The fidelity of prototype and testing environment in usability tests

Andreas Uebelbacher

from Austria

Doctoral Thesis presented to the Faculty of Philosophy

University of Fribourg (CH)

January 2014

Approved by the Faculty of Philosophy, University of Fribourg,
at the recommendation of Prof. Dr. Jürgen Sauer (Supervisor, Fribourg, CH),
and Prof. Dr. Kai-Christoph Hamborg (2nd Reviewer, Osnabrück, D)

Fribourg, 29th of September 2014
Prof. Dr. Marc-Henry Soulet, Dean

Summary

This doctoral thesis investigated what setup of a usability test can best support valid test outcomes. Several aspects of contextual fidelity were manipulated in experimental usability studies, to examine their impact on test results. The first study demonstrated that the medium of prototype presentation has effects on test outcomes, which have not been found in previous research. Using a more hypothesis-driven approach, it was shown that participants exhibited more reading activity when using a paper-based as compared to a computer prototype presented on screen. This resulted in better performance, if task success required reading a short paragraph of text. Consequently, the medium of prototype presentation needs to be considered to avoid that respective usability problems go undetected. A second study demonstrated that additional observers may cause stress for test participants, which can be measured at the physiological level. Some performance indicators were affected, but only in interaction with perceived developmental stage of the test system. A third study investigated the effects of a work or leisure context on the outcome of a usability test. No effects were found for the type of usage context, but even short response time delays proved to be relevant for performance and emotions. Relevant factors for the validity of usability test outcomes were identified and theoretical and practical implications are discussed.

Keywords: Usability test; paper prototype; fidelity; observer presence; work and leisure domain; system response time; heart rate variability; validity.

Acknowledgements

"... we are typically not freed up at all by technology but rather made passive—and if we are freed up it is only to have time for more technology. In this downward spiral, we become consumers, increasingly disengaged from things and from each other. Technology tends to seduce us toward a focus upon material goods, quantitative thinking, commodities, and disposability, where any kind of guidance from considering issues of the good and the excellent is left out."
Fallman (2010; p.56)

In connection with this thesis, I want to thank Prof. Jürgen Sauer for his valuable support of this work. Special thanks are due to Marisa Bützberger, who supported this endeavour a lot from a completely different angle. Thanks go to colleagues, external collaborators and students who offered their cooperation carrying out the empirical work, such as Dr. Javier Bargas-Avila, Caroline Biewer, Eric Bourquard, Jean-Pierre Guenter, Dr. Daniel Felix, Klaus Heyden, Felix Hürlimann, Hans Rudolf Kocher, Amadeus Petrig, Manuela Pugliese, Dr. Andreas Sonderegger, Max von Schlippe, and Claudia Vonlanthen. I also appreciated the support and acceptance I received for doing this doctorate from Markus Riesch at the Foundation «Access for all», and thanks to Roberto Bianchetti, the final version of this thesis is accessible according to PDF/UA. And finally, thanks are due to those unknown reviewers who offered helpful comments to submitted journal articles.

This thesis, in more than one way, is dedicated to my children Aimo and Avin.

Contents

Figures

Tables

Abbreviations

bpm	Beats per minute
CUE	Comparative usability evaluation
df	Degree of freedom
DSL	Digital subscriber line
F	F statistic
HCI	Human-computer interaction
HRV	Heart rate variability
HCD	Human-centred design
HF	High frequency band
IEI	Interaction efficiency index
LAN	Local area network
LF	Low frequency band
M	Mean
NASA-TLX	NASA task load index
ns	Not statistically significant
p	p value
PANAS	Positive and Negative Affect Schedule
PSSUQ	Post-Study System Usability Questionnaire
SD	Standard deviation
SRT	System response time
t	t value
TD	Task difficulty
UCD	User-centred design
χ^2	Chi-square

1. Introduction

1.1. The relevance of usable technology

Interactive technology has become ubiquitous in modern societies. While from the 1960s onwards, mainframes introduced business computing and the respective experiences at the work place, in the 1980s personal computers invaded people's homes and leisure time (Grudin, 2005). Today, smartphones outsell personal computers (Perez, 2011), and they provide constant access to impressive processing power and, by being connected to the internet, practically unlimited information, not just in the office or at home but wherever we are. But they are just the tip of the iceberg of interactive technology being present in all areas of our everyday lives, from walk-up-and-use touch screen ticket machines to home entertainment systems. Irrespective of a dystopian or a more optimistic view of these technological achievements, it can be stated that 'we are increasingly experiencing the world with, through, and by information technology' (Fallman, 2009, p.60).

However, even for the intended users technology is often quite complicated to interact with, and the problem of non-usable interfaces becomes a highly relevant one, for various reasons (Maguire, 2001). First, it reduces productivity, as a badly designed system does not allow users to operate it effectively. For computer usage, research reveals disturbingly high percentages of lost time, due to either trying to fix problems or redoing lost work, ranging between studies from 27% to as high as 53% (Ceaparu, Lazar, Bessiere, Robinson, & Shneiderman, 2004; Hertzum, 2010; Lazar, Jones, & Shneiderman, 2006). In the work context, suboptimal usability can thereby severely damage the potential productivity benefits of enterprise software, easily resulting in costs in the millions (Dalton, Manning, Mines, & Dorsey, 2002). Second, non-optimal usability can cause increased training requirements for users and higher support costs (Dray & Karat, 1994). Third, as the internet becomes an increasingly important distribution channel, low usability of the respective interfaces translate into increasingly more relevant numbers of 'lost' sales, if potential customers are not able to find products or fail in the process of buying them (Bias & Mayhew, 2005). Forth, especially as usability is routinely mentioned in easily available online product reviews, a badly designed system may damage the reputation of a supplier (Jokela, 2004). Fifth, usability problems can lead to increased error rates. This is especially disturbing in the case of appliances that are critical for safety and health, where bad usability may even result in accidents or fatalities. A survey among surgeons in Germany identified ergonomic deficiencies of devices as one of the primary causes of errors in the

operating theatre (Matern & Koneczny, 2007). An exemplary usability study of two common defibrillator models in the US identified a failure rate as high as 50% for tasks that are directly relevant for patients' health, even among paramedic personnel familiar with the tested devices (Fairbanks, Caplan, Bishop, Marks, & Shah, 2007).

In product evaluation and design, the term 'usability' has been introduced to describe a system with respect to how complicated or how easy it is to use. It is a core concept in the fields of human-computer interaction (HCI) and cognitive ergonomics, and it characterises a system in terms of its ease of use, and whether it is designed such that it is possible for users to operate it in intended ways and with acceptable effort (Seffah, Donyaee, Kline, & Padda, 2006). One of the most prominent methods used in system development to make sure the final product meets the needs of the intended user group is the usability test. This method involves real users working on realistic tasks with the system that is to be tested, and observers recording the user behaviour and identifying the relevant usability problems encountered by participants during a test session (Dumas & Redish, 1999). Usability tests provide product designers with highly valuable behavioural user data, which is difficult to collect with other methods, and which is key to evaluate a product's usability (Wichansky, 2000). However, in spite of the method being widely used in industry and its benefits being rarely doubted, it is less clear which factors affect the outcomes of usability testing. Some studies show substantial inconsistencies across usability tests with regard to the usability problems identified (J. R. Lewis, 2006). E.g. Molich et al. (2004) report a study in which nine teams independently conducted usability tests of the same system, and of a total of 310 identified usability problems as high as 75% were reported by one team only, showing a remarkably low overlap of test outcomes. Other studies found a slightly higher consistency between testing teams (Kessner, Wood, Dillon, & West, 2001) but still conclude that the method of usability needs to be considerably improved. These results raise concerns about the reliability of usability testing. In addition, there are factors inherent in the way usability tests are traditionally conducted which can directly affect the validity of test outcomes. As usability tests in industry are always working under constraints with respect to time and budget, the fidelity of the test situation to the intended usage context of a product is usually limited. This can be the case either with the test system itself (e.g. a paper-based prototype being used instead of a computer system; e.g. Meszaros & Aston, 2006), or with regard to the test setting (e.g. a mobile device is tested in the laboratory instead of the field; e.g. Alsos & Dahl, 2008). Other limitations are also possible, e.g. participants being recruited based on availability instead of representativeness of the end users (e.g. Young-Corbett, Nussbaum, & Winchester III, 2010). Therefore, it is questionable whether usability test results produced under such at

least in part artificial circumstances correspond to the actual usability problems that will occur when the system will be used in a realistic situation. Furthermore, there is currently only limited knowledge what factors affect the results of usability tests and how the reported variability in test results can be explained.

This publication-based doctoral thesis addresses some factors that may impact on the effectiveness of usability testing. It is structured as follows. First, the basic concepts of usability and user experience are introduced and the process model of user-centred design is described, which outlines how usable products can be developed. Then, the method of usability testing is presented in detail, which is the focus of this thesis, and it is distinguished from other usability evaluation methods. The 'four-factor framework of contextual fidelity' by Sauer, Seibel, and Rüttinger (2010) is introduced, which helps to organise the different variables that may impact on usability test outcomes. Three empirical studies are presented in the form of three scientific journal articles. The first investigated the fidelity of the prototype and examined the impact of a test system, which was presented to test participants either on paper or on screen, in connection with different task types. The second article addressed the developmental stage of a prototype in usability testing as perceived by test participants and analysed this variable in combination with the presence of additional observers in the test situation. The third study focused on dual-domain products, which can be used at work as well as for leisure purposes. It experimentally manipulated these two usage contexts in a usability laboratory and examined whether different levels of usability of a product had different effects in these two contexts. In the final chapter, a general discussion integrates the main findings of the studies, and presents recommendations for the four-factor framework, and for usability testing research and practice.

1.2. The concept of usability

There is a number of disciplines that contribute research to the design and evaluation of interactive systems, such as e.g. cognitive ergonomics, human-computer interaction (HCI) and human factors. While the research scopes and traditions of these disciplines may be different, they share a lot of common ground in their study of human-technology

interaction (Karvonen, Saariluoma, & Kujala, 2010).[1] They agree in their interest to foster a human-centred instead of a technology-driven approach in the design and evaluation of systems, to make sure that user requirements are sufficiently met. On the conceptual level, as far as interactive technology is concerned, referring to aspects of either the technological system or of the human users is both not sufficient to capture the quality of the interaction between them. Therefore, the concept of usability has been introduced in the late 1970s and early 1980s (e.g. Helander, 1981; Minow, 1978), to describe the quality of use of interactive systems (Bevan, 1995). There are various different and sometimes conflicting definitions of usability in the literature (e.g. Bevan, 2001; Kurosu, 2007; Nielsen, 1993; Quesenbery, 2003; Shackel, 1981). In this thesis, the concept 'usability' is used as it is described in the most widely accepted definition, which is the international standard of ISO 9241-11 (ISO, 1998). There, usability is defined 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' (p.2). The definition focuses on three main outcome variables of system usage, which are effectiveness ('the accuracy and completeness' of goal achievement), efficiency ('resources expended' in relation to the degree of goal achievement) and subjective satisfaction ('freedom from discomfort, and positive attitudes towards the use of the product' (p.2)). As all of these core aspects of usability are directly measureable, usability becomes quantifiable, and various indicators are described in the ISO 9241-11 standard (e.g. task completion rate). The context of use is an integral part of the definition, as the usability of any given product depends on all aspects Shackel (1991, p.22) describes as the 'four principal components of any user-system situation': the user, the task, the system, and the environment. The implication is that usability is not a characteristic of an interactive system per se, but always depends on the context within which it is used (Bevan & MacLeod, 1994), or in other words, a product may be usable in one context but not so in another. With respect to the aspects which make a product usable or not, the definition in ISO 9241-11 takes a 'black box approach', as it does not give any specifics that might be used in system design (Bevan, 2001, p.540). Other definitions specify components of usability, such as learnability, user retention over time, or error tolerance (J. Nielsen, 1993; Quesenbery, 2003), making it more evident that usability goes beyond the scope of simple 'ease of use'. The 'quality in use integrated measurement

[1] Differentiating between disciplines as cognitive ergonomics, human-computer interaction, human factors, etc. is definitely of value when taking a historical perspective on the respective research involved. However, it is questionable whether such a differentiation makes the research more accessible, as it introduces a kind of high-level information architecture layer, which is difficult to understand. This is especially the case for the practitioner facing a specific system design problem, for which findings from all of these fields may be relevant.

model' by Seffah et al. (2006) goes still further, as it decomposes usability into factors, criteria and specific metrics. It gives a catalogue of detailed aspects such as minimal memory load, consistency, self-descriptiveness, feedback, etc., to support the application of the usability concept in practice.

1.3. The concept of user experience

One of the main criticisms of the usability concept is that it focuses too much on performance-oriented outcome variables, with its core aspects of effectiveness and efficiency, thereby neglecting technology usage dimensions such as fun and joy of use (Hassenzahl, 2004). This is hardly surprising, as ergonomics and HCI research traditionally focused on the study of technology at the workplace (Carroll, 2001; Hendrick, 2000). However, as technology is ubiquitous today, usability has become relevant for consumer products not primarily used in a productivity-oriented context. In addition, a lot of research has shown that previously neglected aspects of systems, such as e.g. aesthetic properties of products, have an impact on usability test outcomes and need to be considered (Kurosu & Kashimura, 1995; Sauer & Sonderegger, 2009; Thüring & Mahlke, 2007; Tractinsky, 2000). Therefore, the research scope has been broadened accordingly (Carroll, 2004). The concept of user experience has been introduced, and it complements the previous focus on instrumental dimensions of goal-achievement (effectiveness, efficiency) with hedonic aspects of system usage, such as joy of use and the self-presentational functions of products for the user (Hassenzahl, 2004). While the objective of the traditional usability approach was the improvement of human performance by preventing usability problems, the user experience perspective strives to improve the user's satisfaction and to create outstanding quality experiences (Bevan, 2009; Hassenzahl & Tractinsky, 2006) Accordingly, there is a shift from work towards leisure with regard to systems and usage contexts being studied (Bargas-Avila & Hornbæk, 2011). However, a closer examination of research within the user experience approach shows that the definition of the concept of user experience is still ongoing, and that the methodologies used are either similar to those of traditional usability research, or raise issues with respect to their validity (Bargas-Avila & Hornbæk, 2011; Law, Roto, Hassenzahl, Vermeeren, & Kort, 2009). Overall, the concept of user experience offers a valuable new direction for the study of human-technology interaction, even if it is not entirely ground breaking and conceptual and methodological issues need further clarification. As the focus of this thesis is the method of usability testing, it mostly refers to usability, although the method is equally valuable in a user-centred design process with the objective to create an optimal user experience.

1.4. User-centred design in product development

Given that system usability is increasingly recognised as a highly valuable outcome of product development, the question is how to shape project processes in a way to ensure that user requirements will be met. There are various different software engineering models which describe how interactive technology should be developed, from the traditional waterfall model (Royce, 1970) to more recent agile methodologies (Abrahamsson, Warsta, Siponen, & Ronkainen, 2003). These models provide guidance how a project should structure the development process to implement a system according to requirements and within the typical constraints given by limited time and budget (Boehm, 1988). International standards in software engineering (e.g. ISO 9126-1; ISO, 2001) list usability as one of the six core software quality characteristics that the product of such a process can be evaluated against. However, in practice, software engineering focuses on the process of implementing the technology and clearly prioritises technical quality aspects such as maintainability, reliability, performance or security (Chung & Leite, 2009; CISQ, 2010). As a consequence, the models listed above do not include specific recommendations how to make sure the system will achieve a sufficient degree of usability for the intended users (Parsons, Lal, Ryu, & Lange, 2007). Unfortunately, many development projects are primarily driven by the respective software engineering model (Düchting, Zimmermann, & Nebe, 2007). But, usability is a highly complex design goal, which does not just happen by chance (Dray, 1995). Therefore, usability engineering models have been suggested to complement the software engineering activities with a user-centred design process, to make sure the developed interactive product is also usable. Most of the basic recommendations for such a process have been communicated already in the 1980s (Good, Spine, Whiteside, & George, 1986; Gould & Lewis, 1985), and various models have been suggested, such as the 'Usability Engineering Life-Cycles' by Nielsen (1993) or Mayhew (1999), 'Contextual Design' by Beyer and Holtzblatt (1999), the 'Integrated User-Centred Design' process of IBM (Vredenburg, Isensee, & Righi, 2002), and the 'Wheel Process Model for Usability Engineering' (Helms, Arthur, Hix, & Hartson, 2006). A generalised model framework is described in the international standard of ISO 9241-210 (ISO, 2010) 'Human-Centred Design Processes for Interactive Systems', which was formerly known as ISO 13407. In this thesis, the term 'user-centred design' (UCD) is used instead of 'human-centred design' (HCD), as the former is more directly supported by the method of usability testing, which is typically concentrating on the users of a product and not all humans potentially affected by it. The framework in ISO 9241-210 describes the basic characteristics of a user-centred design process, which are the basis of most of the more

specific process models mentioned above. The presented process model is not meant to replace software engineering models, but to complement them with user-centred activities. At its core, ISO 9241-210 identifies four general principles that are independent from development stages, and five essential activities of user-centred design. The principles include active user involvement and an understanding of the requirements of users and tasks, an appropriate allocation of functions between users and technology, the iteration of design solutions, and multidisciplinary design teams. The activities described in the standard are the planning of the user-centred design process, understanding and specifying the context of use, specifying the user and organisational requirements, production of design solutions, and the evaluation of these designs against the requirements (see Figure 1).

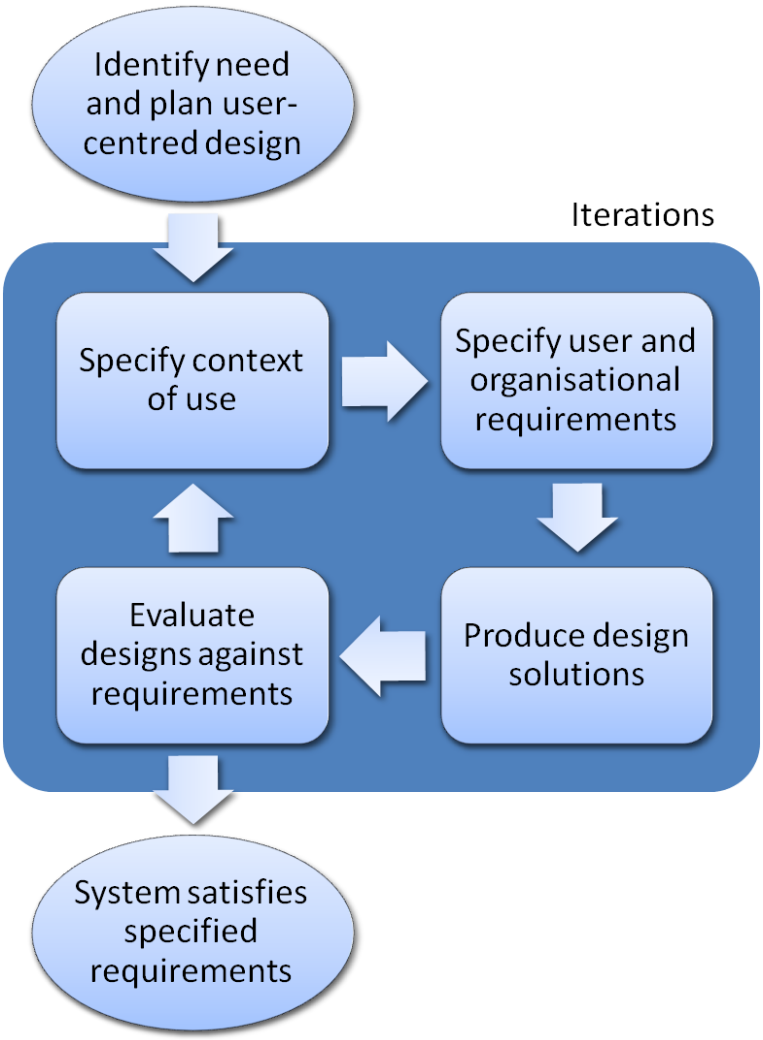

Figure 1. The UCD process model (adapted from ISO 9241-210; ISO, 2010)

In this iterative cycle, the evaluation activity is very important, as it can show to what extent the user requirements have been met, and provide specific information for further improvement (Maguire, 2001). ISO 9241-210 recommends conducting evaluations as early as possible, repeatedly and throughout the development process. It is very important to carry out evaluations early in a project, to be able to act on the feedback before changes become prohibitively expensive or compromise the project schedule (Gould & Lewis, 1985). While early evaluations aim for feedback that can improve the product (formative evaluation), later evaluations may assess the achievement of requirements on a more global level (summative evaluation). The object of the evaluation can either be a prototype in the early stages of design, or the implemented system later on. For both cases, there are various methods available to evaluate the design of an interactive technology.

1.5. Usability evaluation methods

The methods to support the evaluation of the usability of interactive technology are commonly referred to as 'usability evaluation methods' (UEM) (Gray & Salzman, 1998). Somewhat confusingly, these researchers and others (e.g. Hartson, Andre, & Williges, 2001) restrict the usage of the term to formative methods only, i.e. when used to identify specific usability problems of a system. This, however, implies that some very prominent usability evaluation techniques (e.g. usability tests) would be considered UEMs when used in a formative context, but not so when used under a summative approach making an overall assessment of a system. In situations with both formative and summative evaluation goals, the applicability of the term 'UEM' would become entirely unclear. To avoid this confusion, in this thesis the concept of UEMs is used in a broad way, to encompass all methods that perform a formative or summative evaluation of the usability of an interactive product or system.

In the 1980s, usability tests quickly became the primary method to evaluate the usability of interactive technology (Hartson et al., 2001). With its methodological roots in experimental psychology, usability tests at that time were set up as research activities with large sample sizes at relatively high cost and time requirements (Barnum, 2002). They were typically conducted only once and at a late stage in development projects (Rosenbaum, 2008). Because of the constraints in software development in the industry with respect to time and budget, researchers suggested cheaper and faster ('discount') usability evaluation methods as alternatives to expensive user testing (J. Nielsen, 1995). Consequently, in the 1990s, expert-based evaluation techniques became more popular, also because they can be used to evaluate systems very early on, even without a functional prototype being available.

One very prominent and often used expert-based UEM is the so-called 'heuristic evaluation' as introduced by Nielsen and Molich (1990), in which usability evaluators review a system without user involvement, according to a very limited list of specified general usability requirements. The goal was to offer an alternative to previously available, but very extensive and therefore unmanageable usability guidelines by originally specifying just nine usability heuristics (Molich & Nielsen, 1990). The heuristics have later been revised and consist of usability requirements such as 'visibility of system status', 'match between system and real world', and 'recognition rather than recall' (Nielsen, 1994, p.155). Various sets of heuristics have been suggested since, e.g. based on cognitive theory (Gerhardt-Powals, 1996), on the dialogue principles of ISO 9241-10 (Sarodnick & Brau, 2011), or on grounded theory categorising a large body of actual usability problems (Petrie & Power, 2012).

Today, a variety of UEMs are available, and they can be described on several dimensions. Empirical methods, which require some sort of data collection with representative users (e.g. usability testing, questionnaires), are distinguished from analytical methods, which are purely expert-based with no user involvement (e.g. heuristic evaluation, cognitive walkthrough) (J. Nielsen & Molich, 1990; Sarodnick & Brau, 2011). In theory, methods can also be categorised as to whether they are descriptive, assessing the current state of an interactive technology and its actual usability problems, vs. predictive and generating recommendations to improve a technology by avoiding specific usability problems (Gediga, Hamborg, & Duntsch, 2002). In practice, however, methods are often used to achieve both goals, identifying usability problems and to derive recommendations to improve upon them. UEMs can further be described according to a variety of dimensions, which may also serve as selection criteria to support the practitioner when choosing adequate methods in a given situation in product development (Mayhew, 1999). To name just a few, these are: the development lifecycle stage during which a method can be applied, the level of effort (costs) required to implement the method, the persuasive power of the results in development teams, and the level of expertise required to conduct the method (Blandford, Hyde, Green, & Connell, 2008; Hartson et al., 2001; Maguire, 2001; J. Nielsen, 1993).

With the aim of supporting usability practice, a large number of studies compared different UEMs, investigating the relative performance of these methods (e.g. Bastien, Scapin, & Leulier, 1999; Desurvire, 1994; Fu, Salvendy, & Turley, 2002; Jeffries & Desurvire, 1992; Jeffries, Miller, Wharton, & Uyeda, 1991; Karat, Campbell, & Fiegel, 1992; Mankoff et al., 2003; Nielsen & Phillips, 1993; Sears & Hess, 1999). The goal of these comparisons was to assess the predictive power of UEMs, to establish which method works best to identify

the largest proportion of existing usability problems in a given system (John & Marks, 1997). However, this strand of research has been criticised for a number of reasons, the most relevant of which are the following. First, some of the most prominent studies were shown not to fulfil the formal requirements of experimental research, as outlined e.g. by Cook, Campbell, and Peracchio (1990), so that the conclusions with regard to the performance of the investigated UEMs do not seem valid or reliable (Gray & Salzman, 1998). Second, the UEMs being investigated are not well defined, and the respective evaluation activity by the practitioner is not strongly prescribed. Consequently, it is not entirely clear what exactly is compared in these studies, and whether it corresponds to actual usability practice (Hornbæk, 2010; Woolrych, Hornbæk, Frøkjær, & Cockton, 2011). Third, as any practical application of a UEM is highly dependent on considerations given by the specific development context in which it is embedded (e.g. team buy-in, constraints in time and resources), the typical research approach comparing UEMs by isolating the methods from this context may be of limited practical significance (Wixon, 2003). And fourth, the aim to identify a single best UEM may not even be relevant for usability practitioners, as they often use a combination of different UEMs (Rosenbaum, Rohn, & Humburg, 2000; Uldall-Espersen, Frøkjær, & Hornbæk, 2008). Not surprisingly, given the problems with this strand of research, Gediga et al. (2002, p.141) state that the 'empirical comparison of software evaluation techniques is a minefield', and more recently, Cockton and Woolrych (2009, p.44) even conclude that 'positive knowledge (...) has remained very elusive in UEM research'.

However, irrespective of these problems affecting direct comparisons between UEMs, some general recommendations with regard to their use seem possible. Most importantly, evaluation goals need to be specified in advance, as these can guide the choice of the respective method and the measured indicators. As a general rule, evaluations should encompass more than just one UEM, and expert-based methods should always be used in combination with subsequent empirical evaluations (Lindgaard, 2006). In the case of formative evaluations, expert-based methods can be valuable in early design stages, when no functional prototype is available that could be tested with users (Hornbæk, Høegh, Pedersen, & Stage, 2007). If identified usability problems can already be fixed before user testing, the benefits of both methods can be maximised (Law & Hvannberg, 2002). If practical project constraints with respect to time and costs are high, expert-based evaluation may be a relatively cheap and fast way to achieve at least some usability improvements (Gediga et al., 2002). Detection rates for usability problems may be increased by having a group of experts evaluate a system instead of single evaluators, and by making the relevant expertise for the application domain available (Sarodnick & Brau,

2011). In cases of evaluating highly novel technology usage situations, expert-based evaluation may be of limited value, as it is more restricted to already well-known problems and the violation of established standards (Bach & Scapin, 2010). In the case of a summative evaluation goal or the comparison of different systems under realistic conditions ('how good is it?' or 'which one is better?'), empirical methods involving users are recommended (Gediga et al., 2002, p.128), as they allow for user performance measurement (Sarodnick & Brau, 2011). Furthermore, if higher persuasive power of evaluation results is required, e.g. when assessing a final product very late in development, usability tests are the method of choice (Blandford et al., 2008).

Overall, the selection of a UEM in a specific situation will always be based on various factors, including practical considerations of costs and benefits as well as ergonomic considerations (Gediga et al., 2002). But so far, a usability test is still the only UEM that allows for an accurate prediction of system usability in its usage context, while an expert-based assessment on its own may be a risky strategy, as these methods only give a rough indication of potential user problems (Bevan, 2007; Cockton & Woolrych, 2001).

2. Usability testing

2.1. Usability tests and their practical relevance

There may be no single best method to evaluate the usability of an interactive technology, but usability testing comes close. Given that usability is defined in terms of the interaction between users and a system, and expert-based UEMs focus on characteristics of the system, they can only draw indirect conclusions about its usability. A usability test, in contrast, focuses on the very core aspects of usability itself, as the user-system interaction is directly observed, and therefore it can provide an accurate prediction of usability in a specific context (Bevan, 2007). By simulating the actual usage situation, it allows for the measurement of user performance and satisfaction when operating an interactive technology (J. Nielsen, 1993; Sarodnick & Brau, 2011). It is generally agreed among researchers and practitioners alike that usability tests are an effective method to identify relevant usability problems and make it possible to improve the design of interactive products considerably (Bailey, 1993; J. R. Lewis, 2006; Richter & Flückiger, 2007; Ruthford & Ramey, 2000; Säde, Nieminen, & Riihiaho, 1998; Sefelin, Tscheligi, & Giller, 2003; Thompson & Haake, 2004). Or, as Wichansky (2000, p.999) puts it, there is 'no substitute for actual user data' when evaluating the usability of an interactive technology. Not surprisingly, several surveys show that usability testing ranks high up among the most often used methods in usability engineering (Fernandez, Insfran, & Abrahão, 2011; Følstad, Law, & Hornbæk, 2012; Mao, Vredenburg, Smith, & Carey, 2005; Venturi & Troost, 2004; Vredenburg, Mao, Smith, & Carey, 2002). Usability testing is also considered to have the highest strategic impact compared to other methods, possibly by bringing development teams closer to the actual users of their products (Rosenbaum et al., 2000). However, there is a common misuse of the term 'usability test', as it is sometimes used to refer to any technique conducted to evaluate the usability of a system (Dicks, 2002; Rubin & Chisnell, 2008). Therefore, the method needs to be defined and clearly distinguished from other UEMs, by specifying the core aspects of a usability test (J. R. Lewis, 2006). Each of the following five criteria needs to be met for an application of the method to be properly termed a 'usability test' (see Table 1), and these are supplemented by recommendations how to obtain useful test results (Dumas & Redish, 1999).

Table 1: Defining aspects of a usability test

Defining aspects of a usability test	
Usability data	Primary outcome is data informing about the usability of the test system.
User involvement	The test involves participants representing real users.
Task orientation	Participants work on tasks using the test system.
Observation	Participant behaviour is observed and/or recorded.
Data analysis	Usability data is analysed according to the pre-defined test goals.

First, the primary outcome of a usability test is data allowing for an assessment of the usability of a system. While some researchers claim that a usability test always aims to improve system usability (e.g. Dumas & Redish, 1999), this is only one possible goal of the application of the method and should not be confused with the definition of the concept itself. When used in a summative approach, providing only global measures of system usability, a usability test does not necessarily give an indication of how to improve a system at all. And, even for formative tests, it is important to stress that the test per se can only identify usability problems and cannot 'test usability into a product' (Wichansky, 2000, p.1002), without subsequent usability engineering activities beyond testing. Second, the method involves test participants taking on the roles of real users operating the system. As a recommendation to obtain valid test results, it is generally agreed that participants should be either directly recruited from the intended user group, or at least be representative of them in terms of the relevant characteristics (e.g. IT experience) (Rubin & Chisnell, 2008). Third, the participants in a usability test work on realistic tasks, using the interactive technology that is evaluated. To reduce the probability of missing important usability problems of a system, it is recommended to cover all relevant functionalities (J. Nielsen, 1993) and the task descriptions should not give any clues about possible solutions by containing terms used by the tested system (Molich et al., 2004). Fourth, the interaction of the participant with the system is observed in some way, either directly by an evaluator being present in the room or by some means of indirect observation (recording, transmission, etc.). Fifth, the data is analysed to extract the results according to the pre-defined goals of the test, e.g. to identify usability problems or to assess more global aspects of system usability, such as performance.

This definition of usability testing allows for very different applications of the method, as it does not prescribe 'what to test, with whom, when to test, where to test, or why to test' (Woolrych et al., 2011, p.952). Using a formative approach, a usability test can be effective in identifying relevant usability problems, while within a summative approach more global indicators of the usability of a system can be assessed, such as success rate or completion time (Sarodnick & Brau, 2011). The method allows for the collection of qualitative data about specific usability problems just as well as of quantitative data on user performance or perceived usability, depending on the sample sizes used in a specific test (Dicks, 2002). Usability tests can be run early in a user-centred design process, utilising interactive prototypes or late in a development project, testing the final system (Hall, 2001). They can be conducted in the laboratory or in the field, to create realistic usage situations for stationary or mobile systems being tested (Kaikkonen, Kallio, Kekäläinen, Kankainen, & Cankar, 2005; Rowley, 1994). Remote versions of the technique allow for larger samples taking part in their natural usage contexts and for potentially lower recruiting costs (Bruun, Gull, Hofmeister, & Stage, 2009; Hammontree, Weiler, & Nayak, 1994). Competitive testing can compare two or more systems in a usability test (Dolan & Dumas, 1999). Tests can be conducted using different levels of formality of the testing procedure (J. R. Lewis, 2006), and covering very different levels of system complexity (Redish, 2007).

Some confusion in the usage of the term 'usability testing' exists with respect to the method of think-aloud (Duncker & Lees, 1945), such that the two terms are sometimes used synonymously (e.g. Jacobsen, Hertzum, & John, 1998; Jaspers, 2009). However, even if in practice think-aloud is routinely used when conducting a usability test, it is misleading not to distinguish between the two. Think-aloud in the context of usability testing refers to the method of having test participants verbalise their thoughts while working on the test tasks to gain insight into their thought processes (C. Lewis & Mack, 1982). This may be a highly valuable technique supporting data collection when conducting a usability test (J. Nielsen, 1993), but it is no necessary component of the method itself. Its use primarily depends on the research goals of the specific implementation of a usability test. Rather, the way in which think-aloud is often conducted in practice may represent one factor that limits the reliability of data coming out of a usability test (Boren & Ramey, 2000).

2.2. Inconsistencies between usability test results

As usability tests inform critical decisions in product design, the results that the method produces need to be justified (Molich et al., 2004). Rooted in experimental psychology and drawing on the respective techniques for studying human behaviour, it has been claimed that the methodological requirements of reliability and validity equally apply to usability testing (e.g. Wenger & Spyridakis, 1989). The concept of reliability in the context of usability testing refers to the level of accuracy with which the measured indicators (e.g. task completion rate, identified usability problems) can be assessed, minimizing the measurement error and maximising the true score (Wenger & Spyridakis, 1989). One common way to establish the reliability of a measurement procedure is retest reliability (Cronbach, 1947). The concept implies, that if a usability test produced 'reliable' results, independent tests of the same evaluated system would have to yield similar outcomes (Hertzum & Jacobsen, 2003). However, several studies show that the latter is not the case. The most prominent research investigating the reliability of usability tests are the 'comparative usability evaluation' ('CUE') studies by Molich and colleagues (Molich & Dumas, 2008; Molich et al., 1999, 2004). These studies consistently found that the overlap of the reported usability problems of the same tested system was surprisingly low. In the first CUE study on usability testing, four professional usability teams independently conducted usability tests of the same calendar management software. Out of a total of 141 usability problems that were identified, 91% were reported by one team only, and only one usability problem was reported by all four teams (Molich et al., 1998). The second CUE study had nine professional usability teams conduct tests of a popular internet e-mail application and resulted in similar findings: out of a total of 310 reported usability problems, 75% were reported by one team only, and not a single issue was identified by all teams (Molich et al., 2004). A subsequent study involved nine teams carrying out usability tests of a hotel website, again reporting low consistency rates, with 67% of the 237 usability problems in total being reported by only one team (Molich & Dumas, 2008). Other researchers reported only slightly higher rates of consistency. In a study involving six professional usability testing teams evaluating an early prototype of a dialog box, 44% of the total of 36 detected usability problems were reported by one team only (Kessner et al., 2001).

However, in spite of these findings, the justification of the method of usability testing may not be as questionable as it first seems. The results can be explained in a number of ways, and the corresponding factors may well act together and cause apparently inconsistent usability test outcomes. First, in the CUE studies, testing teams received little

guidance on how to carry out the evaluation, and consequently, there was substantial variation in terms of how the tests were conducted (Molich et al., 2004). For example, different tasks were given to participants, and in turn, different parts of the system were covered in the tests. This makes the inconsistent results less surprising and, methodologically, they become much less of a problem for the method of usability testing, as these independent tests do not represent retests at all. Second, although there is a common general understanding what constitutes a usability test, it is clearly not the case that the method prescribes the evaluation procedure to an extent that evaluator decisions become irrelevant (Hornbæk, 2010). As usability professionals differ in various relevant aspects, e.g. their practical knowledge about usability methods, their prior experience with the tested system and their opinions about it, they may set up and conduct a usability test differently and, in consequence, may be able to identify some problems but not others (Hertzum & Jacobsen, 2003). Third, there is no commonly agreed set of criteria what constitutes a usability problem, an objective threshold indicating when a difficulty would need to be classified as a usability problem or on what level of analysis these should be reported. Consequently, there is a lot of room of interpretation on behalf of test evaluators in identifying these (Hertzum & Jacobsen, 2003; Hertzum, Molich, & Jacobsen, 2013; Hornbæk, 2010). Specific inconveniences caused by the interface of an interactive technology can be reported by one evaluator and not by another, or alternatively, be reported in different ways, which can make it very difficult to match these reports in a reliability study (Hornbæk & Frøkjær, 2008; Lavery, Cockton, & Atkinson, 1997). Fourth, empirical evidence shows there is substantial between-subjects variability in terms of the performance in a usability test session (J. Nielsen, 1993). As this adds error variance which does not reflect effects of usability aspects of the evaluated system, it lowers the reliability of the usability test results, especially since the typical sample in a usability test is quite small. In contrast, if the usability test has the goal to identify usability problems of a single system, inter-individual differences may even be an advantage, which can help finding additional weaknesses of a system. In the case of a comparison of two versions of a system, or if quantitative performance data are to be collected, test subject variance becomes a problem for the reliability of the results (Sarodnick & Brau, 2011). Finally, as any interface with some degree of complexity potentially suffers from a very large number of usability problems, it cannot be achieved by a (typically very limited) usability test to capture them all (Molich et al., 2004). Low overlap between independent tests may therefore be expected and might not reflect low reliability in problem detection, but rather indicate a huge problem space, in which additional tests simply add more useful knowledge about further problems of the system in question (Lindgaard, 2006).

Considering all these factors, it might not be overly surprising, that results of independent usability tests of the same system turn out to be rather inconsistent. This is in support of the view that usability tests, strictly speaking, do not represent a 'complete method' prescribing evaluator activities in detail. Rather, they may be conceived of as a 'broad approach', a 'collection of resources' that need to be configured and adapted in every instance of their practical application (Woolrych et al., 2011). This would imply limited direct relevance of methodological requirements such as reliability and validity for the practice of usability testing (but not for scientific studies investigating these approaches) (Lindgaard, 2006). In addition, practitioners may not even be too worried about inconsistent results of parallel usability tests that usually will not take place, given the typical constraints with respect to time and budget in development projects. For them, the far more relevant question may concern that of the validity of their test results, i.e. whether the usability test outcomes reflect user problems that would occur later on when using the final product in its real-world context of use. For research, the findings demonstrate that the method of usability testing is by no means trivial and a lot more work is required investigating potential factors that affect test outcomes.

3. Contextual fidelity in usability tests

3.1. Predictive and ecological validity of usability test results

While Gray and Salzman (1998) discuss several types of validity that are critical in scientific experiments investigating the effectiveness of different UEMs, these may not be as relevant in the case of the practical application of these methods (Lindgaard, 2006). For the practitioner using usability tests as a diagnostic tool to evaluate an interactive technology, two types of validity seem crucial: predictive and ecological validity (Sauer & Sonderegger, 2009). *Predictive validity* is a form of criterion validity and refers to the correspondence between an earlier measurement and the predicted criterion, which is assessed at a later point in time (Cronbach & Meehl, 1955). The results of a usability test can be considered to have predictive validity, if e.g. the usability problems detected with a prototype in development also show up in the usage of the final system later on (or would have, had the design not been improved). *Ecological validity* refers to the extent that a measurement in an experimental setting corresponds to the one in a real-world context (Hoc, 2001). The results of a usability test can be considered ecologically valid, if e.g. the user problems identified in the usability laboratory also occur in the real-world usage within the realistic context of use (Hertzum & Jacobsen, 2003). As the goal of a usability test typically lies in the prediction of usability problems that users would have in their real world usage of a system, predictive and ecological validity of this method practically become inseparable. To increase both types of validity, a usability test aims to simulate the realistic usage situation of the evaluated system (Hertzum, 1999). The various aspects that determine representativeness of the usability test situation for the intended real-world usage are presented in a systematic way by the four-factor framework of contextual fidelity (Sauer et al., 2010).

3.2. The four-factor framework of contextual fidelity

The degree to which the situation in a usability test is representative of the real-world usage context of the evaluated system determines the ecological validity of the test results. Relevant differences between the two situations may reduce the possibility to generalise test results, therefore reduce the value of a usability test (Bevan, 2007). In the context of usability tests, the term 'fidelity' was traditionally used to refer to the test system or prototype only, and not to characterise the test situation overall (Rudd, Stern, & Isensee,

1996). In a similar way, previous models of fidelity in usability testing have concentrated on the similarity of the test system to the final version of the product (e.g. McCurdy, Connors, Pyrzak, Kanefsky, & Vera, 2006; Virzi, Sokolov, & Karis, 1996). However, the situation in a usability test can be representative to the real-world usage context of a system in many other ways, beyond the test system under evaluation. Therefore, the concept of 'fidelity' is used here in a broad way to refer to the degree of similarity of the test situation to the one of real usage, including technology-related, social and environmental aspects of the context. To offer guidance in detecting potential causes of reduced validity of test results, the 'four-factor framework of contextual fidelity' is taking into account the wider context of usability testing (see Figure 2; from Sauer et al., 2010).

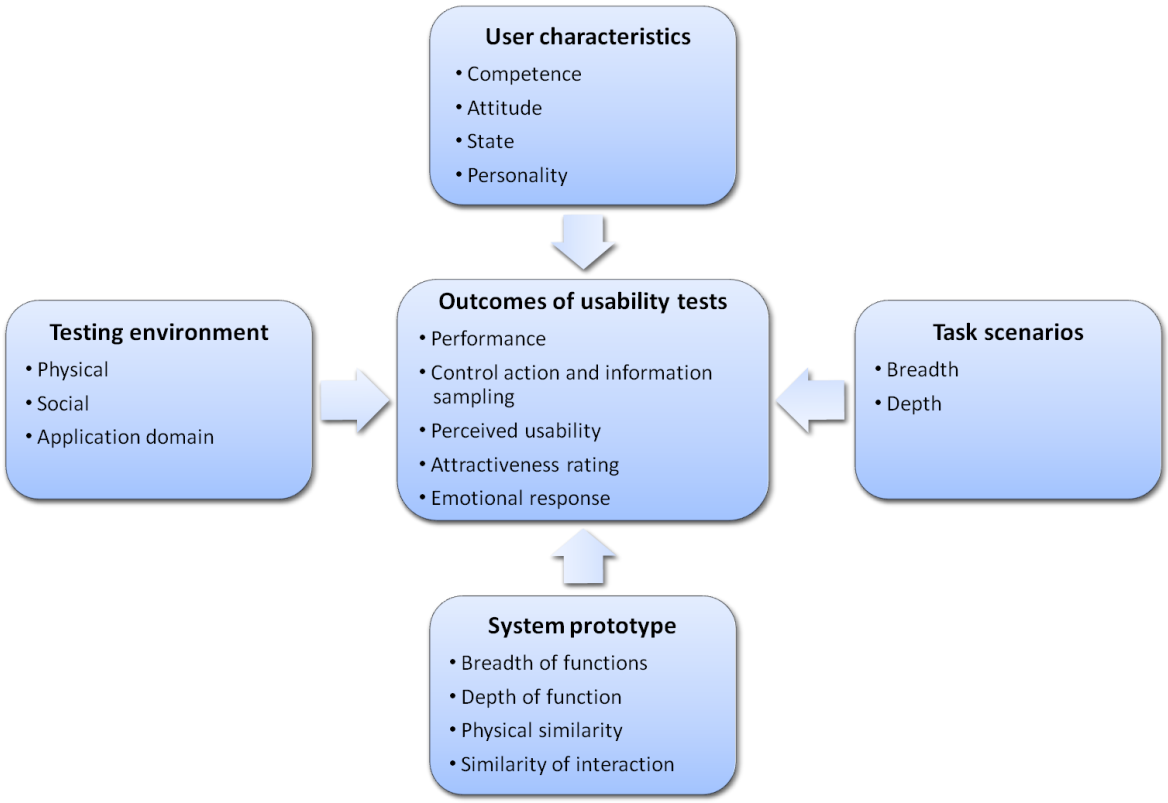

Figure 2. The four-factor framework of contextual fidelity (from Sauer et al., 2010, p.132)

The framework respects the multi-dimensionality of the fidelity of the test situation, incorporating all four components of any user-system interaction (Shackel, 1991): user ('user characteristics'), task ('task scenarios'), system ('system prototype'), and environment ('testing environment'). These correspond to the four dimensions suggested by Thomas and Kellogg (1989), on which a usability test situation can differ from the real-

world usage, which they describe as 'ecological gaps'. For each of these four factors, the framework lists a number of subordinate factors, which indicate aspects that need consideration when conducting a usability test.

The participants in a usability test have to be representative of the users of the final system in terms of *user characteristics*, so that the test can simulate adequately how the system will be used in practice (Sarodnick & Brau, 2011). If test participants are different from the intended users in terms of competence (e.g. domain expertise) and attitudes (e.g. preferences), the ecological validity of the test results may become questionable (Thomas & Kellogg, 1989). Other factors such as state (e.g. level of fatigue), or personality (e.g. conscientiousness) may also affect test outcomes, e.g. by how participants are willing and able to follow the test procedure. In consequence, relevant usability problems may not be detected, measured task performance may not reflect realistic usage, or task motivation may help overcome usability difficulties that would not be accepted by the target population. One common cause of non-representative participants is recruiting users that are easily available instead of those with characteristics that correspond to the intended user group of a system (Sjøberg et al., 2002).

The *task scenarios* in a usability test have to be as representative as possible of the real usage context of the evaluated system (J. Nielsen, 1993). Since the typical usability test cannot cover all parts of a complex system, some functionalities are not evaluated and the respective usability problems cannot be detected. Consequently, the ecological validity of the test results are limited, as they cannot be generalised to realistic system usage (Thomas & Kellogg, 1989). The four-factor framework of contextual fidelity presents two subordinate factors, which may be relevant. Breadth of the task scenario refers to the complexity of the natural task environment that is simulated in the usability test. An example would be whether the concurrent use of multiple devices (e.g. computer and mobile phone) for different tasks, which is a common situation in real-world technology usage (Karlson, Meyers, Jacobs, Johns, & Kane, 2009), would be modelled in the usability test. The depth of a task scenario refers to how detailed a task can be completed in the test situation, and weather all elements of a task are represented.

As the empirical part of this thesis concentrates on the two factors 'system prototype' and 'testing environment', these are presented in more detail in the following two sections.

3.3. Prototype fidelity in usability tests

In the context of the development of interactive technology, usability tests are typically working within tight constraints given by available budget and time. In modern software engineering projects adopting fast-paced agile development processes, constraints with respect to time may become even more relevant (L. Nielsen & Madsen, 2012). In addition, practitioners usually need to present findings from usability tests long before an implemented version of the developed system is available, to be able to maximise their impact on designs that are still work in progress (Gould & Lewis, 1985). Furthermore, if working with an iterative design process of repeated testing cycles, the test system needs to be easily changed or recreated in light of previous test findings, to allow further investigation of the suggested redesign. This is necessary, as changes inspired by usability test results can potentially lead to new problems in an interface (Kaspar, Hamborg, Sackmann, & Hesselmann, 2010). Consequently, usability tests are often conducted using prototypes that require little effort and time to create and adapt, and prototyping is found to be one of the most often used approaches in user-centred design (Vredenburg, Mao, et al., 2002). Various techniques have been suggested to devise prototypes for usability testing, from paper-prototyping (Snyder, 2003) to automated software tools for prototype creation (Hosseini-Khayat, Hellmann, & Maurer, 2010). The resulting prototypes can take many forms, from simple screen drafts presented on paper to clickable slide mock-ups on a tablet, and hardware interfaces can equally be modelled using quickly adjustable alternative materials. At the functional level, a prototype can offer some functionalities of interest in detail ('vertical prototype') or the presentation of a broad range of features, omitting the depth of these ('horizontal prototype') (J. Nielsen, 1993). The specific prototype design depends on the research questions for which user feedback is required (Houde & Hill, 1997).

However, as prototypes are not implemented systems with full functionality and polished interface design, they only represent the final system to a certain degree. Therefore, the question of adequate fidelity is important, i.e. whether a prototype is sufficiently representing the characteristics of the finished system to obtain valid usability test results (Hall, 2001). If the test artefact is not sufficiently representative of the final system, the ecological validity of the test results can become questionable (Thomas & Kellogg, 1989). Traditionally, a distinction is made between low- and hi-fidelity test prototypes, with the former only offering limited similarity to the final product with respect to visual design or functionality. In turn, they are quicker to develop at lower costs, and can be used very early on in a development project. Hi-fidelity prototypes, on the other hand,

offer more faithfulness in representing the look and feel of a system at higher costs, often with the advantage to test detailed interaction, which may not be available in a low-fidelity prototype (Rudd et al., 1996).

Previous models of fidelity in usability testing have concentrated on prototype fidelity only. These models typically include about five dimensions (McCurdy, Connors, Pyrzak, Kanefsky and Vera, 2006; Snyder, 2003; Virzi, Sokolov and Karis, 1996). The dimensions covered are breadth of functionality (number or percentage of features which are available in the prototype), depth of functionality (degree to which the details of operation of a specific feature is available), look (aesthetic refinement of the visual design), interaction (similarity and richness of interactive elements), and the data model (richness of the domain-specific data that is available). In the four-factor framework of contextual fidelity, this is represented by the factor 'system prototype', which includes the subordinate factors breadth of functions, depth of function, physical similarity, and similarity of interaction (Sauer et al., 2010).

Specifically, for the aspect of physical similarity (i.e. the medium of prototype presentation), there is only a limited number of studies investigating the effects in usability testing (see Sauer, Franke and Ruettinger, 2008 for an overview). These studies consistently failed to find differences with respect to the identified number or severity of usability problems, interaction efficiency, or for subjective usability evaluation due to the medium of prototype presentation (Catani & Biers, 1998; Sauer & Sonderegger, 2009; Sefelin et al., 2003; Virzi et al., 1996; Walker, Takayama, & Landay, 2002). There were some interesting findings nonetheless, but as these emerged in single studies, they need to be investigated further. These results indicate that when using a computer prototype, participants make more comments about the system (Walker et al., 2002), show more explorative activity (Hamborg, Klassen and Volger, 2009) and generally prefer working with it (Sefelin et al., 2003), as compared to a paper-based prototype. Based on these results, it is generally accepted that low-fidelity prototypes are equally capable to find valid usability problems as more elaborated versions of the test system, and there is the general recommendation to choose fidelity and medium of prototype presentation according to practical considerations in projects (e.g. costs or availability).

However, this body of research has some limitations, which show there is a need for further investigation. First, the number of studies having investigated the important technique of prototyping is still rather small. Second, some studies confounded prototype fidelity with medium of presentation (e.g. Catani & Biers, 1998; Virzi et al., 1996), which makes the results of these studies questionable. Third, a review of the tasks that were used in these studies shows that these have consistently not been selected according to any

hypotheses. Therefore, the first empirical study (study I) presented in this thesis adds to this research, but takes a more theory-driven approach and avoids the confusion of medium and fidelity level of prototype presentation (section 5). The second empirical study (study II) investigates levels of perceived prototype fidelity and the influence of the social testing environment (section 6).

3.4. Fidelity of the usability testing environment

The physical and social environment in which a system is used is described as one of the main aspects of context of use, as defined in the standard process of user-centred design (ISO 9241-210, former ISO 13407; ISO, 2010). However, while the other factors of context of use (users, tasks and equipment) are usually described in detail in usability testing research, the wider physical and social environment is regularly neglected (Alonso-Rios, Vazquez-Garcia, Mosqueira-Rey, & Moret-Bonillo, 2010). To take these aspects into account, the four-factor framework of contextual fidelity includes the factor 'testing environment', and lists as subordinate factors physical and social environment, and the application domain (e.g. work vs. leisure) (Sauer et al., 2010).

To achieve high ecological validity of results, the situation in the usability test needs to reflect the relevant aspects of the real-world context that have an impact on the usability of the final system (Thomas & Kellogg, 1989). But usability tests have traditionally been conducted in a laboratory environment, which allows for the experimental control of interfering variables and the collection of various data about participants' behaviour when using the test system (Seffah & Habieb-Mammar, 2009). With respect to the factor 'physical testing environment', the setting can thereby differ significantly from the intended usage context of the final product, e.g. if the system is a mobile application typically not used in a stationary setting at all (Kjeldskov & Skov, 2003). For example, Nielsen, Overgaard, Pedersen, Stage, and Stenild (2006) directly compared a usability evaluation of a mobile system in two settings, and they reported significantly more usability problems being found in the field as compared to the laboratory. Interestingly, the researchers found problems with interaction style and cognitive load only in the field setting and not in the laboratory. Dahl, Alsos, and Svanæs (2010) were able to identify potential usability problems of a mobile device for hospital workers in the laboratory, since they designed the testing environment with a sufficient level of fidelity to the real-world setting. In their study, the inclusion of the work uniform worn in the target context proved relevant for testing how the mobile device would be used in the work setting. It seems likely that these type of

problems could easily be missed, if usability tests are conducted in the laboratory with less effort being put into the creation of adequate levels of fidelity of the testing environment.

The factor 'social testing environment' is concerned with the social context of use, e.g. whether observers or co-workers are present. For example, in a laboratory usability study, the presence of silent observers had negative effects on physiological parameters and on some performance indicators (Sonderegger & Sauer, 2009). Other studies found some indication for higher error rates when the social test setup was designed as being more intrusive in terms of observer presence and monitoring equipment (Grubaugh, Thomas, & Weinberg, 2005; Harris, Weinberg, Thomas, & Gaeslin, 2005).

Of course, it is possible to simulate specific usage contexts effectively in the laboratory, e.g. to have test participants walk while operating a mobile technology (Kjeldskov & Skov, 2007). And ideally, in a user-centred design process, usability test design at later stages of a project could be based on an analysis of the context of use conducted at earlier stages (Mayhew, 1999). But, even if relevant aspects of the real-world usage context could be simulated in the laboratory, the question remains whether all such aspects were identified and considered. Consequently, there are reservations with respect to the ecological validity of usability tests, if the test setting does not correspond to the intended usage environment.

The third subordinate factor 'application domain' refers to the usage environment in terms whether the interactive technology is used at home or in a work context. The authors of the four-factor framework set up the hypothesis that this factor may moderate the influence of other aspects in a test situation (Sauer et al., 2010). The third presented empirical study in this thesis investigates this assumption, i.e. whether usability proves to be more relevant in a work than in a leisure context, and whether for aesthetical properties of a test system the opposite might be true (section 7). The empirical studies are presented in the following sections.

4. Overview of the three empirical studies

The three studies presented in this doctoral thesis focus on the fidelity of prototypes and testing environment in usability tests. All studies investigate factors which are included in the four-factor framework of contextual fidelity (Sauer et al., 2010).

Specifically, *study I* revisits the debate on low- vs. high-fidelity prototypes, but using a more hypothesis-driven approach. Based on an analysis of the affordances of the medium of paper, it presents the assumption that paper prototypes might elicit more reading behaviour by test participants, and in consequence, performance differences may emerge only for tasks requiring reading activities.

Study II investigates potential effects of perceived prototype fidelity, specifically with respect to the developmental state of the system. This instruction is examined together with a manipulation of the social testing environment, i.e. the presence of silent observers. It replicates some of the experimental conditions of a previous study of observer presence (Sonderegger & Sauer, 2009).

Study III focuses on the factor 'application domain', for which a usability test is set up. In a laboratory-based approach, it compares outcomes of a usability test in a work context to one set up in a leisure context. A dual-domain product is used as a test system, which is equally relevant in both contexts. In addition, the usability of the tested system is manipulated by introducing system response time delays. The goal of the study was to investigate whether usability would prove to be more relevant in a work context than in a leisure context.

5. Study I: The fidelity of the prototype (paper vs. computer-simulated)

Prototype fidelity and test tasks in usability testing: The impact of paper vs. computer prototypes revisited

Andreas Uebelbacher*, Andreas Sonderegger, Klaus Heyden,

Jürgen Sauer

Department of Psychology, University of Fribourg, Rue de Faucigny 2, CH-1700 Fribourg, Switzerland

(to be submitted)

Abstract

Previous research comparing various media of prototype presentations has not found differences with respect to user performance or the identification of usability problems. The present study investigated whether effects of medium of presentation would emerge if different task types were separated. Employing a 2 x 3 mixed design, an experimental comparison was made between two different forms of prototype presentation, being used to complete three different task types (exploration task, reading task, navigation task). 65 participants were observed completing these tasks with a web application, either presented as a paper prototype or on a computer screen. Performance data, subjective measures and physiological parameters (heart rate variability) were recorded. For the exploration task, the results showed less active system viewing of participants using the paper prototype than the one presented on screen. For the reading task, there was a clear performance advantage for the paper prototype over the screen. In the navigation task, there was no performance difference for prototype conditions. Finally, there was no impact of presentation media on subjective measures such as perceived usability, system attractiveness, or participants' emotion.

Key words: usability test; paper prototype; prototype fidelity; paper vs. screen; test tasks; reading; perceived usability; user performance

* Corresponding author. E-mail: Andreas(@)uebelbacher.ch.

5.1. Introduction

In the process of developing interactive products that should meet user needs it is vital to apply methods which provide user feedback at early stages of a project, to avoid higher costs of later product changes (Bias & Mayhew, 2005). One of the implications for the usability practitioner is that usability tests need to be run long before a finished system is available for testing. Consequently, it is very common to use prototypes as representations of the system to be tested (Vredenburg, Mao, et al., 2002). Prototypes as simulations of a system to be developed may differ from the final system with regard to interactivity, design or functionality. Therefore, it becomes a key issue for usability practitioners to understand what characteristics of prototypes are important for a valid usability test and how these characteristics can influence test outcomes (Lim, Pangam, Periyasami, & Aneja, 2006).

Several studies have addressed this question, directly comparing rough paper-based prototypes to fully functional systems presented on screen to investigate effects on test outcomes (e.g. Catani & Biers, 1998; Virzi, Sokolov, & Karis, 1996; Walker, Takayama, & Landay, 2002). However, none of the previous studies offered an analysis of the characteristics of paper and screen, to be able to derive hypotheses about specific differences that might emerge for the respective prototypes. More specifically, none of these studies put forward any hypotheses for what kind of tasks such differences might emerge. The presented study aims to fill this gap and takes a more hypotheses-driven approach to prototype media in usability testing and the relevance of task types in this context.

5.1.1. Prototypes in usability testing

Within the context of software development projects, usability tests are a core method to gain feedback about a user interface, whether intended users can understand and operate it and what potential usability problems exist (J.R. Lewis, 2006). As a usability test always requires an interface to be tested, early in a project the system which is not yet available is usually represented by a prototype (Hall, 2001). Prototype design faces a number of conflicting requirements (Sauer, Seibel, & Rüttinger, 2010). Practical constraints of limited time and budget clearly favour rapid and low-cost prototyping techniques, which allow for quick adjustments in iterative cycles when critical user feedback is received. However, a prototype also needs to be sufficiently representative of the final product to draw valid conclusions from user comments. The question arises what the optimal fidelity of the test

system would be, given the trade-off between economical production and sufficient representation of the final product (Hall, 2001).

Research on prototypes in usability testing has mostly focused on the question of prototype fidelity. Initially, there was the simplified view that prototypes could be classified as either low- or high-fidelity, and their respective advantages and disadvantages have been discussed (Rudd, Stern, & Isensee, 1996). More recently, the concept of fidelity has been presented to comprise about five dimensions (McCurdy, Connors, Pyrzak, Kanefsky, & Vera, 2006; Snyder, 2003; Virzi, Sokolov, & Karis, 1996). These cover breadth of functionality (number or percentage of features which are available in the prototype, equivalent to the term 'horizontal prototype'), depth of functionality (the degree to which the details of operation of a specific feature is available, equivalent to the term 'vertical prototype'), look (the aesthetic refinement of the visual design), interaction (the similarity and richness of interactive elements), and the data model (the richness of the domain-specific data that is available for display and manipulation). Since the dimensions are considered to be independent from one another, prototypes can score high on one dimension but low on another (Petrie & Schneider, 2006).

One aspect of prototypes that often gets confused with fidelity is the medium in which the prototype is presented to test participants. While fidelity refers to the extent to which a prototype is similar to the final system ('high-fidelity' prototypes being more so than 'low-fidelity' ones), the medium of prototype presentation describes in which form a prototype is presented to test participants (e.g. on paper, on screen, or as an interactive voice response system, etc.). As Walker, Takayama, and Landay (2002) pointed out, in some studies paper prototypes are equated with low-fidelity prototypes and therefore the potential differences between these aspects are neglected. Some studies directly compared low-fidelity paper to high-fidelity prototypes on screen (e.g. Catani & Biers, 1998; Virzi et al., 1996), thereby confounding the two aspects. Although paper prototypes are normally low-fidelity prototypes in practice, as they offer limited options on most dimensions of prototype fidelity, on some of these dimensions they can be just as or even more elaborated than computer prototypes on screen, e.g. on the level of visual refinement and aesthetics. And various computer-based tools specifically support the creation and presentation of prototypes on low levels of fidelity (Hosseini-Khayat, Hellmann and Maurer, 2010), making clear that fidelity and medium of presentation are two different aspects of prototypes.

Given that prototyping is considered to be one of the most widely used methods in user-centred design (Vredenburg, Mao, et al., 2002), it is surprising how little research can be found that compares prototypes of different fidelity levels (see Sauer, Franke & Rüttinger, 2008 for an overview). Furthermore, although paper-based prototyping is still

widely used in practice (e.g. Johannsson & Hvannberg, 2004; Jokela, Koivumaa, Pirkola, Salminen, & Kantola, 2006; Kieffer, Lawson, & Macq, 2009; Meszaros & Aston, 2006; Olmsted-Hawala, Romano, & Murphy, 2009), it is equally surprising that only very few studies specifically addressed the medium of paper or screen presentation of prototypes in usability testing. Our literature review revealed that most studies that did address this question could not find any differences due to the medium of prototype presentation. Virzi et al. (1996) compared two quite different actual products (an electronic encyclopaedia on an e-book reader and an interactive voice response system) each to paper prototype versions and found no differences in the number or type of usability problems found. Catani and Biers (1998) compared paper- and computer-based versions of a library search system in a usability test and found no differences in number or severity of usability problems discovered. While both of these studies confounded the medium of presentation with fidelity, other studies that avoided these methodological problems came to the same results. Walker et al. (2002) conducted a usability study testing an online banking application with a 2x2 between-subjects design, specifically separating the effects of medium of presentation (paper vs. on screen) from prototype fidelity (low vs. high). Again, they did not find any differences in number or severity of usability issues identified by participants due to medium or fidelity either, but found that users made significantly more comments about the product when they were using a computer prototype on screen as compared to paper. Sefelin, Tscheligi and Giller (2003) tested a calendar system and the touch screen of a ticket vending machine on paper and on a computer screen and compared participants' suggestions for improvement. There was no effect of medium of presentation on number of suggestions, but almost all participants stated they preferred working with the computer prototype on screen to using the paper prototypes. When testing a standard mobile phone either on paper or simulated on screen, Sauer and Sonderegger (2009) found no effects for medium of presentation on interaction efficiency, subjective usability evaluation, emotions or rated attractiveness of the system. Hamborg, Klassen and Volger (2009) found no differences for the medium of presentation for the same mobile phone simulation with respect to the identified number, type and severity of usability problems, but they clearly found less exploration for the paper-based prototype. Summing up the findings of previous research, it can be stated that almost none of the studies could find a difference in usability test results as a function of medium of presentation.

However, some issues seem surprising. First, none of the studies mentioned above have considered the specific characteristics, or more specifically affordances, of the medium paper as opposed to screens. This would allow for specific hypotheses in what way different user reactions to the two types of stimuli material could be expected. Second, since such

expectations and hypotheses were missing, the usability test tasks in the mentioned studies were not selected in a way to make it likely for specific differences to emerge. Third, none of these studies referred to the extensive body of research in other fields (e.g. research comparing reading on paper vs. on screen, cf. Noyes & Garland, 2008) that specifically addressed the differences between paper-based and on screen stimuli material and which regularly found effects on task performance. As this research might offer valuable guidance for studying media effects in prototype research, relevant findings are presented below.

5.1.2. Medium of prototype presentation

When analysing the media of paper and screens and possible effects they might have on user behaviour in usability tests, the concept of affordances may be helpful. It refers to 'the perceived and actual properties of the thing, primarily those fundamental properties that determine just how the thing could possibly be used' (Norman, 1988; p.9). Affordances are directly relevant to actions, as they offer information about objects that can be acted upon (Gaver, 1991) and they 'suggest' specific actions or how an artefact is to be used as a standard response (Johnson, 1995). With respect to the media of paper and screens, Sellen and Harper (2001) investigated what the specific affordances of paper might be, taking an empirical approach by studying how people actually use those two media. In several ethnographic diary and laboratory studies, they found that office workers consistently prefer to use paper when reading material thoroughly, even when electronic (screen-based) alternatives are available.

There is a long tradition of research directly comparing reading on screen to reading on paper. Generally, results confirm an advantage of paper over screens (Ackerman & Goldsmith, 2011; Dillon, 1992; Noyes & Garland, 2008). Early studies focused on isolated performance indicators and found better results for paper compared to screens in terms of reading speed, accuracy and subjective preference (e.g. Heppner, Anderson, Farstrup, & Weiderman, 1985; Wilkinson & Robinshaw, 1987; for an overview see Dillon, 1992). Ziefle (1998) showed that some of these differences can be explained by the poor quality of screen displays at the time, which caused more strain on the eyes. She reported experimental evidence that performance in reading tasks can be improved by using higher screen resolution displays. However, O'Hara and Sellen (1997) argued that when analysing realistic reading tasks, paper offers specific advantages which are independent of technological advances of the respective hardware, as e.g. screen resolution. Based on laboratory observations, they identified several such advantages of paper with respect to reading. These were the effortless support of annotation without being detracted from

reading, the quick navigation through a document and the possibility of laying out pages in space to gain a sense of document structure. While manipulating paper can be considered a highly trained activity in our culture which usually happens automatically and without significant cognitive demand (Dillon, 1992), the same is not true for text on electronic devices (Tashman & Edwards, 2011). Consequently, some studies confirm higher cognitive demands for tasks on screen as compared to paper (Noyes, Garland, & Robbins, 2004). Wästlund, Reinikka, Norlander, and Archer (2005) found that a reading comprehension task resulted in lower performance and higher levels of experienced tiredness and stress when text was presented on screen as compared to paper. They explain these findings in terms of a cognitively demanding dual-task situation when reading on screen: in addition to the reading task itself, users are required to operate the computer, which reduces available cognitive resources.

Recent studies, comparing paper with current screen technology, still find advantages for paper in terms of reading performance, if compared to the standard office setup of vertical screens (Kerr & Symons, 2006; Mangen, Walgermo, & Brønnick, 2013; Morris, Brush, & Meyers, 2007). In contrast, preliminary evidence for new devices with horizontally positioned screens, such as computing surfaces, tablet computers and electronic paper displays, shows equal performance in reading tasks for screens as for paper (Chang, Chou, & Shieh, 2013; Morris et al., 2007). But even for these devices, studies find strong preferences by users to read text on paper instead of screens (e.g. Holzinger et al., 2011). Tashman and Edwards (2011) found in their dairy study that although computing devices were used more frequently than paper for reading tasks, users still preferred paper for reading. Reasons were problems with manipulating electronic text, primarily more difficult navigation and annotation than on paper. Overall, research suggests that 'paper is the medium of choice for reading' (Sellen & Harper, 2001; p.81).

In contrast, screens do not seem to elicit reading as a standard response. In a field study which logged on screen website browsing behaviour of their participants over a week, Hawkey and Inkpen (2005) found an average number of visited websites over a week of roughly 1800, with frequent 'bursts of browsing', during which several pages were loaded within a minute, occurring about 37 times a day. The authors concluded that the speed at which their participants were browsing was 'at times staggering' (p.1446), and the reported frequency of visited web pages and the speed of browsing seems to rule out extensive reading behaviour on the majority of sites visited. These results are consistent with the reported phenomenon, which was named 'the paradox of the active user' by Carroll and Rosson (1987). When people use interactive technology they often seem to have a motivational 'production bias', which causes them to exhibit a lot of activity and to avoid

reading of any kind (whether in available paper manuals or on screen). As a consequence, they are not taking into account available information that might help them be more efficient during their usage.

In sum, there is a large body of research that shows different effects of stimuli presentations on paper and on screen. As these results provide evidence for different affordances of the two media, the respective standard responses might also occur when participants are shown a prototype on paper or on screen in the context of a usability test. For these differences to emerge in a test, however, the test tasks and the requirement of reading text to perform well may prove crucial.

5.1.3. The present study

The goal of the present study was to investigate the potential effects of the medium of prototype presentation on the outcomes of a usability test. To allow for specific effects of the affordances of paper and screens to emerge, different task types were examined some of which would benefit by participants reading text. A usability test was conducted in which the medium of prototype presentation was experimentally manipulated. The prototype was presented in either of the two media (paper-based or on computer screen). As a second independent variable, task type was manipulated by giving three different tasks to participants (exploration task, reading task and navigation task). As a prototype, the website of an internet service was used, which allowed a realistic implementation of the different task types in a standard usability test setting.

As dependent measures performance, subjective evaluation and heart rate were recorded. Task performance was assessed by success rate and number of interactions. Self-report data was collected for emotion and perceived usability. Heart rate was measured during a resting phase before the experiment started and during task completion to allow for the analysis of heart rate variability.

Our hypotheses were as follows: (a) On a task which required more reading to perform well, participants using the prototype on paper would perform better than those working on the computer screen. For the navigation task, no differences would occur. The rationale behind this was that due to the different affordances of the two media, the medium of paper would provoke more reading activity than the presentation on screen. (b) In the exploration task, test participants would show less activity when using the prototype on paper as compared to the computer screen. (c) Due to expected increased reading activity and better comprehension on paper, participants would remember more

website content after the exploration task when the prototype was presented on paper as opposed to the screen.

5.2. Method

5.2.1. Participants

A total of 65 participants (53.8% female) took part in the experiment. They were recruited among the general public in Zurich, where the test laboratory was situated, and received a payment of CHF 25 (€ 20) for taking part. It was a strict selection criterion that participants were not familiar with the internet service which was used as the test system. Participants were aged between 17 and 47 years (M = 28.4 yrs; SD = 7.9).

5.2.2. Design

A 2 x 3 mixed design was used, with the medium of presentation as a between-participant variable and task type as a within-participants variable. Two conditions of presentation medium were implemented (paper, computer screen), and task type was manipulated at three levels (exploration, reading, navigation).

5.2.3. Measures and instruments

5.2.3.1. Performance

The following indicators of user performance were recorded: (a) task completion rate (percentage of successfully completed tasks); (b) number of user interactions (number of inputs to complete a task); (c) task completion time (s). Given the strong impact of facilitator behaviour on task completion time with a paper prototype, this indicator was only considered meaningful for the reading task, which could be completed on a single page without facilitator support in both prototype conditions.

5.2.3.2. Emotion

We used the PANAS scale ('Positive and Negative Affect Schedule'; Watson, Clark, & Tellegen, 1988) to measure positive and negative emotions. The German language version was shown to have good psychometric properties (Cronbach's $\alpha > 0.80$; Krohne, Egloff, Kohlmann, & Tausch, 1996). The 20 adjectives of the scale describe different affective states

(e.g. interested, exciting, strong), and their intensity is rated on a 5-point Likert scale ('very slightly or not at all', 'a little', 'moderately', 'quite a bit', 'extremely').

5.2.3.3. Perceived usability

Two instruments were used to measure perceived usability. First, we administered the PSSUQ ('Post Study System Usability Questionnaire'; Lewis, 1995), which was specifically developed for usage in laboratory-based usability tests. In previous research, it proved to be a valid instrument with very good psychometric properties (Cronbach's $\alpha > 0.90$; Lewis, 2002). It was translated into German and slightly modified to be relevant for the test system in question. To indicate that only the software and not the entire device (i.e. including hardware) was to be judged, the term 'system' was replaced by 'software' throughout. On a 7-point Likert scale (ranging from 'strongly agree' to 'strongly disagree') participants rated 19 items (e.g. 'The information was effective in helping me complete the tasks and scenarios'). As a second measure of perceived usability, participants rated a visual analogue scale (0-100; ranging from 'not at all' to 'very much') for an assessment of an overall estimation of perceived usability ('The software is usable'). Previous research showed that such single-item measures of usability can be valuable and reliable instruments (Christophersen & Konradt, 2011).

5.2.3.4. Perceived attractiveness

To measure perceived attractiveness, the 'User Experience Scale' developed by Lavie and Tractinsky (2004) was translated into German. Two subscales were used in the analysis: classic aesthetics (assessing aspects of clean, symmetric design) and expressive aesthetics (assessing creative, original aspects of design). The other subscales of the questionnaire (usability, pleasurable interaction, and service quality) were not included as they measure aspects of perceived usability that were already covered by the PSSUQ. On a 7-point Likert scale (ranging from 'strongly disagree' to 'strongly agree') participants rated 5 items in each subscale. Previous research proved the psychometric properties of the instrument to be good, with good internal consistency reported for both subscales: classic aesthetics ($\alpha > .80$) and expressive aesthetics ($\alpha > .80$; Lavie & Tractinsky, 2004).

5.2.3.5. Psychophysiology

As an indicator of participants' stress response, heart rate variability (HRV) was measured. The frequency bands for the analysis of HRV were determined in line with previous research and as recommended by the European Task Force of the Society of Cardiology and The North American Society of Pacing and Electrophysiology (1996) (high: 0.15-0.4 Hz; low: 0.04-0.15 Hz; very low: 0.003-0.04 Hz). Consistent with these recommendations, we only analysed HRV recordings of at least 5 min in length, taking the first 5 min of task completion. To control for outliers, 3 participants with LF band values lying outside a range of ± 2 *SD* from the mean were excluded from the analysis. The analysis concentrates on the low frequency (LF) band, which previous research has shown to be an indicator of mental effort and stress response in situations of potential social stress (Pruyn, Aasman, & Wyers, 1985; Sonderegger & Sauer, 2009; Uebelbacher, Sonderegger, & Sauer, 2013). For the analysis, we used the Kubios™ HRV 2.0 software (Tarvainen & Niskanen, 2008) for Windows XP™. Possible artefacts were corrected with the artefact correction level 'medium' and the default Fast Fourier transformation was used for time interval calculations.

5.2.3.6. Previous internet experience

On a visual analogue scale (ranging from 0 to 100, labelled 'not experienced' and 'very experienced'), participants reported a rather high level of self-rated internet experience (*M* = 69.1; *SD* = 17.7). They indicated using the internet 13.8 times on average during a day (*SD* = 22.3). These two indicators were used as covariates in the analysis.

5.2.4. Tasks

Test participants had to accomplish the following three tasks, which they were presented with as written instructions: (a) *Exploration task:* Participants were asked to learn about the service offered on a website, and to decide which of the subscriptions would be interesting to them (see Figure 3). Participants were not given any time limit for the task, it was left for them to decide when they had accessed enough information. The number of visited pages was recorded as an indicator of explorative activity. When participants stopped the task, the display of the website was blocked from their view. Then they were asked to respond to 10 open-ended questions about the web service to assess how well they had processed the information on display (e.g. 'How much data can you store with a free account?'). Participants had not been informed previously about these questions.

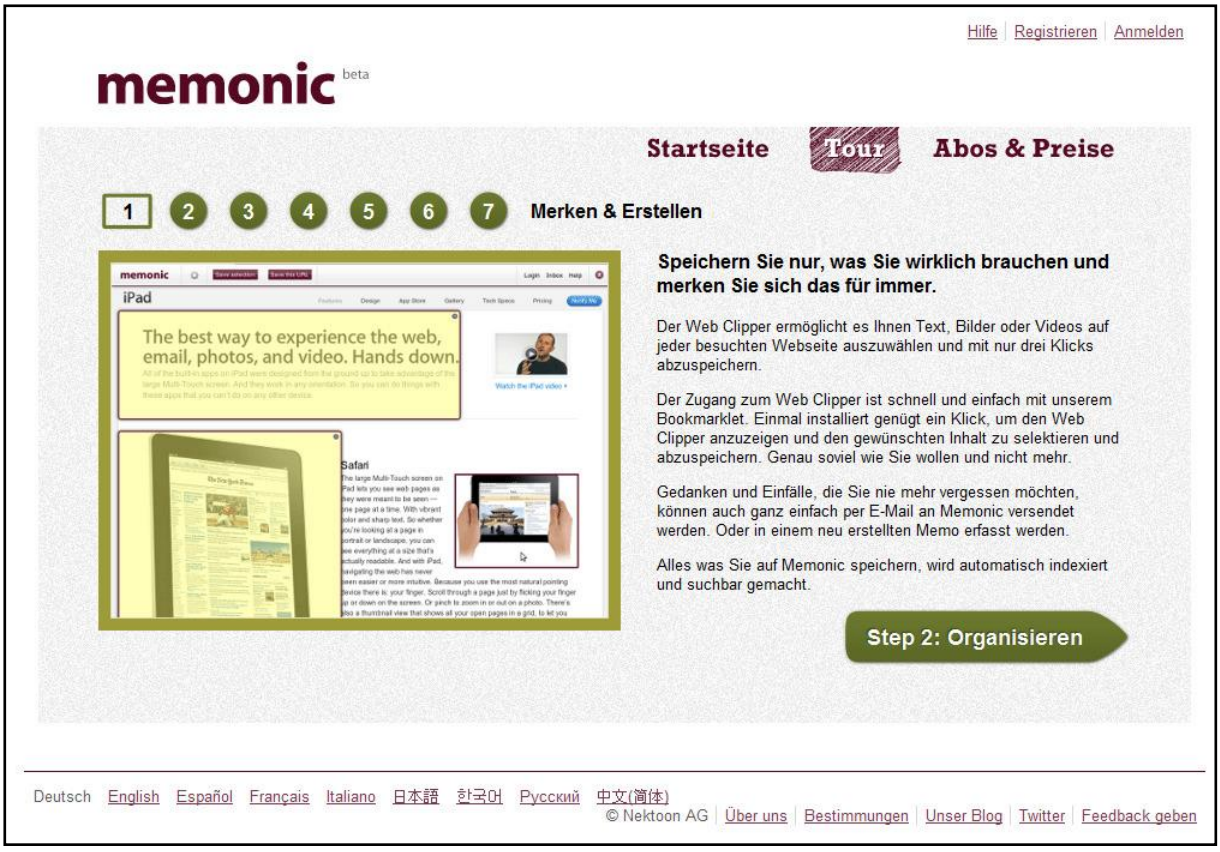

Figure 3. Information page of the prototype, giving details about the web service.

(b) *Reading task:* Participants were instructed to sign up for a free account with the web service. This was possible on a single registration page, and included specification of a personal password for the account. The password requirements were manipulated for all participants: a short paragraph of three lines of text was introduced just above the password form field (see Figure 4). The text informed that the password was required to consist of at least five characters and had to start and finish with a number. The task was successfully completed if participants selected the password accordingly, indicating whether they had read the paragraph.

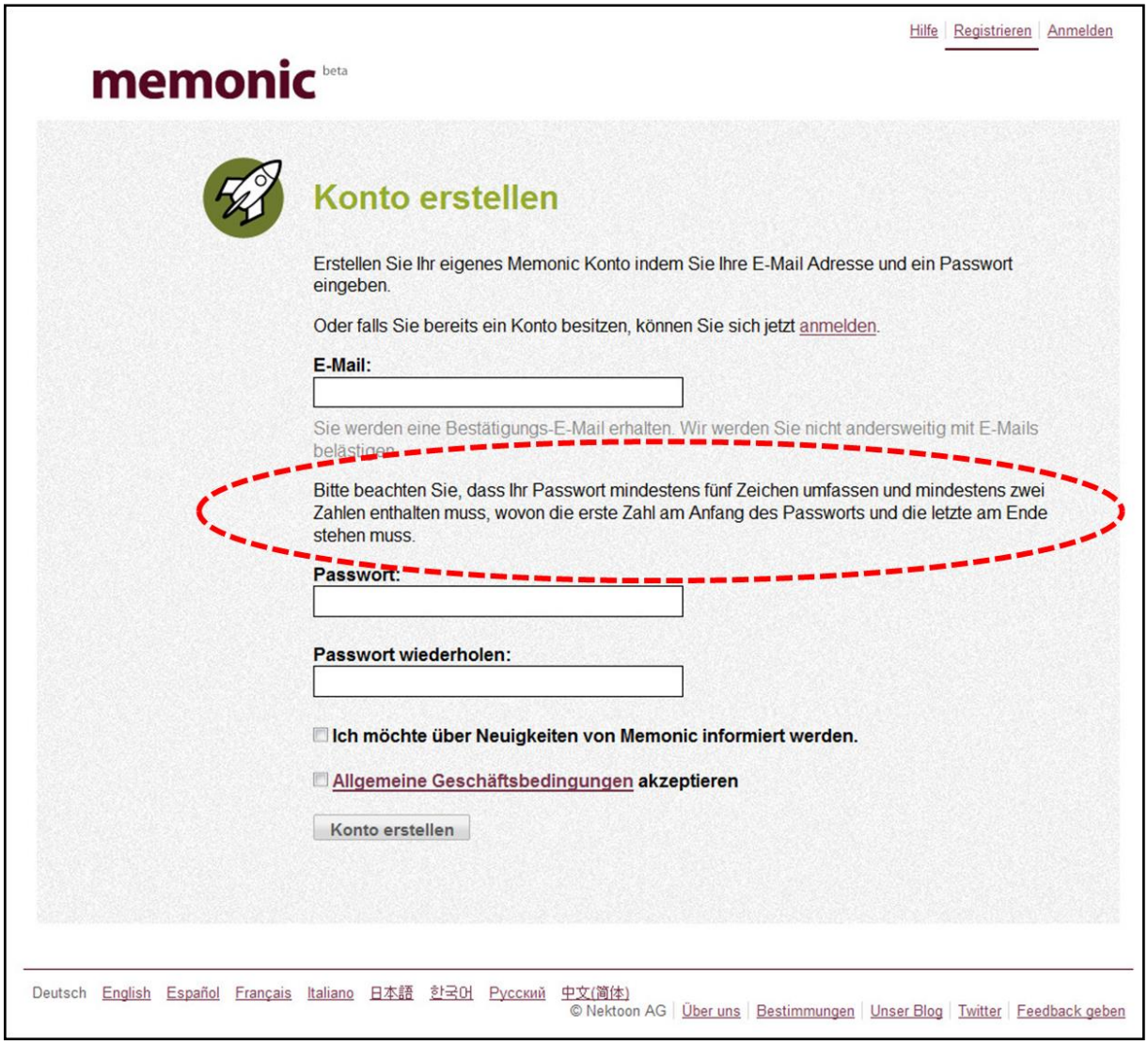

Figure 4. Registration page of the prototype system for the reading task, with specific instructions indicating password requirements.

(c) *Navigation tasks:* First, participants were instructed to create a group of web notes called 'news' and assign a (previously created) web note to this group. Then, a second group of notes had to be created, and another previously created web note had to be assigned to two groups at once. Finally, a group of web notes had to be deleted. All of these tasks could be solved by clicking on icons and web links with short labels that did not require extensive reading of text (see Figure 5).

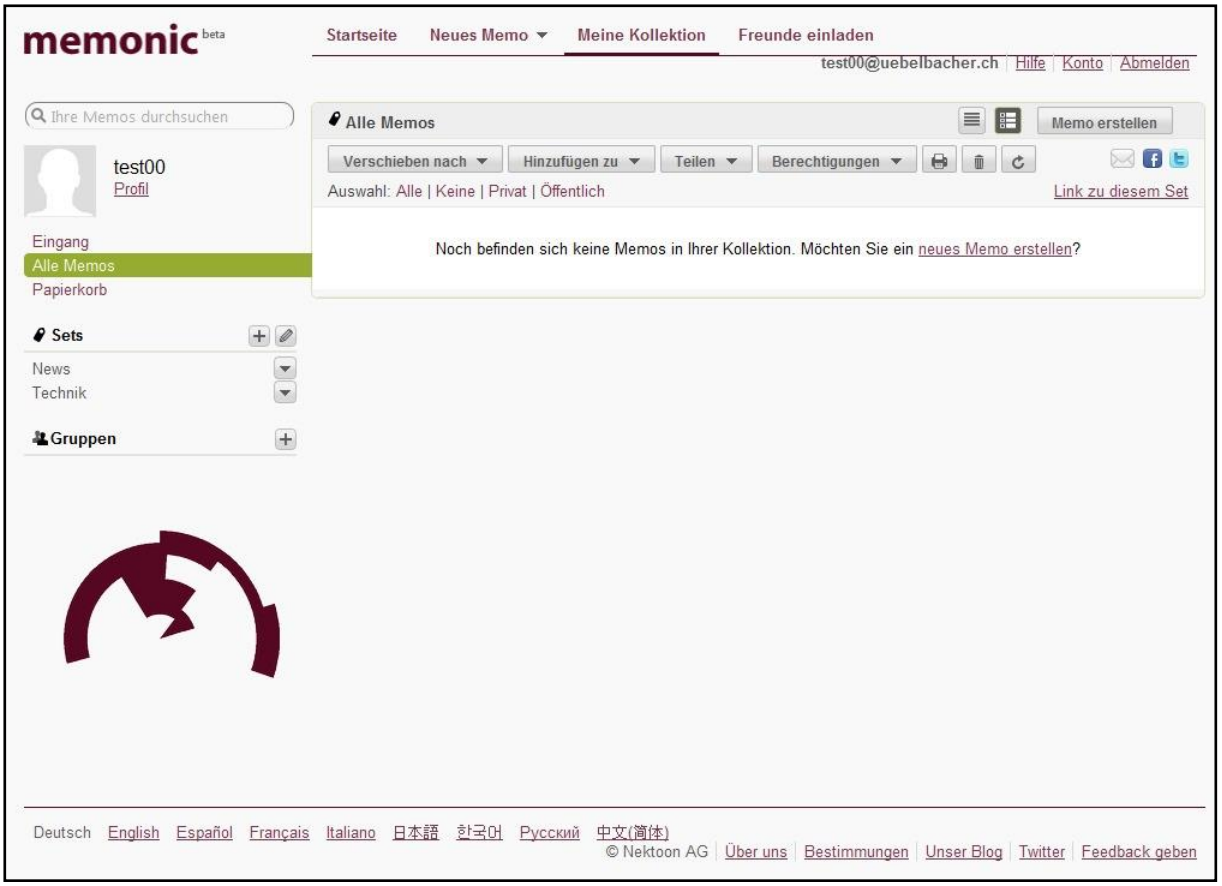

Figure 5. Prototype page used in the navigation task with little requirement of extensive reading of text.

All tasks represent typical tasks of internet users. Task sequence was not varied between test participants, as it represented a meaningful order, which could not be changed without sacrificing the practical meaningfulness of task goals. While it is acknowledged that there is some degree of overlap between tasks (e.g. navigation tasks contain elements of reading, reading task contains elements of web navigation), there are clear differences between tasks with regard to the activity that is crucial for task success. While for navigation tasks it is critical to scan pages for specific links, identifying them as relevant for goal achievement and selecting them, this is not the case for the reading task. While for the reading task, it is crucial to read and understand a paragraph of text, for the navigation tasks this is not the case as on the respective pages there are none. While for the exploration task it is crucial to move across several pages, this is not required for the other two task types.

5.2.5. Materials

5.2.5.1. Test devices and data recording hardware

For the paper prototype, the web pages were colour printed on A3 sheets and were presented within a frame with browser controls (back button, etc.) allowing for realistic web surfing manipulation. The printout and frame sizes made sure that the visible section was equivalent to the displayed part of a page on the computer screen. The computer was a HP Compaq™ DC7800 MiniTower. The desktop screen was a 17-inch Philips™ 170S_2 LCD monitor, set to a 768-by-1024 pixel resolution, and the browser was operated with a standard computer mouse and keyboard. The screen was set up on the desk in front of participants, at a normal usage distance of about 0.5 m. A video camera (Sony™ Handycam HDR-HC9E) was positioned directly above the desk on which the paper prototype was presented, outside the participants' field of vision, recording all prototype manipulations at a 90 degrees vertical angle. For heart rate measurement a Polar™ RS800 heart rate monitor was used, which participants were wearing for the full test session.

5.2.5.2. Software

The desktop computer was running on Windows XP™, and Techsmith Morae™ (v3.1.1) was used to record all prototype manipulations directly on screen.

As a test prototype, a copy of the www.memonic.com website (v1.1.2) was used, with the only adjustment of adding the short paragraph of text informing about password requirements on the registration page. Memonic is a web-based note-taking service that allows to copy parts of or whole web pages into an account for later reference (Figures 3, 4, and 5). The website describes its functionalities on information pages (e.g. Figure 3), while there is considerably less text presented on the pages that allow the management of web-notes (Figure 5). The website was accessed using the Safari™ browser (v5.0.1), of which the browser controls were printed on the frame in which the paper prototype was presented.

5.2.6. Procedure

The test sessions took place in a laboratory of a usability consultancy in Zurich, Switzerland. Participants were randomly assigned to the two conditions of prototype presentation. The experimenter (test facilitator) welcomed the participants and explained that the purpose of the test was to assess the usability of a website. After a short synopsis of the procedure of the experiment, the test participants filled in a short questionnaire asking for biographic

information and internet usage behaviour. Then, the PANAS was administered to assess the emotional baseline. Afterwards, the heart rate measurement was introduced and the participants were guided to a rest room where they could put on the heart rate monitoring device. When they came back, the Polar™ RS800 watch was attached to the participant's wrist and the transmission was checked to work properly. Then the participants were asked to relax and not to move so that the heart rate baseline could be measured for the following 10 min, during which the experimenter left the room.

Then the participants were led to the test laboratory and were seated at a desk with the respective prototype in front of them. In both prototype presentation conditions, a test page of a well-known Swiss newspaper website (www.nzz.ch) was presented to explain how participants could operate the prototype (e.g. point with their fingers on the paper prototype to click, etc.). The tasks were described, and participants were invited to keep working on them until they would be prompted to pass on to the next task, except for the exploration task. Participants were allowed to work on each task for a maximum of 5 min, after which they were asked to move on to the next task.

The participants were informed that no help could be given by the facilitator. To avoid interference with the heart rate variability parameters, participants were instructed not to talk during the psychophysiological measurement, but were given the opportunity to ask questions before the testing began.

Then participants started to work on the tasks, which were presented in written form, and the screen capturing and video recording were started. Participants' interaction with the desktop computer was registered by direct screen recording, manipulation of the paper prototype was filmed with the video camera mounted vertically above the prototype, but outside the participants' field of vision. Task presentation was held constant across all participants, to create a realistic sequence of a usability test. Participants worked first on the exploration task, to inform themselves about the web service, and they could decide when they would move on to the next task. Then the prototype was blocked from the participants' view (screen switched off or paper prototype put aside), and the questions probing the processing of the content of the visited web pages were presented on paper. Participants had not been informed about these questions previously. Then, the reading task (registering for a test account of the web service), a non-recorded task to create a web note with the service, and the navigation task working with the created web note were administered. After task completion, the experimenter stopped the video recording and the post-test instruments were administered (PANAS, single-item scale for usability, PSSUQ and the User Experience Scale). All participants were then debriefed and paid.

5.2.7. Data analysis

For performance measures, one-factorial ANCOVAs were computed, examining user performance for each task separately to test the effects of prototype presentation medium. A separate analysis for each task was necessary since performance measures were not comparable across tasks. For each subjective measure, a two-factorial ANCOVA was used.

For all ANCOVAs, self-reported internet experience, daily internet usage and gender were entered as covariates into the analysis in order to control for any influence. The analysis showed that none of the covariates had a significant influence on the reported findings.

As there was only one reading task which could either be completed successfully or not (depending on correct password definition), a non-parametric χ^2 test was used to analyse the binary distribution of the results.

5.3. Results

5.3.1. User behaviour and performance

5.3.1.1. Reading task

Task completion. Task completion rates for the different task types are presented in Table 2. For the task where reading a small paragraph of text on the web page was critical for success (i.e. correct password definition), the analysis confirmed the expected effect of prototype presentation on successful task completion (χ^2 (1) = 8.31, $p < 0.01$). While only 19.5% of participants using the computer prototype on screen selected a correct password, 54.2% did when using the paper-based prototype.

Task completion time. The analysis showed a significant effect of presentation medium on task completion time ($F(1,63) = 4.98$, $p < 0.05$) (see Table 2). Participants using the paper-based prototype spent significantly more time working on the reading task (M = 117.5; SD = 63.6) than those using the computer prototype on screen (M = 84.7s; SD = 53.1).

Table 2. User behaviour and performance as a function of prototype presentation.

	Paper-based prototype Mean (*SD*)	**Desktop computer** Mean (*SD*)	**Overall** Mean (*SD*)
Reading task			
Task completion rate (%)	54.2% (0.51)	19.5% (0.40)	32.3% (0.47)
Task completion time (s)	117.5 (63.6)	84.7 (53.1)	96.82 (58.9)
Navigation tasks			
Task completion rate (%)	49.2% (35.9)	64.2% (25.2)	59.1% (29.8)
Number of user interactions	5.00 (2.06)	6.29 (1.84)	5.89 (1.98)
Exploration task			
Number of user interactions	4.13 (3.08)	8.76 (4.07)	7.05 (4.34)
Content retention (%)	28.3% (13.0)	33.7% (11.3)	31.7% (12.1)

5.3.1.2. Navigation tasks

Task completion rate. Overall success rate across both prototype conditions was 59.1% for the three navigation tasks (see Table 2). Our analysis showed no significant association between the medium of prototype presentation and completion rate of the three navigation tasks ($F(1,60) = 3.68$, *ns*).

Number of user interactions. There was a significant effect of prototype medium on the number of interactions participants needed to accomplish the navigation tasks ($F(1,56) = 5.69$, $p < 0.05$) (see Table 2). On average, participants needed significantly more interactions per navigation task when working with the computer prototype ($M = 6.29$; $SD = 1.84$) than when working with the paper prototype ($M = 5.00$; $SD = 2.06$).

5.3.1.3. Exploration task

Number of user interactions. As an indicator for participant behaviour during the exploration task the number of user inputs was compared between prototype conditions (see Table 2). This equals the number of pages visited, since all options led to a new page. As expected, the ANCOVA revealed a significant effect of prototype presentation on the number of pages visited ($F(1,63) = 23.24$, $p < 0.001$). Participants using the paper-based prototype viewed significantly fewer pages ($M = 4.13$; $SD = 3.08$) than those using the computer prototype on screen ($M = 8.76$; $SD = 4.07$).

Content retention. After completing exploration, participants answered ten questions about the content of the pages. The number of correctly answered questions was analysed as an indicator of how well the information on the website was processed (see Table 2). The one-way ANCOVA revealed no effect of prototype presentation on content retention ($F(1,61) = 3.10$, *ns*), and it made no difference whether the number of pages visited during exploration was controlled for. Participants across both conditions were able to answer 31.7% of the questions correctly ($SD = 12.1$).

5.3.2. Subjective ratings

5.3.2.1. Perceived usability

Our indicator for perceived usability was the PSSUQ, of which the data are presented in Table 3. The analysis showed no significant difference in usability ratings between the paper-based and the computer prototype conditions ($F < 1$). A separate analysis of the PSSUQ subscales yielded the same results, therefore these are not reported in detail.

5.3.2.2. Perceived attractiveness

Data of the User Experience Scale assessing perceived attractiveness are presented in Table 3. There were no differences between paper-based and computer prototype conditions on either the subscale classic aesthetics ($F(1,63) = 1.86$, *ns*), nor for expressive aesthetics ($F < 1$).

5.3.2.3. Emotion

To evaluate the change of emotional state of participants as a consequence of using either the paper-based or the computer prototype, our analysis compared the baseline measurement before task completion with the second measurement after prototype usage (see Table 3). The analysis showed a significant increase of positive affect from pre- ($M = 2.90$; $SD = 0.55$) to post-task measurement ($M = 3.08$; $SD = 0.62$; $F(1,62) = 4.75$, $p < 0.05$), but no change for negative affect ($F < 1$). Prototype presentation condition had no impact on either positive affect ($F(1,63) = 1.13$, *ns*) or on negative affect ($F < 1$).

Table 3. Subjective measures as a function of prototype presentation.

	Paper-based prototype Mean (*SD*)	**Computer prototype** Mean (*SD*)	**Overall** Mean (*SD*)
Perceived usability (1-7)	5.11 (1.05)	4.99 (1.01)	5.03 (1.02)
Perceived classic aesthetics (1-7)	4.73 (1.10)	4.36 (1.04)	4.50 (1.07)
Perceived expressive aesthetics (1-7)	3.48 (1.05)	3.68 (1.21)	3.60 (1.15)
Positive affect (Δ: pre - post)	0.08 (0.47)	0.23 (0.59)	0.17 (0.55)
Negative affect (Δ: pre - post)	0.03 (0.32)	-0.03 (0.30)	-0.01 (0.30)

Δ: values represent changes from pre- to post-task measurement, on PANAS scale (1-5)

5.3.3. Physiological measures

5.3.3.1. Heart rate

Physiological data is presented in Table 4. To assess effects of task completion on participants' heart rate we compared heart rate changes between resting phase and task completion phase. There was an increase in heart rate from the resting phase (M = 75.36; SD = 10.44) to task completion phase (M = 77.47; SD = 11.35), which was statistically significant ($F(1,36) = 5.26$, $p < 0.05$). This increase, however, did not differ across prototype conditions ($F < 1$).

Table 4. Changes in physiological parameters from baseline to task completion phase, as a function of prototype presentation and task type.

	Paper-based prototype Mean (*SD*)	**Computer prototype** Mean (*SD*)	**Overall** Mean (*SD*)
Δ Heart rate (bpm[a])	+1.23 (5.41)	+3.00 (5.94)	+2.12 (5.68)
Δ LF[b] power (ms^2)	-680.8 (1173.9)	-94.6 (740.1)	-379.3 (1005.3)

Δ: all values represent changes from baseline (resting phase) to task completion phase, with positive values denoting an increase in the parameter.
[a] bpm: beats per minute
[b] LF: low frequency band

5.3.3.2. Heart rate variability

As an indicator of participants' stress response and mental effort, we analysed HRV in the LF band (0.1 Hz component), comparing baseline levels as measured during an initial resting phase with measurements during task completion phase. The analysis confirmed an expected decrease during task completion phase for the power in the LF band, indicating higher levels of mental effort ($F(1,33) = 5.53$, $p < 0.05$). For prototype presentation, there was no significant effect on LF band levels ($F(1,33) = 3.16$, *ns*).

5.4. Discussion

The aim of the study was to investigate effects of usability test prototypes presented on paper and on screen, but with a more theory-driven approach than it was done in previous research. Specifically, based on an analysis of the affordances of the media of paper and screens, it was tested whether tasks requiring reading text would benefit by being presented on paper to participants. In support of our main hypothesis, participants performed better on paper than on screen if reading was crucial for task success. In practice, this could lead to a serious underestimation of usability problems on web pages, if paper prototypes were used in testing. For other task types, no such performance differences for the media of prototype presentation were found.

The major outcome of the study lies in the relevance of task requirements for the investigation of prototype presentation effects in usability testing. As long as a standard navigation or exploration task was presented to participants either on paper or on screen, no performance differences were identified between the presentation media. This finding is in agreement with previous research which consistently failed to find such differences (Sauer et al., 2010; Virzi et al., 1996; Walker et al., 2002). However, none of these studies selected the test tasks according to a hypothesis about a specific impact of the medium of prototype presentation. Very different results emerge, if tasks are selected which require reading of a short paragraph of text for task success. The present study found significantly better performance if participants were using a paper prototype in this usage situation. With the paper-based prototype, about 46% of the participants ignored password requirements in a short paragraph of text when defining their password on the same page. In contrast, with the computer prototype presented on screen, more than 80% of test users ignored the three lines of text and, consequently, defined passwords of an incorrect format. Participants obviously paid more attention to text on a web page if the same prototype was presented on paper as compared to screen.

There are several possible explanations for this finding. First, as hypothesised, paper might exhibit affordances that make reading or generally paying attention to content of prototype interfaces more likely. In support of this assumption, studies consistently find a preference of users to read on paper as compared to screens (Annand, 2008; Tashman & Edwards, 2011; Woody, Daniel, & Baker, 2010). In addition, paper can be considered an almost perfect display with respect to visual ergonomics requirements (i.e. it provides high contrast, but no screen glare, reflections or flicker) (Holzinger et al., 2011), which makes reading from paper usually visually less strenuous than from screens. These differences between paper and screens may equally apply to the presentation of prototypes in a usability test situation, even if previous studies have not used test tasks suited to find these effects (e.g. Virzi et al., 1996; Walker et al., 2002). Second, given the interactivity that users expect of an implemented web page, participants using the computer prototype on screen might have relied on error messages, which would inform them about mistakes when filling in the form. As error messages are a standard website functionality, this may cause participants to select a trial-and-error approach, ignoring potentially unnecessary information at first, thereby effectively reducing attention resources. Bargas-Avila, Oberholzer, Schmutz, de Vito, and Opwis (2007) found very similar behaviour of their participants even ignoring specific error messages for form field entries. The authors explained this finding by a 'completion mode', in which users are when filling in a form, which might cause them to focus on completing all form fields, to ignore information on screen and to tolerate potential mistakes. Third, due to the social situation in a usability test with facilitators being present, participants in the paper prototyping condition might be more wary not to cause them inconvenience or unnecessary effort, e.g. by the need to display additional pages with error messages. This may prompt participants to pay more attention to the stimulus material, not to miss relevant instructions. In addition, this would explain the finding that participants in the paper prototype condition were visiting much fewer pages, which is consistent with previous research (Hamborg et al., 2009; Sefelin et al., 2003). However, as this would equally affect task types that were tested in previous studies, it cannot explain the presentation media effect of the present study.

The practical implications of the different results for paper-based and computer prototypes are serious. Using a paper prototype in usability testing may lead to an overestimation of the accuracy with which content of a web page is read. This may result in a lower detection rate for specific types of usability problems, if tests are conducted with a paper prototype. As a consequence, relevant interface problems may not be found, especially if tests include only few participants, as is often the case in practice. This means computer prototypes would be better suited for testing, and the typical recommendation to

base the decision about the prototype medium purely on practical considerations (e.g. Rudd et al., 1996) needs to be revised. Test results may be overestimating users' performance and miss specific usability problems when the prototype is presented on paper, and if the tasks require considerable reading. This needs to be taken into account by usability professionals who decide on prototype design.

Interestingly, participants using the paper prototype did not only inspect the page more thoroughly and read more content, they also needed significantly more time when manipulating the paper prototype in the reading task as compared to usage on screen. This means that participants using the paper prototype chose a strategy that prioritised accuracy over speed, while those working on screen prioritised speed of task completion over accuracy (Proctor & Vu, 2009). However, physiological parameters did not show different levels of mental effort participants experienced when using the paper-based as compared to the computer prototype. There was no significant difference between these conditions in HRV suppression for the LF band. This indicator did show a significant difference from resting phase to task completion, confirming it represents a promising indicator of mental effort in usability testing (Izsó & Láng, 2000; Rowe, Sibert, & Irwin, 1998; Sonderegger & Sauer, 2009).

Contrary to expectations, content retention following the exploration task did not support our hypothesis of a generally better information processing on paper than on screen. There were no differences for prototype medium, even when controlling for the lower number of pages participants in the paper condition were exposed to, due to their more limited activity compared to those working on screen. Various explanations exist for this finding. First, with the content retention task, various other factors come into play besides the attention that participants pay to the information on display. Given that there was a slight delay between the exposure to the stimulus and the questioning, various memory processes were involved, and the outcome could be affected by recall skills (Dillon, 1992). In comparison, memory processes were much less of an issue for the reading task, during which the relevant information was still on display on the page when defining the password according to requirements. Second, the questions addressing content retention might have been too difficult to detect potential differences, as the success rate was generally quite low.

The findings with respect to subjective measures were consistent with previous studies, some of which compared prototypes of more different fidelity levels (e.g. Sauer et al., 2010). As the paper prototype pages in our study consisted of screen printouts with no reduction of visual fidelity, no difference could be expected. This confirms the general

assumption that with paper prototypes valid and representative results can be achieved, at least with regard to subjective indicators of system evaluation.

The findings presented have several implications for research and practice. First, it becomes clear from these results that, contrary to previous recommendations, the medium of prototype presentation does make a difference and needs to be considered when planning user testing. If reading text on a system interface is crucial for task performance, paper prototyping cannot be recommended, since it might lead to an overestimation of system usability and the failure to detect relevant interface problems. Second, if free navigation through a number of information pages of a system (e.g. a website) needs to be tested or logged during a test, the usage of paper-based prototypes should be avoided, as they seem to inhibit user activity due to social constraints. Third, users generally seem to pay little attention to instructions on web pages, even if paragraphs of texts are quite short, which is consistent with other studies (Callahan & Koenemann, 2000). The presented findings should be a strong reminder to take this into consideration when designing web interfaces.

With respect to future research, several fields of interest can be identified. First, the question remains, whether the presented prototype effects would equally occur for the comparison of paper-based and prototypes which are presented on tablet computers or electronic paper systems. As some studies presented very similar preference ratings for this recent technology as for paper with respect to reading (Morris et al., 2007), the respective prototype effect may well disappear. Second, different task designs would need to be studied, to investigate shown effects of on screen reading requirements for other usage situations than online form completion. Finally, and on a more methodological level, the results call for a more theory-driven approach when investigating potential influences in usability testing.

Acknowledgements

The authors are very grateful to the Swiss National Science Foundation for their financial support of the study (research grant No. 100014/122490). Thanks are also due to Dr Daniel Felix (ergonomie & technologie GmbH), Felix Hürlimann (Nektoon), and to Caroline Biewer for their support in completing this study.

References

Ackerman, R., & Goldsmith, M. (2011). Metacognitive regulation of text learning: on screen versus on paper. *Journal of Experimental Psychology: Applied, 17*(1), 18–32.

Annand, D. (2008). Learning efficacy and cost-effectiveness of print versus e-book instructional material in an introductory financial accounting course. *Journal of Interactive Online Learning, 7*(2), 152–164.

Bargas-Avila, J. A., Oberholzer, G., Schmutz, P., de Vito, M., & Opwis, K. (2007). Usable error message presentation in the World Wide Web: do not show errors right away. *Interacting with Computers, 19*(3), 330–341.

Bias, R. G., & Mayhew, D. J. (2005). *Cost-justifying usability: an update for the internet age* (2nd ed.). San Francisco, CA: Morgan Kaufman.

Callahan, E., & Koenemann, J. (2000). A comparative usability evaluation of user interfaces for online product catalog. In *Proceedings of the 2nd ACM Conference on Electronic Commerce* (pp. 197–206). New York: ACM.

Carroll, J. M., & Rosson, M. B. (1987). Paradox of the active user. In J. M. Carroll (Ed.), *Interfacing thought: cognitive aspects of human-computer interaction* (pp. 80–111). Cambridge, MA: MIT Press.

Catani, M. B., & Biers, D. W. (1998). Usability evaluation and prototype fidelity: users and usability professionals. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (pp. 1331–1335). Human Factors and Ergonomics Society.

Chang, P.-C., Chou, S.-Y., & Shieh, K.-K. (2013). Reading performance and visual fatigue when using electronic paper displays in long-duration reading tasks under various lighting conditions. *Displays, 34*(3), 208–214.

Christophersen, T., & Konradt, U. (2011). Reliability, validity, and sensitivity of a single-item measure of online store usability. *International Journal of Human-Computer Studies, 69*(4), 269-280.

Dillon, A. (1992). Reading from paper versus screens: a critical review of the empirical literature. *Ergonomics, 35*(10), 1297–1326.

European Task Force of the Society of Cardiology and The North American Society of Pacing and Electrophysiology, T. (1996). Heart rate variability. *European Heart Journal, 17*, 354–381.

Gaver, W. W. (1991). Technology affordances. In S. P. Robertson, G. M. Olson, & J. S. Olson (Eds.), In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems – CHI '91* (pp. 79–84). New York: ACM.

Hall, R. R. (2001). Prototyping for usability of new technology. *International Journal of Human-Computer Studies, 55*(4), 485–501.

Hamborg, K.-C., Klassen, A., & Volger, M. (2009). Zur Gestaltung und Effektivität von Prototypen im Usability-Engineering. In H. Wandke, S. Kain, & D. Struve (Eds.), *Mensch & Computer 2009: Grenzenlos frei!?* (pp. 263–272). München: Oldenbourg Verlag.

Hawkey, K., & Inkpen, K. (2005). Web browsing today: the impact of changing contexts on user activity. In *CHI '05 Extended Abstracts on Human Factors in Computing Systems* (pp. 1443–1446). New York: ACM.

Heppner, F. H., Anderson, J. G. T., Farstrup, A. E., & Weiderman, N. H. (1985). Reading performance on a standardized test is better from print than from computer display. *Journal of Reading, 28*(4), 321–325.

Holzinger, A., Baernthaler, M., Pammer, W., Katz, H., Bjelic-Radisic, V., & Ziefle, M. (2011). Investigating paper vs. screen in real-life hospital workflows: performance contradicts

perceived superiority of paper in the user experience. *International Journal of Human-Computer Studies, 69*(9), 563–570.

Hosseini-Khayat, A., Hellmann, T. D., & Maurer, F. (2010). Distributed and automated usability testing of low-fidelity prototypes. In *Proceedings of the 2010 Agile Conference* (pp. 59–66). IEEE Computer Society.

Izsó, L., & Láng, E. (2000). Heart period variability as mental effort monitor in human computer interaction. *Behaviour & Information Technology, 19*(4), 297–306.

Johannsson, H., & Hvannberg, E. T. (2004). Integration of air traffic control user interfaces. In *Proceedings of the 23rd Digital Avionics Systems Conference, Vol. 1* (pp. 1–11). Piscataway, NJ: IEEE Computer Society.

Johnson, J. A. (1995). A comparison of user interfaces for panning on a touch-controlled display. In I. R. Katz, R. Mack, L. Marks, M. B. Rosson, & J. Nielsen (Eds.), *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems – CHI '95* (pp. 218–225). New York: ACM.

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(6), 345–355.

Kerr, M. A., & Symons, S. E. (2006). Computerized presentation of text: effects on children's reading of informational material. *Reading and Writing, 19*(1), 1–19.

Kieffer, S., Lawson, J.-Y. L., & Macq, B. (2009). User-centered design and fast prototyping of an ambient assisted living system for elderly people. In *Sixth International Conference on Information Technology: New Generations,* 1220–1225.

Krohne, H. W., Egloff, B., Kohlmann, C.-W., & Tausch, A. (1996). Untersuchungen mit einer deutschen Version der "Positive and Negative Affect Schedule" (PANAS). *Diagnostica, 42*(2), 139–156.

Lavie, T., & Tractinsky, N. (2004). Assessing dimensions of perceived visual aesthetics of web sites. *International Journal of Human-Computer Studies, 60*(3), 269–298.

Lewis, J. R. (1995). IBM computer usability satisfaction questionnaires: psychometric evaluation and instructions for use. *International Journal of Human-Computer Interaction, 7*(1), 57–78.

Lewis, J. R. (2002). Psychometric evaluation of the PSSUQ using data from five years of usability studies. *International Journal of Human-Computer Interaction, 14*(3), 463–488.

Lewis, J. R. (2006). Usability testing. In G. Salvendy (Ed.), *Handbook of human factors and ergonomics* (pp. 1275–1316). Hoboken, NJ: Wiley.

Lim, Y.-K., Pangam, A., Periyasami, S., & Aneja, S. (2006). Comparative analysis of high- and low-fidelity prototypes for more valid usability evaluations of mobile devices. In A. Mørch, K. Morgan, T. Bratteteig, G. Ghosh, & D. Svanaes (Eds.), *Proceedings of the 4th Nordic Conference on Human-Computer Interaction: Changing Roles (NordiCHI '06)* (pp. 291–300). New York: ACM.

Mangen, A., Walgermo, B. R., & Brønnick, K. (2013). Reading linear texts on paper versus computer screen: Effects on reading comprehension. *International Journal of Educational Research, 58*, 61–68.

McCurdy, M., Connors, C., Pyrzak, G., Kanefsky, B., & Vera, A. (2006). Breaking the fidelity barrier. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems – CHI '96* (pp. 1233–1242). New York: ACM.

Meszaros, G., & Aston, J. (2006). Adding Usability Testing to an Agile Project. In *Proceedings of AGILE '06 Conference* (pp. 289–294). IEEE Computer Society.

Morris, M. R., Brush, A. J. B., & Meyers, B. R. (2007). Reading revisited: evaluating the usability of digital display surfaces for active reading tasks. In *Second Annual IEEE*

International Workshop on Horizontal Interactive Human-Computer Systems – TABLETOP'07 (pp. 79–86). IEEE.

Norman, D. A. (1988). *The design of everyday things.* New York: Basic Books.

Noyes, J. M., & Garland, K. J. (2008). Computer- vs. paper-based tasks: Are they equivalent? Ergonomics, 51(9), 1352–1375.

Noyes, J. M., Garland, K., & Robbins, L. (2004). Paper-based versus computer-based assessment: is workload another test mode effect? *British Journal of Educational Technology, 35*(1), 111–113.

O'Hara, K., & Sellen, A. (1997). A comparison of reading paper and on-line documents. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems – CHI '97* (pp. 335–342). New York: ACM.

Olmsted-Hawala, E. L., Romano, J. C., & Murphy, E. D. (2009). The use of paper-prototyping in a low-fidelity usability study. In *IEEE International Professional Communication Conference* (pp. 1–11).

Petrie, J., & Schneider, K. A. (2006). Mixed-fidelity prototyping of user interfaces. In G. Doherty & A. Blandford (Eds.), *Proceedings of the 13th International Conference on Interactive Systems: Design, Specification, and Verification – DSVIS '06* (pp. 199–212). Berlin, Heidelberg: Springer.

Proctor, R. W., & Vu, K.-P. L. (2009). Human information processing: an overview for human-computer interaction. In A. Sears & J. A. Jacko (Eds.), *Human-computer interaction. Fundamentals* (pp. 19–38). Boca Raton, FL: CRC Press.

Pruyn, A., Aasman, J., & Wyers, B. (1985). Social influences on mental processes and cardiovascular activity. In J. F. Orlebeke, G. Mulder, & L. J. P. Van Dornen (Eds.), *The psychophysiology of cardiovascular control (models, methods, and data).* (pp. 865–877). New York: Plenum Press.

Rowe, D. W., Sibert, J., & Irwin, D. (1998). Heart rate variability: indicator of user state as an aid to human-computer interaction. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems – CHI '98* (pp. 480–487). New York: ACM.

Rudd, J., Stern, K., & Isensee, S. (1996). Low vs. high-fidelity prototyping debate. *Interactions, 3*(1), 76–85.

Sauer, J., Franke, H., & Ruettinger, B. (2008). Designing interactive consumer products: utility of paper prototypes and effectiveness of enhanced control labelling. *Applied Ergonomics, 39*(1), 71–85.

Sauer, J., Seibel, K., & Rüttinger, B. (2010). The influence of user expertise and prototype fidelity in usability tests. *Applied Ergonomics, 41*(1), 130–140.

Sauer, J., & Sonderegger, A. (2009). The influence of prototype fidelity and aesthetics of design in usability tests: effects on user behaviour, subjective evaluation and emotion. *Applied Ergonomics, 40*(4), 670–677.

Sefelin, R., Tscheligi, M., & Giller, V. (2003). Paper prototyping - What is it good for? A comparison of paper- and computer-based low-fidelity prototyping. In *CHI '03 Extended Abstracts on Human Factors in Computing Systems* (pp. 778–779). New York: ACM.

Sellen, A. J., & Harper, R. H. R. (2001). *The myth of the paperless office* (1st ed.). Cambridge, MA: MIT Press.

Snyder, C. (2003). Paper prototyping: the fast and easy way to design and refine user interfaces. San Francisco: Morgan Kaufman.

Sonderegger, A., & Sauer, J. (2009). The influence of laboratory set-up in usability tests: effects on user performance, subjective ratings and physiological measures. *Ergonomics, 52*(11), 1350–1361.

Tarvainen, M. P., & Niskanen, J.-P. (2008). *Kubios HRV Version 2.0 User's guide*. University of Kuopio.

Tashman, C., & Edwards, W. K. (2011). Active reading and its discontents: the situations, problems and ideas of readers. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems – CHI '11* (pp. 2927–2936). New York: ACM.

Uebelbacher, A., Sonderegger, A., & Sauer, J. (2013). Effects of perceived prototype fidelity in usability testing under different conditions of observer presence. *Interacting with Computers, 25*(1), 91–101.

Virzi, R. A., Sokolov, J. L., & Karis, D. (1996). Usability problem identification using both low- and high-fidelity prototypes. In M. J. Tauber (Ed.), In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems – CHI '96* (pp. 236–243). New York: ACM.

Vredenburg, K., Mao, J.-Y., Smith, P. W., & Carey, T. (2002). A survey of user-centered design practice. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems – CHI '02 (pp. 471–478).* New York: ACM.

Walker, M., Takayama, L., & Landay, J. A. (2002). High-fidelity or low-fidelity, paper or computer? Choosing attributes when testing web prototypes. In *Proceedings of the Human Factors and Ergonomics Society 46th Annual Meeting* (pp. 661–665). Human Factors and Ergonomics Society.

Wästlund, E., Reinikka, H., Norlander, T., & Archer, T. (2005). Effects of VDT and paper presentation on consumption and production of information: psychological and physiological factors. *Computers in Human Behavior, 21*(2), 377–394.

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: the PANAS scales. *Journal of Personality and Social Psychology, 54*(6), 1063–1070.

Wilkinson, R. T., & Robinshaw, H. M. (1987). Proof-reading: VDU and paper text compared for speed, accuracy and fatigue. *Behaviour & Information Technology, 6*(2), 125–133.

Woody, W. D., Daniel, D. B., & Baker, C. A. (2010). E-books or textbooks: students prefer textbooks. *Computers & Education, 55*(3), 945–948.

Ziefle, M. (1998). Effects of display resolution on visual performance. *Human Factors, 40*(4), 554–568.

6. Study II: The perception of fidelity (perceived developmental stage)

Effects of perceived prototype fidelity in usability testing under different conditions of observer presence

Andreas Uebelbacher*, Andreas Sonderegger, Jürgen Sauer

Department of Psychology, University of Fribourg, Rue de Faucigny 2, CH-1700 Fribourg, Switzerland

(published in Interacting with Computers, Vol. 25, No. 1, 2013, p.91-101)

Abstract

The study investigated the influence of perceived prototype fidelity in usability tests by comparing two prototypes that differed with respect to their perceived proximity to the final system. The impact of the perceived developmental stage of the product was examined for participants' performance, perceived usability, emotions and psychophysiology. 80 participants were tested, operating an electronic city guide on a mobile phone. In a 2 x 2 x 2 mixed design, the system was either presented as an early prototype or as the final system. In addition, observer presence (no observers vs. three observers) and task difficulty (high vs. low) were experimentally manipulated. Overall, the findings did not indicate major differences for perceived prototype fidelity. However, an interaction between observer presence and prototype fidelity indicated that observer presence had a more negative impact on performance when testing a final system than an early prototype. Furthermore, observer presence resulted in a psychophysiological stress response. The findings suggest that test outcomes are quite robust against different prototype perceptions but that observer presence needs careful consideration.

Key words: usability test; prototype fidelity; perceived usability; heart rate variability; observer presence; developmental stage

* Corresponding author. E-mail: Andreas(@)uebelbacher.ch.

6.1 Introduction

The importance of usability as a goal in product development is increasingly acknowledged and the benefits of usability testing as a core method in this endeavour are hardly controversial (Lewis, 2006). One of the method's advantages is its flexibility, which allows an application at various stages of product development. At early stages, usability testing typically uses prototypes of the system to conduct a formative evaluation, identifying specific usability improvements (Gediga et al., 2002). At later stages, the final product is available for tests with users, and a summative approach is often applied taking global measures of performance and making an overall assessment of usability (e.g. using standardised questionnaires such as SUMI, PSSUQ, etc.) (Tullis & Albert, 2008). The tested prototypes can be very different from the final product with respect to various dimensions of their fidelity, and so can the respective test outcomes (e.g. a specific interaction pattern can only be tested if the prototype offers the required richness of interaction) (McCurdy et al., 2006; Virzi et al., 1996). However, prototype fidelity as perceived by participants in a usability test can be very different even when testing the same system (e.g. it may depend on task instructions). Therefore, the question arises whether the perceived prototype fidelity might have an impact on test outcomes. For usability practitioners, conducting prototype tests at various stages of product development is a reality and it is of high importance to know how a prototype should be presented to participants to avoid any undesired side-effects.

The present study aims to compare the influence of the perceived proximity of the current prototype to the final system on usability test outcomes such as performance, perceived usability, emotions and psychophysiology. To investigate the effects of participants' perception of the fidelity of a prototype compared to the final system, the presentation of the test system was systematically manipulated. The test system was either introduced as an early prototype in a formative testing context or as a final system that was evaluated with a summative approach. In addition, the previously shown effect of observer presence in usability testing (Sonderegger & Sauer, 2009) is investigated in both testing conditions.

6.1.1. Perceived prototype fidelity

In the development of interactive products it is important to gain user feedback on variants of system design as early as possible, to avoid the high costs of product changes after implementation of a system (Mantei and Teorey, 1988). Therefore, the usability

practitioner often faces the situation that user data needs to be collected even before a working system is available for testing. Prototype testing then becomes the method of choice, and surveys among usability practitioners prove that the method of iterative user testing with early prototypes is very common (Vredenburg et al., 2002). The main requirements for prototypes are low cost of production and sufficient similarity to the final product to reach valid test outcomes. Therefore, a very important characteristic of prototypes is their 'fidelity' to the final product, including aesthetic refinement, similarity of interaction and breadth of functions (McCurdy et al., 2006; Sauer et al., 2010; Virzi et al., 1996).

One aspect of prototype fidelity that has hardly been discussed in the scientific literature is how test participants subjectively perceive a system's developmental stage. This perception may not be without influence, as it might make a difference to participants' willingness to criticise a presented system (e.g. if a system is to be released soon, participants may hold back critical issues since they do not wish to disappoint the system developer).

Participants' perceptions of a system's developmental stage may well differ from its objective fidelity, as the association of objective prototype characteristics and how they are perceived may not be a simple one for several reasons. First, some dimensions of fidelity may have a more prominent impact on perceptions than others. A prototype scoring high on the dimensions of refinement of visual design and richness of interactivity but with a very limited data model might be judged by participants to be much closer to the final product than a prototype for which only the design of the front-end was very rough. This is in agreement with the understanding that for the user the interface effectively *is* the system (Mayhew, 1999). Second, the instructions of a test facilitator by which a system is introduced to users may also have an influence on the perception of fidelity or developmental stage, especially if the visual design of the interface does not reflect well how elaborate the prototype is with respect to other fidelity dimensions. Empirical work showed that information in the instructions of the participants prior to testing the system influenced the perceived usability directions in the expected direction (Hartmann et al., 2008; Raita and Oulasvirta, 2011). Of particular interest is the study of Bentley (2000), which showed (though in a rather small sample of $N = 24$) that participants gave higher usability ratings when they received information that the system had already undergone previous usability tests and was close to market introduction than when they were led to believe that the system had not been previously tested and was at an early stage of development.

The question of perceived developmental stage as an aspect of prototype fidelity is also relevant to the distinction between formative and summative evaluation, which represent two important types of usability evaluation approaches (e.g. Hix and Hartson, 1993; Nielsen, 1993). *Formative usability evaluation* typically takes place during early product development and is broadly defined to be 'anything that helps improve design within the user interface development process' (pp. 21, Hix et al., 1994). It comprises all usability engineering methods, which aim to identify usability problems and to improve the design on the basis of an understanding of the causes of these problems (Redish et al., 2002).

In contrast, *summative usability evaluation* aims to make an overall assessment of product qualities (e.g. ISO 9241-11 in ISO, 1998). It assesses global aspects of a system with respect to user needs and examines whether defined usability requirements have been met (Jokela, 2002). A summative usability test typically is applied late in the product development process, when the development of a product is almost complete (Hix and Hartson, 1993).

The comparison of these two forms of usability evaluation reveals that one important difference between the two is the stage in the product development at which the evaluation typically takes place (i.e. early for formative testing and late for summative testing). Whether a usability test is conducted early or late in the product development cycle may have different consequences if the test participants become aware of the product's development stage. In a test conducted early in the product development cycle (such as in formative testing), the participants may be happy to give critical feedback about the weaknesses of the product because there is plenty of time to remedy any such shortcomings. In contrast, giving such open and critical feedback in a usability test conducted late in the product development cycle (such as in summative testing) may have more dramatic consequences if the test outcomes require a delay in product launch (Karat, 1994). Therefore, the test participants may feel that the social desirability of achieving positive test results would be higher in a late test than in an early one.

6.1.2. Observer presence in usability testing

In addition to the perceptions of product development stages, there are further factors that may influence the test participants' perception of the testing situation. The set-up of the laboratory represents such a factor in usability testing, which may also affect user behaviour during the test. Very few studies have investigated the influence of observer presence as an important social factor during the testing process. For example, Sonderegger

and Sauer (2009) compared three conditions of observer presence in user testing (facilitator present with two observers, only facilitator present, test subject working alone in the room) and found that the presence of two additional observers had a negative impact on participants' heart rate variability (HRV), performance, and emotions. Harris et al. (2005) found some evidence for effects of observer presence, as their work showed higher error rates for complex tasks when a facilitator was present than when the participant was alone. Grubaugh et al. (2005) also found higher error rates in usability testing when the lab set-up was more intrusive in terms of monitoring equipment used.

These findings can be explained by social facilitation effects, which have been extensively investigated in social psychology. A large body of research shows that an individual's performance and levels of physiological arousal are affected by the presence of others, even if they do not directly interfere, compete, or interact with a person (Guerin, 1993). More precisely, for simple or well learnt tasks (automated processing) the presence of others generally improves a person's performance due to increased effort expenditure, while performance for difficult or unfamiliar tasks (controlled processing) is impaired, due to attention overload and distraction by others (Bond and Titus, 1983; Manstead and Semin, 1980). Guerin (1986) found in his meta-analysis of over 100 studies on social facilitation that this effect is strongest when individuals feel watched and evaluated rather than under mere presence of others. These issues are of particular relevance in the context of usability testing since the presence of observers may exert a stronger social pressure for good performance and restraint in criticising the system when it is close to market introduction.

6.1.3. The present study

The main goal of our study was to investigate whether developmental stage of a prototype as perceived by test participants in usability testing had an impact on test outcomes. Therefore, we conducted laboratory-based testing sessions which were instructed either as taking place at an early product development stage (as in formative testing of a prototype under development) or taking place at a late stage (as in summative pre-launch testing to decide whether a product launch would be advisable). In the present study, the terms 'early prototype testing' and 'final system testing' are used to refer to the different developmental stages of the system in usability testing.

To investigate whether a previously demonstrated effect in usability tests (i.e. impact of observer presence) would equally occur in early prototype testing as in final system testing, we implemented two experimental conditions, which proved in previous

work by Sonderegger and Sauer (2009) to have an effect on test outcomes. In one condition, a test facilitator (introduced as a university researcher) and two non-interacting observers (introduced as representatives of the product developer) were present in the test room throughout the test session. While the facilitator explained the procedure and answered participants' questions during the instruction phase, during task completion no interaction between participant and facilitator was allowed. In the other condition, no observers and no facilitator were present in the test room.

As a test system a modern smartphone was used. A variety of quantitative measures typically used in usability tests were recorded to assess the impact of the experimental conditions. Performance was assessed on several parameters (e.g. task completion time). Self-report measures were taken for participants' emotions, perceived usability and mental load. Since heart rate and its variability were shown to be reliable indicators of mental effort and stress (Izsó and Láng, 2000; Rowe et al., 1998), these physiological measures were chosen as indicators for physiological arousal. They proved to be suitable measures especially in the context of social situations (Pruyn et al., 1985).

Our hypotheses were as follows: (a) In the final system condition, we predicted better performance than in the early prototype condition. This was expected due to social desirability effects generating higher pressure to show good performance, because of the more severe impact usability problems would have in the pre-launch condition (i.e. delayed product launch). (b) We expected higher perceived usability ratings of the system in the early prototype condition than in final system testing. Since all participants were using the same fully operational application, it was expected that those in the early prototype condition would be more positively surprised (given the test system was introduced to them as an early prototype), as compared to the final system condition where participants were told that the design was complete. (c) The different laboratory settings represent different levels of social stressors and we predicted stronger effects on the dependent variables in the condition with three observers, that is, decreased heart rate variability, lower performance for difficult tasks but not for simple ones, increased negative emotion and decreased positive emotion.

6.2 Method

6.2.1. Participants

80 participants (70% female) took part in the experiment, aged between 17 and 65 years (age: M = 27.9; SD = 10.2). They were recruited from the general public and among

students, using a test participant pool from the Universities of Basel and Fribourg. They all had no prior interaction with the experimenter or the observers being present during the experiment. Prior to the experiment, it was checked that participants did not own the specific device to be used in the experiment and were excluded if this had been the case. They were paid 25 Swiss francs (approx. €20) for participation.

6.2.2. Design

A 2 x 2 x 2 mixed design was used, to investigate the following independent variables: as between subjects variables (a) developmental stage: early prototype testing vs. final system testing; and (b) observer presence: three observers vs. no observers; as within-subjects variable (c) task difficulty: high vs. low.

6.2.3. Measures and instruments

6.2.3.1. Performance

Three performance measures were taken: (a) task completion rate (percentage of successfully completed tasks), (b) task completion time (s), (c) efficiency of interaction (minimum number of pages to be viewed for task completion divided by actual number of pages viewed).

6.2.3.2. Perceived usability

Perceived usability was measured by two instruments. The first was the PSSUQ ('Post Study System Usability Questionnaire', Lewis, 1995), which was translated into German and slightly modified to be relevant for the test system in question (the term 'system' was replaced by 'software' throughout, to stress that only the software and not the device was to be judged). The scale consists of 19 items and uses a 7-point Likert scale (from strongly agree to strongly disagree), and had very good psychometric properties (Cronbach's $\alpha > 0.90$). The questionnaire was specifically developed for usage in usability tests in a laboratory setting and proved to be a valid instrument in previous research (Lewis, 2002).

Additionally, we used a visual analogue scale (0-100; ranging from 'not at all' to 'very much') to measure an overall estimation of perceived usability ('The software is usable'). Single-item measures of usability proved to be valuable and reliable in previous research (Christophersen and Konradt, 2011).

6.2.3.3. Emotions

To assess short-term changes in emotion during the test procedure we used the PANAS scale ('Positive and Negative Affect Schedule'), which allows the assessment of two independent dimensions of mood: positive and negative affect (Watson et al., 1988). The German language version (Krohne et al., 1996) was shown to have good psychometric properties (Cronbach's $\alpha = 0.84$). The scale uses 20 adjectives to describe different affective states (e.g. interested, exciting, strong), the intensity of which is rated on a 5-point Likert scale ('very slightly or not at all', 'a little', 'moderately', 'quite a bit', 'extremely').

6.2.3.4. Task load

A German version of the well established NASA task load index (TLX) by Hart and Staveland (1988) was used to assess task load on the six dimensions mental demands, physical demands, temporal demands, own performance, effort and frustration. The weighting procedure was not used, so that each single dimension was given the same weight. Our data indicated that psychometric properties were sufficient for the translated scale (Cronbach's $\alpha = 0.78$).

6.2.3.5. Psychophysiology

Heart rate variability (HRV) was used as an indicator for participants' stress response. We determined the frequency bands for analysis of HRV in line with previous research and as recommended by the Task Force of the European Society of Cardiology (1996) (high: 0.15-0.4 Hz; low: 0.04-0.15 Hz; very low: 0.003-0.04 Hz). Since previous research has shown the low frequency (LF) band to be specifically relevant for measuring mental effort and physical stress response in situations of potential social stress (Pruyn et al., 1985; Sonderegger and Sauer, 2009), the subsequent analysis concentrates on this HRV indicator. For the analysis, we used the Kubios HRV 2.0™ software (Tarvainen and Niskanen, 2008) for Windows XP™. Possible artefacts were corrected with the artefact correction level 'medium' and the default Fast Fourier Transformation was used for time interval calculations.

6.2.3.6. Additional measures

As a control variable, previous experience with the specific test device was assessed by means of a visual analogue scale (0-100; ranging from 'no experience at all' to 'a great deal

of experience') to rule out an impact of different knowledge and skill levels with respect to the hardware and software in question. As a manipulation check we introduced four visual analogue scales, questioning participants for their estimation of the test system in terms of developmental state and distance to market introduction, and whether they felt observed and how much they felt disturbed by observation (see Section 6.2.7).

6.2.4. Materials

6.2.4.1. Test device and data recording hardware

As a mobile phone test device a black Apple Inc. iPhone™ 3G was used with 16GB memory and a touch screen with a 480-by-320 pixel resolution. The mobile phone screen was transferred wirelessly to a nearby laptop (Siemens Fujitsu LifeBook™ T3010), where all mobile phone manipulations were recorded. Heart rate was logged with a Polar™ RS800 heart rate monitor which participants were wearing for the full testing session. A video camera (Panasonic™ NV-MS5EG) was positioned in one corner of the room facing in the direction of the test participants' work space.

6.2.4.2. Software

The test device was running on iOS2.2™. Veency™ (v.1.0.4) was installed which allowed for displaying the mobile phone screen directly on a nearby laptop computer wirelessly. On the laptop, TightVNC™ (v1.3.10 for Windows XP™) was installed to support this connection.

As a test system the cityscouter™ Berlin application (v2.01) for the Apple iPhone™ was used, which is an electronic travel guide for the city of Berlin (Figure 6). The software offers tourist information on Berlin city sights, restaurants, hotels and information on city transport. The application is fully menu driven so that no touchscreen keyboard usage was necessary. All the data necessary for the test tasks was available offline in the application.
As a screen capturing tool on the portable computer the software CamStudio™ (v.2.0) was used to record all mobile phone screen manipulation during the test sessions.

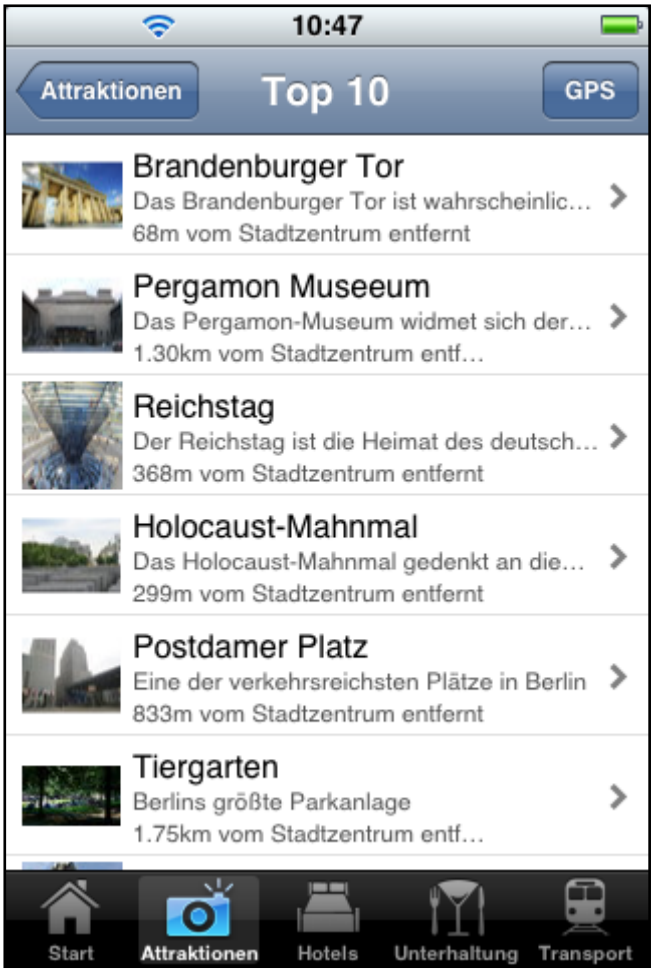

Figure 6. Interface of the city guide.

6.2.5 Tasks

Test participants had to accomplish the following six tasks of finding specific information with the Berlin travel guide, which were given on paper: (a) opening hours for the Berlin Reichstag; (b) admission information for the Berlin TV tower; (c) the telephone number for a specific restaurant; (d) details for public transport connection to the Holocaust monument; (e) a vegetarian restaurant near a specific shopping centre; and (f) public transport connection to the Kaiser Wilhelm memorial church. Participants had to note down the results of their search on a task sheet. Tasks (a) to (c) were easier than tasks (d) to (f) because they required fewer interactions, respective navigation options were easier to understand and the solutions were more directly supported by the functionality of the application.

6.2.6. Procedure

The test sessions were conducted in a laboratory at the University of Fribourg. The participants were randomly assigned to one of the four testing conditions (resulting from the combination of the two instructed developmental stages and the two observer presence set-ups). The experimenter (test facilitator) welcomed the participants and led them to the preparation room, where he gave an overview of the subsequent procedure of the experiment.

The purpose of the usability test was explained and varied according to the two experimental conditions for early prototype vs. final system testing. To create an early prototype testing situation, the system status was described as a 'prototype in development', which only served the purpose to run the test and to identify usability problems which were planned to improve upon during the following project stages. There would be ample time to redesign the application, since the launch was only planned 5 months later. In the final system testing condition the test-goal was described as generating the basis for the decision whether the application should be launched as planned or not. The system was described as a final product with a launch date in the upcoming weeks.

Then the heart rate measurement procedure was introduced and the participant was asked to put on the heart rate monitoring device in a rest room next door. Afterwards the Polar RS800™ watch was attached to the participant's wrist and it was checked for functional transmission. Then the participant was seated on a sofa and asked to relax and not to move so that the heart rate baseline could be recorded for the following 10 min. During that time the experimenter was not in the room.

Afterwards, the test participant filled in the PANAS questionnaire to measure the emotional baseline and was then guided to the usability laboratory. In the observers present condition, the participant was introduced to the two observers, who were presented as representatives of the IT development company responsible for the application that was about to be tested and who would like to have a first-hand insight into how their system would work. The two observers (one female, aged 27, one male, aged 63 years) were confederates of the experimenter, dressed in casual business style and were not interacting with the participant during the experiment. They were seated at a table about 2m behind the participant, outside the participants' field of vision.

Then the participant was seated on an office chair at a table and was introduced to the test device, specifically the touch screen, the home button and the pinch-zoom functionality, and the tasks were described. In case of difficulties, the participant was invited to keep working on the tasks until a message would automatically be displayed on

the mobile phone screen prompting to pass to the next task (which was manually triggered after 5 min). The tasks were described as all having a solution but showing a variation in difficulty, and the participants were given the opportunity to ask questions.

In the observer condition, the participants were instructed that no help would be given by the facilitator and observers during task completion. This was to create a realistic usage condition. Then the test tasks were given to the participant in written form and the video recording was started. In the no-observer condition, the test facilitator left the room. Participants' full interaction with the system was registered by direct screen recording. To avoid interference with the heart rate variability parameters, participants were instructed not to talk during the psychophysiological measurement. After task completion, the video recording was stopped and the post-test instruments were administered (PANAS, NASA-TLX, single-item scale for usability, PSSUQ and manipulation check). Finally, open feedback about the system and the test procedure was requested. All participants were then debriefed and were given their incentive.

6.2.7. Manipulation check

A manipulation check consisting of four items was used to test whether the manipulation of the independent variables was successful.

For the early prototype vs. final system conditions the manipulation check confirmed that the instructions had the intended effects on participants for the following two aspects of the instruction. On a visual analogue scale (0-100; ranging from 'rough prototype' to 'final system'), they judged the test system as being significantly closer to a final product in the final system condition (M = 83.0) than in the early prototype condition (M = 68.4), $t(76) = 4.35$, $p < 0.001$. On another visual analogue scale asking about the perceived distance from market introduction (0-100; ranging from 'very far away' to 'not very far away'), participants judged the system as being closer to the market introduction under the final system condition (M = 84.7) as compared to the early prototype condition (M = 74.9), $t(76) = 2.65$, $p < 0.01$.

For the manipulation of observer presence, the manipulation check also confirmed the desired effects. On a visual analogue scale ranging from 'not felt observed at all' to 'felt strongly observed' (0-100), participants reported significantly stronger feelings of being observed (M = 36.5) when observers and test facilitator were present in the same room, compared to when they were working on their own (M = 26.7), $t(76) = 1.72$, $p < 0.05$. There was also an association of age with the level of feeling observed across all conditions, with older participants indicating to have felt less observed than younger ones

(r = -.26, p < 0.05). When asked about how much they felt disturbed by observation (0-100; ranging from 'not felt disturbed by observation at all' to 'felt strongly disturbed by observation'), they reported feeling significantly more disturbed by observation when observers and test facilitator were present in the room during task accomplishment (M = 21.6) compared to when working alone (M = 13.4), t(68.5) = 1.72, p < 0.05.

6.3. Results

We controlled for the influence of several variables (e.g. previous experience with mobile phones, daily mobile phone usage, gender and age) by including them as covariates in our analysis of variance. Since none of them had any impact on the main results, this analysis is not reported here. For all analyses, the alpha level was set to 5%.

6.3.1. User performance

6.3.1.1. Task completion rate

The performance data is presented in Table 5. The data analysis showed that neither developmental stage of the system nor observer presence had a significant impact on task completion rate (both F's < 1). Furthermore, there was no interaction between the two factors (F < 1). As expected, task difficulty had a strong influence on completion rate, with easy tasks showing a significantly higher completion rate than difficult tasks (F(1,75) = 79.94, p < 0.001).

6.3.1.2. Task completion time

As reported above for task completion rate, there was no significant effect of experimental conditions on task completion time (see Table 5). Neither developmental stage of the system (F < 1) nor observer presence (F(1,78) = 1.07, *ns*) had any significant impact on completion times, and also the respective interaction proved non-significant (F(1,78) = 1.01, *ns*). As expected, task difficulty determined task completion time, with easy tasks being accomplished much faster than difficult tasks (F(1,75) = 438.53, p < 0.001).

Table 5. Measures of user performance as a function of perceived developmental stage, observer presence and task difficulty (TD).

	Early prototype testing		**Final system testing**		
	No observers present Mean (*SD*)	Observers present Mean (*SD*)	No observers present Mean (*SD*)	Observers present Mean (*SD*)	**Overall** Mean (*SD*)
Task completion rate (%)	79.8 (20.5)	77.5 (20.4)	81.7 (17.9)	80.0 (18.4)	79.7 (19.0)
low TD	95.0 (12.2)	93.3 (13.7)	96.7 (10.3)	100 (0)	96.3 (10.6)
high TD	63.2 (36.7)	61.7 (32.9)	66.7 (32.4)	60.0 (36.8)	62.9 (34.2)
Task completion time (s)	139.1 (38.2)	139.3 (38.3)	122.7 (38.3)	140.3 (38.2)	135.3 (38.2)
low TD	59.3 (41.4)	70.0 (40.0)	49.4 (38.2)	60.3 (42.7)	59.8 (40.5)
high TD	222.0 (54.6)	208.6 (60.7)	196.0 (59.3)	220.2 (58.6)	211.6 (58.2)

6.3.1.3. Efficiency of user-product interaction

An analysis of the efficiency of task completion revealed overall high levels of efficiency, as indicated by the task efficiency index (minimum number of pages to be viewed / actual number of pages viewed during task completion) presented in Figure 7. As the data show, there were no main effects of developmental stage of the test system or of observer presence (both *F's* < 1). However, the corresponding interaction between the two experimental conditions was significant ($F(1,78) = 4.37$, $p < 0.05$). This was because more efficient performance occurred in the final system testing than in the early prototype condition, when no observers were present ($F(1,38) = 4.48$, $p < 0.05$), while with observers being present, participants in the two developmental stage conditions performed no different ($F < 1$). Finally, an expected main effect of task difficulty occurred, with participants performing significantly more efficiently on easy ($M = 0.71$) than on difficult tasks ($M = 0.48$) ($F(1,65) = 61.59$, $p < 0.001$). No further effects were recorded.

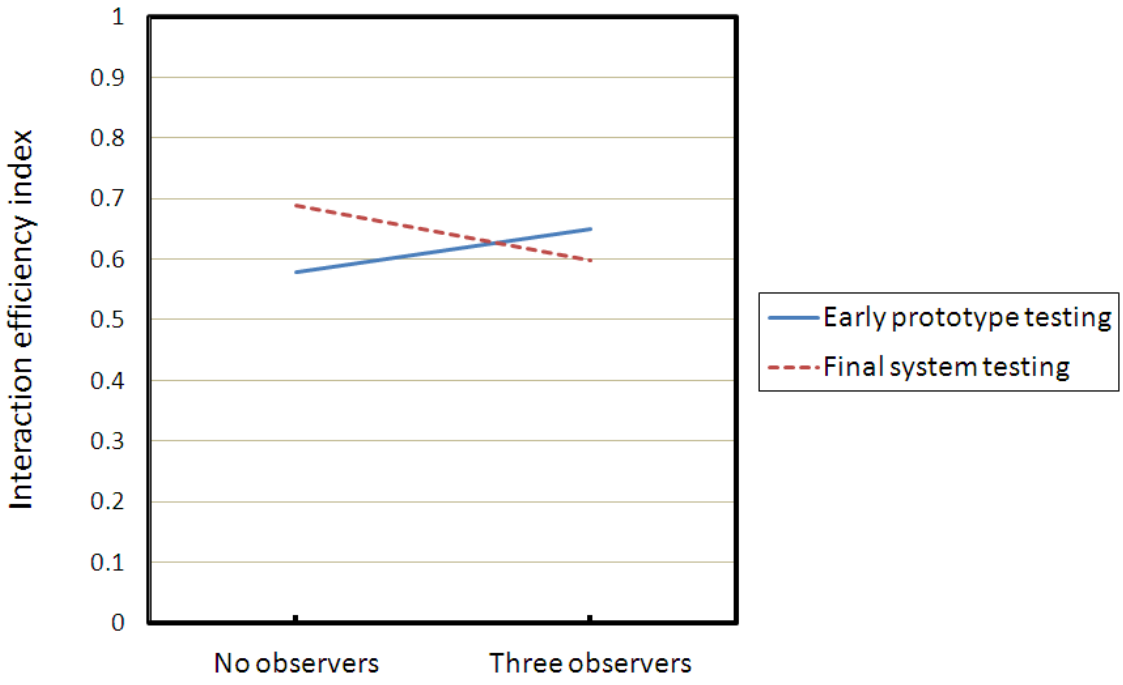

Figure 7. Efficiency of user-product interaction (minimum number of pages to be viewed / number of pages viewed) as a function of testing approach and observer presence.

6.3.2. Subjective ratings

6.3.2.1. Perceived usability

The data for perceived usability is presented in Table 5. In contrast to our expectations, in the early prototype condition participants rated the system's usability on the one-item scale significantly less positively (M = 60.5) than in the final system testing condition (M = 71.0) ($F(1,75) = 6.64$, $p < 0.05$). However, for the overall PSSUQ score, there was no such significant effect of developmental stage ($F(1,78) = 2.37$, *ns*).

Observer presence did not have an impact on perceived usability on the one-item scale or for the PSSUQ overall rating (both F's < 1). Overall, the correlation between the usability one-item scale and the total PSSUQ score was $r = 0.71$ ($p < 0.001$). Interestingly, age showed a significant effect on PSSUQ overall score, with older participants rating the system more negatively ($F(1,74) = 9.78$, $p < 0.01$).

Table 6. Measures of perceived usability, emotions and mental load as a function of perceived developmental stage and observer presence.

	Early prototype testing		**Final system testing**		
	No observers present Mean (*SD*)	Observers present Mean (*SD*)	No observers present Mean (*SD*)	Observers present Mean (*SD*)	**Overall** Mean (*SD*)
Perceived usability on one-item scale (0-100)	60.0 (18.8)	61.0 (16.3)	72.9 (17.5)	69.1 (18.0)	65.9 (18.2)
PSSUQ[a] (1-7)	5.1 (1.01)	5.0 (0.68)	5.4 (0.70)	5.2 (0.78)	5.2 (0.80)
Positive affect (Δ: pre - post)	0.50 (0.55)	0.22 (0.62)	0.38 (0.61)	0.21 (0.51)	0.33 (0.58)
Negative affect (Δ: pre - post)	0.14 (0.34)	0.16 (0.33)	-0.03 (0.20)	0.15 (0.44)	0.11 (0.34)
NASA-TLX (1-20)	10.1 (2.5)	10.5 (2.4)	9.4 (3.4)	10.4 (3.3)	10.1 (2.9)

Δ: all values represent changes from baseline (resting phase) to task completion phase, on average PANAS scale (1-5)
[a] PSSUQ: Post-Study System Usability Questionnaire

6.3.2.2. Task load

Task load data is presented in Table 6. There was no effect of developmental stage ($F < 1$) or of observer presence ($F(1,78) = 1.23$, *ns*) on experienced task load, and there was no interaction effect for the two conditions ($F < 1$). A separate analysis of the subscales of the NASA-TLX also provided no significant effects.

6.3.2.3. Emotions

Table 6 presents the data for participants' emotions. Developmental stage had no significant effect on the change of reported positive affect from before to after task completion ($F < 1$), and neither had observer presence ($F(1,78) = 3.06$, *ns*). There was no significant interaction ($F < 1$). Similarly, for negative affect none of the effects were significant, showing the same pattern of results and are therefore not reported in detail. Interestingly, taking part in the experiment had both an impact on participants' positive and negative affect. Participants reported significantly more positive affect after the test ($M = 3.21$) than before ($M = 2.88$) ($F(1,78) = 25.66$, $p < 0.001$). For negative affect there was also an increase from pre- to post-test measurement with participants reporting higher

negative affect after (M = 1.34) than before task completion (M = 1.24) (F(1,78) = 7.78, p < 0.01).

6.3.3. Physiological measures

6.3.3.1. Heart rate

We compared participants' heart rate changes between resting phase and task completion phase. Overall there was a significant increase in heart rate from the resting phase (M = 73.3) to task completion phase (M = 76.8), which was statistically highly significant (F(1,72) = 26.62, p < 0.001, see Table 7). This increase differed across observer presence conditions. When observers were present, there was a much stronger increase in mean heart rate from resting phase to task completion phase (+5.37bpm) than when participants were working on their own (+1.57bpm) (F(1,72) = 7.96, p < 0.01). There was no main effect for developmental stage of the test system (F(1,72) = 1.08, *ns*), and no interaction effect (F(1,72) = 1.33, *ns*).

Table 7. Changes in physiological parameters from baseline to task completion phase, as a function of perceived developmental stage and observer presence.

	Early prototype testing		**Final system testing**		
	No observers present Mean (*SD*)	Observers present Mean (*SD*)	No observers present Mean (*SD*)	Observers present Mean (*SD*)	**Overall** Mean (*SD*)
Δ Heart rate (bpm[a])	+1.65 (5.2)	+3.89 (5.5)	+1.50 (5.3)	+6.84 (6.7)	+3.55 (6.0)
Δ LF[b] power (ms^2)	+216.9 (678.6)	-346.4 (612.2)	-0.1 (645.5)	-114.2 (548.8)	-77.5 (638.2)

Δ: all values represent changes from baseline (resting phase) to task completion phase, with positive values denoting an increase in the parameter
[a] bpm: beats per minute
[b] LF: low frequency band

6.3.3.2. Heart rate variability

Using HRV in the LF band (0.04-0.15 Hz) as a sensitive indicator for participants' stress response and mental effort, we compared baseline levels as measured during an initial resting phase with measurements during task completion phase. To control for outliers, we

excluded from the analysis 8 participants with LF band values lying outside a range of ± 2 standard deviations from the mean. Our analysis showed an impact of observer presence on these difference values (see Table 7). There was a decrease during task completion phase for the power in the LF band when observers were present, indicating higher stress levels. In contrast, when working alone, participants showed a significant increase of power in the LF band during task completion phase ($F(1,64) = 4.81$, $p < 0.05$). For developmental stage of the system ($F < 1$) or for the interaction between developmental stage and observer presence there were no such effects ($F(1,64) = 2.12$, *ns*).

6.4. Discussion

The primary research question of our study was to investigate the impact of perceived prototype fidelity on central outcome variables and whether observer presence had the same effects in both testing conditions. In contrast to our hypotheses, the results did not provide a great deal of support for our assumptions that perceived prototype fidelity in the form of early or final stages in product development had a significant effect on usability test outcomes. However, there was some indication that observers had a more negative impact on participants' performance in final system testing than in the early testing condition. As expected, observer presence caused higher stress levels in participants during the task completion phase.

A major outcome of the present study is that for both developmental stages very similar results were recorded for the vast majority of dependent variables, including performance, psychophysiology, emotion and perceived usability. Despite this general pattern, there were selected indications of differences between developmental stages. For example, there was an interesting interaction between developmental stage and observer presence with respect to performance. Non-observed participants in final system testing condition performed better than those in the early prototype one in terms of efficiency of accessing information. This effect may be explained by the expectation to perform well, which might have been higher in the final system testing condition, especially when observers were present. Two mechanisms might account for this. First, consequences of bad test results were more dramatic in the final system condition, because it may lead to a delay in product launch. Second, in the final system testing condition observers may be perceived as being more negatively affected by poor user performance because as product developers they would have to take the blame for poor product usability (in the early prototype condition, there would still be the chance to rectify usability problems).

On the one-item measure of usability, participants judging a final system rated its usability more positively than when judging an early prototype. This was in contrast to our expectation that in the final system condition users would be less tolerant and have higher expectations towards the finished system, resulting in lower ratings of usability. However, the results for the single-item measure are consistent with the assumption that expectations towards a more favourable evaluation of the test system were higher in the final system testing condition.

In contrast, for the PSSUQ no significant difference occurred. Since the tested system's usability was the same, the fact that the different developmental stages (presenting the system as an early prototype vs. a final system) had no significant impact on the PSSUQ ratings indicates that this more elaborate, multi-item instrument is less affected by social desirability than an overall one-item measure. This is in line with concerns raised elsewhere that an overall measure of usability might have insufficient psychometric qualities, compared to a multi-item scale (Hornbaek and Law, 2007). However, other studies have found effects of positive system information being given to participants in that judgments of perceived usability increased after testing (Bentley, 2000; Raita and Oulasvirta, 2011). One possible explanation for these different results might be that those studies manipulated information with respect to core aspects of quality of use (e.g. whether the system had already been usability tested; positive/negative usability ratings of system features). This is in contrast to the present study, which referred to developmental stage (rather than directly addressing the system's usability), representing a more peripheral aspect of the system.

The results provided evidence that the presence of additional observers in a usability test setting caused higher stress levels among test participants. Participants with observers present during task accomplishment reported to have felt more disturbed by observation, an effect which was mitigated by age, as older participants indicated that they generally felt less observed than younger ones. When observers were present, participants also showed changes at the physiological level, in the form of decreased HRV and increased heart rate when working on tasks. These results on the physiological level confirm findings of a previous study (Sonderegger and Sauer, 2009), providing further evidence for the HRV in the LF band to be a sensitive indicator for social stress in usability testing and even beyond. However, we could not confirm the previous study's (Sonderegger and Sauer, 2009) findings concerning a negative impact of observer presence on performance. Participants with observers present generally performed no worse than those working alone in a room. The question arises why we could not replicate the findings of impaired performance under observation, although we implemented the same conditions and used

very similar performance indicators that should have been equally sensitive. One possible explanation for this difference in results between the two studies might be the gender of the observers. In contrast to the previous study, which used an all-male group of observers, we had one female observer. This may have mitigated the effect of observer presence by reducing the evaluative characteristics of the social situation, as suggested by research on gender differences (Leary et al., 1994). This work suggests lower self-presentation concerns of participants in social situations when all interaction partners were female rather than male. Research on gender stereotypes also suggests that females are expected to show behaviours as taking care of the well-being of others in a social context, while men are expected to act task-oriented, even at the expense of the well-being of others (Eagly and Karau, 2002; Leon, 2005). This may have also contributed to a female observer not being perceived as socially threatening as a male observer.

There are implications of our findings for usability practitioners and researchers alike. First, usability test outcomes may be more robust against different instructions given to participants, as long as these do not directly concern aspects of quality of use of a system, such as information about previous usability tests or usability ratings (cf. Bentley, 2000; Raita and Oulasvirta, 2011). Second, observer presence has an impact on participants in a usability test, which has now repeatedly been confirmed (e.g. Grubaugh et al., 2005; Harris et al., 2005; Sonderegger and Sauer, 2009). Which aspects of the social situation in a usability test cause or moderate these effects, however, is not yet fully understood. There is a need for further research investigating factors such as the age and gender of participants and observers, and how observers are introduced to test participants. A more qualitative approach exploring participants' perceptions of different aspects of the test situation could provide valuable insights in this respect. There are also important implications for usability engineers. Whenever possible, observers who are not directly involved in running a usability test should not be in the same room as participants, because it may put the latter under stress. When infrastructure does not allow for a separation of observers and participants in a usability test, special care must be taken to make test participants feel at ease, since the findings provided evidence that even subtle differences in the user's perception of the testing situation can have considerable effects on the test outcomes.

Acknowledgement

The authors are very grateful to the Swiss National Science Foundation for their financial support of the study (research grant No. 100014/122490). Thanks are also due to Dr. Javier

Bargas-Avila, Jean-Pierre Guenter, Hans-Rudolf Kocher, Amadeus Petrig, and Manuela Pugliese for their support in completing this study.

References

Bentley, T. (2000). Biasing web site user evaluation: a study. In *Proceedings of the Australian Conference on Human-Computer Interaction – OZCHI 2000* (pp. 130–134). IEEE Computer Society.

Bevan, N. (2006). Practical issues in usability measurement. *Interactions, 13*(6), 42-43.

Bond, C.F., Titus, L.J. (1983). Social facilitation: a meta-analysis of 241 studies. *Psychological Bulletin, 94*(2), 265-292.

Christophersen, T., Konradt, U. (2011). Reliability, validity, and sensitivity of a single-item measure of online store usability. *International Journal of Human-Computer Studies, 69*(4), 269-280.

Eagly, A.H., Karau, S.J. (2002). Role congruity theory of prejudice toward female leaders. *Psychological Review, 109*(3), 573-598.

Gediga, G., Hamborg, K.-C., Düntsch, I. (2002). Evaluation of Software Systems. In Kent, A., Williams, J.G., (Eds.), *Encyclopedia of computer science and technology, Vol. 45 (*pp. 127-153). New York: Marcel Dekker, Inc.

Grubaugh, B., Thomas, S., Weinberg, J. (2005). The effects of the testing environment on user performance in software usability testing. In M.H. Hamza (Ed.), *Proceedings of the IASTED International Conference on Human-Computer Interaction* (pp. 39-43). Anaheim: ACTA Press.

Guerin, B. (1986). Mere presence effects in humans: a review. *Journal of Experimental and Social Psychology, 22*(1), 38-77.

Guerin, B. (1993). *Social facilitation.* Cambridge, England: University Press.

Harris, E., Weinberg, J., Thomas, S., Gaeslin, D. (2005). Effects of social facilitation and electronic monitoring on usability testing. In: *Proceedings of the Usability Professionals Association Conference*.

Hart, S.G., Staveland, L.E. (1988). Development of NASA-TLX (task load index): results of empirical and theoretical research. In P.A. Hancock, N. Meshkati (Eds.), *Human mental workload* (pp. 139-183). Amsterdam, The Netherlands: Elsevier Science Publishers.

Hartmann, J., De Angeli, A., Sutcliffe, A. (2008). Framing the user experience: information biases on website quality judgments. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems – CHI '08* (pp. 855–864). New York: ACM Press.

Hix, D., Hartson, H.R. (1993). Developing user interfaces: ensuring usability through product and process. New York: John Wiley & Sons.

Hix, D., Hartson, H.R., Nielsen, J. (1994). A taxonomy for developing high impact formative usability evaluation methods. *SIGCHI Bulletin, 26*(4), 20-22.

Hornbaek, K., Law, E.L.-C. (2007). Meta-analysis of correlations among usability measures. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems – CHI '07* (pp. 617-626). New York: ACM.

ISO (1998). ISO 9241, Ergonomic requirements for office work with visual display terminals (VDTs) - Part 11, Guidance on usability, International standard.

Izsó, L., Láng, E. (2000). Heart period variability as mental effort monitor in human computer interaction. *Behaviour & Information Technology, 19*(4), 297-306.

Jokela, T. (2002). Making user-centred design common sense: striving for an unambiguous and communicative UCD process model. In *Proceedings of the NordiCHI* (pp. 19-26). New York: ACM.

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*(6), 345-355.

Karat, C.-M. (1994). A business case approach to usability cost justification. In R.G. Bias, D.J. Mayhew (Eds.), *Cost-justifying usability* (pp. 45-70). New York: Academic Press.

Krohne, H.W., Egloff, B., Kohlmann, C.-W., Tausch, A. (1996). Untersuchungen mit einer deutschen Version der "Positive and Negative Affect Schedule" (PANAS). *Diagnostica, 42*(2), 139-156.

Leary, M.R., Nezlek, J.B., Downs, D., Radford-Davenport, J., Martin, J., McMullen, A. (1994). *Journal of Personality and Social Psychology, 67*(4), 664-673.

Leon, G.R. (2005). Men and women in space. *Aviation, Space, and Environmental Medicine, 76*(6), 84-88.

Lewis, J.R. (1995). IBM computer usability satisfaction questionnaires: Psychometric evaluation and instructions for use. *International Journal of Human-Computer Interaction, 7*(1), 57-78.

Lewis, J.R. (2002). Psychometric evaluation of the PSSUQ using data from five years of usability studies. *International Journal of Human-Computer Interaction, 14*(3&4), 463-488.

Lewis, J.R. (2006). Usability testing. In G. Salvendy (Ed.), *Handbook of human factors and ergonomics* (pp. 1275-1316). Hoboken, NJ: Wiley.

Manstead, A.S.R., Semin, G.R. (1980). Social facilitation effects: mere enhancement of dominant responses? *British Journal of Social and Clinical Psychology, 19*(2), 119-136.

Mantei, M.M., Teorey, T.J. (1988). Cost/benefit analysis for incorporating human factors in the software lifecycle. *Communications of the ACM, 31*(4), 428-439.

Mayhew, D. (1999). The usability engineering lifecycle: a practitioner's handbook for user interface design. San Diego, CA: Academic Press.

McCurdy, M., Connors, C., Pyrzak, G., Kanefsky, B., Vera, A. (2006). Breaking the fidelity barrier. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems – CHI '*06 (pp. 1233-1242). New York: ACM.

Nielsen, J. (1993). *Usability engineering.* San Diego, CA: Academic Press.

Pruyn, A., Aasman, J., Wyers, B. (1985). Social influences on mental processes and cardiovascular activity. In J.F. Orlebeke, G. Mulder, L.F.P. van Doornen (Eds.), *Psychophysiology of cardiovascular control (models, methods, and data)* (pp. 865-877). New York: Plenum Press.

Raita, E., Oulasvirta, A. (2011). Too good to be bad: favorable product expectations boost subjective usability ratings. *Interacting with Computers, 23*(4), 363-371.

Redish, J., Bias, R., Bailey, R., Molich, R., Dumas, J., Spool, J.M. (2002). Usability in practice: Formative usability evaluations - Evolution and revolution. In L.G. Terveen, D.R. Wixon, (Eds.), *CHI '05 Extended Abstracts on Human Factors in Computing Systems* (pp. 885-890). New York: ACM.

Rowe, D.W., Sibert, J., Irwin, J. (1998). Heart rate variability: indicator of user state as an aid to human-computer interaction. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems – CHI '98* (pp. 480-487). New York: ACM.

Sauer, J., Seibel, K., Rüttinger, B. (2010). The influence of user expertise and prototype fidelity in usability tests. *Applied Ergonomics, 41*(1), 130-140.

Scriven, M. (1967). The methodology of evaluation. In R. Tyler, R. Gagne, M. Scriven (Eds.), *Perspectives of curriculum evaluation* (pp. 39-83). Chicago, IL: Rand McNally.

Seffah, A., Habieb-Mammar, H. (2009). Usability engineering laboratories: limitations and challenges toward a unifying tools/practices environment. *Behaviour & Information Technology, 28*(3), 281-291.

Sonderegger, A., Sauer, J. (2009). The influence of laboratory set-up in usability tests: effects on user performance, subjective ratings and physiological measures. *Ergonomics, 52*(11), 1350-1361.

Task Force of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology (1996). Heart rate variability. Standards of measurement, physiological interpretation, and clinical use. *Circulation, 93,* 1043-1065.

Tarvainen, M.P., Niskanen, J.-P. (2008). *Kubios HRV Version 2.0 User's guide.* University of Kuopio.

Tullis, T., Albert, W. (2008). Measuring the user experience: collecting, analyzing, and presenting usability metrics. Burlington, MA: Morgan Kaufmann.

van den Haak, M., De Jong, M., Schellens, P.J. (2003). Retrospective vs. concurrent think-aloud protocols: testing the usability of an online library catalogue. *Behaviour & Information Technology, 22*(5), 339-351.

Virzi, R.A., Sokolov, J.L., Karis, D. (1996). Usability problem identification using both low- and high-fidelity prototypes. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems – CHI '96* (pp. 236-243). New York: ACM.

Vredenburg, K., Mao, J.-Y., Smith, P.W., Carey, T. (2002). A survey of user-centered design practice. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems – CHI '02* (pp. 471-478). New York: ACM.

Watson, D., Clark, L.A., Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: the PANAS scale. *Journal of Personality and Social Psychology, 54*(6), 1063-1070.

7. Study III: The fidelity of the test situation (dual-domain products in usability testing)

Usability testing of dual-domain products in the context of work and leisure

Andreas Uebelbacher*, Jürgen Sauer, Andreas Sonderegger

Department of Psychology, University of Fribourg, Rue de Faucigny 2, CH-1700 Fribourg, Switzerland

(to be submitted)

Abstract

The present study examines the influence of usage domain on the utilisation of dual-domain products. A comparison is made between work and leisure domain by modelling these in a laboratory. Using a 2 x 2 experimental design, in addition to usage domain (work vs. leisure), usability of the appliance was manipulated (normal vs. delayed response time). 60 participants were tested while completing tasks with an internet application on a modern smartphone. Several performance parameters and a range of subjective measures were recorded (e.g. emotion, perceived usability, and task load). Overall, there was little evidence for an influence of usage context on outcomes of usability tests, suggesting that it would be sufficient to test dual-domain products in either usage domain. System response time showed the expected effects on performance and on user emotion, whereas, unexpectedly, no influence on perceived usability was observed.

Key words: usability test; system response time; usage domain; perceived usability; user performance

* Corresponding author. E-mail: andreas(@)uebelbacher.ch.

7.1. Introduction

7.1.1. Context of use in usability testing

There is a long tradition in psychology to conceive behaviour as being highly dependent on context (Lewin, 1926). This also applies to the field of ergonomics, where context is considered one of the main determinants of user behaviour when operating interactive systems (Bevan & MacLeod, 1994). The importance of the concept is reiterated by the fact that it is also part of ISO 9241-11, which defines context as covering users, tasks and equipment, and the specific social and physical environment in which a product is used (ISO, 1998). It is therefore not meaningful to describe a product's usability per se, without taking into account the context in which the product will be used (Maguire, 2001b). In other words, a system that may be usable in one context may not be in another.

While users, tasks and equipment are routinely specified in usability studies and there seems to be little disagreement that these need to be taken into account, the environmental aspects of context are rarely considered (Alonso-Rios et al., 2010). However, research has provided evidence that a number of environmental factors might influence the outcomes of usability tests, such as lab vs. field set-up (C.M. Nielsen et al., 2006), observer presence (Sonderegger & Sauer, 2009), or the use of electronic recording equipment (Harris et al., 2005). One important characteristic of the usage environment which received little attention in previous research is the distinction of the usage domains of leisure and work context.

7.1.2. Work vs. leisure domain

Up into the 1980s, most people experienced interactive technology almost exclusively in the workplace (Grudin, 2005). Since then personal computers have reached people's homes, and mobile phones have become mobile computing devices. Today, the pervasiveness of interactive technology in all areas of people's lives, including leisure, is a reality (den Buurman, 1997). In this process, the distinction between technology at work and technology for leisure use has become increasingly fuzzy, as devices often cannot anymore be described as clearly being one or the other. Many of today's technical devices are dual-domain products, as they can equally be used in a work context as in a leisure context (e.g. mobile phones and laptops).

Since research in ergonomics traditionally concentrated on interactive technology in the work domain, the discipline was primarily concerned with performance and to

provide more usable interfaces to increase efficiency at work (Carroll, 2001). Together with a recent shift of attention towards 'user experience' as a concept which explicitly encompasses aspects of joy of use (Bargas-Avila & Hornbæk, 2011), ergonomic studies have started to investigate more leisure-oriented technology, such as portable digital audio players (Thüring & Mahlke, 2007). For dual-domain products, however, both domains are equally relevant and different requirements may result from these contexts, which might need to be considered during product evaluation. In order to identify the domain-specific requirements, the differences between work and leisure need to be analysed. There is one previous study that empirically compared work and leisure domains but found little differences between them (Sauer, Uebelbacher, Pugliese, & Sonderegger, submitted). However, this may be due to the fact that in addition to usage domain, product aesthetics was manipulated as a second factor. It was assumed that aesthetics would play a more important role in a leisure than in a work context, which was not confirmed. Since system usability is a more direct determinant of effectiveness and efficiency of use than aesthetics its influence in different usage domains is worth examining.

In addition to the lack of empirical research in that field, there have also been difficulties in establishing a clear theoretical distinction between work and leisure so that no widely accepted definition of the concepts has yet been proposed. Three approaches to distinguish between the two domains have been discussed (Beatty & Torbert, 2003). (a) The purely time-based or 'residual' definition of leisure is most commonly used. According to this approach, leisure is when people do not do paid or unpaid work, do not complete personal chores, and do not fulfil obligations. (b) An activity-based approach distinguishes between work and leisure by means of specific behaviours people show in each domain. (c) The third approach conceives work and leisure by the attitudes people have towards their activities. Beatty and Torbert (2003) argue that the third approach is considered to be most promising to distinguish leisure from other domains, and there is also some empirical evidence in support of this approach. Several studies confirmed that people described work in terms of goal-directed and performance-oriented behaviour and connected with external rewards while leisure was associated with intrinsic satisfaction, enjoyment, novelty and relaxation (Rheinberg, Manig, Kliegl, Engeser, & Vollmeyer, 2007; Tinsley, Hinson, Tinsley, & Holt, 1993).

This third approach distinguishes between work and leisure in a way that also allows us to derive domain specific requirements of interactive technology. If users perceive a work context as more goal- and performance-oriented than a leisure context, and if usability is about improving performance of a system (i.e. effectiveness and efficiency),

usability might be perceived by users as a more important requirement in a work than in a leisure context.

7.1.3. Response time as a facet of system usability

One aspect of system usability which previous research has shown to be directly relevant for various outcome variables is system response time (SRT). SRT is defined as the time it takes from a user input to the moment the system starts to display the response (Shneiderman, 1984). Although one might think that delayed SRT would be less of a problem with today's much increased processing power, delays are still a relevant problem in human-computer interaction, given that modern multi-tasking systems usually run various software at the same time (virus scanners, update checks, etc.) (Szameitat, Rummel, Szameitat, & Sterr, 2009). In addition, a strong increase of mobile internet usage and a mobile network infrastructure that currently cannot keep up with the sharp increase in data connection demands make delayed SRT an everyday experience of the mobile user (Roto & Oulasvirta, 2005).

Negative effects of SRT delays have been shown at several levels. First, there is evidence that response time delays have a negative effect on user satisfaction with a system (Barber & Lucas, 1983; Dellaert & Kahn, 1999; Hoxmeier & Di Cesare, 2000; Rushinek & Rushinek, 1986). Systems with delayed responses are generally perceived as being less usable and more strenuous to operate, which also extends to web pages with long download times being judged to be less interesting (Ramsay, Barbesi, & Preece, 1998). Second, user performance has been shown to be impaired by SRT delays (Butler, 1983; Galletta, Henry, McCoy, & Polak, 2004; Shneiderman, 1984; Szameitat et al., 2009). Third, system delays have resulted in impaired psychophysiological well-being, increased anxiety, frustration and stress, and were even found to reduce job satisfaction (Barber & Lucas, 1983; Guynes, 1988; Polkosky & Lewis, 2002; Selvidge, Chaparro, & Bender, 2002; Trimmel, Meixner-Pendleton, & Haring, 2003).

Various moderators of the effects of system response delays have been identified in the context of internet usage, such as webpage properties (Jacko, Sears, & Borella, 2000), user expectations (Hui & Tse, 1996), and delay information displays (Branaghan & Sanchez, 2009). For example, users were less willing to accept download delays when websites were highly graphical compared to plain text documents (Jacko et al., 2000). It also emerged that information about the duration of the download had a positive effect on user evaluation (Hui & Tse, 1996), and progress bars as delay indicators performed best in terms of user preference and judged reasonableness of the wait (Branaghan & Sanchez, 2009). To our

knowledge, only one study has researched SRT in a work context (Barber & Lucas, 1983). Conducting a field study in a large telephone circuit utility observing professionals in their work domain, they found that increased SRT not only impaired performance but also system evaluation and even job satisfaction. No study was found in the literature which investigated SRT delays in a leisure context.

7.1.4. The present study

The main goal of the present study was to investigate the requirements that result from the two domains of work and leisure with respect to usability. For this purpose a usability test was conducted in which the two domains of work and leisure were experimentally modelled. The two types of testing context were created by lab room design (office vs. living room), task wording (work related vs. leisure related) and a priming task which directed participants' attention towards their own work or leisure activities, respectively. As a second independent variable, system usability was manipulated through SRT delays.

As a test system, an internet site was specifically set up for the experiment, which was designed to offer realistic tasks for both contexts. Care was taken that the tasks for the two experimental conditions were comparable in terms of mental demands but only differed in type of context. The tasks used were information search tasks that required navigating through various levels of a menu hierarchy.

As dependent measures performance and subjective evaluation were recorded. Task performance was assessed by success rate, page inspection time, and efficiency of task completion. Self-report data was collected for emotion and perceived usability.

Our hypotheses were as follows: (a) Test participants in the work context perform better and report higher perceived task demands than those in the leisure context, since the work context is perceived as more goal- and performance-oriented. (b) Performance is lower when working in the condition with delayed SRT compared to working with normal SRT. (c) Perceived usability of the system and emotional reactions are less positive in the delay condition than with normal SRT, since the reduced system usability is reflected in participants' evaluation and emotion. (d) At work, delayed SRT causes a stronger decrease in perceived usability and in emotion than in the leisure context, since the negative impact of system delay on performance is perceived as more relevant in the goal- and performance-oriented work context.

7.2. Method

7.2.1. Participants

The sample of the experiment consisted of 60 participants, aged between 19 and 44 years (*M* = 22.6 yrs; *SD* = 3.3), the majority of which were female (60.3%). Participants were recruited among students at the University of Fribourg, and it was made sure that none of them had the specific mobile phone model which was used in the experiment. To motivate participants to take part in the study, they could enter a draw to win an MP3 music player (worth 50 EUR). Psychology students were also given credit points for their participation.

7.2.2. Design

A 2 x 2 between-subjects design was used to investigate the two independent variables usage domain and usability. Usage domain was varied at two levels (work vs. leisure context), and so was usability (normal vs. delayed system response time).

7.2.3. Measures and instruments

7.2.3.1. Performance

The following three measures of user performance were recorded: (a) task completion rate (percentage of successfully completed tasks); (b) page inspection time (time a user stays on a page); (c) efficiency of task completion (minimum number of interactions needed for task completion divided by actual number of interactions). Participants were allowed to work on each task for a maximum of 5 min, after which a task was recorded as failed and participants moved on to the next task. All analyses of performance data took into account the shorter overall time participants had available in the delay condition.

7.2.3.2. Emotion

The PANAS scale ('Positive and Negative Affect Schedule', Watson *et al.,* 1988) was used to measure short-term emotional changes before and after task completion. The scale allows the assessment of two independent dimensions of mood: positive and negative affect. It is available in German (Krohne *et al.,* 1996) and was shown to have good psychometric properties (Cronbach's $\alpha = 0.84$). The scale uses 20 adjectives to describe different affective

states (e.g. 'interested', 'exciting', 'strong'), for which the intensity is rated on a 5-point Likert scale ('very slightly or not at all', 'a little', 'moderately', 'quite a bit', 'extremely').

7.2.3.3. Task load

To assess task load the German version of the well-established NASA task load index (TLX) was used (Hart and Staveland, 1988). It measures the following six dimensions: mental demands, physical demands, temporal demands, performance, effort and frustration. In the subsequent analysis, each dimension was given the same weight. Based on our data, psychometric properties were shown to be satisfactory for the translated scale (Cronbach's $\alpha = 0.72$).

7.2.3.4. Perceived usability

Perceived usability of the test system was measured by two instruments. First, we used a 100mm visual analogue scale to measure an overall evaluation of perceived usability ('This website is usable'). The same scale was already used in previous work (Sonderegger & Sauer, 2009). The use of one-item scales to evaluate technical systems was found to be appropriate, as other work has shown (e.g. Christophersen & Konradt, 2011). Second, the PSSUQ ('Post Study System Usability Questionnaire'; Lewis, 1995) was applied, which was translated into German and slightly modified to be relevant for the test system in question (the term 'system' was replaced by 'software' to make sure only the software and not the device was judged). The scale consists of 19 items and uses a 7-point Likert scale (ranging from 'strongly agree' to 'strongly disagree'). The questionnaire was specifically developed for usage in usability tests in a lab setting, and Lewis (1995) reports very good psychometric properties (Cronbach's $\alpha > 0.90$).

7.2.3.5. Previous mobile phone experience

Previous mobile phone experience was assessed by a visual analogue scale on which participants reported an intermediate self-rated mobile phone experience of 5.0 on a scale ranging from 0 to 10 (labelled 'not experienced' and 'very experienced'). They indicated using their devices 12.6 times on average during a day. Mobile phone experience and daily usage were used as covariates in the analysis.

7.2.4. Materials: mobile phone, server and software

The test device was a black Motorola Milestone/Droid™, which had a touch screen with 854-by-480 pixel resolution, and which was running on Android™ OS (version 2.0). To restrict usage of the test device to direct touch screen manipulation of the web application, the hardware buttons of the test device and the WebKit™ browser address bar were covered by black tape. This also prevented participants from accessing the test device's hardware keyboard. The web application was used with the phone's Android WebKit™ HTML5 browser, and was accessed over a wireless LAN (local area network) connection. The application was running on a nearby Apple MacBook™ 2.1 and as a server software XAMPP™ (Mac™ OS X Version 1.7.3) was used. This server was connected by LAN to a Netopia™ 3347W DSL router, setting up the wireless network, which was password protected and exclusively accessed by the test device. In the delay condition, a PHP script was running on the server and generated a random system response delay of between 0s and 3s (1.3s on average) whenever a new page was requested. A server log recorded the pages viewed, the time at which the page was accessed, the duration during which the page was displayed and the size of the delay.

The web application that was used for task completion was specifically set up for the experiment. It consisted of a hierarchical navigation system (as shown in Figure 8), offering a number of categories at each level and detailed pages at the deepest level. Navigation options were 'return to the previous page', 'return directly to the home page', or selecting one of the displayed categories. Scrolling was necessary for some of the pages, which had a larger number of categories than the screen could display. Category labels were deliberately named such that it was not always obvious in which the target page would be found so that a trial and error approach to target search became necessary (e.g. a specific Asian restaurant was located under the category 'Japanese', while other categories available included 'Asian', 'Chinese', 'German', 'Greek', 'Indonesian', and 'Italian'). A message on the target page stated clearly that the task had been solved and requested that the user directly went back to the home page.

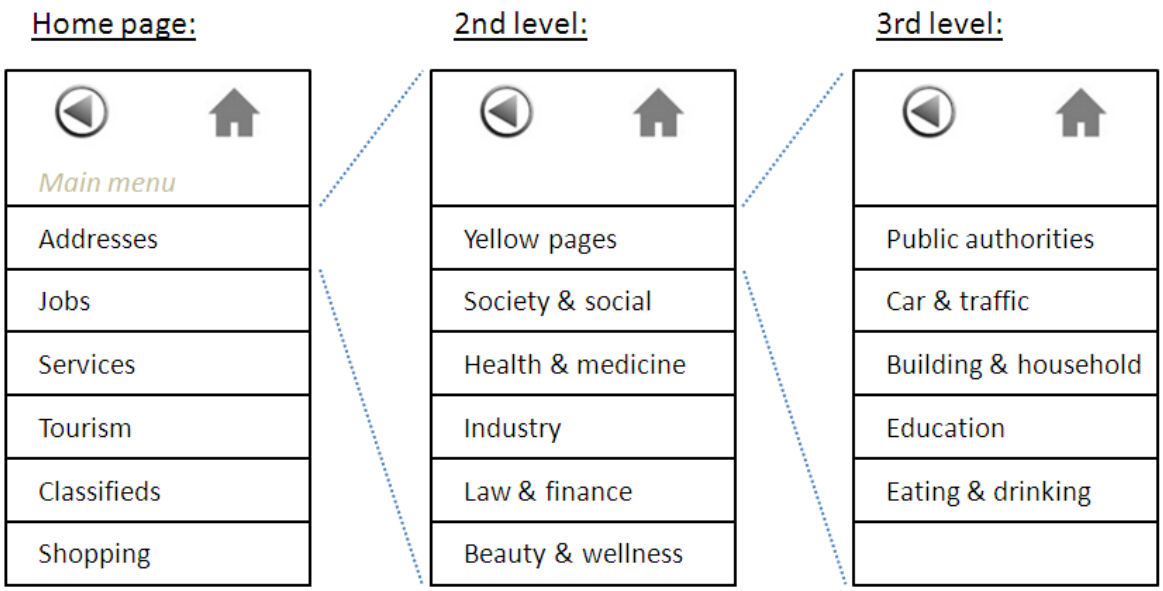

Figure 8. Home page and two subordinate pages of the system, illustrating the navigation options available.

7.2.5. Procedure

Participants were randomly assigned to one of the four testing conditions (resulting of the combination of leisure vs. work context, and delay vs. no delay conditions). The testing sessions were conducted in a usability laboratory at the University of Fribourg. In the leisure condition, the laboratory was set up like a living room, containing a sofa (on which the participant was seated), wooden furniture with travel books, a (switched off) TV set, plants on the window sill, and pictures on the wall. In the work condition, the laboratory contained several desks, a (switched off) computer, a desk lamp, some folders and typical office stationery (stapler, etc.).

The experimenter described the purpose of the experiment as testing the usability of a web application for smartphones, giving an overview of the experimental procedure. Participants filled in the PANAS and the questionnaire measuring previous mobile phone experience. The experimenter presented the test device, showed all functions of the web application and explained how to operate it (e.g. choosing categories, home, back, scrolling). Afterwards he explained the error messages which would be shown if participants diverted too much from the correct dialogue path. Participants completed a practice trial to become familiar with the web application. They were instructed to tell the experimenter whenever they finished a task. They were given the opportunity to ask questions, and then instructions concerning the usage context followed. In the work context condition, they were asked to put themselves in the position of being at work, to imagine they would be

working the following two days and to think about what they would have to do during these days. A respective instruction was given in the leisure context. After about 1 min the experimenter gave the first task to the participant. They had 5 min for each task, but were not informed about this time constraint, and after that time the experimenter thanked them and gave them the next task. After the last task, the participants were given the PANAS again, then the NASA-TLX, the subjective usability questionnaires (one-item scale and PSSUQ), and finally the manipulation check. The participants were debriefed and could leave their e-mail address to take part in the draw to win a MP3 music player. The duration of a testing session was about 45 min.

7.2.6. Manipulation check

The manipulation check consisted of a visual analogue scale (0-100; ranging from 'rather leisure-oriented' to 'rather work-oriented'), on which participants judged the situation in which they completed the tasks. Results confirmed a significant impact of the context manipulation, as participants indicated to have experienced the situation significantly more work-related in the work context condition (M = 56.8), compared to the leisure context condition (M = 27.5; $t(58) = 4.92$; $p < 0.001$).

7.3. Results

Self-reported mobile phone experience, daily mobile usage and gender were entered as covariates into the analysis in order to control for any influence. However, the analysis showed that none of the covariates had a significant influence on the reported findings.

7.3.1. User performance

7.3.1.1. Task completion rate

Our analysis showed significant differences in the number of completed tasks as a function of SRT (see Table 8). When working with a system with a delayed response, participants solved significantly fewer tasks (M = 83.3%) than when response time was not delayed (M = 93.3%; F = 5.28; df = 1, 56; $p < 0.05$). Testing context (work vs. leisure) had no effect and there was no interaction between SRT and testing context (both $F < 1$).

Table 8. Measures of user performance as a function of testing context and system.

	Leisure context		**Work context**		
	No delay Mean (*SD*)	Delayed Mean (*SD*)	No delay Mean (*SD*)	Delayed Mean (*SD*)	**Overall** Mean (*SD*)
Task completion rate (%)	96.4 (9.1)	82.8 (21.8)	90.6 (15.5)	83.9 (18.6)	88.3 (17.5)
Page inspection time (s)	5.34 (1.04)	5.93 (1.19)	5.32 (0.75)	5.73 (0.54)	5.58 (0.94)
Efficiency of task completion (minimum No of interactions needed / actual No of interactions)	0.41 (0.08)	0.39 (0.15)	0.39 (0.13)	0.38 (0.12)	0.39 (0.12)

7.3.1.2. Page inspection time

As the data in Table 8 indicates, participants in the delay condition stayed significantly longer on a page (M = 5.84s) than those working with a non-delayed system (M = 5.33s). This difference was statistically significant (F = 4.46; df = 1, 56; p < 0.05). With regard to the other independent factors, there was neither a significant effect of testing context (F < 1) nor an interaction (F < 1).

7.3.1.3. Efficiency of task completion

An important indicator of user efficiency is determined by the calculation of the ratio of the actual number of user inputs and the optimal number of user inputs. The data in Table 8 indicate overall a medium level of efficiency of about M = 0.4. This efficiency index shows that 40% of the user inputs contributed towards task completion, whereas the remaining inputs were not directly leading to the task goal or were part of a less direct path towards task completion. As the data in Table 8 suggest, there was little difference between conditions, which was confirmed by analysis of variance (all F < 1).

7.3.2. Subjective ratings

7.3.2.1. Emotion

Data for emotions is presented in Table 9. For the analysis of the emotional state of the user as a consequence of using the product, a comparison was made between the baseline measurement (i.e. prior to task completion) and a second measurement taken after task

completion. This analysis revealed a change in positive affect as a function of SRT. While participants reported an increase in positive affect after task completion when working with a non-delayed system (M = 0.12), lower positive affect was reported when working with a delayed system (M = -0.17; F = 4.67; df = 1, 56; $p < 0.05$). For the changes in negative affect, it appeared that in the delayed SRT condition negative affect decreased more strongly (M = -0.55) than in the no-delay condition with nearly stable levels (M = 0.08). However, this effect was not statistically significant ($F < 1$). As the data in Table 9 show, testing context had no effect on the change of positive affect levels and there was no interaction either (both $F < 1$). Equally, there was no effect on the change of negative affect (F = 2.98; df = 1, 56; *ns*), nor was there an interaction ($F < 1$).

Table 9. Measures of emotions, task load, and perceived usability.

	Leisure context		**Work context**		
	No delay Mean (*SD*)	Delayed Mean (*SD*)	No delay Mean (*SD*)	Delayed Mean (*SD*)	**Overall** Mean (*SD*)
Emotions (1-5) positive affect (Δ: pre - post)	0.09 (0.42)	-0.15 (0.48)	0.14 (0.56)	-0.20 (0.59)	-0.03 (0.53)
negative affect (Δ: pre - post)	-0.29 (2.23)	-1.31 (2.50)	0.44 (3.05)	0.21 (2.08)	-0.25 (2.55)
Task load (1-20)	8.3 (1.5)	7.4 (2.7)	8.0 (3.0)	7.8 (3.0)	7.8 (2.6)
Perceived usability (1-7)	4.7 (0.88)	4.5 (1.33)	5.0 (0.94)	5.0 (0.90)	4.8 (1.03)

Δ: all values represent changes from baseline (pre) to task completion phase (post); a positive value denotes an increase

7.3.2.2. Task load

The data for the overall NASA-TLX score are presented in Table 9. While this indicates overall a low task load score, there was generally very little difference between experimental conditions. This was confirmed by the analysis of variance, with neither a main effect for the two independent factors nor an interaction between them (all $F < 1$). To check whether any differences could be found at the single item level, a separate analysis the NASA-TLX items demands, physical demands, temporal demands, performance, effort, and frustration was carried out. Analysis of variance confirmed the absence of such effects.

7.3.2.3. Perceived usability

The data for perceived usability, as measured by the PSSUQ, are presented in Table 9. Interestingly, the expected effect of SRT was not observed, with ratings being nearly identical for both conditions ($F < 1$). Usability ratings appeared to be higher in the work domain than in the leisure domain but this difference failed to reach significance levels ($F = 2.01$; $df = 1, 52$; *ns*). No interaction between the two factors was found ($F < 1$). An additional analysis examined the PSSUQ subscales separately but found by and large the same pattern of results as for the overall scale. Finally, the results are reported for the one-item usability scale, which was included to evaluate its utility in research contexts. The 100mm visual analogue scale showed an overall rating of $M = 52.2$ but there was little difference between the four experimental conditions (all $F < 1$), herewith confirming the results pattern found for the PSSUQ.

7.4. Discussion

The aim of the study was to investigate the influence of usage domain on the outcomes of a usability test and whether any such influence would be mediated by poor system usability in the form of SRT delays. The findings showed that, contrary to expectations, usage domain did not have the expected impact, with none of the measures showing differences between domains. This suggests that it is sufficient to test dual-domain products in either usage domain. In contrast, system response time showed the expected effects on performance and on user emotion whereas, surprisingly, no influence on perceived usability was observed.

Given that context of use has been considered an important determinant of usability (Maguire 2001) and that the two domains of work and leisure have been associated with different perceptions and behaviour (Rheinberg *et al.* 2007), we expected that testing a product in one domain would produce differences in usability test results compared to the other domain. The manipulation check showed very clearly that participants perceived the leisure domain differently from the work domain. Despite this successful manipulation of context (involving different usability lab set-ups, domain-specific task instructions, and a priming task), there were no differences in usability test results, neither for performance nor for subjective measures. A major implication of this finding is that there may be no need for practitioners to test dual-domain products in both usage domains. The domains of work and leisure may not require specific consideration in test set-ups, as long as the relevant use cases are covered in the test. The absence of an interaction between usage domain and system usability strengthens this argument, suggesting that even under conditions of

impaired system usability the work domain provides test results that are no different from the leisure domain.

One previous study comparing work and leisure domains also found little difference between them (Sauer *et al.* under review). However, it did not manipulate the usability of the technical device but its aesthetical features. Taken together, this study and the present work provide evidence that across a range of conditions (i.e. different levels of product aesthetics and of product usability) the influence of usage domains appears to be of smaller magnitude than expected. The findings are in support of (Lindroth & Nilsson, 2001) claim that environmental aspects of usage context are generally not an important issue in usability testing as long as stationary technology usage is concerned (which was the case in the present study as the smartphone was operated like a desktop device). While Lindroth and Nilsson did not empirically test their proposition, the present work provides some empirical evidence to support it. This may have been partly due to the general difficulty to distinguish between the concepts of work and leisure (Beatty & Torbert, 2003; Haworth & Veal, 2004), which might have led to less unequivocal interpretations of the usage domain by participants. Although the successful manipulation check indicated that a distinction was made by participants in this study, some concerns may remain with regard to the extent to which motivational processes associated with the work domain could be appropriately reproduced in the lab. However, it has to be noted that this problem would affect all lab-based usability testing, independently of the domain, and previous research has shown that lab-based testing often provides no worse results than conducting tests in the field (Kjeldskov & Stage, 2004). Finally, there are increasingly elements of work to be found in leisure time, that is, people are virtually permanently contactable, they make use of electronic organisers, and they also manage their appointments in a business-style manner (Ling & Haddon, 2003). These changes in how leisure time is organised and experienced may have also led to a waning difference between the two domains. As a qualification, it has to be added that a dual-domain product may well be used for quite different purposes in different contexts so that a usability evaluation of such a product would still need to consider different requirements as a function of usage domain. This is especially important, as task coverage was found to be an important determinant of usability test outcomes (Lindgaard & Chattratichart, 2007). However, given the reported findings, the respective tasks could be tested within a single context in the laboratory.

While usage domain had little impact on the results of usability testing, a number of effects of poor system usability were found, confirming several of our research hypotheses. First, it emerged that poor system usability had the hypothesised negative effect on task performance. When SRT was delayed, task completion rate was lower and participants

spent more time on a page, compared to participants working with a system without delays. These findings are consistent with an extensive body of research showing a negative impact of delayed system response on performance (e.g. Barber & Lucas, 1983; Galletta et al., 2004). One explanation for this effect is that users adapt their speed of task completion to SRT and work faster when the system responds more promptly (Boucsein, 2009). An alternative explanation for longer page inspection times under delayed SRT could be that participants adjusted their strategy, moving from a trial and error approach to a more reflective one, thus reducing the number of delayed system responses. Previous work has shown that even short SRT delays made participants consider their actions more carefully (Guynes, 1988; Teal & Rudnicky, 1992).

Second, poor usability had a negative effect on participants' emotions, consistent with our hypothesis. When working with a delayed system, participants showed a stronger reduction in positive affect than when working with a non-delayed system. This finding is consistent with an extensive body of research, showing negative effects of delayed SRT on various aspects of emotions, such as frustration, anxiety, stress and impatience (e.g. Guynes, 1988, Selvidge *et al.,* 2000, Polkosky & Lewis, 2002). The present study adds to these findings by showing that such effects on emotion may occur, even if such SRT delays are very short.

Third, although poor system usability had a negative effect on performance and affected participants' emotional state, no such effects were found for perceived usability. This observation is of particular interest since other work found a substantial positive association between performance and preference (Nielsen & Levy, 1994). While users generally provide a more positive evaluation when systems are more usable, Nielsen and Levy also cite some cases in their meta-analysis, in which users prefer systems, with which they perform worse. These systems, however, had rather short SRT delays and performance impairments did not reach critical levels. The magnitude of the delay in our study might have been below that critical level and therefore did not have an effect on perceived usability. An alternative explanation for the observed finding is that participants did perceive such delays but the SRT delay was not associated with the application but with the server from which the pages were downloaded. Similar observations were made in other work where users of internet-based software attributed the cause of delayed response to internet connection rather than the software itself (Rose, Meuter, & Curran, 2005). Overall, although we employed very short SRT delays, most of our hypotheses were confirmed, which highlights the importance of paying attention to even short delays during system design as it may affect performance and user emotion.

The findings presented have several implications for research and practice. First, the findings support the notion that results of a usability test having either been conducted in a work or a leisure domain are transferable to the other domain. This facilitates usability testing of dual-domain products for practitioners since several testing contexts would not have to be covered so that they only need to ensure that the relevant tasks are included in the test set-up. Second, even rather short SRT delays can have an effect on performance as well as on user emotion, suggesting that careful consideration should be given to SRT in product design and evaluation.

Acknowledgements

The authors are very grateful to the Swiss National Science Foundation for their financial support of the study (research grant No. 100014/122490). Thanks are also due to Eric Bourquard, Klaus Heyden and Max von Schlippe for their support in completing this study.

References

Alonso-Ríos, D., Vázquez-García, A., Mosqueira-Rey, E., and Moret-Bonillo, V. (2010). A context-of-use taxonomy for usability studies. *International Journal of Human-Computer Interaction, 26*(10), 941–970.

Barber, R.E. and Lucas, H.C. (1983). System response time, operator productivity, and job satisfaction. *Communications of the ACM, 26*(11), 972-986.

Bargas-Avila, J.A. and Hornbaek, K. (2011). Old wine in new bottles or novel challenges? A critical analysis of empirical studies of user experience. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems – CHI '11* (pp. 2689-2698). New York: ACM.

Beatty, J.E. and Torbert, W.R. (2003). The false duality of work and leisure. *Journal of Management Inquiry, 12*(3), 239-252.

Bevan, N. and MacLeod, M. (1994). Usability measurement in context. *Behaviour & Information Technology, 13*(1-2), 132-145.

Boucsein, W. (2009). Fourty years of research on system response times - What did we learn from it? In C.M. Schlick (Ed.), *Industrial Engineering and Ergonomics* (pp. 575-593). Berlin: Springer.

Branaghan, R.J. and Sanchez, C.A. (2009). Feedback preferences and impressions of waiting. *Human Factors, 51*(4), 528-538.

Butler, T.W. (1983). Computer response time and user performance. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems – CHI '83* (pp. 58-62). New York: ACM.

Carroll, J.M. (2001). Community computing as human-computer interaction. *Behaviour & Information Technology, 20*(5), 307-314.

Christophersen, T. and Konradt, U. (2010). Reliability, validity, and sensitivity of a single-item measure of online store usability. *International Journal of Human-Computer Studies, 69*(4), 269-280.

Dellaert, B.G.C. and Kahn, B.E. (1999). How tolerable is delay?: Consumers' evaluations of internet web sites after waiting. *Journal of Interactive Marketing, 13*(1), 41-54.

den Buurman, R. (1997). User-centred design of smart products. *Ergonomics, 40*(10), 1159-1169.

Galletta, D.F., Henry, R., McCoy, S. and Polak, P. (2004). Web site delays: how tolerant are users? *Journal of the Association for Information Systems, 5*(1), 1-28.

Grudin, J. (2005). Three faces of human-computer interaction. IEEE Annals of the History of *Computing, 27*(4), 46-62.

Guynes, J.L. (1988). Impact of system response time on state anxiety. *Communications of the ACM, 31*(3), 342-347.

Harris, E., Weinberg, J., Thomas, S., and Gaeslin, D. (2005). Effects of social facilitation and electronic monitoring on usability testing. In *Proceedings of the Usability Professionals Association Conference,* 27 June - 1 July 2005 Quebec, published on CD.

Hart, S.G. and Staveland, L.E. (1988). Development of NASA-TLX (task load index): results of empirical and theoretical research. In P.A. Hancock and N. Meshkati (Eds.), *Human mental workload* (pp. 139-183). Amsterdam, The Netherlands: Elsevier Science Publishers.

Haworth, J.T. and Veal, A.J. (Eds.). (2004). *Work and leisure.* London: Routledge.

Hoxmeier, J. and DiCesare, C. (2000). System response time and user satisfaction: an experimental study of browser-based applications. In *Proceedings of the Americas Conference on Information Systems* (pp. 140-145). Long Beach, CA: Association for Information Systems.

Hui, M.K. and Tse, D.K. (1996). What to tell consumers in waits of different lengths: an integrative model of service evaluation. *The Journal of Marketing, 60*(2), 81-90.

ISO (1998). ISO 9241, Ergonomic requirements for office work with visual display terminals (VDTs) - Part 11, Guidance on usability, International standard.

Jacko, J.A., Sears, A. and Borella, M.S. (2000). The effect of network delay and media on user perceptions of web resources. *Behaviour & Information Technology, 19*(6), 427-439.

Kjeldskov, J. and Stage, J. (2004). New techniques for usability evaluation of mobile systems. *International Journal of Human-Computer Studies, 60*(5-6), 599-620.

Krohne, H.W., Egloff, B., Kohlmann, C.-W. and Tausch, A. (1996). Untersuchungen mit einer deutschen Version der "Positive and Negative Affect Schedule" (PANAS). *Diagnostica, 42*(2), 139-156.

Lewin, K. (1926). Untersuchung zur Handlungs- und Affektpsychologie I. Vorbemerkungen über die psychischen Kräfte und Energien und über die Struktur der Seele. *Psychologische Forschung, 7*, 294-329.

Lewis, J.R. (1995). IBM computer usability satisfaction questionnaires: psychometric evaluation and instructions for use. *International Journal of Human-Computer Interaction, 7*(1), 57-78.

Lindgaard, G. and Chattratichart, J. (2007). Usability testing: what have we overlooked? In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems – CHI '07* (pp. 1415-1424). New York: ACM.

Lindroth, T. and Nilsson, S. (2001). Mobile usability: rigour meets relevance when usability goes mobile. In S. Bjørnestad, R.E. Moe, A.I. Mørch and A.L. Opdahl (Eds.), *Proceedings of the 24th Information Systems Research Seminar in Scandinavia – IRIS '24* (pp. 641-654). IRIS Association.

Ling, R.S. and Haddon, L. (2003). Mobile telephony, mobility and the coordination of everyday life. In J.E. Katz (Ed.), *Machines that become us: the social context of personal communication technology* (pp. 245-266). New Brunswick, NJ: Transaction Publishers.

Maguire, M.C. (2001). Context of use within usability activities. *International Journal of Human-Computer Studies, 55*(4), 453-483.

Nielsen, C.M., Overgaard, M., Pedersen, M.B., Stage, J. and Stenild, S. (2006). It's worth the hassle! The added value of evaluating the usability of mobile systems in the field. In *Proceedings of the NordiCHI* (pp. 272-280). New York: ACM.

Nielsen, J. and Levy, J. (1994). Measuring usability: preference vs. performance. *Communications of the ACM, 37*(4), 66-75.

Polkosky, M.D. and Lewis, J.R. (2002). Effect of auditory waiting cues on time estimation in speech recognition telephony applications. *International Journal of Human-Computer Interaction, 14*(3-4), 423-446.

Ramsay, J., Barbesi, A. and Preece, J. (1998). A psychological investigation of long retrieval times on the World Wide Web. *Interacting with Computers, 10*(1), 77-86.

Rheinberg, F., Manik, Y., Kliegl, R., Engeser, S. and Vollmeyer, R. (2007). Flow bei der Arbeit, doch Glück in der Freizeit. *Zeitschrift für Arbeits- und Organisationspsychologie, 51*(3), 105-115.

Rose, G.M., Meuter, M.L. and Curran, J.M. (2005). On-line waiting: the role of download time and other important predictors on attitude toward e-retailers. *Psychology & Marketing, 22*(2), 127-151.

Roto, V. and Oulasvirta, A. (2005). Need for non-visual feedback with long response times in mobile HCI. In *Proceedings of the WWW '05: Special Interest Tracks and Posters of the 14th International Conference on World Wide Web* (pp. 775–781). New York: ACM.

Rushinek, A. and Rushinek, S.F. (1986). What makes users happy? *Communications of the ACM, 29*(7), 594-598.

Sauer, J., Uebelbacher, A., Pugliese, M. and Sonderegger, A. (submitted). The influence of aesthetics in usability testing: the case of dual-domain products.

Selvidge, P.R., Chaparro, B.S. and Bender, G.T. (2002). The world wide wait: effects of delays on user performance. *International Journal of Industrial Ergonomics, 29*, 15-20.

Shneiderman, B., 1984. Response time and display rate in human performance with computers. *Computing Surveys, 16* (3), 265-285.

Sonderegger, A. and Sauer, J. (2009). The influence of laboratory set-up in usability tests: effects on user performance, subjective ratings and physiological measures. *Ergonomics*, *52*(11), 1350-1361.

Szameitat, A.J., Rummel, J., Szameitat, D.P. and Sterr, A. (2009). Behavioural and emotional consequences of brief delays in human-computer interaction. *International Journal of Human-Computer Studies, 67*(7), 561-570.

Teal, S.L. and Rudnicky, A.I. (1992). A performance model of systems delay and user strategy selection. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems – CHI '92* (pp. 295-305). New York: ACM.

Thüring, M. and Mahlke, S. (2007). Usability, aesthetics and emotions in human-technology interaction. *International Journal of Psychology, 42*(4), 253-264.

Tinsley, H.E.A., Hinson, J.A., Tinsley, D.J. and Holt, M.S. (1993). Attributes of leisure and work experiences. *Journal of Counseling Psychology, 40*(4), 447-455.

Trimmel, M., Meixner-Pendleton, M. and Haring, S. (2003). Stress response caused by system response time when searching for information on the internet. *Human Factors, 45*(4), 615-621.

Watson, D., Clark, L.A. and Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: the PANAS scale. *Journal of Personality and Social Psychology, 54*(6), 1063-1070.

8. General discussion

The main goal of the presented thesis was to investigate factors that may impact on the effectiveness of usability testing. Based on the 'four-factor framework of contextual fidelity' (Sauer, Seibel, & Rüttinger, 2010), two prominent aspects of usability testing were focused on. The first factor was the fidelity of the prototype, which was addressed by two of the presented studies. One of these compared prototype presentations on paper and on screen, while the other manipulated the developmental stage of a prototype in usability testing as perceived by participants. The second investigated factor was the testing environment, which was also addressed by two out of the three studies. One experimental study manipulated the social testing context and analysed the effects of the presence of two additional observers. The third study specifically focused on the usage domain of products, which had not been addressed in previous research. The main findings of this thesis strongly suggest that both factors of contextual fidelity, the prototype and the social testing environment, can have a relevant effect on outcomes of usability tests. The results are discussed and interpreted in the following chapters.

8.1. Summary and interpretation of the main results

8.1.1. Prototype fidelity in usability tests

The results of this thesis demonstrate that prototype fidelity can have a considerable effect on usability test outcomes, but these effects occur only under specific test conditions. In contrast to previous research (e.g. Catani & Biers, 1998; Sauer & Sonderegger, 2009; Sefelin et al., 2003; Virzi et al., 1996; Walker et al., 2002), the first empirical study showed that the medium of prototype presentation, i.e. whether the test system is presented on paper vs. on screen, can elicit very different participant behaviour and test performance. Task completion was significantly better for participants working with paper prototypes, but the finding only occurred if successful task accomplishment required reading a short paragraph of text. To achieve this improved performance, however, participants using the paper prototype needed significantly more time for the reading task than those working with the computer system. For other, mainly navigational tasks, these differences did not emerge. Consistent with previous research (e.g. Hamborg et al., 2009), the presented study confirmed that paper prototypes lead to less activity navigating different system parts as compared to a computer prototype. These results demonstrate that prototype fidelity is a

relevant factor in usability testing. As prototype presentation on paper caused an increase in active reading behaviour by test participants, this threatens the validity of test results. If system usage requires a certain amount of reading, choosing paper over on-screen presentation can lead to an overestimation of system usability and makes the detection of specific interface problems less likely.

The second study showed, that participant's perception of prototype fidelity seems to be quite robust against the impact of specific instructions by test facilitators. Even though the two groups of test subjects in the study were informed about very different developmental stages of the test system (early prototype vs. final system), there was little impact on the behavioural level. There were no main effects of this very explicit instruction on performance, and only for interaction efficiency as one out of three performance measures, there was an interaction effect in connection with observer presence. The finding indicated, that presenting the system to be at a very late developmental stage caused the social testing environment to have more of an impact on test participants' stress levels and behaviour, compared to the system being presented as an early prototype (see chapter 8.1.2.).

The outcome that study I did reveal main effects of prototype fidelity, while study II did not, can be explained by several factors. First, the theoretical approach of study I (i.e. the analysis of the affordances of the media of prototype presentation) made sure that a relevant aspect of prototype fidelity was manipulated, in relation to respective tasks and behavioural outcome measures. In contrast, the developmental stage of the prototype as instructed in study II might not be a core aspect of prototype fidelity. System usability might be more relevant in this respect, but even for this aspect other studies found effects primarily on the level of subjective evaluation (Bentley, 2000; Raita & Oulasvirta, 2011). Second, as the stimulus itself (i.e. the test system) was not a highly ambiguous artefact, which would have required a lot of interpretation on behalf of test participants, an instruction about its prototype fidelity might only have been a peripheral information, with limited impact on test outcomes. The primary determinant of participant behaviour under these circumstances can be assumed to be the stimulus itself, i.e. the test prototype, which was manipulated in study I, but not in study II. In addition, as the prototype in study II was a fully working system, the instruction about its early developmental stage could have contradicted participants' own perception, further reducing the impact of the respective manipulation.

For the subjective perception of usability, there were hardly any effects of objectively very different fidelity levels of the presented prototype (paper vs. on screen), nor for instructed stage of development (early prototype vs. final system). This is consistent

with previous studies in which subjective usability ratings remained surprisingly unaffected even by very different levels of prototype fidelity (Sauer et al., 2010; Sauer & Sonderegger, 2009). The only effect on the subjective level which occurred in the two studies was that the single-item measure of perceived usability was affected by the instructed developmental stage of the prototype. If the system was presented as an early prototype, its usability was rated significantly lower than if the system was instructed to be at a late developmental stage. A multi-item instrument to assess perceived usability remained unaffected. Overall, this demonstrates that perceived usability can be evaluated using prototypes of quite different fidelity levels, but that measurement instruments need to be selected carefully.

8.1.2. Testing environment in usability tests

With respect to the testing environment in usability tests, two subordinate factors in the four-factor framework of contextual fidelity have been investigated, social aspects of the testing context and the application domain. The second study presented in this thesis demonstrated that the presence of two additional observers in the test room during task completion caused higher stress reactions among participants. Consistent with previous research (Pruyn, Aasman, & Wyers, 1985; Sonderegger & Sauer, 2009), heart rate measurements proved to be valid indicators to distinguish not only between resting phase and task completion, but also showed significantly different cardiovascular activity due to the social testing environment. First, a higher increase in participants' mean heart rate from resting phase to task completion was registered, if two additional observers were present. Second, the analysis of heart rate variability showed there was a decrease of power in the low frequency band (LF) indicating higher stress levels, but only under the condition of observer presence. These results clearly demonstrate the impact of the social testing environment on test participants at the level of physiological processes.

As most usability tests in practice do not record physiological indicators of participants, the question remains whether the recorded stress levels would also cause differences at the level of data typically collected in usability tests, therefore threatening the validity of test results. While a previous study with the same laboratory set-up (Sonderegger & Sauer, 2009) found direct negative effects of observer presence on task completion rate, completion time and on participants' emotions, no main effects on any performance, emotion or subjective usability measures were found. However, in study II there was an interaction effect, which demonstrated that the social testing environment can have an impact on test outcomes. The presence of two additional observers caused lower

performance in terms of interaction efficiency, in comparison to participants working alone, but this effect only occurred if participants were instructed that the system was at a final developmental stage. If the system was introduced as an early prototype, there was no such effect. This interaction effect shows that the social testing environment is not independent of participants' perception of the test system: if the system is perceived as being fully developed, the costs of unfavourable test results are much higher. This, in turn, causes higher social pressure for good test results. Consequently, additional observer presence becomes more relevant and can impair performance during task completion. This finding can be explained by social facilitation theory (Guerin, 1993), which predicts lower performance when others are present for unfamiliar tasks, which participants faced in the test sessions of study II. The effect was shown in numerous studies to be most pronounced when participants feel watched and evaluated (Guerin, 1986), which must be presumed to be the case in a usability test, especially if it is a final system test as opposed to an early prototype test. The explanatory mechanisms are attention overload and distraction caused by others (Bond & Titus, 1983; Manstead & Semin, 1980).

The third presented study investigated a subordinate factor of testing environment in the 'four-factor framework of contextual fidelity', which had been neglected by previous research, the application domain. The study created two different testing environments, a work-oriented context and a leisure-oriented context, by manipulating the lab room set-up and the task wordings. In addition, participants were given a priming task, in which they were concentrating on either their work or leisure time, and the manipulation check confirmed the respective test conditions were experienced by participants as intended. In contrast to expectations, the two domains of work and leisure, as implemented in the experiment, had no impact on any of the recorded measures of performance, perceived usability, emotions, or task load. An independent manipulation of system usability by means of introducing response time delays obviously worked as expected, as there were negative effects on performance and participants' emotions. But contrary to expectations, there were no interaction effects with the application domain, which means that usability did not prove to be of higher importance in a more goal-oriented work domain as compared to a leisure context. One explanation of these results would be to distinguish between relevant and peripheral aspects of the usability testing context. Aspects of the test room design, which were unrelated to task requirements, and which had no distracting impact on test participants, could not be demonstrated to be relevant in usability testing. In addition, modifications of the wording of the test tasks with no impact on the comprehensibility of the tasks or on task requirements did not prove relevant for usability test outcomes. This means that the outcomes of usability tests may be quite robust to changes of peripheral

aspects of the testing procedure. However, on a more conceptual level, the notion of the 'application domain' and its meaning as an aspect of the testing environment needs to be revisited in light of these findings, and as this has implications for the four-factor framework of contextual fidelity, it will be elaborated in the next section.

8.2. Implications for the four-factor framework of contextual fidelity

The presented studies provided evidence in support of most of the investigated factors in the framework, i.e. demonstrating the significance of the system prototype and of the social testing environment. But, as there were some findings in contrast to the framework and its assumptions, these need to be discussed with respect to their implications.

As the presented empirical work could not find any effects of the subsidiary factor 'application domain', it might be helpful to analyse the concept in more detail. In the original publication introducing the four-factor framework (Sauer et al., 2010; p.131), the authors do not define this aspect of the testing environment, but refer to examples of different domains. They associate the concept with the work as opposed to the domestic domain, and with the public domain, in which 'walk-up and use products' play a role. As an exemplary dual-domain product, a mobile phone is mentioned, the use of which 'may be more strongly dominated by performance-oriented goals than in a leisure context, in which the joyful experience of the user with the product is of greater importance'. However, there are theoretical problems with the concept of the 'application domain' in the framework, as it cannot be distinguished from the other included factors. With the reference to goals for the work domain, the distinction from the factor 'task scenarios' becomes unclear, since in a usability test the tasks that are given to participants define the goals that are relevant in the situation. And it seems that all mentioned domains (work, leisure, and public) could be fully characterised by specific combinations of the other factors in the framework. E.g. a 'walk-up-and-use' scenario would include people from the general public not being trained for using a system (factor 'user characteristics'), other people watching or cuing for it ('social testing environment'), specific goals such as buying a train ticket ('task scenario'), with the specific touch screen system being specified ('system prototype'), etc. In previous research, which was able to find relevant usability problems due to different application domains, this can equally be explained in terms of the other factors of contextual fidelity. E.g. when a system was transferred without adaptations between different work domains, in this case from an office environment to a hospital setting (Bardram, 2005), the findings can be explained in terms of task scenarios not being adequately considered. And if Dahl et al.

(2010) report that only the inclusion of work domain specific aspects in a laboratory setting, specifically the work uniform, allowed them to identify specific usability requirements, this can be explained in terms of the physical testing environment. So the question remains, what it is exactly that characterises the construct 'application domain' independently of user characteristics, task scenarios, system prototype and physical and social testing environment. A potential conceptual distinction might be possible to conceive the application domain in terms of the attitudes which people have towards the specific setting or activities, as Beatty and Torbert (2003) suggest for the distinction between work and leisure. However, this definition of 'application domain' would either be concerned with aspects which are already covered by the factor 'user characteristics' and the respective subsidiary factor 'attitudes'. Or, it would imply the concept to be on a different theoretical level, which would require it to be a separate factor in the framework altogether, concerned with the cognitive interpretation of the testing situation by test participants (cf. Dahl et al., 2010; Paige & Morin, 2013). Further theoretical considerations are required for the concept 'application domain'.

The empirical study on the medium of prototype presentation demonstrated that some factors moderate the effects of others in the framework. The medium of prototype presentation (paper vs. on screen) had an effect on performance, but only for a specific task type. For a task that required reading of a small paragraph of text on a webpage, about 54% of participants working with the paper prototype were successful, while only just under 20% of those working with the computer prototype on screen were able to solve it. For other tasks not requiring reading activities, there was no such difference. In other words, task requirements were demonstrated to be crucial for prototype characteristics to have an impact on test outcomes. Since this interdependence of the effects of the factors in the framework was not considered in the design of previous studies investigating the impact of prototype presentation in usability testing (e.g. Virzi et al., 1996; Walker et al., 2002), the respective effects could not be found. In addition, a similar moderation can be expected for other factors as well. For example, task requirements may cause specific user competencies to become relevant or not for testing outcomes. And specific user characteristics (e.g. social anxiety) might make a difference, whether the social testing environment has an effect on test results. The authors of the four-factor framework already acknowledged moderating effects between the factors (Sauer et al., 2010), and the presented findings highlight the relevance of taking these into account.

In addition, the presented empirical work demonstrated that task aspects are relevant for the outcomes of usability tests, which are not reflected in the original framework. As it was only possible to find the different impacts of prototype presentations

on paper vs. on screen by introducing specific task requirements, the factor 'task scenarios' in the framework needs to be extended. Currently, the subsidiary factors 'breadth' and 'depth of task scenario' are included, but these do not fully describe task requirements, e.g. whether reading is necessary for solving a task. We therefore suggest to introduce the subsidiary factor 'task requirements', to complement the framework accordingly.

8.3. Strengths and limitations of the empirical studies

Several strengths and limitations of the reported studies can be identified. A strength of this thesis lies in the methodological approach which was used. Surprisingly, it was not possible to find previous studies investigating the fidelity of prototypes in usability tests, which selected test tasks according to firm hypotheses. In contrast to previous studies, which found no effects of the medium of prototype presentation (e.g. Sefelin et al., 2003; Virzi et al., 1996; Walker et al., 2002), the more theory-driven approach in study I proved successful to demonstrate the importance of this factor for usability test outcomes. In addition, every study which was presented used a multi-method approach. A consistent effort was made to use multiple measurements on different levels, including several objective performance indicators, scales for self-reported perceived usability, emotion or task load, and physiological indicators of participants' stress response. In contrast to very small sample sizes of less than 25 participants, which are often used in research on usability tests (e.g. Andreasen, Nielsen, Schrøder, & Stage, 2007; Bentley, 2000; Hartson, Castillo, Kelso, Kamler, & Neale, 1996; Thompson & Haake, 2004), the presented studies included between 60 and 80 participants in each of the usability experiments.

It can also be considered a strength of the reported studies, that in some of them, it was possible to realise a co-operation with private sector companies, to run usability tests on existing products and make test findings available to these partners. As it is definitely a challenge to set up tests which fulfil both requirements, those of conducting methodologically strict experimental research and obtaining relevant evaluation results of real products, some of the presented studies prove the approach is feasible.

On the other hand, some limitations of the reported studies need to be addressed. First, the research approach was focusing on quantitative measures for usability test outcomes only (e.g. performance, physiological indicators), and no analysis of qualitative usability problems was carried out. Further research is required to complement the presented work in this respect. In addition, in one of the studies the approach of recruiting participants was taking a pragmatic approach, thereby not fulfilling requirements to include test subjects from outside the university student population. This could partly

explain the results of the study, comparing work and leisure contexts, which would have required participants with more work experience than the average student can offer.

8.4. Implications for usability research and practice

Several implications for further research and practice in the field of usability testing can be identified based on the presented work.

Consider effects of the medium of prototype presentation. While it was previously recommended to choose between prototype presentation on paper vs. on screen purely based on practical considerations (e.g. Rudd et al., 1996), the presented work demonstrates this is not justified. As the presentation of prototypes in usability testing on paper can cause more reading activities by test participants, this effect may impact on test outcomes in research and practice. It is recommended to avoid testing with paper altogether, if the evaluated interface offers instructive texts, which are relevant for task success. Otherwise, system usability may be overestimated and relevant user problems may go undetected.

Use a hypothesis-driven experimental approach. Although it is standard procedure to conduct usability research with a firm set of hypotheses, some very core aspects of studies are commonly not designed according to any theoretical basis. One crucial factor in conducting research on usability tests was demonstrated in this work to be the test tasks. In contrast to previous studies, study I was able to identify substantial effects of the medium of prototype presentation on performance, but only because the test tasks have been carefully selected according to specific hypotheses. A more hypothesis-driven approach proved to be successful and is recommended for research in the field of usability studies.

Reduce social stress for test participants. The presented data of study II demonstrated that test participants show physiological stress reactions due to the social situation in a usability test. Therefore, it needs to be taken into account that the test situation may be a potentially stressful experience for test participants, especially if additional observers are present during the test. It is recommended to avoid that additional observers are present in the same room as participants, and care has to be taken to make test participants feel at ease.

Inclusion of qualitative indicators. Usability testing in practice is mostly formative (Bevan, 2006), but usability research is often taking a summative approach. The studies reported in this thesis were consistently recording quantitative data only, which was a successful approach to identify relevant effects, e.g. the impact of prototype presentation on paper as opposed to on screen. However, relevant questions remain unanswered, e.g.

whether there would be a different detection rate for specific types of usability problems. Therefore, this research needs to be complemented by studies including qualitative data.

For the four-factor framework of contextual fidelity (Sauer et al., 2010), the following recommendations can be given.

Moderating effects of factors. The presented work has demonstrated the moderating effects of some factors on the effects of others in the framework. Possible other moderations would need to be empirically tested, e.g. what the impact of task requirements on outcomes of usability tests would be, depending on user competencies, or whether specific user characteristics make the social testing context more relevant. More research is required to analyse in detail the complex interconnections between these aspects in usability testing.

Inclusion of task requirements. In the four-factor framework of contextual fidelity, test tasks are represented by the factor 'task scenarios', including the dimensions of 'breadth' and 'depth' as subordinate factors. However, relevant aspects of tasks in usability tests may not be captured well by this factor. As the first presented study showed, specific task characteristics, e.g. the requirement to read text, play an important role for test outcomes, and these go beyond task scenarios and their breadth and depth. On the basis of our research, a conceptual adaptation of the framework is recommended to include a more general factor 'task characteristics' instead of 'task scenarios'. Task scenario breadth and depth should be kept at the level of subordinate factors.

Theoretical foundation of the concept of 'application domain'. Currently, the concept of 'application domain' in the four-factor framework of contextual fidelity is difficult to distinguish from other factors (e.g. from 'task scenario' and 'social testing environment'). More theoretical and subsequently empirical work is needed to clarify the concept and to decide whether it should be kept in the framework.

8.5. Conclusion

Contextual fidelity in usability testing is highly relevant for valid test outcomes. The medium of prototype presentation can have substantial effects on participant behaviour and task performance, depending on task requirements. For tasks which involve reading short paragraphs of instructive text, paper prototypes should be avoided altogether in tests of system usability, as they may lead to an overestimation of participant performance and the non-detection of relevant usability problems. The general recommendation to decide about the medium of prototype presentation on basis of practicality needs to be revised.

Bibliography

Abrahamsson, P., Warsta, J., Siponen, M. T., & Ronkainen, J. (2003). New directions on agile methods: a comparative analysis. In *Proceedings of the 25th International Conference on Software Engineering (Vol. 6)* (pp. 244–254). IEEE Computer Society.

Alonso-Rios, D., Vazquez-Garcia, A., Mosqueira-Rey, E., & Moret-Bonillo, V. (2010). A context-of-use taxonomy for usability studies. *International Journal of Human-Computer Interaction, 26*(10), 941–970.

Alsos, O. A., & Dahl, Y. (2008). Toward a best practice for laboratory-based usability evaluations of mobile ICT for hospitals. In *Proceedings of the 5th Nordic Conference on Human-Computer Interaction Building Bridges - NordiCHI '08* (pp. 3–12). New York: ACM.

Andreasen, M. S., Nielsen, H. V., Schrøder, S. O., & Stage, J. (2007). What happened to remote usability testing? An empirical study of three methods. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems – CHI '07* (pp. 1405–1414). New York: ACM.

Bach, C., & Scapin, D. L. (2010). Comparing inspections and user testing for the evaluation of virtual environments. *International Journal of Human-Computer Interaction, 26*(8), 786–824.

Bailey, G. (1993). Iterative methodology and designer training in human-computer interface design. In *Proceedings of the INTERACT '93 and CHI '93 Conference on Human Factors in Computing Systems* (pp. 198–205). New York: ACM.

Bardram, J. E. (2005). The trouble with login: on usability and computer security in ubiquitous computing. *Personal and Ubiquitous Computing, 9*(6), 357–367.

Bargas-Avila, J. A., & Hornbæk, K. (2011). Old wine in new bottles or novel challenges? A critical analysis of empirical studies of user experience. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems – CHI '11* (pp. 2689–2698). New York: ACM.

Barnum, C. M. (2002). *Usability testing and research.* New York: Longman.

Bastien, J. M. C., Scapin, D. L., & Leulier, C. (1999). The ergonomic criteria and the ISO/DIS 9241-10 dialogue principles: a pilot comparison in an evaluation task. *Interacting with Computers, 11*(3), 299–322.

Beatty, J. E., & Torbert, W. R. (2003). The false duality of work and leisure. *Journal of Management Inquiry, 12*(3), 239–252.

Bentley, T. (2000). Biasing web site user evaluation: a study. In *Proceedings of the Australian Conference on Human-Computer Interaction - OZCHI 2000* (pp. 130–134). IEEE Computer Society.

Bevan, N. (1995). Usability is quality of use. In Y. Anzai, K. Ogawa, & H. Mori (Eds.), *Proceedings of the 6th International Conference on Human Computer Interaction (Vol. 20)*, (pp. 349–354). Amsterdam: Elsevier.

Bevan, N. (2001). International standards for HCI and usability. International *Journal of Human-Computer Studies, 55*(4), 533–552.

Bevan, N. (2006). Practical issues in usability measurement. *Interactions, 13*(6), 42–43.

Bevan, N. (2007). Usability evaluation in industry. In D.L. Scapin & E.L.-C. Law (Eds.), *Review, report and refine usability evaluation methods (R3 UEMs). Proceedings of the COST294-MAUSE 3rd International Workshop* (pp. 15–18). Athens, Greece.

Bevan, N. (2009). What is the difference between the purpose of usability and user experience evaluation methods? Proceedings of the *UXEM'09 Workshop, INTERACT 2009*. New York: ACM.

Bevan, N., & MacLeod, M. (1994). Usability measurement in context. *Behaviour & Information Technology, 13*(1), 132–145.

Beyer, H., & Holtzblatt, K. (1999). Contextual design. *Interactions, 6*(1), 32–42.

Bias, R. G., & Mayhew, D. J. (2005). *Cost-justifying usability: an update for the internet age* (2nd ed.). San Francisco, CA: Morgan Kaufman.

Blandford, A. E., Hyde, J. K., Green, T. R. G., & Connell, I. (2008). Scoping analytical usability evaluation methods: a case study. *Human-Computer Interaction, 23*(3), 278–327.

Boehm, B. W. (1988). A spiral model of software development and enhancement. *Computer, 21*(5), 61–72.

Bond, C. F., & Titus, L. J. (1983). Social facilitation: a meta-analysis of 241 studies. *Psychological Bulletin, 94*(2), 265–292.

Boren, M. T., & Ramey, J. (2000). Thinking aloud: reconciling theory and practice. *IEEE Transactions on Professional Communication, 43*(3), 261–278.

Bruun, A., Gull, P., Hofmeister, L., & Stage, J. (2009). Let your users do the testing: a comparison of three remote asynchronous usability testing methods. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems – CHI '*09 (pp. 1619–1628). New York: ACM.

Carroll, J. M. (2001). Community computing as human-computer interaction. *Behaviour & Information Technology, 20*(5), 307–314.

Carroll, J. M. (2004). Beyond fun. *Interactions, 11*(5), 38–40.

Catani, M. B., & Biers, D. W. (1998). Usability evaluation and prototype fidelity: users and usability professionals. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (pp. 1331–1335). Human Factors and Ergonomics Society.

Ceaparu, I., Lazar, J., Bessiere, K., Robinson, J., & Shneiderman, B. (2004). Determining causes and severity of end-user frustration. *International Journal of Human-Computer Interaction, 17*(3), 333–356.

Chung, L., & Leite, J. C. S. P. (2009). On non-functional requirements in software engineering. In A. T. Borgida, V. K. Chaudhri, P. Giorgini, & E. S. Yu (Eds.), *Conceptual Modeling: Foundations and Applications (Vol. 5600)*, (pp. 363–379). Berlin, Heidelberg: Springer-Verlag.

CISQ. (2010). *CISQ annual report of progress 2010.* Consortium for IT Software Quality.

Cockton, G., & Woolrych, A. (2001). Understanding inspection methods: lessons from an assessment of heuristic evaluation. In A. Blandford, J. Vanderdonckt, & P. D. Gray (Eds.), *People and computers XV: interactions without frontiers: joint proceedings of HCI 2001 and IHM 2001* (pp. 171–191). London: Springer-Verlag.

Cockton, G., & Woolrych, A. (2009). Comparing usability evaluation methods: strategies and implementation. In E. L-C. Law, D. L. Scapin, G. Cockton, M. Springett, C. Stary, & M. Winckler (Eds.), *Maturation of usability evaluation methods: retrospect and prospect. Final Report of COST294-MAUSE Working Group 2* (pp. 18–82). Toulouse: IRIT Press.

Cook, T. D., Campbell, D. T., & Peracchio, L. (1990). Quasi experimentation. In M.D. Dunnette & L.M. Hough (Eds.), *Handbook of industrial and organizational psychology, Vol.1* (2nd ed., pp. 491–576). Palo Alto, CA: Consulting Psychologists Press.

Cronbach, L. J. (1947). Test 'reliability': its meaning and determination. *Psychometrika, 12*(1), 1–16.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281–302.

Dahl, Y., Alsos, O. A., & Svanæs, D. (2010). Fidelity considerations for simulation-based usability assessments of mobile ICT for hospitals. *International Journal of Human-Computer Interaction, 26*(5), 445–476.

Dalton, J. P., Manning, H., Mines, C., & Dorsey, M. (2002). Packaged apps fail the usability test. *TechStrategy Research* (pp. 1–19). Forrester Research.

Desurvire, H. W. (1994). Faster, cheaper!! Are usability inspection methods as effective as empirical testing? In J. Nielsen & R. L. Mack (Eds.), *Usability inspection methods* (pp. 173–202). New York: John Wiley & Sons.

Dicks, R. S. (2002). Mis-usability: on the uses and misuses of usability testing. In *Proceedings of the 20th Annual International Conference on Computer Documentation - SIGDOC '02,* (pp. 26–30). New York: ACM.

Dolan, W. R., & Dumas, J. S. (1999). A flexible approach to third-party usability. *Communications of the ACM, 42*(5), 83–85.

Dray, S. M. (1995). The importance of designing usable systems. *Interactions, 2*(1), 17–20.

Dray, S. M., & Karat, C.-M. (1994). Human factors cost justification for an internal development project. In R.G. Bias & D. J. Mayhew (Eds.), *Cost-justifying usability* (pp. 111–122). San Diego, CA: Academic Press.

Düchting, M., Zimmermann, D., & Nebe, K. (2007). Incorporating user centered requirement engineering into agile software development. In J. A. Jacko (Ed.), *Proceedings of the 12th International Conference on Human-Computer Interaction: Interaction Design and Usability - HCI'07* (pp. 58–67). Berlin, Heidelberg: Springer-Verlag.

Dumas, J. S., & Redish, J. C. (1999). *A practical guide to usability testing.* Norwood, NJ: Ablex.

Duncker, K., & Lees, L. S. (1945). On problem-solving. *Psychological Monographs, 58*(5).

Fairbanks, R. J., Caplan, S. H., Bishop, P. A., Marks, A. M., & Shah, M. N. (2007). Usability study of two common defibrillators reveals hazards. *Annals of Emergency Medicine, 50*(4), 424–432.

Fallman, D. (2010). A different way of seeing: Albert Borgmann's philosophy of technology and human–computer interaction. *AI & Society, 25*(1), 53–60.

Fernandez, A., Insfran, E., & Abrahão, S. (2011). Usability evaluation methods for the web: a systematic mapping study. *Information and Software Technology, 53*(8), 789–817.

Følstad, A., Law, E. L.-C., & Hornbæk, K. (2012). Analysis in practical usability evaluation: a survey study. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems – CHI '*12 (pp. 2127–2136). New York: ACM.

Fu, L., Salvendy, G., & Turley, L. (2002). Effectiveness of user testing and heuristic evaluation as a function of performance classification. *Behaviour & Information Technology, 21*(2), 137–141.

Gediga, G., Hamborg, K.-C., & Duntsch, I. (2002). Evaluation of software systems. In A. Kent & J. G. Williams (Eds.), *Encyclopedia of Computer Science and Technology (Vol. 45*), (pp. 127–153). New York: Marcel Dekker, Inc.

Gerhardt-Powals, J. (1996). Cognitive engineering principles for enhancing human-computer performance. *International Journal of Human-Computer Interaction, 8*(2), 189–211.

Good, M., Spine, T. M., Whiteside, J., & George, P. (1986). User-derived impact analysis as a tool for usability engineering. In M. Mantei & P. Orbeton (Eds.), *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems – CHI '*86 (pp. 241–246). New York: ACM.

Gould, J. D., & Lewis, C. (1985). Designing for usability: key principles and what designers think. *Communications of the ACM, 28*(3), 300–311.

Gray, W. D., & Salzman, M. C. (1998). Damaged merchandise? A review of experiments that compare usability evaluation methods. *Human-Computer Interaction, 13*(3), 203–261.

Grubaugh, B., Thomas, S., & Weinberg, J. (2005). Effects of the testing environment on user performance in software usability testing. In M. H. Hamza (Ed.), *Proceedings of the IASTED International Conference on Human-Computer Interaction* (pp. 39–43). Anaheim: ACTA Press.

Grudin, J. (2005). Three faces of human-computer interaction. *IEEE Annals of the History of Computing, 27*(4), 46–62.

Guerin, B. (1986). Mere presence effects in humans: a review. *Journal of Experimental and Social Psychology, 22*(1), 38–77.

Guerin, B. (1993). *Social facilitation.* Cambridge, England: University Press.

Hall, R. R. (2001). Prototyping for usability of new technology. *International Journal of Human-Computer Studies, 55*(4), 485–501.

Hamborg, K.-C., Klassen, A., & Volger, M. (2009). Zur Gestaltung und Effektivität von Prototypen im Usability-Engineering. In H. Wandke, S. Kain, & D. Struve (Eds.), *Mensch & Computer 2009: Grenzenlos frei!?* (pp. 263–272). München: Oldenbourg Verlag.

Hammontree, M., Weiler, P., & Nayak, N. (1994). Remote usability testing. *Interactions, 1*(3), 21–25.

Harris, E., Weinberg, J., Thomas, S., & Gaeslin, D. (2005). Effects of social facilitation and electronic monitoring on usability testing. In *Proceedings of the Usability Professionals Association Conference*, 27 June - 1 July 2005 Quebec, published on CD.

Hartson, H. R., Andre, T. S., & Williges, R. C. (2001). Criteria for evaluating usability evaluation methods. *International Journal of Human-Computer Interaction, 13*(4), 373–410.

Hartson, H. R., Castillo, J. C., Kelso, J., Kamler, J., & Neale, W. C. (1996). Remote evaluation: the network as an extension of the usability laboratory. In M. J. Tauber (Ed.), *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems – CHI '96* (pp. 228–235). New York: ACM.

Hasan, L., Morris, A., & Probets, S. (2012). A comparison of usability evaluation methods for evaluating e-commerce websites. *Behaviour & Information Technology, 31*(7), 707–737.

Hassenzahl, M. (2004). The interplay of beauty, goodness, and usability in interactive products. *Human-Computer Interaction, 19*(4), 319–349.

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

Helander, G. A. (1981). Improving system usability for business professionals. *IBM Systems Journal, 20*(3), 294–305.

Helms, J., Arthur, J., Hix, D., & Hartson, H. (2006). A field study of the Wheel - a usability engineering process model. *Journal of Systems and Software, 79*(6), 841–858.

Hendrick, H. (2000). The technology of ergonomics. *Theoretical Issues in Ergonomics Science, 1*(1), 22–33.

Hertzum, M. (1999). User testing in industry: a case study of laboratory, workshop, and field tests. In A. Kobsa and C. Stephanidis (Eds.), *User interfaces for all. Proceedings 5. ERCIM Workshop* (pp. 59–72).

Hertzum, M. (2010). Frustration: a common user experience. In M. Hertzum & M. Hansen (Eds.), *DHRS2010: Proceedings of the 10th Danish Human-Computer Interaction Research Symposium, Vol. 132, Computer Science Research Reports* (pp. 11–14). Roskilde: Roskilde Universitet.

Hertzum, Morten, & Jacobsen, N. E. (2003). The evaluator effect: a chilling fact about usability evaluation methods. *International Journal of Human-Computer Interaction, 15*(1), 183–204.

Hertzum, Morten, Molich, R., & Jacobsen, N. E. (2013). *Behaviour & Information Technology*. Advance online publication. DOI: 10.1080/0144929X.2013.783114.

Hoc, J.-M. (2001). Towards ecological validity of research in cognitive ergonomics. *Theoretical Issues in Ergonomics Science, 2*(3), 278–288.

Hornbæk, K. (2010). Dogmas in the assessment of usability evaluation methods. *Behaviour & Information Technology, 29*(1), 97–111.

Hornbæk, K., & Frøkjær, E. (2008). Comparison of techniques for matching of usability problem descriptions. *Interacting with Computers, 20*(6), 505–514.

Hornbæk, K., Høegh, R. T., Pedersen, M. B., & Stage, J. (2007). Use case evaluation (UCE): a method for early usability evaluation in software development. In *Proceedings of the 11th IFIP TC 13 International Conference on Human-Computer Interaction - INTERACT'07* (pp. 578–591). Berlin, Heidelberg: Springer-Verlag.

Hosseini-Khayat, A., Hellmann, T. D., & Maurer, F. (2010). Distributed and automated usability testing of low-fidelity prototypes. In *Proceedings of the 2010 Agile Conference* (pp. 59–66). IEEE Computer Society.

Houde, S., & Hill, C. (1997). What do prototypes prototype? In M. G. Helander, T. K. Landauer, & P. V. Pradhu (Eds.), *Handbook of human-computer interaction* (2nd ed., pp. 367–381). Amsterdam: North-Holland.

ISO. (1998). ISO 9241, Ergonomic requirements for office work with visual display terminals (VDTs) - Part 11, Guidance on usability, International standard.

ISO. (2001). ISO 9216-1, Software engineering. Product quality - Part 1, Quality model.

ISO. (2010). ISO 9241, Ergonomics of human-system interaction - Part 210, Human-centred design for interactive systems.

Jacobsen, N. E., Hertzum, M., & John, B. E. (1998). The evaluator effect in usability tests. In CHI '98 Conference Summary on Human Factors in Computing Systems (pp. 255–256). New York: ACM.

Jaspers, M. W. M. (2009). A comparison of usability methods for testing interactive health technologies: methodological aspects and empirical evidence. *International Journal of Medical Informatics, 78*(5), 340–353.

Jeffries, R., & Desurvire, H. (1992). Usability testing vs. heuristic evaluation: was there a contest? *SIGCHI Bulletin, 24*(4), 39–41.

Jeffries, R., Miller, J. R., Wharton, C., & Uyeda, K. (1991). User interface evaluation in the real world: a comparison of four techniques. In S. P. Robertson, G. M. Olson, & J. S. Olson (Eds.), *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems – CHI '*91 (pp. 119–124). New York: ACM.

John, B. E., & Marks, S. J. (1997). Tracking the effectiveness of usability evaluation methods. *Behaviour & Information Technology, 16*(4/5), 188–202.

Jokela, T. (2004). When good things happen to bad products: where are the benefits of usability in the consumer appliance market? *Interactions, 11*(6), 28–35.

Kaikkonen, A., Kallio, T., Kekäläinen, A., Kankainen, A., & Cankar, M. (2005). Usability testing of mobile applications: a comparison between laboratory and field testing. *Journal of Usability Studies, 1*(1), 4–16.

Karat, C.-M., Campbell, R., & Fiegel, T. (1992). Comparison of empirical testing and walkthrough methods in user interface evaluation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems – CHI '*92 (pp. 397–404). New York: ACM.

Karlson, A. K., Meyers, B. R., Jacobs, A., Johns, P., & Kane, S. K. (2009). Working overtime: patterns of smartphone and PC usage in the day of an information worker. In H. Tokuda, M. Beigl, A. Friday, A. J. B. Brush, & Y. Tobe (Eds.), *Proceedings of the 7th International Conference on Pervasive Computing* (pp. 398–405). Berlin, Heidelberg: Springer.

Karvonen, H., Saariluoma, P., & Kujala, T. (2010). A preliminary framework for differentiating the paradigms of human-technology interaction research. In 3rd International Conference on Advances in Computer-Human Interactions (pp. 7–12). IEEE Computer Society.

Kaspar, K., Hamborg, K.-C., Sackmann, T., & Hesselmann, J. (2010). Die Effektivität formativer Evaluation bei der Entwicklung gebrauchstauglicher Software - eine Fallstudie. *Zeitschrift für Arbeits- und Organisationspsychologie, 54*(1), 29–38.

Kessner, M., Wood, J., Dillon, R. F., & West, R. L. (2001). On the reliability of usability testing. *CHI '01 Extended Abstracts on Human Factors in Computing Systems* (pp. 97-98). New York: ACM.

Kjeldskov, J., & Skov, M. B. (2003). Creating realistic laboratory settings: comparative studies of three think-aloud usability evaluations of a mobile system. In *Proceedings of the 9th IFIP TC13 International Conference on Human-Computer Interaction, Interact* (pp. 663–670). Zurich: IOS Press.

Kjeldskov, J., & Skov, M. B. (2007). Studying usability in sitro: simulating real world phenomena in controlled environments. *International Journal of Human-Computer Interaction, 22*(1), 7–36.

Kurosu, M. (2007). Concept of usability revisited. In J. Jacko (Ed.), *Human-Computer Interaction, Part I, HCII 2007, LNCS 4550* (pp. 579–586). Berlin/Heidelberg: Springer-Verlag.

Kurosu, M., & Kashimura, K. (1995). Apparent usability vs. inherent usability. In *Conference Companion on Human Factors in Computing Systems - CHI '95* (pp. 292–293). New York: ACM.

Lavery, D., Cockton, G., & Atkinson, M. P. (1997). Comparison of evaluation methods using structured usability problem reports. *Behaviour & Information Technology, 16*(4-5), 246–266.

Law, E. L.-C., & Hvannberg, E. T. (2002). Complementarity and convergence of heuristic evaluation and usability test: a case study of UNIVERSAL Brokerage Platform. In *Proceedings of the 2nd Nordic Conference on Human-Computer Interaction - NordiCHI '02* (pp. 71–80). New York: ACM.

Law, E. L.-C., Roto, V., Hassenzahl, M., Vermeeren, A. P. O. S., & Kort, J. (2009). Understanding, scoping and defining user eXperience: a survey approach. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems – CHI '09* (pp. 719–728). New York: ACM.

Lazar, J., Jones, A., & Shneiderman, B. (2006). Workplace user frustration with computers: an exploratory investigation of the causes and severity. *Behaviour & Information Technology, 25*(3), 239–251.

Lewis, C., & Mack, R. (1982). Learning to use a text processing system: evidence from "thinking aloud" protocols. In *Proceedings of the 1982 Conference on Human Factors in Computing Systems* (pp. 387–392). New York: ACM.

Lewis, J. R. (2006). Usability testing. In G. Salvendy (Ed.), *Handbook of human factors and ergonomics* (pp. 1275–1316). Hoboken, NJ: Wiley.

Lindgaard, G. (2006). Notions of thoroughness, efficiency, and validity: are they valid in HCI practice? *International Journal of Industrial Ergonomics, 36*(12), 1069–1074.

Maguire, M. (2001). Methods to support human-centred design. *International Journal of Human-Computer Studies, 55*(4), 587–634.

Mankoff, J., Dey, A. K., Hsieh, G., Kientz, J., Lederer, S., & Ames, M. (2003). Heuristic evaluation of ambient displays. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems – CHI '*03 (pp. 169–176). New York: ACM.

Manstead, A. S. R., & Semin, G. R. (1980). Social facilitation effects: mere enhancement of dominant responses. *British Journal of Social and Clinical Psychology, 19*(2), 119–136.

Mao, J.-Y., Vredenburg, K., Smith, P. W., & Carey, T. (2005). The state of user-centered design practice. *Communications of the ACM, 48*(3), 105–109.

Matern, U., & Koneczny, S. (2007). Safety, hazards and ergonomics in the operating room. *Surgical Endoscopy, 21*(11), 1965–1969.

Mayhew, D. J. (1999). *The usability engineering lifecycle - a practitioner's handbook for user interface design.* San Diego, CA: Morgan Kaufman.

McCurdy, M., Connors, C., Pyrzak, G., Kanefsky, B., & Vera, A. (2006). Breaking the fidelity barrier. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems – CHI '06* (pp. 1233–1242). New York: ACM.

Meszaros, G., & Aston, J. (2006). Adding usability testing to an agile project. In *Proceedings of AGILE '06 Conference* (pp. 289–294). IEEE Computer Society.

Minow, M. (1978). Minicomputer timesharing performance and usability. *SIGMINI Newsletter, 4*(3), 4–16.

Molich, R., Bevan, N., Curson, I., Butler, S., Kindlund, E., Miller, D., & Kirakowski, J. (1998). Comparative evaluation of usability tests. In *Proceedings of the Usability Professionals Association 1998 Conference (Vol. 2)* (pp. 189–200). UPA.

Molich, R., & Dumas, J. S. (2008). Comparative usability evaluation (CUE-4). *Behaviour & Information Technology, 27*(3), 263–281.

Molich, R., Ede, M. R., Kaasgaard, K., & Karyukin, B. (2004). Comparative usability evaluation. *Behaviour & Information Technology, 23*(1), 65–74.

Molich, R., & Nielsen, J. (1990). Improving a human-computer dialogue. *Communications of the ACM, 33*(3), 338–348.

Molich, R., Thomsen, A. D., Karyukina, B., Schmidt, L., Ede, M., van Oel, W., & Arcuri, M. (1999). Comparative evaluation of usability tests. In *CHI '99 Extended Abstracts on Human Factors in Computing Systems* (pp. 83–84). New York: ACM.

Nielsen, C. M., Overgaard, M., Pedersen, M. B., Stage, J., & Stenild, S. (2006). It's worth the hassle!: the added value of evaluating the usability of mobile systems in the field. In *Proceedings of the 4th Nordic Conference on Human-Computer Interaction Changing Roles - NordiCHI '06* (pp. 272–280). New York: ACM.

Nielsen, J. (1993). *Usability engineering.* San Diego, CA: Academic Press Prof., Inc.

Nielsen, J. (1994). Enhancing the explanatory power of usability heuristics. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems – CHI '94* (pp. 152–158). New York: ACM.

Nielsen, J. (1995). Applying discount usability engineering. *IEEE Software, 12*(1), 98–100.

Nielsen, J., & Molich, R. (1990). Heuristic evaluation of user interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems – CHI '*90 (pp. 249–256). New York: ACM.

Nielsen, J., & Phillips, V. L. (1993). Estimating the relative usability of two interfaces: heuristic, formal, and empirical methods compared. In *Proceedings of the INTERACT '93 and CHI '93 Conference on Human Factors in Computing Systems - CHI '93* (pp. 214–221). New York: ACM.

Nielsen, L., & Madsen, S. (2012). The usability expert's fear of agility: an empirical study of global trends and emerging practices. In *Proceedings of the 7th Nordic Conference on Human-Computer Interaction - NordiCHI '12* (pp. 261–264). New York: ACM.

Paige, J. B., & Morin, K. H. (2013). Simulation fidelity and cueing: a systematic review of the literature. *Clinical Simulation in Nursing, 9*(11), 481–489.

Parsons, D., Lal, R., Ryu, H., & Lange, M. (2007). Software development methodologies, agile development and usability engineering. In *Proceedings of the 18th Australasian Conference on Information Systems - ACIS*. Paper 32.

Perez, S. (2011). Smartphones outsell PCs. *New York Times*, Feb 8, 2011.

Petrie, H., & Power, C. (2012). What do users really care about? A comparison of usability problems found by users and experts on highly interactive websites. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems – CHI '*12 (pp. 2107–2116). New York: ACM.

Pruyn, A., Aasman, J., & Wyers, B. (1985). Social influences on mental processes and cardiovascular activity. In J. F. Orlebeke, G. Mulder, & L. J. P. Van Dornen (Eds.), *The psychophysiology of cardiovascular control (models, methods, and data)* (pp. 865–877). New York: Plenum Press.

Quesenbery, W. (2003). Dimensions of usability. In M. Albers & B. Mazur (Eds.), *Content and complexity: information design in technical communication* (pp. 81–102). Mahwah, NJ: Lawrence Erlbaum.

Redish, J. (2007). Expanding usability testing to evaluate complex systems. *Journal of Usability Studies, 2*(3), 102–111.

Richter, M., & Flückiger, M. (2007). *Usability Engineering kompakt: Benutzbare Software gezielt entwickeln.* München: Spektrum Akademischer Verlag.

Rosenbaum, S. (2008). The future of usability evaluations: increasing impact on value. In Effie Lai-Chong Law, E. Hvannberg, & G. Cockton (Eds.), *Maturing usability: quality in software, interaction and value* (pp. 344–378). London: Springer.

Rosenbaum, S., Rohn, J. A., & Humburg, J. (2000). A toolkit for strategic usability: results from workshops, panels, and surveys. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems – CHI '00* (pp. 337–344). New York: ACM.

Rowley, D. E. (1994). Usability testing in the field: bringing the laboratory to the user. In B. Adelson, S. Dumais, & J. Olson (Eds.), *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems – CHI '94* (pp. 252–257). New York: ACM.

Royce, W. W. (1970). Managing the development of large software systems: concepts and techniques. In *Technical Papers of Western Electronic Show and Convention - WesCon* (pp. 1–9). IEEE Computer Society.

Rubin, J., & Chisnell, D. (2008). *Handbook of usability testing* (2nd ed.). Indianapolis, IN: Wiley Publishing Inc.

Rudd, J., Stern, K., & Isensee, S. (1996). Low vs. high-fidelity prototyping debate. *Interactions, 3*(1), 76–85.

Ruthford, M. A., & Ramey, J. A. (2000). The design response to usability test findings: a case study based on artifacts and interviews. In *IEEE International Professional Communication Conference* (pp. 315–323). Piscataway, NJ: IEEE Computer Society.

Säde, S., Nieminen, M., & Riihiaho, S. (1998). Testing usability with 3D paper prototypes - Case Halton system. *Applied Ergonomics, 29*(1), 67–73.

Sarodnick, F., & Brau, H. (2011). *Methoden der Usability Evaluation* (2nd ed.). Bern: Hans Huber.

Sauer, J., Franke, H., & Rüttinger, B. (2008). Designing interactive consumer products: utility of paper prototypes and effectiveness of enhanced control labelling. *Applied Ergonomics, 39*(1), 71–85.

Sauer, J., Seibel, K., & Rüttinger, B. (2010). The influence of user expertise and prototype fidelity in usability tests. *Applied Ergonomics, 41*(1), 130–140.

Sauer, J., & Sonderegger, A. (2009). The influence of prototype fidelity and aesthetics of design in usability tests: effects on user behaviour, subjective evaluation and emotion. *Applied Ergonomics, 40*(4), 670–677.

Sears, A., & Hess, D. J. (1999). Cognitive walkthroughs: understanding the effect of task-description detail on evaluator performance. *International Journal of Human-Computer Interaction, 11*(3), 185–200.

Sefelin, R., Tscheligi, M., & Giller, V. (2003). Paper prototyping - What is it good for? A comparison of paper- and computer-based low-fidelity prototyping. In *CHI '03 Extended Abstracts on Human Factors in Computing Systems* (pp. 778–779). New York: ACM.

Seffah, A., Donyaee, M., Kline, R. B., & Padda, H. K. (2006). Usability measurement and metrics: A consolidated model. *Software Quality Journal, 14*(2), 159–178.

Seffah, A., & Habieb-Mammar, H. (2009). Usability engineering laboratories: limitations and challenges toward a unifying tools/practices environment. *Behaviour & Information Technology, 28*(3), 281–291.

Shackel, B. (1991). Usability - context, framework, definition, design and evaluation. In B. Shackel & S. Richardson (Eds.), *Human factors for informatics usability* (pp. 21–37). Cambridge: University Press.

Sjøberg, D. I. K., Anda, B., Arisholm, E., Dybå, T., Jørgensen, M., Karahasanovic, A., … Vokác, M. (2002). Conducting realistic experiments in software engineering. In *Proceedings of the 2002 International Symposium on Empirical Software Engineering* (pp. 17–26). IEEE Computer Society.

Snyder, C. (2003). *Paper prototyping: the fast and easy way to design and refine user interfaces.* San Francisco: Morgan Kaufman.

Sonderegger, A., & Sauer, J. (2009). The influence of laboratory set-up in usability tests: effects on user performance, subjective ratings and physiological measures. *Ergonomics, 52*(11), 1350–1361.

Thomas, J. C., & Kellogg, W. A. (1989). Minimizing ecological gaps in interface design. *IEEE Software, 6*(1), 78–86.

Thompson, K. E., & Haake, A. R. (2004). Here, there, anywhere: remote usability testing that works. *Proceedings of the 5th Conference on Information Technology Education* (pp. 132–137). New York: ACM.

Thüring, M., & Mahlke, S. (2007). Usability, aesthetics and emotions in human-technology interaction. *International Journal of Psychology, 42*(4), 253–264.

Tractinsky, N. (2000). What is beautiful is usable. *Interacting with Computers, 13*(2), 127–145.

Uldall-Espersen, T., Frøkjær, E., & Hornbæk, K. (2008). Tracing impact in a usability improvement process. *Interacting with Computers, 20*(1), 48–63.

Venturi, G., & Troost, J. (2004). Survey on the UCD integration in the industry. In *Proceedings of the 3rd Nordic Conference on Human-Computer Interaction - NordiCHI '04* (pp. 449–452). New York: ACM.

Virzi, R. A., Sokolov, J. L., & Karis, D. (1996). Usability problem identification using both low- and high-fidelity prototypes. In M. J. Tauber (Ed.), *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems – CHI '*96 (pp. 236–243). New York: ACM.

Vredenburg, K., Isensee, S., & Righi, C. (2002). *User-centered design: an integrated approach.* Upper Saddle River, NJ: Prentice Hall.

Vredenburg, K., Mao, J.-Y., Smith, P. W., & Carey, T. (2002). A survey of user-centered design practice. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems – CHI '02* (pp. 471-478). New York: ACM.

Walker, M., Takayama, L., & Landay, J. A. (2002). High-fidelity or low-fidelity, paper or computer? Choosing attributes when testing web prototypes. In *Proceedings of the Human Factors and Ergonomics Society 46th Annual Meeting* (pp. 661–665). Human Factors and Ergonomics Society.

Wenger, M. J., & Spyridakis, J. H. (1989). The relevance of reliability and validity to usability testing. *IEEE Transactions on Professional Communication, 32*(4), 265–271.

Wichansky, A. M. (2000). Usability testing in 2000 and beyond. *Ergonomics, 43*(7), 998–1006.

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

Woolrych, A., Hornbæk, K., Frøkjær, E., & Cockton, G. (2011). Ingredients and meals rather than recipes: a proposal for research that does not treat usability evaluation methods as indivisible wholes. *International Journal of Human-Computer Interaction, 27*(10), 940–970.

Young-Corbett, D. E., Nussbaum, M. A., & Winchester, W. W. (2010). Usability evaluation of drywall sanding tools. *International Journal of Industrial Ergonomics, 40*(1), 112–118.

Résumé of Andreas Uebelbacher

Employment Since Nov. 2012	**Foundation Access for all, Zurich** Head of accessibility services.
Employment / doctorate Jan. 2009 - Feb. 2012	**Cognitive Ergonomics, University of Fribourg** Researcher in project on usability methodology, funded by the Swiss National Science Foundation.
Employment Jan. 2006 - Sept. 2008	**Swisscom Fixnet AG / Swisscom AG, Zurich** Team leader Usability, Design & Concept (usability consulting and interaction design).
Employment March 2002 - Dec. 2005	**Bluewin AG, Zurich** Usability Consultant and leader of the usability team.
Employment Aug. 2000 - May 2001	**Employment Service, Sheffield (UK)** User researcher in an internet project directed at jobseekers.
Internships Aug. 1996 - April 1998	**University of Bern / Employment Service, Sheffield (UK)** Various research projects in Switzerland and the UK, e.g. for National Swiss Railways (analysis of different shift rotation systems of train drivers), NOK (job analysis in a Swiss nuclear power plant).
Education Oct. 1991 - July 2000	**University of Bern** Swiss licentiate in Psychology (lic. phil. hist.), awarded with 'summa cum laude'. Specialisation in Work & Organisational Psychology, Social Psychology and Philosophy of Science.
Diploma thesis	Quantitative-statistical thesis on resources of functional autonomy in old age, using structural equation modelling.
Education 1986 – 1989	Grammar school Muttenz, BL
Compulsory education 1982 – 1986	Secondary school Münchenstein, BL
Compulsory education 1977 – 1982	Elementary school Münchenstein, BL

Conference presentations, publications and jury memberships (selection)

Jury membership Since 2011	Best of Swiss Web Awards (CH) Jury member in the category 'usability'
Presentation Oct. 2011	Annual Meeting of the Human Factors and Ergonomics Society Europe Chapter 2011, Leeds (UK) 'The medium of prototype presentation in usability testing: effects on performance, psychophysiology and subjective evaluation.'
Presentation Sept. 2011	Monthly Event of the Software Ergonomics/SwissCHI, Zurich 'Die Königsmethode im Usability Engineering: Aktuelle Forschungsergebnisse zu Einflussfaktoren in Usability-Tests.'
Presentation Oct. 2010	Annual Meeting of the Human Factors and Ergonomics Society Europe Chapter 2010, Berlin (D) 'Formative vs. summative usability testing: the effects on test participants´ performance, subjective evaluations and physiology.'
Presentation Aug. 2009	Congress of the Swiss Psychological Society 2009, Neuchâtel 'The influence of design aesthetics in usability testing: effects on user performance and perceived usability.'
Presentation Sept. 2005	Mensch & Computer 2005, Linz (A) ‚Nachhaltige organisatorische Verankerung von Usability im Unternehmen – ein Fallbeispiel aus der Schweiz.'
Publication	Uebelbacher, A., Sonderegger, A. & Sauer, J. (2013). Effects of perceived prototype fidelity in usability testing under different conditions of observer presence. *Interacting with Computers, 25*(1), 91-101. [Included in this doctoral thesis].
Publication	Sonderegger, A., Zbinden, G., Uebelbacher, A. & Sauer, J. (2012). The influence of product aesthetics and usability over the course of time: a longitudinal field experiment. *Ergonomics, 55*(7), 713-730.
Publication	Perrig-Chiello, P., Perrig, W.J., Uebelbacher, A. & Staehelin, H.B. (2006). Impact of physical and psychological resources on functional autonomy in old age. *Psychology, Health & Medicine, 11*(4), 470-482.

Candidate's declaration of authenticity

I hereby certify that this thesis has been written by me, that I have not received any undue assistance from third parties, and that it has not been submitted in any previous application for a higher degree.

Andreas Uebelbacher