someone; searching for a specific movie; asking a Tribler friend to donate bandwidth to speed up downloading. Tasks were designed to make users use basic downloading functionality, as well as features that distinguish Tribler from other competing file sharing software (e.g., recommendations, donating bandwidth to friends). Users were asked to think aloud while the facilitator was sitting next to them. The facilitator took an active listener approach [9] and asked questions at appropriate times to probe the user's understanding of the software's functionality, e.g., *Have you noticed the checkmark / exclamation mark in the top right corner? What does it mean to you?* Retrospective interviews were conducted to further discuss some of the events or problems that came up during sessions and users were asked to fill in the Attrakdiff[8] questionnaire on paper.

Expert review

In the expert review five experts were provided with the Tribler software to review it. They were asked various questions assessing their opinions on the software. They were also asked to estimate what problems users would have with the software, and to fill in the Attrakdiff questionnaire imagining what the users would answer. Communication with the experts was solely through email.

Measurements in the Different Studies

The various types of measurements taken are discussed using the UX framework explained in section "UX Framework", focusing on compositional, meaning and aesthetics UX aspects. It is assumed here that in case of software for voluntary use its attractiveness largely determines whether or not and to what extent the software will be used. The attractiveness of using the software is closely related to the emotional response to (the exposure to and interaction with) the software. In addition, factors like the user's context, a user's pre-dispositions and constraints play an important role (e.g., familiarity and availability of other software with similar functionality, previous knowledge of P2P file sharing software, technicalities and compatibility of the user's system).

The sense-making process of the UX framework (depicted in the outer circle) implies that there are *temporal issues* involved. One has to take into account the user's experience in relation to the product prior to or at the very start of the actual interaction (e.g., anticipation and connecting), as well as during and after the interaction (i.e., the other elements of the sense-making process).

Data was gathered on usage and on the attractiveness of the software, in addition to data that related to the different UX aspects. Temporal issues were addressed in different ways in the three types of studies. In the field trial data gathering on usage was automatically done over the five week study period. In case of the laboratory study, users were asked UX and usage related questions after their initial experiences with Tribler, as well as during the whole session and afterwards, imagining future usage. In the expert review experts were asked to provide their expectations on the UX after their first confrontation with the software, as well as after further inspection and trial.

Usage

In the field study, TUMCAT's automated logging facilities made it possible to monitor actual usage of Tribler at the level of UI events (e.g., shifts in input focus, key strokes) and the abstract interaction level (e.g., providing values in input fields) [10], providing insight into how usage developed over the five week time period. In the laboratory study test participants were asked to imagine whether and how they would use Tribler at home in three different ways*: 1)* after a short exploration of the software: *what kind of things do you think you would use Tribler for at home; 2)* after each task: *would this be something you can picture yourself doing at home? (scale 1-5) and 3)* at the end of the session: *Would you consider starting to use Tribler at home? How frequently? What would you use it for? Under what circumstances?* In the expert review, the experts were asked similar questions: *1)* after a quick exploration of Tribler: *Do you think that the target group may indeed consider downloading, or actually download the software once they are aware of its existence? What do you think users would want to use Tribler for in particular? 2)* in relation to Tribler's friends and recommendations facilities: *Do you think that (in the long term) Tribler users will actively engage in using this facility?*

Attractiveness

One of the ways of measuring attractiveness was by having the test participants fill in the Attrakdiff questionnaire. This questionnaire is based on a theoretical model in which pragmatic and hedonic aspects of experiences are thought to affect the attractiveness of a product. A number of questions assess the product's attractiveness for the user. In the field test, participants who had actually used Tribler were asked to fill in the questionnaire in the fourth week of their use. In the laboratory study participants filled in the questionnaire after their task performance. In the expert view, the experts were asked to fill in what they thought users would fill in.

In addition to this questionnaire, participants were also asked about their appreciation of specified Tribler features on a 5-point scale. In the field study, these questions were asked as experience sample questions in response to the use of the specified function (e.g., after the 1st and every 5th time of using the function). In the laboratory study all

[8] For more information see: http://www.attrakdiff.de/

participants were asked to answer the questions after having performed a task related to that function. In the expert view, experts were instructed to answer the questions imagining how they thought users would appreciate the specified functionality.

Finally, insight into the users' emotions in reaction to the product were received through spontaneous (written) feedback by the participants (field test) as well as through the retrospective interviews and observed (verbal and no-verbal) reactions during task performance (laboratory study).

Compositional aspects, meaning and aesthetics

Compositional aspects relate to the pragmatic aspects of interactions, including usability problems, effectiveness etc. In all three studies measurements included those questions of the Attrakdiff questionnaire that related to the software's pragmatic quality. In the field study, the logged and sensed data in combination with spontaneous user feedback shed a light on pragmatic issues. In the laboratory study such data were gathered by observing task performance and through retrospective interviews. In the expert review, the experts were given the tasks used in the laboratory study as suggestions to structure their search for usability problems in the software and were also asked to provide some reasons on why they thought a problem would occur.

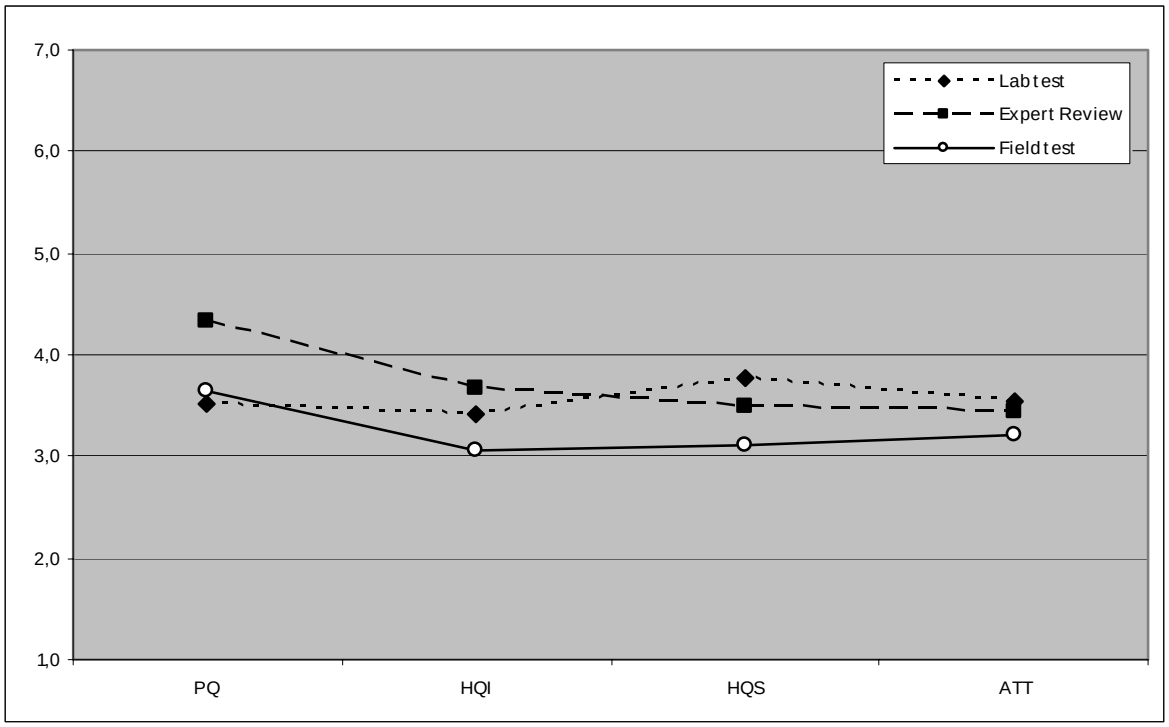

Figure 4. Attrakdiff results (PQ: pragmatic quality, HQI: hedonic quality (identification), HQS: hedonic quality (stimulation), ATT: attractiveness. 1 indicates low score, 4 indicates neutral. Lab test (n=11), Expert view (n=5), Field test (n=6)).

In addition, the experts were explicitly asked about whether they thought users would understand the logic behind specified functions. *Aspects of meaning* were measured by the questions in the Attrakdiff questionnaire that related to the software's hedonic qualities (all three studies). It was also expected that in all three studies spontaneous feedback from users and explanations on expected future usage would provide some insights on these aspects. *Aesthetic aspects* were not explicitly included in any of the measurements of the three studies. However, spontaneous remarks or feedback could provide some data on such aspects.

PRELIMINARY RESULTS AND ANALYSIS

A preliminary analysis of the data was conducted, which allows drawing some tentative conclusions on how the findings from the three studies relate to each other in terms of UX framework elements.

From the field study it became clear that 15 of the 39 test participants had started using Tribler during the five week period; these were given the questionnaire. Only 6 of them filled it in, being active users in the sense that on occasional days they spent some hours on using the software. From questions asked to the 15 'real' users and

through informal communication it became clear that reasons for not using Tribler had to do with software packages crashing, the combination of software making computers too slow, as well as them being used to competing software. Statements participants in the laboratory study made, predicted that utility and usability problems (compositional aspects) would prevent some people from using the software. In the expert review similar views were expressed. A superficial analysis of the logging and sensing data gathered in the field study could not provide a clear view on compositional aspects like usability problems; it was too difficult to trace back what users were trying to do from mere log data. The data from the laboratory study, as well as from the expert review provided a rich view on problems and their possible causes. This was the kind of data that designers in the Tribler development team could most readily use in their attempts to redesign Tribler. Aspects of meaning were found mainly in the laboratory study through spontaneous think aloud utterances, as well as in retrospective interviews. This related to issues like terminology being too dull for them, not wanting to use social software, not valuing recommendations of files based on popularity, issues in relation to (appreciating or not appreciating) illegal downloads and (laughing about or feeling offended by) adult content. In the expert review two experts commented on issues of meaning, only in relation to not appreciating illegal content and feeling offended by adult content. In the field study only two users commented on issues of meaning. They did so in the same way as the experts in the expert review.

As to the aesthetic aspects, three experts mentioned issues of graphical design and layout in their comments. Generally they indicated they liked the graphical design, although also once 'bad layout' was mentioned, as well as being disturbed about the software not showing thumbnails in the files view. In the field test only two participants commented on aesthetic aspects, mentioning they disliked the library and files view. In the laboratory study a rich mix of comments was given (by 7 participants) on aesthetic issues. Opinions here were more mixed in the sense of valuing the design or not, but also of the level of design detail they commented on, ranging from comments on specific icons to an opinion on the general looks of the software. Many of the comments were spontaneous exclamations when confronted with a new screen.

From the Attrakdiff questionnaire we found that pragmatic aspects in the field trial and the laboratory test were more or less the same and scored more negatively than in the expert review. Hedonic identification with, and stimulation by Tribler scored lowest in the field test, followed by the laboratory test and the experts' opinions. Attractiveness scored similar over the different studies, see figure 3 for an overview.

DISCUSSION AND CONCLUSIONS

Analysis of these studies is still preliminary and ongoing. In the following we describe our first findings. We can only gain insight in the actual usage of Tribler and its' functions by using tools such as logging and sensing over a longer period of time. Laboratory tests and expert reviews give weak predictions about usage which do not agree with the results from logging and sensing. Logging and sensing do not provide any explanation for the actual usage. We found lower scores in the field test for both meaning and hedonics (from the Attraktdiff questionnaire) than from the laboratory study or the expert reviews. Actual usage makes people aware of the match between product and their higher order goals (or meaning). This may account for this result.

Laboratory tests and expert reviews give a detailed and rich insight in compositional, pragmatic issues such as the usability. Logging and sensing do not provide this detailed insight. Logging tools used, monitored user activity on the levels of UI events (e.g., shifts in input focus, key events) and the abstract interaction level (e.g., providing values in input fields). To generate information about usability issues, the data needs to be transformed to higher levels of abstraction such as domain or task related levels (e.g., providing address information) or to goal and problems related levels (e.g., placing an order) and compared to predefined or automatically identified sequences of user activity within these levels [10].

Attractiveness scores were slightly negative on the Attrakdiff questionnaire in all three studies. Though the score was slightly negative we doubt if this is a valuable predictor for the (non-)usage as measured during the field test (it seems too bold that a slightly negative score can result in such overwhelming non-usage). The influence of attractiveness on usage might be dependent on people's motivation to use a product (externally motivated or internally motivated product interaction). Aesthetic aspects are difficult to measure for interactive products. We found attractiveness as well as aesthetics are best measured through direct interaction with participants. For experts it's difficult to formulate an opinion about attractiveness and aesthetics from a user's viewpoint.

FUTURE WORK

Based on the conclusions above, we identified several issues for future work on improving UX measurement in long-term field studies:

1. Tools or methods to raise the level of logging data from the UI events and the abstract interaction level to higher levels such as domain and task or even goal and problem levels and means to analyze the data on these higher levels (detecting sequences and interpreting these sequences);

2. Approaches for automated gathering of data that provide (a) insight into reasons of (non-) usage at the level of products or product features, (b) insight into why or how the product succeeds (or not) in making the user attribute meaning to it; (c) rich and detailed data on usability issues, especially those that relate to longer term usage and are highly affected by the personal situation of the user. Especially for topic 2a more detailed theoretical knowledge in the area of UX would help to relate the various aspects in the framework to each other and to a product's attractiveness and (non-) usage in real-life.
3. A practical approach for assessing the aesthetic aspects of a product.

ACKNOWLEDGEMENTS

We would like to especially thank the following people for their contributions to these studies: Gilbert Cockton, Effie Law, Jens Gerken, Hans-Christian Jetter and Alan Woolrych from the COST294 network. Ashish Krishna and Anneris Tiete for their contribution in executing and analysing the study and the preliminary results. The test participants for their active contribution and feedback. Leon Roos van Raadshoven, Paul Brandt, Jan Sipke van der Veen and Armin van der Togt for their technical contributions.

REFERENCES

1. Vermeeren, A.P.O.S. and J. Kort, Developing a testbed for automated user experience measurement of context aware mobile applications, in User eXperience, Towards a unified view, E. Law, E.T. Hvannberg, and M. Hassenzahl, Editors. 2006, COST294-MAUSE: Olso. p. 161.
2. Fokker, J.E., A.P.O.S. Vermeeren, and H. de Ridder, Remote User Experience Testing of Peer-to-Peer Television Systems: a Pilot Study of Tribler, in EuroITV'07, A. Lugmayr and P. Golebiowsky, Editors. 2007, TICSP, Tampere: Amsterdam, The Netherlands. p. 196-200.
3. Kort, J., A.P.O.S. Vermeeren, and J.E. Fokker, Conceptualizing and Measuring UX, in Towards a UX Manifesto, COST294-MAUSE affiliated workshop, E. Law, et al., Editors. 2007, COST294-MAUSE: Lancaster. p. 83.
4. McCarthy, J. and P. Wright, Technology as Experience. 2007: The MIT Press. 224.
5. Pals, N., et al., Three approaches to take the user perspective into account during new product design. International Journal of Innovation Management, 2008. **In press**.
6. Desmet, P. and P. Hekkert, Framework of Product Experience. International Journal of Design, 2007. **1**(1): p. 10.
7. Hekkert, P., Design Aesthetics: Principles of Pleasure in Design, Delft University of Technology, Department of Industrial Design: Delft. p. 14.
8. Pouwelse, J.A., et al., Tribler: A social-based peer-to-peer system. Concurrency and computation: Practice and experience, 2008. **20**(2): p. 127-138.
9. Boren, T.M. and J. Ramey, Thinking aloud: reconciling theory and practice. IEEE Transactions on Professional Communication, 2000. **43**(3): p. 261-277.
10. Hilbert, D.M. and Redmiles D.F, Extracting usability information from user interface events. ACM Computing Surveys (CSUR). 2000. **32**(4): p. 384 – 421.

Evaluating Migratory User Interfaces

Fabio Paternò
ISTI-CNR
Via G. Moruzzi, 1
56124 Pisa
fabio.paterno@isti.cnr.it
+39 050 315 3066

Carmen Santoro
ISTI-CNR
Via G. Moruzzi, 1
56124 Pisa
carmen.santoro@isti.cnr.it
+39 050 315 3053

Antonio Scorcia
ISTI-CNR
Via G. Moruzzi, 1
56124 Pisa
antonio.scorcia@isti.cnr.it
+39 050 315 3127

ABSTRACT
Migratory user interfaces are user interfaces that allow the users to change device and continue the task from the point they left off. In this paper we discuss aspects that are important in evaluating migratory interfaces and report on a usability test we carried out for this purpose.

Author Keywords
Migratory user interfaces, Usability, Adaptation, Consistency

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

INTRODUCTION

Migratory user interfaces provide users with the ability to change the interaction device and still continue their tasks through an interface adapted to the new platform. We use the concept of 'platform' to mean those environments that have similar interaction capabilities (such as the form-based graphical desktop, the vocal device, the digital TV). Migration can involve devices belonging to different platforms. In some cases, migration can be used to improve the user's experience by switching to a better suited devices (bigger screen, more graphical power, …) or to a more efficient communication channels that can guarantee better QoS (shorter delays, higher bandwidth).

The increasing availability of various types of interactive devices has raised interest in model-based approaches useful for logically specifying relevant user interface information in appropriate models which are often described by using XML-based languages. In our migration environment we refer to TERESA¬XML language [1].

VUUM2008, June 18, 2008, Reykjavik, Iceland.

In the paper, we first introduce some relevant design dimensions for migratory interfaces, and next we introduce the software architecture of our environment for supporting such interfaces. We also discuss usability aspects that characterise the migration and report on some user tests. Lastly, some concluding remarks along with indications for future work are provided.

MIGRATION DIMENSIONS

Different dimensions have been identified regarding the migratory user interfaces. Such dimensions can help the designer in understanding the characteristics of the different migrations and also better identify possible relevant aspects as far as the migration evaluation is concerned. Amongst them we identified the following ones:

- Activation Type: the agent triggering the migration (the user or the system or a mixed initiative) is triggered.
- Type of Migration: This dimension analyses the 'extent' of migration, as there are cases in which only a portion of the interactive application should be migrated.
- Number/Combinations of Migration Modalities This dimension analyses the modalities involved in the migration process.
- Type of Interface Activated: This dimension specifies how the user interface is generated in order to be rendered on the target device(s) (pre-computed or dynamically generated or a mix of the two options).
- Granularity of Adaptation: The adaptation process can be affected at various levels: the entire application can be changed depending on the new context or the user interface components (presentation, navigation, content).
- How the UI is Adapted: Several strategies can be identified regarding how to adapt user interfaces after a migration process occurs.
- The Impact of Migration on Tasks: The impact of migration on tasks depends on how the UI is adapted (e.g.: the device characteristics might

enable reduction/increase on the range of tasks supported by each device).

- Context Aspects. During adaptation of the user interface the migration process can consider the context in terms of description of device, user and environment.
- Implementation Environment. The migration process can involve different types of applications (Web, Java, .NET, …).
- Architecture of the migration infrastructure: client/server; and peer to peer,
- Usability Issues: The adaptation and generation of user interfaces should be performed taking into account the main principles of usability.

OUR ARCHITECTURE FOR MIGRATION

Starting with a pre-existing Web version for the desktop platform, our environment is able to dynamically generate interfaces for different platforms and modalities exploiting semantic information, which is derived through reverse engineering techniques. The environment currently supports access to Web applications and is able to generate versions for PDAs, digital TV, various types of mobile phones and vocal devices and support migration through them. Our solution migrates the presentations as they are needed. There is no transformation of the entire application in one shot, which would be very expensive in terms of processing and time. Consequently, the migration process takes only a few seconds from when it is triggered to when the user interface appears on the target device. The application implementation languages currently supported include XHTML, XHTML Mobile Profile, VoiceXML and Java. The tool has been tested on a number of Web applications (e.g. Movie rental, Tourism in Tuscany, Restaurant reservations, on-line shopping…). Our environment assumes that the desktop Web site has been implemented with standard W3C languages and filters out some dynamic elements of the page (e.g. advertising and animations).

Our migration platform is composed of a number of phases. The process starts with the environment discovering the various devices available in the current context, which also implies saving related device information, e.g. device interaction capabilities. Whenever a user accesses an application, the proxy server included in our migration environment receives the Web page, and modifies it in order to include scripts that allow the collection of information about its state. Afterwards, it sends the modified page to the web browser of the source device. When the migration server receives the request for migration (which specifies the source device, the target device) from the source device, it triggers the following actions:

- the migration server detects the state of the application modified by the user input (elements selected, data entered, …) and identifies the last element accessed in the source device. At the same time, the server gets information about the source device and, depending on such information it builds the corresponding logical descriptions by invoking a reverse engineering process. The result of the reverse engineering process, together with information about source and target platforms is used as input for carrying out a semantic redesign phase in order to produce a user interface for the target platform.
- afterwards, the migration server identifies on the target device the logical presentation to be activated, and it consequently adapts the state of the concrete user interface with the values that have been saved previously.
- Finally, the generation of the final user interface from such a logical description for the target platform is delivered, and the resulting page is sent to the browser of the target device in order to be loaded and rendered.

It is worth pointing out that in our environment both the system and the user can trigger the migration process, depending on the surrounding context conditions.

USABILITY EVALUATION

The usability evaluation of migratory interfaces should consider their two main components: continuity and adaptation. In addition, it should also consider the usability of the migration client, which is used by the users to trigger migration and select the target device.

Continuity includes the ease with which users are able to immediately orient themselves in the new interface and therefore are able to recognise and have the feeling of a "continuous" interaction. Indeed, it implies that users can easily recognise that the new user interface is the follow-up of previous interactions through another device while carrying out the same task. If users do not recognise such a situation, they may even be confounded by the user interface presented on the target device. Different factors can affect such recognition and they might make continuing the interaction in a seamless way problematic from the user's point of view. For instance, factors might include i) whether long time has passed since the last interaction, which therefore might be difficult to remember, or/and ii) whether the adaptation process has changed the user interface rendered on the target device in such a way that the users do not recognise that it enables them to logically *continue* the performance of their tasks from the point where they left off in the source device. For instance, the new user interface should clearly highlight the elements that were changed during the interaction with the previous

device. Due to screen size limitations, such highlighting may not be immediately evident.

In the first case, adequate and (adaptable) feedback messages should be identified in order to take into account the aspects related to the time passed [2]. For instance, if a long time has passed, the user interface activated on the target device could enrich the quantity of feedback information initially provided to the users. This should allow them to better contextualise the interaction and remember the action(s) already done on the user interface of the source device (e.g.: which sub-tasks they already accomplished on the device, and possibly which overall task they were about to complete, ..). Regarding time, another aspect that can affect the user experience is the time necessary for the migration to take place: a migration that takes a long time to complete may compromise again the feeling of a continuous interaction and have a negative impact on the user's experience.

In the case of the adaptation process, it should be a trade-off between two sometimes conflicting aspects: on the one hand the devices are different and thus need an interface adapted to the varying interaction resources, on the other hand users do not want to change their logical model of how to interact with the application at each device change. Therefore, suitable mechanisms should be identified in order to make the users easily recognise the features for supporting interaction continuity.

Further aspects that should be taken into account regard the users' model of the application and how to find information in it. In particular, the user's familiarity in using the same application through different devices can affect the usability of migratory user interfaces. Indeed, if the users already have some familiarity in using the same application through different devices (but without having experienced the migration capabilities), on the one hand they might more easily recognise the effect of adaptation on the user interface of the target device and therefore they should feel more confident with the adapted user interface. However, on the other hand they are likely to notice the changes that have occurred on the user interface of the target device(s) as an effect of the migration process, which might become a potential source of disorientation for them because they might expect a different presentation and/or navigation.

Another factor that can impact the usability of migratory user interfaces is the *predictability* of the effects of the migration's trigger, namely the ability of the user to understand the effect of triggering a specific type of migration. To this regard, the migration client should be designed in such a way to effectively allow the users to understand the effects that a migration trigger will provoke on the target device(s). The predictability aspect is particularly relevant especially when different options for migration are offered to the users. Therefore, it also depends on the number of different migration options the migration client can offer to them, and to what extent such options were designed in such a way to be able to effectively communicate to the user the result that will be achieved by activating each of them. This might be especially relevant when the migration process involves more than two devices (and/or more than one user), although different options can be also available with only one user and two devices involved (partial/total migration). For example, users should be able to easily understand to what actual devices correspond to the list of devices that can be selected as migration targets. In case of partial migration, users should be able to easily predict what part of the interface will migrate and on which device the results of the interactions will appear. An aspect connected with the predictability is the *learnability*, which is the easiness with which the users become familiar to the migration features and therefore they are capable of controlling its features and related effects. If the system is easy to be learned, even occasional users should not have difficulties to use it.

USER TESTS

In order to understand the impact of migration on users and the usability of migratory user interfaces, some user tests were performed in order to evaluate some of the abovementioned usability-related aspects. The trans-modal migration functionality, from graphic to vocal, has been tested in the first version of the migration service.

Trans-modal Migration Test

The first test was performed on the "Restaurant" application [3], which allows users to select a restaurant, accessing its general information and make a reservation. The interfaces of the test application for desktop, PDA and vocal platforms differ both in the number of tasks and their implementation. For example, the date insertion is a text field in the desktop version and a selection object in the PDA version, while the insertion of free comments was removed from the vocal interface. When migrating on the vocal platform, it might happen that there are some vocal inputs which are automatically disabled after migration because the user already provided the corresponding values through the graphical device. This type of support is provided in order to facilitate the user in efficiently completing the application form. Since we were interested in considering multi-device environments, both a desktop PC and a PDA were used as graphic source platforms. The 20 users involved were divided into two groups. The first one started with migration from PDA to vocal platform and then repeated the experiment starting with the desktop. The second one started with the desktop and repeated the test using the PDA. The scenario proposed to the users was that they start booking a table at a restaurant by using their graphic device (the PDA is supposed to be used if they are on the move, the desktop if they are at home or at office) and load the "Restaurant" application. At some point, due to

some external conditions (e.g.: the low battery level of the PDA), they had to ask for migration towards the available vocal device and then complete the Restaurant Reservation task. After the session the users filled in the evaluation questionnaire.

Users were recruited in the research institute community. The average user age was 33.5 years (min 23 - max 68). Thirty percent of them were females, 65% had at least undergraduate degrees and 55% had previously used a PDA. Users had good experience with graphic interfaces but far less with vocal ones: on a scale of 1 to 5, the average self-rating of graphic interface skill was 4.30 and 2.05 for vocal interfaces. For each migration experiment, users were asked to rate from 1 (bad) to 5 (good) the parameters shown in Table 1.

Parameters	Desktop to vocal	PDA to vocal
Migration client interface clearness	3.4	3.9
Interaction continuity easiness	4.35	4.65
Initial vocal feedback usefulness	4.1	4.2
Vocal feedback usefulness	4.25	4.25

Table 1. User rating for trans-modal migration attributes.

Vocal feedback was provided via both the initial message, recalling the information inserted before migration, and a final message at the end of the session about the information inserted after migration. We chose this solution as the most likely to reduce user memory load but in any case, after the test, we asked the users if they would have preferred only an overall final feedback instead. Also, we asked whether they noticed any difference between the graphic and vocal interface with the aim of finding out whether they could perceive the different number of supported tasks. The numeric test results were interpreted taking into account the answer justifications and free comments left in the questionnaire and considering user comments while performing the test.

The service in itself was appreciated by users. Many judged it interesting and stimulating. The users had never tried any migration service before and interacted with it more easily in the second experiment, thus, showing it was easy to learn through practise, once the concepts underlying migration were understood. Interaction continuity received a slightly higher score in the PDA-to-vocal case.

Parameters	Desktop to vocal	PDA to vocal
Only final vocal feedback preferred	Yes 20% - No 80%	Yes 20% - No 80%
Noticed different task set	Yes 25% - No 75%	Yes 20% - No 80%

Table 2. User preferences and salience of task differences

This might be justified by the fact that the PDA and the vocal versions are more similar in terms of tasks supported than the desktop and the vocal ones. In any case, the difference in ease of continuity between the two platforms (desktop and PDA) is small, thus the interaction continuity ease is influenced, but not compromised. Both the initial and the overall feedback through the vocal application were judged positively (Table 1). The vocal feedback design was appreciated and 80% of the users would have not changed its style. One concern was the potential user disorientation in continuing interaction, not only due to the change of modality, but also due to the different range of possible actions to perform. Only 20-25% noticed the difference and it was perceived more in the desktop-to-vocal case (Table 2). This first study provided some useful suggestions to bear in mind when designing user interfaces intended for trans-modal migration. The modality change does not cause disorientation but must be well supported by proper user feedback balancing completeness while avoiding boredom. The differences in interaction objects used to support the same task on different platforms were not noticed at all, while the difference in the number of tasks supported was. This has to be carefully designed in order to reduce as much as possible any sudden disruption in the user's expectation.

Test on the New Version of the Migration Environment.

The first test confirmed our choices concerning the support for trans-modal migration and gave useful suggestions for improving the migration environment. The work undertaken after the first test resulted in the new version of the migration environment discussed in this article and we performed a new test with a different application: the "Virtual Tuscany Tour". The goal was to have further empirical feedback on the migration concept and the new solution supporting it. Since very few changes concerned the trans-modal interface transformation while many more affected the unimodal (graphic) migration, the new test considered the (unimodal) desktop-to-PDA migration, in order to evaluate the new functionality.

The test application is the Web site of a tourism agency specialized in trips around Tuscany, an Italian region. Interested users can request material and get detailed information about trips around the sea and the beach or around the countryside and the mountains; it is also possible to get overviews of good quality places to sleep. From this Web site it is also possible to get information about the main social events, entertainment, places to taste good food and wine, sports activities and other useful information, such as the weather forecast and public transportation. During the test, users could freely interact with the application and at a certain point they were asked to fill in a form requesting tourist information about Tuscany. The form had to be filled in partially, choosing some fields in any preferred order (see for example Figure

1). Then, users had to interact with the desktop migration client interface to require migration towards the PDA and continue the registration on the new device. They also were asked to check previously inserted information and perform some more tasks on the PDA among those supported by the application indicated above (e.g.: completing the form or going to another section of the web site). Lastly, users were asked to fill in an evaluation questionnaire, whose questions were identified in such a way to collect user comments on migration features .

Figure 1. The Test Application

The experiment involved 20 users whose average age was 32 years (min 22 - max 58). Forty percent of them were females, 85% were graduated or had an advanced degree. Users were experienced accessing Web applications through the desktop platform and far less experienced through the PDA: on a scale of 1 (minimum value) to 5 (maximum value), the average self-rating of Web access skill through the desktop platform was 4.35, while PDA usage ability was 1.95. The evaluation questionnaire was divided into two parts, one concerning migration and one related to the result of the automatic page redesign from the desktop to PDA platform. The migration related part asked for a 1 to 5 score to the characteristics shown in Table 3 along with the average values collected.

Migration characteristics	Average Score
Migration Service Usefulness	4.1
Migration Client Interface Clearness	4.1
Easiness in Continuing Task Performance After Migration	4.4
Easiness in Retrieving Information Inserted Before Migration	4.75

Table 3. Migration service evaluation

Users were also asked to say if they would have changed anything in the way migration is performed and 45% of them answered positively providing useful suggestions for improving the migration client interface.

Characteristics	Average Score
Visual Impact of PDA Redesigned Pages	3.9
Functional Navigation Structure on PDA	4.1
Interaction Difficulties due to Redesign Result	1.6
Interaction Difficulties due to PDA device	1.75

Table 4. Redesign transformation evaluation

In Table 4 we show the average scores collected for the redesign part (which performed the adaptation to the target device) of the questionnaire and the characteristics we asked the participants to evaluate, still on a 1 to 5 score scale. A final question concerning the redesign part was about whether participants noticed changes in the redesigned pages compared to the original ones. 40% of the participants answered positively, and in the next paragraphs we discuss what this means.

It is worth noting that the evaluation questionnaire was not only a device for a mere collection of numeric values, because users were also asked to provide justifications for each answer and free comments were also always allowed.

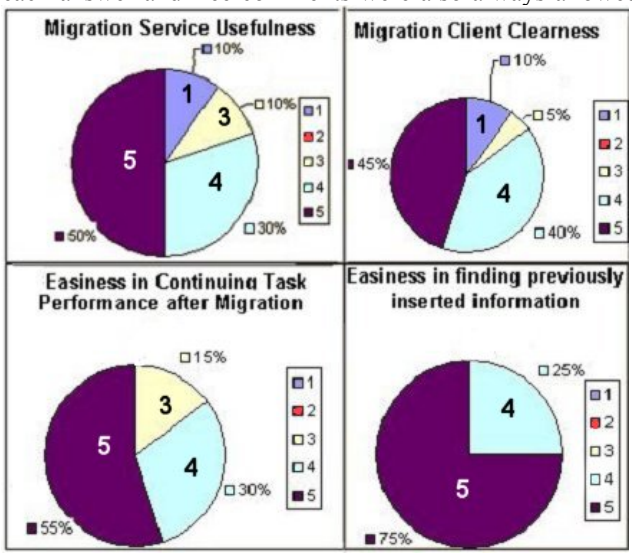

Figure 2. Migration evaluation results

This gave us the ability to better understand the numeric results. Looking at Table 3, we can see that all the average scores concerning migration evaluation were fully positive; moreover, as you can see in Figure 2, only 10% or less of the participants rated any of the parameters a score of 1. Some of the people who did not find the migration client useful said that they simply do not use a PDA.

The migration client interface was deemed quite clear since it was considered basic and easy to use, some of the difficulty that was found was because this user interface had never seen before, as some users commented, at a second usage they would have known what to do. In addition, at a second usage, since the testers already have an idea of what would have happened, they are likely to be more focused on the technical evaluation of the features of the application and then less affected by the novelty of the prototype. The values about task performance continuity and access to information inserted before migration can be considered quite satisfactory. Regarding the 45% of users who would have changed something in the way the migration was performed, this most often referred to the list of IP numerical addresses, which was used to indicate the possible target devices (instead of using logical names, with more significance to the user). We understand such criticism since the migration client interface was at prototype stage and it has been improved to this respect.

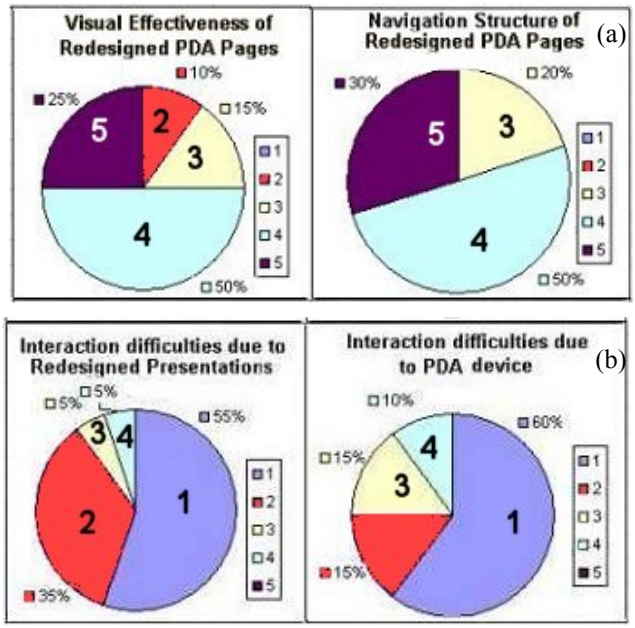

Figure 3(a)- (b). Evaluation of the Redesign Transformation

Concerning the evaluation results obtained for the redesign part (see Table 4 and Figure 3a), one issue about visual effectiveness of the redesigned page was that current browsers for PDAs have limitations with respect to those for desktop systems in terms of implementation constructs that are supported. While the navigation structure of the redesigned pages was appreciated, some difficulties were found while interacting with the PDA (see Table4), but it must be said that most users had never used one before and many difficulties arose simply because of the lack of knowledge of this platform.

Also, users said that they noticed changes between the desktop and the PDA presentations. This question was made in order to understand if changing the specific interaction object implementing a given basic task (i.e. a desktop pull down menu with hundreds of choices transformed to a PDA text input field) when changing platform would have disturbed the users. The comments provided in the questionnaire outlined that the changes that had been noticed were the ones related to the fact that a single desktop page had been split into multiple PDA pages and images where reduced in size. They did not make any observation concerning the different implementations of some tasks, which was a signal of the fact that they did not notice any particular difference that disturbed them. About difficulties of interacting on the PDA platform (see Figure 3b), the numbers must be interpreted in a converse way, since the parameters are related to a negative feature. Indeed, the vast majority of users did not find any major difficulty or only some very small problems in interacting with the redesigned user interface.

CONCLUSION

In this paper we have discussed usability evaluation of migratory interfaces in multi-device environments. We also report on usability tests carried out to better understand the usability of migration and the interfaces resulting from our semantic redesign transformation. The results are encouraging and show that migration is a feature that can be appreciated by end users because it provides them with more flexibility in emerging multi-device environments.

The test provided useful results especially regarding how the user experience is affected by the different adaptation strategies employed for supporting migratory user interfaces. For instance, the test showed that the users are more sensitive to changes affecting the set of available *tasks* when changing devices during migration, rather than to possible changes occurring on the implementation of different interaction objects supporting the same task on different platforms.

In addition, the test gave us the opportunity to verify that the migration features are easily to be familiarised, since the users interacted more easily during the second experiment of the test. The test results have also shown that the proposed redesign transformations are able to automatically generate adapted presentations that allow users to continue their task performance in the target device without being disoriented by the platform change.

Future work will better evaluate the issue of continuity and to what extent the time factor (eg: the time passed between the last interaction on the source device and the first interaction on the target device, time needed for executing the migration) can affect the user experience. Future work will be also dedicated to identifying more suitable evaluation methods and metrics in order to better quantify the benefits of the migration from the user's perspective.

REFERENCES

1. Mori G., Paternò F., Santoro C. Design and Development of Multidevice User Interfaces through Multiple Logical Descriptions. IEEE Transactions on Software Engineering August 2004, Vol. 30, No 8, IEEE Press, pp.507-520.
2. Denis, C., and Karsenty, L., Inter-Usability of Multi-Device Systems−A Conceptul Framework. In "Multiple User Interfaces: Cross-Platform Applications and Context-Aware Interfaces", H. Javahery and A. Seffah (Eds.), 373-385, Wiley and Sons, 2003.
3. Bandelloni R., Berti S., Paternò F., Analysing Trans-Modal Interface Migration, Proceedings INTERACT '05, Roma, September 2005, Lecture Notes Computer Science 3585, pp. 1071-1074, Springer Verlag.

Assessing User Experiences within Interaction: Experience As A Qualitative State and Experience As A Causal Event

Mark Springett
Interaction Design Centre
Middlesex University, Town Hall, Hendon,
London, NW4 4BT, UK
m.springett@mdx.ac.uk

ABSTRACT
This paper discusses issues in assessing the qualitative affective elements of human interaction with software products. Evaluating qualitative responses seems inherently problematic due to the personal and intangible nature of experience. The paper argues that the 'soft' discipline of user-experience evaluation is not significantly softer than more traditional 'measurement' of HCI phenomena. It then considers the key phenomena of qualia, tacit knowledge and behaviour, considering what formative and comparative assessment needs to know and the availability of that information. In particular it considers the causes and effects of qualitative states in interaction and their significance. The implications for selection of evaluation approaches is then discussed using cameo examples of systems that have a key affective element in use.

Author Keywords
User-experience, qualia, tacit, evaluation

INTRODUCTION
This paper considers the nature of the relationship between qualitative experience, tacit causality and evaluation approaches for artefacts in use. The question addressed is about how we can appropriately measure qualitative factors in interactive products. In this context we use a wide-scope interpretation of the term 'measurement'. Orthodox measurement theory implies quantification or proof [16]. A more generous definition of measurement means the ability to assess against criteria or assess progress towards a goal or target. In the first section it is argued that HCI evaluation is typically a soft discipline in which persuasive evidence is used to inform design. This is no less true of user-experience enquiry than of traditional HCI. The nature of qualia, or the first person experience of product use is then considered and its relationship to tacit factors causing qualia, and the consequences of qualitative reactions. The argument is that whilst qualia are first person felt states and therefore in themselves explicitly expressible, crucial causal relationships involving those states are tacit, and only measurable in less direct ways.

VUUM2008, June 18, 2008, Reykjavik, Iceland.

PROOF AND PERSUASION IN EMPIRICAL RESEARCH
Compared to traditional HCI enquiry, user-experience can initially seem softer, less tangible and less available to measurement. The best we can hope for it seems is to get some weak opinion or reaction data, and see if it is persuasive. However, even where something more like 'hard science' is brought to bear on traditional HCI phenomena, it is persuasion that is ultimately the test of its worth. Experimental evidence may provide a proof of a theory pertinent to an HCI problem, but this is evidence to support persuasion. The proponents must then argue for the significance of the experimental finding in reference to the target HCI phenomena. The proof is a phase on this journey but it is never the entire journey, or its essence.

Various approaches have been developed to assess experience. For example, Product reaction cards [1] prompt users to identify 'attributes' present in a product, some of which are user experience attributes, pairing their emotional responses with a subset of preset terms. The reactions imbue the product with descriptors reflecting the user's experience of them. Variants on that (admittedly used predominantly on children) allow a non-verbal expression of mood by selection from a range of cards bearing alternative facial expressions [8]. The arguable position on this is that verbalisation or expression of a qualitative state is a translation of the tacit into the explicit. The potential flaws in such an exercise are similar to issues in requirements analysis, where expertise that is essentially skill-based and well practised tends to be poorly and inaccurately captured though direct methods and explicit expression. Similarly, the concern is that a verbal deconstruction of felt states would be prone to the same sort

of negative filtering effect, potentially rendering the information unreliable. The qualitative dimension does not easily lend itself to articulation through language, particularly if we are looking for trend data. The selection of descriptions for one's reaction seems to be dependent on shared perception of meanings for the terms used. Where multiple data sets are compared with a view to analysing trends, the need to make assumptions about shared meanings could render any conclusions hazardous. Arguably, the greater the complexity of the elicitation (e.g. multiple possible reaction terms) the greater the risk of bad data. In turn, the simpler and less explicit approach (the Smileycards used in [8]) could be criticised as not expressing enough, although they have the advantage of avoiding the linguistic filter.

The notion of proof (over simply persuasion) does not appear to be significantly stronger or flimsier in user-experience evaluation than in traditional HCI evaluation. This sounds an odd claim when we consider something like an 'efficiency' evaluation, for example using the Keystroke Level Model [2]. The targets that are set against the quality goal are concise and quantifiable. The experiments are rigid and formal, samples carefully selected and potential sources of bias carefully removed. What emerges is a contingent proof. However, many things have not been 'proved' in such an analysis. The technique involves the assumption that the operators performing the tasks are consummate experts and will not make errors. The claim that these contrived findings are useful in assessing their real use in real work situations must be put with reason and persuasion. Also, it involves the assumption that the identified evaluation criteria (e.g. speed of expert performance) are indeed key to assessing the efficacy of the system. Whether or not those concerned accept or reject the evaluation findings' validity is dependent on persuasion and the balance of probability. Similarly, evidence from protocol studies can be used to persuade us about causes of user problems and help specify suitable fixes [3]. Measures of performance rate are usually reliant on very strong assumptions about conditions, environment and operator skill. Equally, assessment of experience or the consequences of experience requires investigation, a degree of rigour, and ultimately persuasion. Self reporting and reaction data can fairly reliably denote that a qualitative state change is occurring in the subject. These, using the classification of Hollnagel [6] can be seen as user-experience 'phenotypes' indicating a 'surface level' event. Inferences about their causality may or may not require a deeper level of investigation in search of 'genotypes', or distant causes and consequences of the reported event. The internal states that are elicited from the user seem to be the key data in user-experience enquiry. The next section considers issues in reliably assessing qualia, articulation of qualia, tacitness and causality.

QUALIA, TACITNESS AND EVALUATION

Qualia are key phenomena that user experience research inevitably deals with are, but the role of qualia in formative evaluation enquiry is inextricably linked to causality. Perhaps the goal of a design is to produce a simple non-instrumental emotional state (joy, fun?). Perhaps the goal is to produce a set of behaviours, and this is a goal that is shared with the user (e.g. persuasive technologies such as diet assistants). Perhaps the qualitative experience of interaction for the user is not the goal in using the system linked to instrumental goals (e.g. e-banking). Whatever the goal of the system is, we want to know what causes quales (tokens of qualia) and what effect the quales have, in order to engineer better products.

Qualia are broadly referred to as an essentially individual first-person experience, a first person experience that is essentially unique. So in this sense one can use descriptor terms reporting a felt state, but it is not possible to confirm that your experience is the same as the experience that others are having. Jackson describes as "...certain features of the bodily sensations especially, but also of certain perceptual experiences, which no amount of purely physical information includes"[7].

A key point about a quale is that it is *experienced*. For something to be experienced presumably it makes little sense to say that it is experienced without the subject being aware of it. First-person awareness of it would seem to be part of its essence. This raises the question of whether qualia are necessarily what we are referring to when we consider affective aspects of system use. If we are held to the claim that qualia must be experienced, and experienced knowingly, then qualia are non-tacit and available to self-reporting. But expressions of qualia do not explain key cause-and-effect phenomena that are critical to evaluating user-experience. In a concurrent protocol session we can ask a subject "how do you feel now?". The subject can tell us this but not so easily tell us why. The subject reports awareness of a qualitative state, but the causal reasoning offered has an inference gap. In some cases (and in service of some evaluation objectives) the inference gap will be trivial. In a game of Battleships the cathartic moment of victory is overwhelming likely to be linked to an explicit report of elation or joy. A sense of unease during an e-commerce encounter is rather harder to fathom.

What is ironic in this debate is that it would at first seem that the quales in their unextended form are as inaccessible an element of mental phenomena as it is possible to imagine. However, further analysis of this concept seems to support the argument that quales (in themselves) are more accessible to probing and evaluation than other HCI evaluation phenomena to which they are relevant. What are rather more elusive are the causal relationships, the elements of an encounter with an artefact that critically effect experience. Whereas simply reporting how one feels

gives an informative account of the felt state, how it was caused and what effect it will have on attitudes, disposition and behaviour may not be available to the reporting subject. This suggests that the essence of this is in the tacit dimension. If this is the case there are implications for its characterisation elicitation and assessment.

DEFINITIONS OF TACITNESS

The most simple, traditional notion of tacitness is knowing that cannot be made explicit through verbalisation. In Polyani's words [10] 'we know more than we can tell' citing examples such as face recognition and bicycle riding where sophisticated ability to perform contrasts with a marked inability to explain. We have natural aptitude for these things that seems to be beyond our conscious understanding. Work on knowledge acquisition and requirements engineering in the 1980's tended to characterise tacitness as problematic, a knowledge reef that must be overcome. A more modern way of thinking is that some mental phenomena are in essence tacit and the project of making them explicit is fundamentally flawed [14]. Polyani's assertion is that knowledge is different from knowing, knowing being more a process than a state. This seems a useful point. We have tacit skills as described above. Domain experts have knowledge much of which provides a great deal of skill to elicit for a variety of reasons. Much of this is described as 'semi-tacit' by Maiden and Rugg [9]. Their examples of semi-tacit knowledge refer to communication failures such as withholding useful information in elicitation interviews because they do not sense the value in imparting it. These types of elicitation problems seem to be different in nature from the tacitness issue in user experience evaluation. Whereas Maiden and Rugg describe tacit expertise, tacit knowing seems to be more fundamentally about human behaviour and self-awareness. Humans do not track the causality of their mental states. Tacit skills, tacit behaviours and tacit processing all seem useful descriptor phrases for processes that are key to understanding what is truly going on when we form an affective reaction to an artefact and the subsequent consequences of that event. If tacit skills, process and dispositions are types, tokens of them are events. First-person attempts to describe and account causally for those events are inevitably flawed and unreliable.

VARIANCE IN THE ROLE AND SIGNIFICANCE OF QUALIA

In this section we look at cameo examples of three significantly different types of system. In each of these the qualitative affective dimension plays a different role. In games it is the main goal of system use. In a personal mentoring system it is the 'coalface' the problem space with which the system is involved for instrumental purposes. In e-banking the affective dimension plays a key role in building and maintaining the relationship between customer and organisation.

Case 1 - Games

Games and entertainment applications could be seen as having the goal of a good user experience, of positive felt states. Therefore the experience is the key goal, and any notion of the instrumental is subsumed within the affective. Certain measures relating to engagement, flow, and excitement can be brought to bear in formative assessment of games. These are terms referring to desirable qualitative states that are caused by using a good game or fun system. It may be hard however to pair a measure with these attributes. The cited user experiences are outcomes, and it is their pairing with design concepts and practice that can result in those that provides assessable phenomena. Success against the experiential outcome attributes can be assessed summatively, but is not helpful feedback for design unless it can be paired with more tangible design concepts, and properties of the artefact. A sense of challenge in a computer-based game is an outcome, experienced by the game player. But could challenge also be something that is identifiable in the artefact itself? Are there design techniques in the games domain that support a sense of challenge?

In a sense game design turns longitudinal design for usability on its head. In most systems the ideal scenario is that access to the system metaphor is seamless, internalisation of system rules of operation facilitated optimally and expertise, once established, exploited and supported. This reflects the description of knowledge-based, rule-based, and skill-based processing using software packages described by Rasmussen [11,12]. Optimal usability may (very broadly) be seen as the facilitation of progress towards skill-based processing, and its support once arrived at. Conversely, game design tantalisingly allows its user to reach a level of skill-based competence at a particular stage in a game. The skill in good games that have progressive levels is to incrementally subvert the expertise of the game-player, taking them to the next level of difficulty. In a sense the (simplified) progression from knowledge-based through rule-based to skill-based processing becomes more like a cycle than a simple progression (n.b. it is acknowledged that this 'progression' is not quite as neat as this description suggests in real cases).

The designers' manipulation of the game player's ability to deploy cognitive resources seems persuasively to have a reasonably tangible link to the qualitative outcome of challenge experienced by the game player. Contrast a well-designed game that maintains the sense of challenge in its user with one that does not. For example, a golf-playing game may have an 'expertise threshold' beyond which the sense of challenge is insufficient to maintain interest. Initially the virtual golfer is on a learning curve. The

delicate skill of lining up and realising the control to play a shot takes time to develop. Judgements on how to use controls to line up and estimate power and drift on the shot resemble the real game. The level of challenge seems authentic, like playing an actual game of golf. The aim eventually is to resemble the performance of a world-class golfer. Once too great a level of expertise is reached (e.g. when one has failed if a chip from the fairway fails to go in) the sense of challenge is diminished.

So challenge (the outcome) has a connection to some tangible design properties fortified by theories of user processing and actual design features. So there is something tangible that we can 'measure' and evaluate against the quality attribute of challenge. However, the apparent presence of these is not sufficient to guarantee that an individual game player will sense challenge, or in a broader sense, will be satisfied with the experience. Other factors may cut across what is described. A potential gamer may find the nature of challenge uninteresting, a domain that is not engaging or fascinating, and not feel challenge for this reason. Conversely, a virtual golfer may derive great satisfaction from being improbably accurate with chip shots from off the green. What is key is that the experience attribute links to design criteria. Where a user or users report a lack of challenge in an evaluation exercise, there is an investigative 'phenotype to genotype' thread, criteria for investigating the space of possible redesigns. However, there seems to be a strong connection between the qualitative sense of challenge and more established usability attributes such as complexity. Controlled progressive introduction of complexity is to some extent measurable, or at least estimable.

Case 2 – The diet companion; A Mobile Mentor for dieting

This cameo product example would be in the spirit of 'Persuasive Technologies' [4]. The concept is that the mobile device is there with the role of helping an individual resist their immediate inclination to eat in service of the longer term goal of for losing weight. The idea is that the device informs its user of calorie targets for the day, provides on-the-spot estimates of the calories of certain items (that the user is contemplating eating). It also keeps a continuous record of the accumulated calorie intake for the day, issuing warnings when the user is contemplating eating an unhealthy meal, or is in danger of exceeding a daily limit. Records are kept over time for reference and reminder. The goal of the user in using this system is to help them to keep to a diet, to lose weight and become healthier. The problem addressed is that short-term urges or careless food choices cut across this longer term goal. The user wants a virtual mentor that helps them with the difficult task of resisting immediate urges and sticking to the diet. Part of the system's role is simply information provision, but its presence, its ability to affect the mental state of its user, persuading them to keep their discipline and providing encouragement in this, is central to its value. The assessment of the mentoring role appears difficult here. The affective element is to make the user self aware at key moments. Also, the sense of having support, 'companionship', sharing their struggle to resist temptation is important. The system variously helps the user recognise, alter and respond to qualitative states. So in a case such as this what properties or qualities give us targets buy which to evaluate this device? Are design criteria available that can be usefully linked to experience outcomes?

This example seems to contrast with games because the outcome is defined, longitudinal and instrumental. The qualitative experience during use is subsumed within the instrumental goal of, say, achieving a target weight. The success of this product in achieving this seems to be in its ability to support a complex interplay of emotions, physical dispositions and critical events. Success of a dietary mentor is subject to the unfolding of events in key situations and personal factors, and factors that are often not in the user's realm of control or awareness.

Success of a product such as the Diet Companion seems to be in awareness rather than understanding of diversity and putting products out there that can cater for it. Whilst evaluations can help engineer by refining components, improving adeptness in individual situations (perhaps through shadowing exercises) its effectiveness is difficult to unpack. The complexity of the problem-space and its elusively personal nature perhaps does not need deep understanding, more a 'black-box' knowledge, and engineering of solutions through feedback from users. As with all engineered solutions knowledge is passed on and feeds formatively into subsequent designs and new products. How should information be presented, should avatars be used, what tone should be used in the alerts? But predicting the overall suitability of a design for a particular user seems elusive. They simply offer choices and opportunities for candidate users to try.

Case 3 – E-Banking

In the case of e-banking the goals and values of participants are separate but linked, and issues of user experience have a subsumed, instrumental role in it. The aim of the organisation is to establish and maintain a trading relationship with the customer. In order to do this they need to establish and maintain a trust relationship with potential clients. Previous work [5,13] demonstrates that tangible as well as intangible factors affect the users perception of whether or not an organisation is worthy of trust. Therefore an 'impenetrable' qualitative dimension at least partly determines the customer's willingness to do business with an organisation. In the tangible zone, an audit of available features such as the Verisign seal, declared guarantees and declared commitments such as

privacy policies contributes to an evaluation of an e-banking site's fitness for purpose. However, assessment of the affective role of factors such as visual design, interactivity, feature layout and clustering is essential to reliably assert this. These are seen as intangible factors [5] and a range of personal and cultural factors may influence them. Qualitatively the user has perhaps a sense of confidence, or a sense of unease, suspicion or resentment that they can report. The reasons for that suspicion could in some cases be satisfactorily explained by tangible factors, and directly articulated.

Whether a prototype design will positively or negatively reinforce trust perceptions and diagnosis of designs that are detected as causing negative trust propagation are key evaluation agenda. But the phenomena of interest where intangible affective factors such as the use of colour, layout, style of written content are not so reliably characterised through direct self reporting. As a consequence, enquiry in this area favours techniques such as eye-tracking [13,15] and the deployment of indirect probing techniques such as card-sorting [5]. Again an accumulated body of craft knowledge is emerging from evaluation and engineering of this type of system.

CONCLUSION

The cameo examples describe the key but essentially causal role of qualia in assessment of three different types of system. In the three cases qualia was variously central to the user's goal, intimately bound to the user's instrumental goal and an underlying factor relevant to the users goal. In all cases it seems clear that causality around the qualitative episode (pre and post) is of key interest. Despite the conscious, self-aware nature of qualia, the critical knowledge about causality relating to qualia is in the tacit dimension to a greater or lesser degree. This suggests that direct approaches to assessment are limited in that they can only denote an agenda for diagnosis and understanding of user experience factors, or at least factors in the experience of use.

Evaluation of qualitative user experience seems dependent on identifying the role of 'qualia events' in the use of an artefact, and their relationship to user goals and properties of the artefact. The measurability of this is dependent on the tangibility of the design factors that contribute to them. In games it seems possible to conceive of a link between complexity measures and design that maintains a feeling of challenge. Contrastingly, achieving a sense of trust in e-commerce is only partly based on tangible factors. How trustworthy someone finds a site in a user-trial can be measured, perhaps on a Likert Scale, but assessment (in support of formative evaluation) of how something like a sense of trust can be engineered is not available to such straightforward measurement.

REFERENCES

1. Benedek,J. and Miner, T; Measuring Desirability: New methods for evaluating desirability in a usability lab setting, In: Proceedings of Usability Professionals Association, Orlando, (2002), July 8-12.
2. Card, S, Moran, T., and Newell, A, The keystroke-level model for user performance with interactive systems, *Communications of the ACM*, 23 (1980), 396-210
3. Ericsson, K. A., & Simon, H. A.. *Protocol analysis: Verbal reports as data*. MIT Press, Cambridge, MA. (1993)
4. Fogg B.J., *Persuasive Technology: Using Computers to Change What We Think and Do*, Morgan Kaufmann Publishers, (2003); ISBN 1-55860-643-2
5. French, T. K. Liu, K. and Springett, M. , 'A Card-Sorting probe of E-Banking Trust Perceptions', *Proceedings HCI 2007*, BCS, (2007) ISBN 1-902505-94-8
6. Hollnagel, E. (1998) *Cognitive Reliability and Error Analysis Method*. Oxford: Elsevier Science Ltd
7. Jackson, F., "Epiphenomenal Qualia", *Philosophical Quarterly*, (1982) vol. 32, pp. 127–36.
8. MacFarlane, S., Sim, G., Horton, M.(2005), Assessing Usability and Fun in Educational Software. *IDC2005*, pp.103-109, Boulder, CO, USA
9. Maiden, N.A.M., and Rugg, G. ACRE: selecting methods for requirements acquisition, Software Engineering Journal (1996) 183–192.
10. Polanyi, M. "The Tacit Dimension". First published Doubleday & Co, 1966. Reprinted Peter Smith, Gloucester, Mass, (1983).
11. Rasmussen, J. Skills, rules, knowledge; signals, signs, and symbols, and other distinctions in human performance models. *IEEE Transactions on Systems, Man and Cybernetics*, (1983). 13, 257-266.
12. Rasmussen, J. Mental models and the control of action in complex environments. In D. Ackermann, D. & M.J. Tauber (Eds.). *Mental Models and Human-Computer Interaction 1* (pp.41-46). North-Holland: Elsevier Science Publishers. (1990). ISBN 0-444-88453-X
13. Riegelsberger, J., Sasse, M.A., and McCarthy, J. Rich media, poor judgement? A Study of media effects on users' trust in expertise. In T. McEwan, J. Gulliksen &. D. Benyon [Eds.]: *People and Computers XIX - Proceedings of HCI 2005*, Springer, (2005), 267-284.
14. Senker,J., 'The Contribution of tacit knowledge to innovation'. In *AI & Society* Volume 7 , Issue 3 (1993) Pages: 208 - 224 ISSN:0951-5666
15. Sillence, E, Briggs, P. Harris, P. and Fishwick, L., A framework for understanding trust factors in web-based health advice, *International Journal of Human-Computer Studies* 64, 8, 2006, 697–713. ISSN:1071-5819.
16. Stevens, S.S. *On the theory of scales and measurement* 1946. Science. 103, 677-680

Only Figures Matter? – If Measuring Usability and User Experience in Practice is Insanity or a Necessity

Jan Gulliksen
Uppsala university, Dept. of IT/HCI
PO Box 337,
75105 Uppsala, Sweden
jan.gulliksen@it.uu.se
+46 - 70 425 00 86

Åsa Cajander
Uppsala university, Dept. of IT/HCI
PO Box 337,
75105 Uppsala, Sweden
asa.cajander@it.uu.se
+46 -70 425 07 86

Elina Eriksson
Uppsala university, Dept. of IT/HCI
PO Box 337,
75105 Uppsala, Sweden
elina.eriksson@it.uu.se
+ 46 70 561 88 58

ABSTRACT
Measurement of usability and user experience is a method much sought after in order to introduce and motivate user-centred design activities. Yet, when measurements are available they have little or no influence on forthcoming decisions or any impact on future development. What is the reason for this?

In this paper we will discuss and question measurement as such, based on previous research and two case studies. In both organisations thorough methods have been deployed to be able to measure usability and user experiences. Data was gathered through an interview study in one of the organisations, and on conversations with two usability experts in the other organisation. One experience from these cases have been that measurements as such are desired and give rise to great expectations, before they have been conducted, but with limited or no impact once the measurements have been made.

Author Keywords
Usability, user experience, UX, case study, metrics, CIF, user-centred design

ACM Classification Keywords
H5.5. Information interfaces and presentation

INTRODUCTION

Measuring usability and more recently user experience (UX)in a meaningful and valid way has received increasing interest recently (Law, Roto et al. 2008). This is fair enough, but when discussing these issues there is a need to clarify what we intend to measure and how. It is pointless discussing and comparing different methods for measurement as long as there is no common agreement upon what we are measuring. Particularly within COST 294[9] there has been a strong focus towards arriving at a unified view on how to define UX (Law, Hvannberg & Hassenzahl, 2006), which mainly shows that there is very little common agreement upon what UX is and even less on how it should be measured. At a Special Interest Group (SIG) presented at the conference Human Factors in Computing Systems (CHI 2008) the goal was to target a shared definition of UX. However, this SIG mainly lead to a decision to further distribute surveys that had been made on people's interpretation and appreciation of the concept.

Despite this the group developing process standards for human-centred design considers that UX now has matured enough to find its way even into international standards. In the revision of ISO 13407 (Human-centred design processes for interactive systems) that in the future will be known as ISO 9241-210, the concept of UX have been introduced. At this moment ISO 9241-210 has only reached a committee draft (CD) level, meaning that the contents is not yet stable and has not undergone an international vote. However, extensive discussions have showed the need for including a definition of UX into the standard as a pleasant UX may be an important additional goal of a human-centred design process. But, when introducing such a concept into an international standard it is a requirement that we arrive at a commonly agreed upon definition. The current CD defines UX in the following way:

"all aspects of the user's experience when interacting with the product, service, environment or facility"

Following the SIG at CHI 2008 the definition might change and not include "when interacting" as the discussion elaborated on the opportunities of having a UX of something before and after the actual interaction. In summary, before we have a commonly agreed upon

VUUM2008, June 18, 2008, Reykjavik, Iceland.

[9] www.cost294.org

definition it will be very difficult to reach any consensus on to how to measure it.

Another relevant discussion is what we intend to achieve by measuring altogether. The focus of this paper is to discuss the effect of introducing measurements of usability, which is exemplified by two different cases in which two organizations have tried to arrive at usability and UX measurements.

Many organizations in Sweden use measurement when assessing and understanding different aspects of their organization. These measurements are essential for management and include for example Customer Satisfaction Index, Productivity Statistics, Employee Satisfaction Index and Environmental Index. In this context a measurement of usability might be used to safeguard the usability of the systems used in the organization. One might argue that a measurement could put usability on the agenda, and that it fits into an organizational culture of objective truths through numbers and measurements. Moreover, it might be a way to overview the problem area like a helicopter scanning large areas of forests looking for fires. The index might locate a fire, but neither understands the reasons for the fire nor can it help the organization when trying to extinguish it.

THEORETICAL BACKGROUND

Measuring usability has been discussed for a long time in the usability engineering community. For example Nielsen & Levy (1994) argue that usability as such cannot be measured but that aspects of usability can. Particularly they distinguish two different types of usability metrics that is the subjective preference-related usability problems in comparison to the objective much more easily quantifiable performance measurements. Nielsen and Levy showed a positive correlation between preference and performance in this sense. On the other hand Frøkjær, Hertzum and Hornbæk (2000) contradict this in their study of to what extent the various components of usability (effectiveness, efficiency and satisfaction) are correlated. Based on their study they claim that there was no correlation between effectiveness and efficiency, but that it was difficult to judge the potential correlation with the more subjective satisfaction criteria.

The Common Industry Format (CIF) is an international standard (ISO 25062, 2006) for documenting and reporting the results of a usability evaluation. It has been developed within ANSI (American National Standards Institute) and subsequently been proposed as and developed into an international standard within ISO (International Organization for Standardization). In addition to providing templates for documenting usability problems it also provides a procedure for quantitatively grading the severity of the findings based on the goals that the user intends to achieve.

Sauro and Kindlund go even further and suggest a method for standardizing all the aspects of usability (according to ISO) into one single score, that they refer to as SUM (summarized usability metric) (Sauro & Kindlund, 2005). These papers, as well as most papers dealing with measurement, are neither critical about why measurement is necessary nor do they cite the major publications that criticize quantitative measurement. Rather the tendency is that if you do make claims without justifying it with quantitative measurements, it is not true science:

"Without measurable usability specifications, there is no way to determine the usability needs of a product, or to measure whether or not the finished product fulfils those needs. If we cannot measure usability, we cannot have usability engineering." (Good, Spine Whiteside & George, 1986)

Such a positivistic and objective view of science has been there for a long time and unfortunately there is a tendency to judge this sort of science to be more mature than subjective, qualitative science. This development is not new, it can be found even in science from more than 100 years ago:

"When you can measure what you are speaking about and express it in numbers, you know something about it. But when you cannot measure it, when you cannot express it in numbers, your knowledge is of an unsatisfactory kind: It may be the beginning of knowledge, but you have scarcely, in your thoughts, advanced to the stage of Science."(Lord Kelvin, 1891)

Another famous quote when it comes to people's tendency to rely on measures and metrics is what is referred to as the McNamara Fallacy (for example published in Broadfoot, 2000). It comes from the American secretary of state, Robert McNamara who discussed the outcomes of the Vietnam war based on body counts:

"The first step is to measure whatever can easily be measured. This is OK as far as it goes.

The second step is to disregard that which can't be easily measured or to give it an arbitrary quantitative value. This is artificial and misleading.

The third step is to presume that what can't be measured easily really isn't important. This is blindness.

The fourth step is to say that what can't be easily measured really doesn't exist. This is suicide."

PURPOSE AND JUSTIFICATION

It is still difficult to incorporate the ideas of user-centred systems design, and usability in practice. It is a complex problem, and our research group has addressed it from many different angels to make usability a part of systems development. In our action research projects we aim at influencing systems development in practice through close

collaboration with different organizations. Several of the organizations we have cooperated with have expressed a need for measurements of usability. Hence, in one of the case studies in this paper (CSN) an IT user index was developed and introduced, measuring how the user perceives the usability of their IT systems and their work environment.

The other case study (Riley[10]) is from a large global product development company that we have been remotely collaborating with for some years. This company have made a usability assessment using the CIF framework.

This paper briefly describes these two case studies addressing the issue of meaningful measurement of usability and user experience. In the paper the cases are described and discussed, and finally some conclusions are made concerning the use and efficacy of usability measurements in practice.

ORGANISATIONAL SETTING OF OUR CASE STUDIES

CSN

The Swedish National Board of Student Aid (CSN) handles financial aids for students. With offices in 13 cities in Sweden, CSN employs more than 1100 employees of whom 350 are situated in Sundsvall at headquarters. Officially, all IT projects are run according to methods provided by the development department, and the IT architecture department. These include methods for acquisition, project management and systems development. In short, these methods provide a common framework, with descriptions of milestones, decision points as well as templates and role descriptions.

The organization of CSN, as many other organizations, has been evolving the last ten years towards an organization that is focused on developing work process and the IT-systems supporting these work processes and the organization is structured accordingly. The system development begins at the development department, where pre-studies and the acquisition process are conducted. The development department initiates development projects and system developers from the IT-department staff the projects. Our research group has had a close collaboration with the organization in an action research project that has run for three years.

Riley Systems, Inc.

Riley Systems Inc is the global leader of product development of a technical device used in a wide range of industrial, commercial and governmental activities. The company also develops a series of PC software applications supposed to support the use of the products. With customers in more than 60 countries they are seen as pioneers of their field, and have developed their products for more than 30 years. The average user of their product is a technically skilled person who uses the device in work related situations. In Sweden Riley has worked with usability and user-centred design for a few years, and they have employed a few usability experts, that frequently need to be supported by usability consultants working with design and development of their products.

METHOD

As a part of our large action research project at CSN we conducted 36 semi-structured interviews based on an interview guide. The interview guide included many different topics, one of which was related to the usability index. The following are examples of issues dealing with the usability index; general opinion of work process when developing the usability index, expectations compared to results, future use, benefits for the development of usability in the organization, utility of the results in the organization, and organizational belongings. However, questions were adapted in accordance with the organizational role of each informant, and their background. The interviews were mostly conducted on site and each lasted for about one hour. Most of the interviews were conducted by two researchers interviewing one person. We took detailed notes on paper and the interviews were audio recorded.

In the case study at Riley we have interviewed one usability designer working at the company, and the external usability expert who was the person working with the key performance index in all phases of the project. These interviews were informal and conducted by one of the authors. Detailed notes were taken during the interviews. The interviews were conducted on site and over the phone and lasted for about an hour, each.

After the interviews, the data was analyzed using mind maps where main themes and ideas were highlighted. Finally, the results were discussed in relation to other findings from the participatory observations, and hence put in a wider organizational context.

A case study is always contextual, and consequently it is not possible to generalize findings from case studies to other contexts. However, even though these studies are case studies, the organization and the findings are not unique, nor unusual, and therefore we hope that the results will contribute to a deepened understanding of the use of usability measurements in general.

RESULTS

CSN

Our research group developed a web based usability index method at CSN that resulted in measurements of usability and UX on three different occasions (Kavathatzopoulos, 2006, 2008a, 2008b). During a trial period the questionnaire

[10] This name is an alias due to reasons of anonymity

gradually improved in which questions were clarified and some even deleted.

The main purpose of the introduction of a usability index in the organisation was to measure usability and find usability problems in their computer systems. However, it was not clearly stated from the beginning which department that should be responsible for it. Finally it was decided that the usability index should be incorporated into the existing Employee Satisfaction Index (ESI) that is distributed by the Human Resource department twice a year to all employees at CSN. This means that the usability index needed to be shortened into only a few questions.

The interviews that were conducted at the end of the collaboration with the organization show that upper management highly support the usability index, both in its longer form and the shortened version that would be incorporated into the ESI. However, upper management states that they have no detailed knowledge about the questionnaire, but that they feel confident in the tool. Overall the interviewees who did have knowledge about the usability index where positive, although the expectations varied. Some believed that the usability index could measure a general level of usability in organizations; others believed that it could be an important tool in the systems development process. However, at the moment, with the incorporation of the short version in ESI, there is no clear indication on how the results could be used in the systems development process. The Human Resource department will initially interpret the results from the questionnaire, but those results that could be used in system development should be shared with the Business Development department. However, in the interviews, the manager for the Business Development department was not aware that the results from the questionnaire should be taken care of by her department.

What is further interesting is that the part of the questionnaire that was mostly appreciated by the interviewees concerning usage in the system development process was the open-ended questions at the end of the questionnaire. These open-ended questions generated an enormous amount of improvement suggestions that the organizations were rather unprepared for at first. However, it is not clear whether these open-ended questions will be kept in the shorter version of the questionnaire, although these where the most appreciated by both developers and case handlers.

Furthermore, the ESI is distributed twice a year, when the workload is at the lowest for the case handlers. This might influence the results of the questionnaire, since the case handlers are not working with all parts of the internal systems in slow periods. As one of the interviewees commented:

"When you fill in the questionnaire, you do it from how things are at the moment. The workload for a case handler is distributed unevenly over the year, with peaks and slow periods. And if you fill in the questionnaire in May, [which a slow period], then they will get a much better result than if you fill in the questionnaire in January [which is a peak]. And the questionnaire has always been distributed in May or November [both slow periods] in order for the results to be as good as possible. At least I have felt it that way."

Riley Systems, Inc.

The purpose of the case at Riley was to establish a key performance index for the field of usability in order to be able to assess the effects of usability activities. The organization had previously experiences with using key performance indexes from electronics, mechanics, project management and so forth and therefore wanted to treat the usability initiatives in the same way. The purpose was not to gather knowledge of the level of usability of the developed product, nor was it to gather information useful for improving the usability of the product. However, this should not be interpreted as if they are not interested in doing so, but rather seen as a positive bi-effect of the assessment of the performance of the organisational unit.

A usability designer was hired as a consultant to perform the assessment of the key performance index for usability, particularly targeting a product that was just about to be launched onto the market. He produced a form based on the CIF and set levels of goal fulfilment and identified tasks the users were supposed to perform. Moreover, he was in charge of user tests and authored an evaluation report.

The first report should be considered as a baseline, which means that they really cannot say much about any effects in the organisation until they have made a follow up study later. However, according to the interviewee there are difficulties concerning dependence associated with the measurement, for example the person who is measuring, who the users are, status of the product on the market, etc. This occurs despite the fact that the goal of the CIF is to provide a methodology that is independent of whoever is in charge.

The usability designer, who is a senior expert with many years in the field, experienced great difficulties in finding appropriate tasks that expressed something about the use of the system, and also to define appropriate levels for each tasks. He confronted renowned experts on CIF about these issues and was told that he should be pragmatic and not be too rigorous when defining the tasks –*"it doesn't need to be that thorough"*.

The usability field at Riley has experienced some problems lately, and it may be because of the difficulties they experience in showing real evidence of the benefits of their work. The organisation lacks the ability to see the benefits beyond what can be quantitatively measured.

DISCUSSION

One idea is that if usability can be measured on a regular basis, the IT department must take usability into account in their development process. The underlying assumption here seems to be that by expressing usability in numbers, it will make things visible and these numbers represent some kind of objective truth about the IT systems. Frequent complaints, suggestions for improvements, reported problems when using the IT systems are seen as subjective and do not seem to represent the same kind of truth.

Usability comprises many different aspects, some of which cannot be easily measured and turned into numbers. Specifying usability and UX as a set of well-defined parameters complies with the need for formal representations in the development projects, but obscures the complex aspects. The difficulties with addressing aspects, such as numbers, in no way make them less important. The question is how to deal with them in an organization that focuses to that extent on measurement and metrics.

What conclusions can be made from the cases?

In both cases there was an outspoken request for some form of metrics to be able to assess the state of the computerised work situation or the state of the commercial product. But, once the usability index was in place the organizations did not pay much attention to it. The reasons for this may be many. First of all they experienced difficulties knowing how to interpret the results and what conclusions to draw.Second, it was problematic to turn the acquired results into action items that could be dealt with in the upcoming organisational development.Third, it turned out to be challenging from an organisational point of view to judge whose responsibility it is to deal with whatever comes out of the study.

What do we measure?

It is difficult to measure since it is very difficult to come up with questions that really target the crucial aspects that you do want to highlight. It turns out that the things people tend to measure are those aspects that are easily measurable. But, do we measure a function, a unit of work or an organisation? Most often measurement is restricted to an application or even to a specific function. That means that the measurements do not capture and holistic view whatsoever. Most workers in an organisation deal with a large number of applications simultaneously, and the usability problems that people experience often occur as a consequence of bad synchronization between the applications rather than within them. Only measuring aspects within one application do not lead to any good tool for managers and/or organizations to base their decision making on. And vice versa, a measurement of an organization is also problematic since it is difficult to know which system or combination of systems that are troublesome.

When do we measure?

If the goal from within the organization is to show good results it will clearly influence the timing of the measurement, hence the data becomes even more unrealistic. In one of the cases we saw a conscious effort to making measurement when it was likely to receive the best possible results. This is clearly misleading. The workload will obviously influence the users' perception of the system and the work situation differently if working under stress.

Why do we want to measure?

Measures may often be considered as eternal truths. Many people do not believe something until it is confirmed by numerical figures. Therefore numerical measures are attributed heavy weight in strategic argumentation.

Decision makers often want measures to base their decisions upon, and figures can be the door opener to receiving budgets for certain activities. This seems to be the case regardless of weather or not the measures are realistic hypothetical, or even meaningful. In some situations measurement can even be what constitutes the basis for the existence of a particular line of work. And even if this is a great risk, it may also be a good opportunity. But it is important with an awareness of the limitations that the measurement induces.

What are the consequences of the measures?

Once there was a numerical assessment of some sort the focus of discussion tended to be on the measurement as such, ignoring or forgetting the complexity behind the work situation in which the measurement was acquired. It is much easier to discuss and trust a numerical value as such, and even more so a change in value from consecutive measures, compared to investigating the reason for a particular measure and what to do about this situation.

CONCLUSION

One might read this paper as arguing against any form of measurement, but that is not how we intended it to come out. Measurements may be very useful and valid in great many situations, but once you do measure things you need to be aware of the consequences these measures have. One thing to do then is to improve the measurement methods to provide a better coverage of the aspects we want to assess. One important part of increasing the credibility of measurement is to clarify what the method measures and how these measures should be interpreted. For example, if one intends to measure user experience (UX) it needs to have a commonly agreed upon definition. Otherwise the results obtained may be considered nonsensical.

Most measurement provides far from the whole truth about a situation and the results must be dealt with caution. Usability provides so many non-measureable aspects that enforcing measurement upon usability in an organisation

may risk comparing with other disciplines that do not have the same reliance upon qualitative effects.

If the goal is to manage to achieve the highest possible level of usability in a system in an organisation, we cannot really see any evidence that measurement techniques and metrics actually have such effects in the organisations we have encountered. Therefore organisations may consider weather it is worthwhile investing in measurement techniques over dedicated usability improvement efforts, such as applying a user-centred iterative process of contextual analysis and system redesign based on formative evaluations.

ACKNOWLEDGMENTS

We thank the usability designers and the interview subjects that have taken part in our interviews for their contribution. A particular thank you to BengtGöransson who read and provided comments on the manuscript based on his experiences of measurement in practice.

REFERENCES

1. Broadfoot, P. (2000) Assessment and intuition. In: T. Atkinson and G. Claxton, Editors, The intuitive practitioner: on the value of not always knowing what one is doing, Open University Press, Buckingham, pp. 199–219.5
2. Frøkjær, E.,Hertzum, M., &Hornbæk, K. (2000). Measuring usability: are effectiveness, efficiency, and satisfaction really correlated? Proceedings of the SIGCHI conference on Human factors in computing systems, p.345-352, April 01-06, 2000, The Hague, The Netherlands.
3. Good, M., Spine, T.M., Whiteside, J., George, P. (1986) User-derivedimpact analysis as a tool for usability engineering. In: Proceedings of the SIGCHI Conference on Human factors in computing systems
4. ISO/IEC 25062 (2006). Software Engineering -Software product Quality Requirements and Evaluation (SQuaRE)-Common Industry Format (CIF) for Usability Test Reports.
5. Kavathatzopoulos, I. (2006): AvI-enkäten. Ettverktygförattmätaanvändbarhet, stress ochnyttaav IT-stöd (in Swedish – The AVI survey – a tool for measuring usability stress and usefulness of an IT support system). Report No. 2006-050). Uppsala universitet: InstitutionenförInformationsteknologi.
6. Kavathatzopoulos, I. (2007): Usability Index. In A. Toomingas, A. Lantz & T. Berns (Eds.) Work with computing systems (p. 160). Stockholm: Royal Institute of Technology and National Institute of Working Life.
7. Kavathatzopoulos, I. (2008a): Ettförbättratverktygförmätningavanvändbarhet, stress ochnytta: Andraförsöketinom CSN.(in Swedish – An improved tool for measuring usability stress and usefulness: The 2nd attempt within CSN) (Report No. 2008-003). Uppsala universitet: InstitutionenförInformationsteknologi.
8. Kavathatzopoulos, I. (2008b): AvI-index: Ettverktygföranvändbar IT, nyttaocharbetsmiljö. (in Swedish – The AVI index – a tool for usable IT, usefulness and work environment) (Manuscript). Uppsala universitet: InstitutionenförInformationsteknologi.
9. Law, E., Roto, V., et al. (2008). Towards a shared definition of user experience. CHI '08 extended abstracts on Human factors in computing systems. Florence, Italy, ACM
10. Law, E., Hvannberg, E. &Hassenzahl, M. (Eds.) (2006) UX 2006 : 2nd International Workshop on User eXperience. 14 October 2006, Olso, Norway. Held in conjunction with NordiCHI 2006.
11. Nielsen, J. and Levy, J. (1994) Measuring usability: Preference vs. performance, Communications of the ACM 37, 4, 66-75.
12. Sauro, J. & Kindlund E. (2005) "A Method to Standardize Usability Metrics into a Single Score." in Proceedings of the Conference in Human Factors in Computing Systems (CHI 2005) Portland, OR.

Measuring the User Experience of a Task Oriented Software

Jonheidur Isleifsdottir
deCODE genetics
Sturlugötu 8, 101 Reykjavik
jonheidur.isleifsdottir@decode.is

Marta Larusdottir
Reykjavik University
Kringlan 1, 103 Reykjavik
marta@ru.is

ABSTRACT
In this paper a study on a web based tool is described that is used to keep track of attendance and work schedules by employees and managers in large companies. Ten users particpated in a think aloud test measuring the usability of a new version of the software and the user experience was measured before and after each user test.

The user experience results show that the group of questions that measure the personal growth of the user got the lowest scores for this product, but pragmatic attributes, hedonic identification and attraction got much higher scores. It is not surprising that pragmatic issues get high scores for a task-oriented software like this one, but it is an interesting result that the users value highly the attraction and hedonic identification.

Author Keywords
User experience, usability, think-aloud, user testing.

ACM Classification Keywords

INTRODUCTION
In the past decade user experience has become a popular field of study within the field of Human Computer Interaction. It challenges the past notion that task related features are the only ones that contribute to usability. User experience focuses on the user entire experience when using a software product, not just the ISO 9241-11 factors, effectiveness, efficiency and satisfaction. User experience introduces new concepts to the quality of software like fun, beauty and pleasure.

The problem with the concept of user experience is how vague it is and that it can be interpreted in many ways. More empirical results are needed to define the concept clearly [6] and these will only be obtained by making models and using them to measure user experience for different products [3]. There are many unanswered questions that need to be addressed. What is beauty? Are beauty and usability related? What contributes to the goodness and beauty of products? What effects how the users summarize experiences during usability evaluations? [5] To address some of these problems Hassenzahl [2] has proposed a model of user experience that divides the attributes of a product into pragmatic and hedonic attributes. Based on this model he has made the AttrakDiff 2 [4] questionnaire that can be used to measure the user's experience of these different attributes in the product. That questionnaire has been translated to Icelandic by Marta Larusdottir.

The goal of this study is to measure the user experience of users participating in a typical think aloud test of a task-oriented software by using the Icelandic version of AttrakDiff 2 questionnaire. Our goal was to compare the measurements of the expectations to the tool and the user experience measured just after taking part in a think aloud test. Secondly we wanted to analyse the Icelandic translation of the questionnaire.

VUUM2008, June 18, 2008, Reykjavik, Iceland.

BACKGROUND
In this section it is explained what is meant by User Experience, a clarification of the different concepts that lie behind the AttrakDiff 2 questionnaire is given and some of the studies that have used it to measure what attributes to attractiveness of products are described.

User Experience
User Experience is a relatively new field within the larger scope of Human Computer Interaction. It proposes a more holistic view of the user's experience when using a product then is usually taken in the evaluation of usability [2]. Until now usability evaluations have primarily focused on task-related issues such as efficiency and effectiveness. Stating that a product that is efficient and effective in allowing the user to solve the tasks needed, to fulfill the user's goals, makes the user satisfied [6].

But is it enough to have a satisfied user? The researchers leaning toward UX say the answer is no [6]. The user needs to experience more that satisfaction with the product for it to be marketable. Hassenzahl [2] proposes a model for the different attributes a product can have and make up a product character. He states that a product has both pragmatic and hedonic attributes. The former being the task related attributes we are used to from the classic usability literature and the later emphasizing the users well being

while using the product. Hassenzahl also introduces in the same chapter three different classes of hedonic attributes; stimulation, identification and evocation. Later Hasssenzahl decided to drop the evocation class from the model and that is not included in AttrakDiff 2.

AttrakDiff 2

AttrakDiff 2 [4] is a questionnaire that measures hedonic stimulation and identity and pragmatic qualities of software products [1]. The questionnaire was originally made in German but has been translated to English. AttrakDiff 2 has four, seven anchor scales, in total 28 questions. These will be described in more details in the following.

Pragmatic Manipulation

Pragmatic attributes are the ones that are associated with how easy the user finds it to manipulate the environment, in this case the product or software. It is what makes us able to fulfill our goals and what we have until now talked about as usability. If we think pragmatically the only requirement from a product to squeeze juice from an orange is that it actually squeezes the juice from the orange and that we can find out how to use it on our own. There is no beauty or design needed to make a product pragmatic.

Hedonic Stimulation

The attributes connected to hedonic stimulation are the ones that encourage personal growth of the user. People want to develop their skills and knowledge further and these are the attributes of the product that allow for that to happen. As an example Hassenzahl provides unused features of a software [2]. Those features that the user does not yet use are not a part of the pragmatic experience but are rather perceived as hedonic as they provide stimulation for further development. Stimulation can also be provided by presenting things in a novel way or by a new interaction style.

Hedonic Identification

Attributes connected to hedonic identity are the ones that make us identify with the product in a social context. What message are we communicating to other socially by using this product? These attributes are connected to the fact that all persons communicate their identity through things they use and own. An example of this would be a personal website where you can communicate who you are to the outside world. If a product communicates what we think to be advantageous to others we might prefer that product. Ipods would be a good example of a product that communicates a strong identity. There are several other mp3 players on the market that work the same but the brand name is so strongly connected to the product and its coolness that everyone has to have an ipod.

Attraction

When we talk about something as being attractive to us, we are usually summarizing the whole experience of the product. We judge the product as a whole and use words like good, bad, beautiful and ugly to describe things. In AttrakDiff 2, attraction is used to measure the global appeal of a product and to see how the other attribute affect this global judgment [3].

Scale Examples

The Pragmatic Quality (PQ) scale has seven items each with bipolar anchors that measure the pragmatic qualities of the product. This includes anchors such as Technical-Human, Complicated-Simple, Confusing-Clear. The Hedonic Quality Identification (HQI) and Stimulation (HQS) scales also have seven anchors each. HQI has anchors like Isolating-Integrating, Gaudy-Classy, Cheap-Valuable. HQS has anchors like Typical-Original, Cautious Courageous and Easy-Challenging. AttrakDiff 2 also has a seven item anchor scale for overall appeal or attraction (ATT) with anchors like ugly-beautiful and bad-good. The anchors are presented on opposite sides of a seven point likert scale, ranging from -3 to 3, where zero represents the neutral value between the two anchors of the scale.

Related Work

Hassenzahl used the AttrakDiff 2 questionnaire to study four different mp3 player skins [3]. He used the questionnaire to explore the effect of pragmatic and hedonic qualities on beauty and goodness. The four different skins had been pre tested and judged ugly or beautiful; two skins were judged beautiful and two ugly. In the first study 33 students were asked to look at each skin and fill out an AttrakDiff 2 questionnaire for each of the skins without using them. Two 2x3 ANOVAs were performed on the result data, one for the beautiful skins and another for the ugly ones. The ANOVAs had skin and attribute group (PQ, HQS, HQI) as within subject factors and score as dependent variable. This revealed that there was one skin in each group that was significantly more stimulating than the other and the other was thought to be significantly more pragmatic. Other results were that the identity communication factors, HQI were the factors that had the highest correlation with the beauty rating, i.e. the HQI scores were significantly higher for the more beautiful skins.

A study on the influence of hedonic quality on attractiveness was done by Schrepp, Held and Laugwitz [7]. They sent out e-mails to students and asked them to look at three different interfaces of business management software. Around 90 people responded and the response rate was 34%. AttrakDiff 2 was used to measure the user experience after the user had looked at 11 different screenshots, with an explanatory text before it, of the interface executing a part of a business scenario. Schrepp et al. expected, since they were testing business software that are in their nature meant

to support people in their work, that pragmatic qualities would have greater influence on attractiveness than hedonic qualities. What they found on the other hand was that both HQI and PQ contributed evenly to the attraction and HQS also contributed significantly.

They also found, as they expected, that more attractive interfaces were preferred over the less attractive ones.

MATERIALS AND METHODS

In this chapter the study using the newly translated AttrakDiff 2 questionnaire is described, first the tool which was evaluated, then the usability tests, the participants and the measurements in the study.

The tool - Workhour

Usability tests were conducted on a new version of software called *Workhour* (Vinnustund in Icelandic) [8]. It is designed by the Icelandic software company Skyrr, see figure 1. An old version had been in use for serveral years, but in the new version the user interface was changed exstensively. There are four main user groups of Workhour; ordinary users that work on shifts and those that work regular hours. The other two main user groups are managers that work on shifts and those that don´t.

The main tasks for ordinary users working on shifts is to check there monthly plan for shifts, ask for a day off and check if they have fullfilled all their work obligations for that month. The main tasks for regular users are asking for holidays and check if they have been too many hours off work. The *Workhour* system is very useful to managers, because they can do much of their organizing work in Workhour like check if all timestamps for their employees are correct, insert information about an employee that is sick and get an overview of how many have been sick over a particular period to name a few.

Participants

Ten individuals participated in the study, eight women and two men. Five of the participants were categorized as managers and the other five as ordinary users. The participants are employees of Landspitali – University Hospital and Financial Management Authority in Iceland, and were divided into two groups but all of them use *Workhour* as a part of their workday. One of the participants only filled out a very small portion of the pre-use questionnaire and none of the post-use so we did not use any data from him in the results.

The usability tests

The ten usability tests were conducted by two usability specialists on the new version of *Workhour* running on a test database two months before it was installed.

Each user solved six or seven tasks in think aloud tests which were adjusted to their ordinary tasks. The total number of tasks in the study was 17. The tasks were made by one of the developers of the user interface that has good connections to the users.

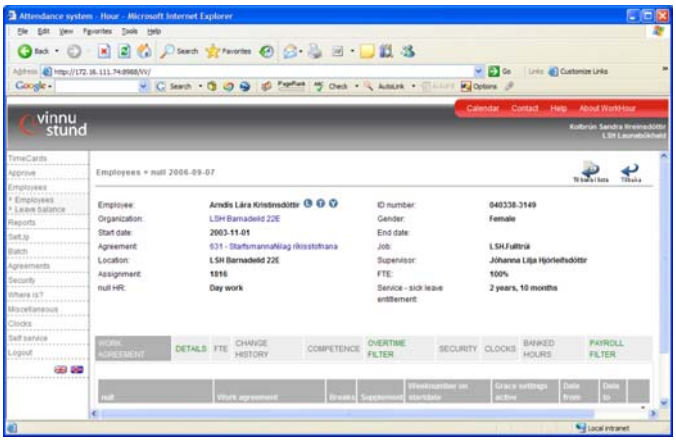

Figure 1. The new version of the software *Workhour*

The user tests were conducted at their ordinary working place, so a lot of contextual information was also gained. Two usability specialists conducted the tests; one was the organiser and one the data recorder. Additionally everything that was said was recorded on tape.

The AttrakDiff 2 questionnaire was administered before and after the think aloud test. First the participants were asked to answer the questionnaire according to their expectations to the new version of Workhour they would be trying in a minute. The users are all familiar with older versions of *Workhour* and were chosen to be in the study as typical users of the system. After the think aloud test was finished the user filled in the questionnaire again and now the participants were asked to base their answers on the experience of using Workhour to solve the given tasks. The reason for measuring both the expectations and the experience is that user experience a subjective factor that changes over time and we were interested to see what affect actually using the tool would have on the measurements of the user experience for the tool.

Unlike studies done by Hassenzahl on skins for mp3 players [3], our study had participants that are actual users of the software being evaluated. The tasks they performed were tasks they know and perform every day. The expectation of the user when answering the AttrakDiff 2 questionnaire before use was purely based on what the user expected of the new version, because most users had not seen the new version before answering the questionnaire.

Measurements

The AttrakDiff 2 questionnaire was translated from English to Icelandic by Marta Larusdottir and this was the first time it was used in the Icelandic format. There was one set of anchors that were left out when the questionnaire was translated. It was the HQI item with anchors Alienating-

Integrating. The translator was not able to find a suitable translation that differed from the translation of another pair of anchors. So the HQI only had six items.

The internal consistency of the HQI, HQS, PQ and ATT scores was measured both before use and after use and unfortunately it was not as high as previously measured by Hassenzahl [3]. Cronbach's α on the pooled values for the different scales before use was: PQ, α = .58; HQI, α = .57; HQS, α = .42; ATT, α = .43. These values for alpha are rather low and that indicates that there is not a high correlation between the answers with in each group. Usually the criteria of internal validity wanted from questionnaires is an α > .70 [3]. After use the scales internal validity was higher in three cases: PQ, α = .86; HQS, α = .55; ATT, α = .70. In the HQI, α = .46 scale the validity was lower than before. Both PQ and ATT were over the, α > .7 mark when measured after use, which is good but the internal validity of HQS is still too low.

RESULTS

In the following chapter the results will be described, first on the user experience and then the analysis of the translation of AttrakDiff 2.0 to Icelandic is described.

The user experience

We calculated the mean score of all user answers for each quality scale (each scale has 7 questions). As mentioned earlier each answer gets a value from -3 to 3, with zero as the neutral value between the anchors of the question.

As can be seen in Figure 2 all the quality scales have means above zero both before and after use of Workhour. The post-use line is also always beneath the pre-use line. A paired T-test that compared the pre and post-use scores for each participant and each category showed that difference in mean scores pre and post-use was significant in HQS(meanDiff = .67,t = 3.56, p = .007) and in ATT(meanDiff = .49,t = 3.43, p = .009) but not in HQI(meanDiff = .35,t = 1.62, p = .115) and PQ(meanDiff = .34,t = 1.78, p = .145). It is also noticeable that HQS score means are much lower than the other means both pre- and post use.

Since the internal validity of the scales was not very high we decided not to do any further analysis of the effects of the different qualities (HQI, HQS and PQ) on ATT as was done in other studies using AttrakDiff 2. One idea was also to test the correlation between different scales. At the AttrakDiff website there is a confidence square diagram made from the data which we did not attempt. There online experiments can be set up in German and in English [4].

Discussion

It is interesting to see that the post-use line (Experience) in Figure 2 is below the pre-use line (Expectations) and consistently so. Even though the difference is not statistically significant (This might be due to the small sample we had). We still believe it is relevant because it is found in all categories. We think that one reason for this is that users are optimistic that a new version is somehow better than the old and therefore have high expectations to its hedonic and pragmatic qualities, that are lowered when the product is used but not considerably. Participants might move their scoring one point closer to the middle.

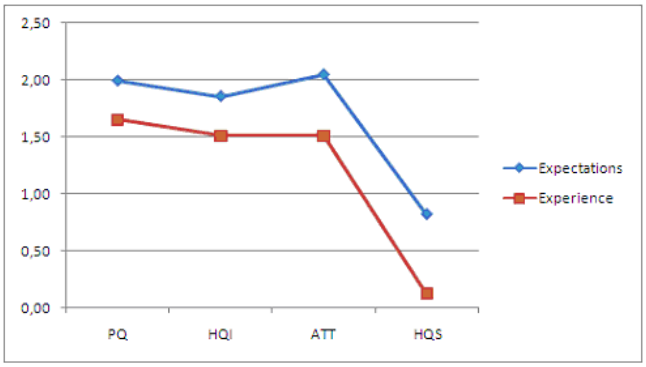

Figure 2. Mean scores for each scale of AttrakDiff 2

Moreover, even if the scores moved closer to zero they were over all pretty good compared with both the mp3 skin study and the business interface study mentioned earlier.

In Figure 2 we can see that HQS gives the lowest score. That category has a mean that is 1-1.5 lower that the other means.

This indicates that the software has less hedonic stimulation qualities than identification and pragmatic ones. This is an interesting result.

We think this is very understandable since we doubt that stimulation was one of the design goals of Workhour. It is software that is used to check on work schedules and to punch in and out of work. We think that the stimulation factor is more important when designing software for creative work rather than support software that most user use only for a short period of time each day and mainly just have to trust that it works.

Translation of AttrakDiff 2

As stated before the study above was also done to test how the Icelandic translation of the questionnaire was working. In the description of the variables and measurements above we showed that the internal validity of the scales was rather poor. Since other studies have shown much better validity scores our first thought was to look closer at how the scores were for each item on the scale. We started with the HQS scale because that gave the lowest internal validity in both the pre- and post-use study and also a much lower mean score than the others.

In Figure 3 we see that the third and sixth items of HQS give a very different mean than the others in this group.

This indicates that these items are not measuring effect that is similar to the others. The third HQS item has the word anchors bold - cautious where markings closer to bold give a higher score. The translation was *ótraust - traust* where *ótraust* gave a higher score. This is a somewhat misleading translation, that indicates that the software is not reliable and secure enough. A better translation would be *djarft-varfærnislegt*. The only problem is that *varfærnislegt* is already used in item four in HQS and we propose that *íhaldsamt* will be used there instead.

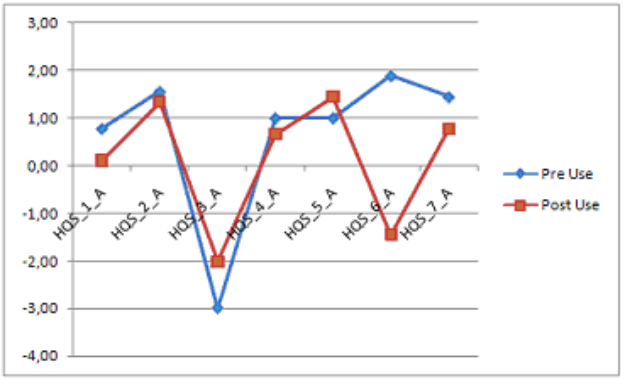

Figure 3. Scores of HQS: Pre- and post-use

The sixth HQS item has the word anchors undemanding - challenging where markings closer to challenging give a higher score. The translation was *auðvelt-krefjandi*. We don't think this translation can be improved and the reasons for the low score are most likely that users don't want software of this type to be challenging.

There was also one item missing from the translation. That was the fifth item in HQI that has the anchors alienating-integrating. The Icelandic translation for integrating was in use in the first HQI item.

Our suggestions are that we change item HQI_1 to *einangrandi-tengjandi* and item HQI_5 which was missing to *fráhverft-sameinandi*.

The other scales did not seem to be suffering from the same problems as the HQS scale. It is therefore our belief that if those changes are made the internal consistency of the scales would improve. If this does not happen it would raise the question whether the HQS scale simply does not apply in the same way as the other scales when measuring the user experience of very practical software. This is certainly a point worth studying further.

CONCLUSION

It was very interesting to see the scores for the different attributes of the AttrakDiff 2 Questionnaire on Workhour. It seems that hedonic stimulation is the least important factor in such software or at least what the participants thought deserved the lowest score.

It was surprising though how high the scores were in HQI considering that identification was probably not a part of the design. The good score in PQ, HQI and ATT is pleasant to see because that indicates that there is overall happiness with the software product.

It is dangerous to draw conclusions about the relationship of hedonic qualities and usability and goodness from the studies that have been done at present. A great deal of software is used every day for extended periods of time and not recreationally. It is questionable, if that kind of software would follow the same patterns as the mp3 player skins in Hassenzahl's study, and what about real users? Even though the people in his study probably use some kind of mp3 players frequently we do not know whether they are to be considered "real" users.

Further empirical studies are needed to be able to draw any conclusions about user experience and how it is affected. Hassenzahl's model is a step in the right direction and hopefully we will see a great increase in studies using that or similar models to evaluate user experience with software and other products. It is also our hope that translating the AttrakDiff 2 questionnaire to Icelandic will inspire more people to use it alone or as an addition to other user testing to gain more knowledge about what factors contribute to how attractive our software is to the user.

ACKNOWLEGDEMENTS

We want to thank the people at Skyrr for their very good co-operation in this project, the users, the developers and the managers.

REFERENCES

1. Hall, M., & Straub, K. (2005, October).
2. Ui design newsletter. Internet Resource http://www.humanfactors.com/downloads/oct05.asp. (Retrieved March 24, 2007)
3. Hassenzahl, M. (2003). The thing and I: Understanding the relationship between user and product. In M. A. Blyth, A. F. Monk, K. Overbeeke, & P. C. Wright (Eds.), Funology: From usability to enjoyment, 1-12 (chap. 3). Kluwer Academic Publishers.
4. Hassenzahl, M. (2004). The interplay of beauty, goodness, and usability in interactive products. Human-Computer Interaction, 19, 319-349.
5. Hassenzahl, M. (2007). AttrakDiff(tm). Internet Resource http://www.attrakdiff.de.
6. Hassenzahl, M., & Sandweg, N. (2004). From mental effort to perceived usability: transforming experiences into summary assessments. In Chi '04: Chi '04 extended abstracts on human factors in computing systems (p. 1283-1286). New York, NY, USA: ACM Press.

7. Hassenzahl, M., & Tractinsky, N. (2006, March-April). User experience - a research agenda. Behavior & Information Technology, 25, 91-97.
8. Schrepp, M., Held, T., & Laugwitz, B. (2006). The influence of hedonic quality on the attractiveness of user interfaces of business management software. Interacting with Computers 18 (5), 1055–1069.
9. Skyrr. (2007). Workhour. Internet Resource http://www.skyrr.is/vorur/vinnustund/.

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

COST is supported by the EU RTD Framework Programme