Influencing Factors in Usability Tests

The testing Situation, the Product Prototype, and the Test User

Sonderegger Andreas

Dissertation zur Erlangung der Doktorwürde an der Philosophischen Fakultät der Universität Freiburg (Schweiz).

Genehmigt von der Philosophischen Fakultät auf Antrag der Professoren Jürgen Sauer (1. Gutachter) und Andy Tattersall (2. Gutachter).
Freiburg, den 2. November 2010 (Datum der Thesenverteidigung).
Dekan Prof. Dr. Thomas Austenfeld

Thank you Jürgen

Merci Sora

Für Nils

CONTENT

1 Background and rationale of thesis

The role technology plays in our life is getting more and more important. It pervades our daily life and activities since many of the products we use every day are operated via a more or less complex interface; the home stereo system, the coffee machine, the micro wave oven, the DVD-recorder, the personal digital assistant (PDA), the computer but also products outside the own household such as the ATM machine or the ticket vending machine. Coming along with the constantly increasing technological progress, products are being continuously upgraded and developed and their inherent features and functions become more complex. Whereas a telephone, at its origin, served exclusively as a device allowing two persons to talk to each other over distance, modern mobile phones can be used for writing emails, watching videos, listening to radio, organizing appointments, playing games, and they even offer the possibility for use as camera and GPS-navigation device. The increasing complexity of products and functionalities may represent a risk that users might be hindered using them in an efficient, flawless, and satisfactory way. For product developers it is hence a particular challenge to design products that are easy and comfortable to use.

To ensure the development of user-friendly products[1], it is important to guarantee that the needs and limitations of the user are taken into account throughout the whole development process (Rubin & Chisnell, 2008). This "user-centred" approach is widely accepted in product design and embodies three main principles of design: a) early focus on users and tasks, b) empirical measurement, and c) iterative design (Gould & Lewis, 1985). This implies that designers should bear the end user in mind throughout the whole design process. For that purpose, the authors propose to include the user very early into the design process. To do so, users should use simulations and prototypes of the product to carry out real tasks while their performance and reactions to the product should be observed, recorded, and analyzed empirically. Based on those empirical findings, the product should be developed further, with the aim to create a new prototype which can be tested with users again – this iterative process of user testing and further development should lead finally to a functional product that corresponds to the

[1] In literature on human-computer interaction, authors use different terms when they refer to the object that should be usable (e.g. interface, system, software or computer). In this thesis, the term product is used, which can include physical objects (e.g. mobile phone) or non-physical designs (layout of a webpage) that have a utilitarian function. In contrast, works of art or non-utilitarian items are excluded from this categorization.

requirements and limitations of the user (Gould & Lewis, 1983). The essential role of usability tests in such a user centred design approach indicates the importance of this evaluation method for the development of usable products (Lewis, 2006).

In a usability test, a typical usage scenario is simulated, in order to assess the usability of a product. In a typical usability test situation, a test user operates a prototype of a product in a laboratory which is especially equipped for this purpose (Rubin & Chisnell, 2008). The simulation of the product usage may hence differ considerably from the scenario of its real use: observers and a camera are present, the product is simulated by a prototype, and the test user might differ from the end user of the product (e.g. with regard to expertise, motivation or culture). Such contextual factors may represent an important issue for the validity of usability tests, since there is no guarantee that effects detected in such an experimental condition prove to be the same in a real life context (Jordan, 1998a). Given the significance of usability testing for product development, it is important for usability practitioners to be aware of these issues and to know in detail how specific contextual factors of usability tests influence its outcomes. Awareness of such influencing factors may provide indications about the adequate interpretation of usability test results and may help to plan and conduct the usability evaluation. Until now however, only little is known about the influence of contextual factors in usability tests on their outcomes.

This thesis is aimed at investigating and analyzing the effect of different contextual factors on usability test outcomes. Furthermore, it aims to give recommendations to usability practitioners about the design and organisation of usability tests as well as about the interpretation of their results. In order to answer to these research goals, four experimental studies have been conducted, based on a framework presented by Sauer, Seibel and Rüttinger (2010) which integrates different contextual factors that are assumed to impinge on usability test results. This framework is presented in chapter 3 of this thesis, followed by an overview of the studies composing this thesis in chapter 4. The four studies are presented in the form of three published papers and one submitted manuscript in chapter 5 to 8. In chapter 9, the findings of the studies are summarised and their implications for usability practice as well as for usability research are discussed. To begin with however, it is important to describe and define the main concepts and notions mentioned in this thesis. Since usability is a topic shaped by various professional domains such as programmers, system developers, product and web designers, engineers, ergonomists and

psychologists, the understanding of the notion *usability* and its measurement are heterogeneous and multifaceted. The following chapter aims to define central concepts of this thesis such as usability and usability testing as well as to differentiate them from other concepts.

2 Usability and the evaluation of usability

2.1 Definition of the term usability

The term usability was coined in the early 1980s as a substitute for the terms *user friendliness* and *ease of use,* which suffered from a vague and unclear definition that focused mainly on the aspect of comfort in product use (Sarodnick & Brau, 2006; Bevan, Kiriakovsky & Maissel, 1991). Miller (1971) might have been one of the first authors who attempted to define usability in terms of measures for *ease of use*, whereas Bennett (1979) further developed this concept to describe usability. According to Sarodnick and Brau (2006), the first fully discussed and detailed formal definition of usability was proposed by Shackel (1981) who defined usability of a product as ‘‘the capability to be used by humans easily and effectively’’ (p. 24). Since then, reams of different usability definitions have been proposed. English (1993) for instance cites twenty different definitions, and differentiated between more than twenty components that were sometimes overlapping and sometimes contradicting. Bevan et al. (1991) summarized four different perspectives that have influenced the way, usability was defined: (i) the product oriented perspective emphasizes that usability can be measured with regard to the ergonomic characteristics of a product. (ii) According to the user-oriented perspective, usability is represented by means of the user’s mental effort in product usage and his or her attitude towards the product. iii) Following the performance oriented perspective, usability is described in terms of the interaction of the user with the product while the (iv) context oriented perspective emphasizes that usability is depending on the user group that is studied, the tasks those users are performing, and the environment in which the task are completed. Bevan et al. (1991) argue that all those views should be considered in a definition of usability. This requirement is satisfied by a definition proposed by the International Standardisation Organisation (ISO) which is often referred to in literature and which is applied in many subsequent related ergonomic standards (Bevan, 2001). In part 11 (*guidance on usability*) of the ISO standard 9241 (*ergonomic*

requirements for office work with visual display terminals), usability is defined as "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). Effectiveness represents the accuracy and completeness with which the user can achieve the goals whereas efficiency represents the resources expended in relation to the accuracy and completeness with which users achieve goals. Satisfaction stands for the freedom from discomfort, and positive attitudes towards the use of the product (ISO 9241-11, 1991). Since the definition of usability in ISO 9241-11 is widely accepted in usability research and usability practice, it is referred to in this thesis for the characterization of the term usability. While the discipline of ergonomics was for a long time focused on the usability of products as defined above, this perspective has been expanded over recent years. Increasingly, the interest of practitioners and researchers began to focus on other subjective experiences arising from the use of products such as pleasure, fun and emotions (cf. Hekkert & Schifferstein, 2008; Jordan & Green, 2002).

2.2 User experience

User experience is an approach in product development that focuses on the physical, sensual, cognitive, aesthetic and emotional experience of product use (Forlizzi & Battarbee, 2004; Hekkert, 2006). It enlarges the classical notion of usability which is focussed on user's tasks and accomplishments (effectiveness, efficiency and satisfaction of the user-product interaction) by a more holistic approach which focuses on aspects such as the user's fun, affect and emotions evoked by the human-product interaction (Hartmann, De Angeli & Sutcliffe, 2008; Jordan, 2000; Hassenzahl & Tractinsky, 2006).

Although the notion of user experience has been widely adopted by practitioners and researchers (Hassenzahl & Tractinsky, 2006), the concept is still elusive, ill defined and lacks of widely accepted and empirically confirmed measures (Väänänen-Vainio-Mattila, Roto, & Hassenzahl, 2008). Several frameworks and theoretical models are discussed in ergonomic literature. Forlizzi and Ford (2000) differentiate four dimensions of user experience: sub conscious, cognition, narrative and story telling. Norman (2004a) distinguishes three subdivisions of experience: the visceral, the behavioural and the reflective. Wright, McCarthy and Meekinson (2003) depict four aspects of experience: emotional, sensual, compositional and spatio-temporal whereas Hassenzahl (2004) distinguishes between pragmatic and hedonic user experience. These

examples for theoretical conceptions of user experience indicate the actual elusiveness of the construct. Aside from these differences, most definitions agree that user experience comprehends more than the usability and utility of a product and that it is affected by the internal state of the user, the context and the product's perceived image (Väänänen-Vainio-Mattila et al., 2008). Väänänen-Vainio-Mattila et al. (2008) mention however that this consensus might not be substantial enough to consider user experience systematically in product development.

Despite the difficulties of definition and measurement, the notion of user experience has become very important in the field of product and system design (Hekkert & Schifferstein, 2008). It has been included in a standard of the international standardisation organization on Human-centred design processes for interactive systems (ISO 13407) and is often considered as the main goal of product development (Earthy, Jones & Bevan, 2001). Some authors even suggest to replace the notion of user centred design with its focus on the development of usable products by the term "experience centred design" (cf. Shedroff, 2001; McCarthy & Wright, 2004) focussing on the development of products that elicit positive user emotions. Referring to the user centred design approach, this implies that user experience should be considered throughout the whole development process and that emotions and fun users experience by using a product should be measured when a product is to be evaluated.

Regardless of whether the focus is set on usability or user experience, the empirical evaluation of the user-product interaction is very important component in the product design process (Gould & Lewis, 1983). A typical and widespread method for conducting empirical product evaluations is the usability test.

2.3 Usability testing in product development

Usability testing employs techniques to collect empirical data of the interaction of representative end users with the product by completing representative tasks in a controlled environment. In a usability test, the usability of a product is hence evaluated by means of user trials, with the aim to analyze or improve the usability of the evaluated product (cf. Dumas & Reddish, 1999; Rubin & Chisnell, 2008). Usability tests can furthermore be a vehicle for improving the co-operation between users and product designers (Buur & Bagger, 1999) to teach product designers about usability (Nielsen, 1993), or even to improve the PR of the company using the methods (Salzman

& Rivers, 1994). With regard to the scope of the usability evaluation, two different approaches can be distinguished: formative and summative usability testing (Scriven, 1967). Formative testing is conducted throughout the whole development phase in an iterative design process with the goal to gather qualitative information about weaknesses and operation problems of a product. Summative testing on the other hand aims to collect quantitative data about the accomplishment of task goals. It is mainly conducted at the end of specific phases in the product development process (e.g. for a comparison of different design alternatives) or at the end of the development process (e.g. for a comparison of the final design with usability requirements defined in the product specifications or with a predecessor or concurrent product) (Lewis, 2006; Rubin & Chisnell, 2008). Depending on their aim, the methods and measures used in usability testing may differ considerably. Common for all of the different approaches however is the use of prototypes of the product.

According to the user centred design approach the usability of a product should be tested very early in development process. This implies that the not yet developed and functional product needs to be simulated for testing. Due to constraints of the design process such as time pressure and budgetary limitations, this is usually done by means of a prototype. Prototypes can be very basic (low-fidelity prototypes) such as drawn images and handwritten features on a piece of paper, or more advanced (medium-fidelity prototypes) such as interactive computer simulations (Engelberg & Seffah, 2002; Vu & Proctor, 2006). Low-fidelity prototypes are often used in usability practice since they are easy, fast and inexpensive to deploy (Snyder, 2003; Newman, Lin, Hong, & Landay, 2003). Similar to that, also medium-fidelity prototypes are very common in usability practice (Engelberg & Seffah, 2002) which may be due to the availability of software tools for the prototype development that are easy to learn and to use (e.g. Dreamweaver, Powerpoint, DENIM, SILK or PatchWork).

Usability testing however is not the only technique applied in product development practice that allows evaluating the usability of a product. Other popular methods are for example cognitive walkthroughs, heuristic evaluations, checklists or interviews and focus groups (Jordan, 1998a; Kuniavsky, 2003; Nielsen, 1993). The *cognitive walkthrough* is an expert usability evaluation method in which the designer guides his or her colleagues through actual user tasks using product specifications or a prototype of that product. They are asked to envision the use of the product from the point of view of a typical user. The investigators try to predict whether or

not a user would encounter problems in completing the tasks with the actual product design. Encountered difficulties and concerns of the team are recorded and used for an improvement of the product (Jordan, 1998a; Kuniavsky, 2003; Wharton, Rieman, Lewis & Polson, 1994). In a *heuristic evaluation*, an expert (e.g. usability specialist) reviews a product according to accepted usability principles (heuristics). Those heuristics are based on professional experience, human factors literature and the body of research (Jordan, 1998a; Nielsen & Molich, 1990). Heuristics are proposed for example by Nielsen (1993), Mankoff, Dey, Hsieh, Kientz, Lederer and Ames (2003) or Somervell, Wahid and McCrickard (2003). *Checklists* are a collection of design properties which are supposed to ensure that a product is usable. An investigator hence checks whether the design of an evaluated product conforms the properties of that list. Usability problems can be expected where the product does not correspond to the properties of the list (Jordan, 1998a). In *Interviews* and *focus groups*, a structured interview is conducted after the presentation of an early prototype to a representative user (interview) or a group of users (focus group). These methods aim at the detection of the audience's desires, experiences and priorities. Furthermore, they help identifying how acceptable the design concepts are and how they might be improved with regard to usability and acceptability (Kuniavsky, 2003; Rubin & Chisnell, 2008). Each of those methods features some advantages and disadvantages (for an overwiev see Jordan, 1998a or Nielsen, 2003), the usability test however is the only method that considers representative end users and provides empirical data as requested in the user centered design principles (Gould & Lewis, 1985). The benefits of usability testing are the minimized costs and risks associated with releasing a product that has usability problems as well as increased sales and customer brand loyalty. Furthermore, usability testing helps to optimize development costs by minimizing documentation and user support costs, by reducing costs of unscheduled updates and product recalls (see e.g. Jordan, 1998a; Nielsen, 1993). All this are reasons why usability testing is one of the most important and most widely applied methods in usability practice (Lewis, 2006). While there is a general agreement on the importance of usability tests for usability practitioners, no such agreement exists in regard to the measures recorded in usability tests.

2.4 Measures in usability tests

Usability cannot be measured directly. But aspects of usability can be assessed through operationalization of the usability construct (Hornbaek, 2006). The type of measures that are

recorded and the methods that are applied in usability tests may differ considerably with regard to the theoretical conceptualisation of usability. But also the aim of the evaluation has an influence on the measures recorded in a usability test (e.g. summative vs. formative evaluation). This chapter reviews typical aspects of usability assessed in usability testing.

2.4.1 Performance data

In summative usability evaluation, product usability is evaluated under controlled conditions with regard to efficiency, effectiveness and satisfaction of user-product interaction (Jordan, 1998a). Effectiveness may be measured by the proportion of users that can complete a given task whereas deviations from the critical path (e.g., number of unnecessary clicks during task completion), error rates (e.g., number of clicks on the home or back button before task completion), and time on task (e.g., time needed to accomplish the task) are typical measures of efficiency (Jordan, 1998a).

2.4.2 Perceived usability

In addition to such objective measures, subjective data on the user-product interaction are usually collected during usability tests, for the most part by means of questionnaires or semi-structured interviews (Jordan, 1998a; Rubin & Chisnell, 2008). A plethora of different usability questionnaires has been developed so far. They differ considerably with regard to length, response format, applied product category, but also with regard to their scope, objective and theoretical background. Most instruments measure product usability in a subjective way (compared to "objective" performance measures) or assess user satisfaction as defined in the ISO-Standrad 9241-11 (ISO-9241-11, 1991). A typical example for a subjective usability scale is the IsoMetrix (Gediga, Hamborg & Düntsch, 1999). Typical examples for questionnaires measuring user satisfaction are the Questionnaire for User Interaction Satisfaction (QUIS, Chin, Diehl, & Norman, 1988), or the Post-Study System Usability Questionnaire (PSSUQ, Lewis, 1995).

2.4.3 User emotions and user experience

Whereas the measurement of usability and user satisfaction is well established in usability practice, the assessment of user experience is less common. Until today, no general agreement upon evaluation method exists, but different approaches assessing user emotions after product usage can be distinguished. One possibility to measure user emotions in usability evaluation is the deployment of classical emotion questionnaires such as for example the Positive and Negative Affect Schedule (PANAS; Watson, Clark & Tellegen, 1988) or the Learning Affect Monitor (LAM; Reicherts, Salamin, Maggiori & Pauls, 2007). Another approach to user experience evaluation consists of the direct measurement of users' emotional responses to product usage (Desmet, Overbeeke & Tax, 2001). Desmet, Hekkert and Jacobs (2000) propose an instrument (Product Emotion Measurement tool – PrEmo) using non-verbal audio-visual cues in the form of manikins for the assessment of 14 different user emotions (disgust, indignation, contempt, unpleasant surprise, dissatisfaction, disappointment, boredom, desire, pleasant surprise, amusement, admiration, inspiration, satisfaction, fascination). Based on a three dimensional theoretical model of emotions proposed by Mehrabian and Russel (1977), the Self-Assessment Manikin questionnaire (SAM; Bradley & Lang, 1994) is another instrument that uses manikins to measure emotions. The instrument, originally developed for the assessment of affect in consumer research, measures three independent dimensions of emotions: pleasure-displeasure, degree of arousal, and dominance-submissiveness. With regard to questionnaires measuring user emotions, only few instruments have been developed so far. The AttracDiff2 (Hassenzahl, Burmester & Koller, 2003) is one example of a questionnaire for product evaluation that considers a broader range of aspects of user experience. Based on a model that distinguishes between pragmatic quality of a product (focussing on instrumental aspects of usability and utility, for example effective and efficient goal-achievement) and hedonic product quality (which is related to the needs and expectations of the user and contains aspects such as stimulation, identification and evocation), the authors proposed a questionnaire to assess perceived pragmatic quality as well as different aspects of perceived hedonic quality of a product. According to Väänänen-Vainio-Mattila et al. (2008) however, the existing instruments measuring user experience are still inadequate, which may be due to the missing generally agreed-upon definition of the concept of user experience.

2.4.4 Physiological measures

A more recent approach is the measurement of physiological data in usability testing as indicators of for example mental workload, stress, frustration or pleasure. Heart rate data (HR), galvanic skins response (GSR) and blood volume pressure (BVP) were for example measured as a reaction to media quality (audio and video) and showed to be correlated with subjective ratings of satisfaction and objective measures of task performance (Wilson & Sasse, 2000). Furthermore, measures of HR, GSR and BVP were influenced by the design of web pages (Ward & Marsden, 2003). The use of poorly designed web pages caused higher levels of arousal compared to the use of well designed web pages. In a further study, GSR showed to be related to task performance and subjective assessments of participant's stress level (Lin, Omata, Hu, & Imamiya, 2005). Analyzing user experience with entertainment systems, Mandryk, Ikpen and Clavert (2006) reported differences in physiological measures of test participants when playing against a computer versus playing against a friend. In the field of human computer interaction (HCI) research, physiological data were used for example as indicators of mental effort (heart rate variability; Rowe, Sibert, & Irwin, 1998; Rani, Sims, Brackin, & Sarkar, 2002) or frustration (electromyography: Partala & Surakka, 2004; GSR and BVP: Scheirer, Fernandez, Klein, & Picard, 2002). These examples reveal how different physiological measures were used with success in the context of usability evaluation and HCI. However, no standardization of task, domain, or measures exists so far. This makes it difficult to compare across different studies as well as to give recommendations on which measure to choose for a specific evaluation objective (Mandryk, Ikpen & Clavert, 2006).

2.4.5 Usability Problems

A usability problem can be described as an "aspect of the system and/or a demand on the user which makes it unpleasant, inefficient, onerous or impossible for the user to achieve their goals in typical usage situations" (Lavery, Cockton, & Atkinson, 1997; p. 254). However, there is no general agreement about what exactly a usability problem consists of and how it can be defined (for a discussion of this see Lavery et al., 1997). Nonetheless, numerous attempts have been conducted to classify usability problems (e.g. with regard to severity, frequency, impact, cost to fix or human error theory; for an overview see Keenan, Hartson, Kafura & Schulman, 1999). In usability practice, different methods usually are applied to detect usability problems. The most common and most widely used method is the think aloud protocol (Nielsen, Clemmensen, &

Yssing, 2002), in which the test user is asked to speak out loud what he or she is thinking while completing specific tasks on the product. But also other methods such as retrospective think aloud (a test user explains what he or she thought by watching a video of the preceding product interaction), co-evaluation (product evaluation in teams of two test users who talk together while interacting with the product), interviews, questionnaires, surveys and also diaries are common methods used in usability practice (Öörni, 2003). Usability problems are hence detected by evaluation of the verbalized thoughts of the test participants and by the observation and analysis of their interaction with the product. It has been shown that both, the detection and the classification of usability problems are prone to evaluator effects and lack satisfactory reliability (Jacobsen, Hertzum, & John, 1998; Kessner, Wood, Dillon, & West, 2001; Molich et al., 1999).

Usability problems are primarily measured in formative usability evaluation in which the focus is set on detecting and eliminating usability flaws and the acquisition of information for a further improvement of the product usability. Results of formative usability evaluation may provide some information about efficiency and effectiveness of user-product interaction (e.g. by calculating an index with regard to number and severity of detected usability problems) as well as about user satisfaction (e.g. by analysing user statements about concerns, dissatisfaction, fun and pleasure), but they provide no objective and quantitative information about the product usability. Since the main focus was rather on a summative evaluation of product usability, formative measures of usability problems such as errors in task completion were not analysed in a qualitative way in the studies presented in this thesis but were only considered quantitatively. Moreover, no additional information about improvement possibilities of the evaluated product was recorded.

2.4.6 Relation between usability measures

As mentioned above, in usability testing several measures are usually recorded as indicators of product usability. Typical measures are objective data of user performance as indicators for efficiency and effectiveness as well as subjective data of preference and satisfaction (Jordan, 1998a; Hornbæk, 2006). Usability research has shown however, that subjective and objective measures of usability are only faintly related (Bailey, 1993; Frøkjær, Hertzum & Hornbæk, 2000; Nielsen & Levy, 1994). A recent meta-analysis of correlations among usability measures, calculated from data of 73 usability studies, revealed only low correlations of .196 (± .064) between measures of efficiency and satisfaction and .164 (± .062) between measures of

effectiveness and satisfaction (Hornbæk & Law, 2007). Such results can primarily be interpreted as indicator of the multidimensionality of usability (Frøkjær et al., 2000) but they might also point at some issues of measurement in usability tests. In usability tests, it is a common phenomenon that test participants, struggling through several tasks, at the end report that the product was easy to use (Dicks, 2002). According to Dicks (2002), this might be due to the unfamiliar environment in usability tests in which participants may make (not accurate) assumptions about the goals and scopes of the evaluation as well as the expectations of the evaluator. Such particular results might however also be due to other specific aspects of usability tests that influence the results of usability evaluation.

3 The four factor framework of contextual fidelity – influencing factors in usability tests

Despite their popularity and wide acceptance in ergonomics, usability tests have a number of inherent shortcomings. According to Rubin (1994), "even the most rigorously conducted formal test cannot, with 100 percent certainty, ensure that a product will be usable when released" (p. 27). The author names for example the artificial testing situation or the missing representativeness of test participants as a reason for this. Because usability tests are simulations of a real usage situation, usability practitioners can not rely on the assumption that the modelled scenario exactly represents the real usage scenario. It depends in fact on the accuracy of the modelled scenario whether the usability of the product is measured accurately. If the simulation is too far from a real usage scenario, the measures can not be considered as containing sufficient external validity (for a detailed definition of issues of validity in psychological measurement see Elmes, Kantowitz & Roediger, 2003) and therefore do not allow generalisation of the findings to the population or the situation of interest (Thomas & Kellog, 1989). Usability practitioners however sometimes have, due to reasons of measurability and efficiency, to accept some simplifications and abstractions when designing usability tests. For an exact measurement of user performance e.g., data collection in the real usage environment may be difficult. Therefore, usability tests are often conducted in usability laboratories (for an overview of the lab vs. field discussion see Sauer & Sonderegger, under review) where the human-product interaction can be observed and recorded in a more detailed and accurate way. Since the real usage scenario can not be simulated with 100% accuracy in usability tests, it is very important for usability practitioners to know exactly about the consequences of such simplifications and abstractions of usability test scenarios. If they prove to have no influence on the results of usability tests (compared to the real usage scenario), the applied scenario can be considered as useful. If however the results of the usability test are influenced by the test scenario, it is important to consider this influence in the design of the usability test or the interpretation of test results. The detailed knowledge of influencing factors of usability test scenarios is hence an important issue for usability practice and usability research. This is because the lack of knowledge about the limitations of usability testing methods to ensure valid and reliable results is a great risk for undermining and trivializing the whole concept of usability testing (Dicks, 2002). Therefore, a detailed analysis of influencing factors in usability test scenarios is of great importance.

According to the human-machine system framework (for an overview see Bennett, 1972, 1979 and Eason, 1981), the user-product interaction consists of four principal components: user, task, tool and environment. Since in usability tests, user-product interactions are evaluated, those four components are important aspects characterizing a usability test scenario. The Four Factor Framework of Contextual Fidelity (4FFCF; Sauer et al., 2010) picks up on these main components by describing four factors *system prototype, testing environment, user characteristics and task scenarios* on which the testing scenario in usability tests may differ from the real usage scenario and therefore be of high (small differences) or low fidelity (severe differences). Referring to the component *tool* of the human-machine system framework, the four factor framework of contextual fidelity proposes the *prototype* as one important factor in usability testing. Prototypes are frequently used in usability tests in place of an operational product because usability tests are often conducted early in the development process when a fully operational product is not yet available (e.g. a mock up of a car cockpit is used to test the arrangement of different displays). Prototypes however might differ considerably from the final product (e.g. with regard to aesthetic design, level of functionality, way of interaction etc.), which might have an influence on the results of usability tests. *Testing environment* is the second factor proposed by the four factor framework on which the usability test scenario can differ from the real usage scenario. Usability tests are for example often conducted in usability laboratories, for reasons of measurability and controllability of disturbing factors (Rubin, 1994). The lab situation presents an environment that can differ compared to the real usage scenario. A third factor influencing the fidelity of usability tests described by the four factor framework are *user characteristics*. Characteristics of usability test participants (such as i.e. competence, attitude or state) may differ from the future user population, which might influence on the results of usability tests. As a fourth factor described by the framework, the *task scenarios* given in usability tests may not be representative or complex enough compared to the real usage situation. Due to a restricted time budget in usability test for example, often only a selection of possible tasks are selected, assuming that if users can successfully complete the selected tasks, they should also be able to complete all other tasks using the product (Dicks, 2002). This assumption however might not prove true, and therefore, the choice of tasks scenarios can have an influence on the outcomes of usability tests.

Furthermore, the authors of the 4FFCF additionally have defined sub-factors for each factor (see figure 1 for an overview of the framework). With regard to the factor *system prototype,* breadth of functions (e.g. prototype of a mobile phone simulates all functions of the future product compared to a limited selection of functions such as editing a text message and changing phone settings), depth of function (e.g. prototype of a mobile phone that actually can be used for making phone calls compared to a prototype that simulates the phone call only), physical similarity (e.g. cardboard mock-up of a mobile phone compared to a realistic prototype made out of plastic) and similarity of interaction (e.g. interaction with a prototype of a mobile phone by physical pushbuttons compared to an interaction using a touch-screen computer) are differentiated. With regard to the factor *testing environment*, two sub-factors are described by the framework: the physical environment (e.g. lab vs. field, noise levels etc.) and the social environment (i.e. domain in which a product is used such as work, domestic or public domain). For the factor *user characteristics,* four sub-factors are mentioned in the framework: competence (e.g expertise of the user), attitude (e.g. user motivation), state (e.g. mood of the user) and personality (e.g. computer anxiety or extraversion). And finally, task scenarios can be differentiated by the sub-factors breadth (e.g. single task scenarios such as using a mobile phone sitting at a table compared to a multiple-task scenario such as using a mobile phone while walking in the street) and depth (e.g. sending a text message that already has been prepared compared to sending a text message that has to be written by the test user). According to the 4FFCF, all these specific usability characteristics of the test context are potential factors influencing the test outcomes.

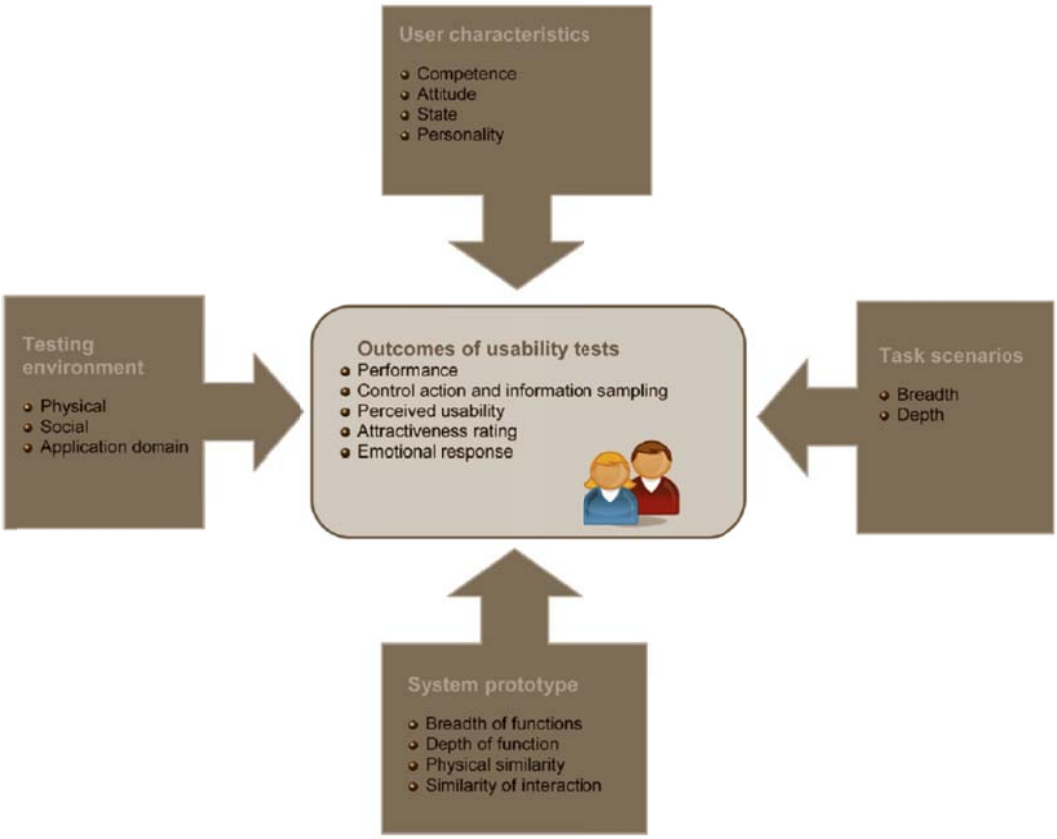


Figure 1: The Four-factor framework of contextual fidelity (Sauer et al., 2010, p. 132)

According to its authors, the 4FFCF requires empirical validation to meet the requirements, which are to make predictions about how different outcome measures in usability tests are influenced by a reduced fidelity level on the four factors. It represents a framework for future research projects that are aimed to assess the influence of each of the four factors, to provide a theoretical justification for the influence as well as to evaluate potential interactions of different factors. The present dissertation represents such a research project. Its aim is to evaluate the influence of some of the four factors presented in the framework on outcomes of usability tests. Furthermore, it aspires to examine the interactions of different factors proposed by the framework.

4 Overview of the four studies presented in this thesis

All the studies presented in this thesis (see table 1 for an overview) address the influence of contextual fidelity in usability testing. The focus of each study is on a specific aspect of one or two influencing factors described by the 4FFCF. The factors were chosen with regard to their importance for usability practice and the extent of existing research literature. Due to the central role prototypes play in product development (see e.g. Mann & Smith, 2006), a focus was set in this thesis on the influence of fidelity aspects of the product prototype in usability tests. A further issue that has often been discussed in usability literature is the influence of the testing environment on the outcomes of usability tests. Various aspects of the setup of usability laboratories have been mentioned as potential source of error (such as the presence of monitoring equipment such as cameras, one-way mirrors or microphones, cf. Nielsen, 1993; Rubin, 1994). Even though the influence of the testing environment has been widely discussed in the literature, the empirical evidence is rather limited. Therefore, specific aspects of the testing environment were considered in this thesis. Obtaining a suitable sample of test users is another key issue in usability practice. It might be a very difficult task to find a current textbook on usability testing that does not emphasise the notion that test users should be as representative as possible compared to the future end users. In usability practice however, time and budget available for usability testing are usually rather limited (Bauersfeld & Halgren, 1996; Marty & Twidale, 2005). Therefore, often ordinary members of the staff of the company are recruited as participants in usability tests, which is why usability evaluation in practice is also said to be simply a study of the reaction of secretaries to design (cf. Thomas, 1996). Furthermore, the importance of user characteristics in usability practice manifests itself through the fact that they are an integral part of the definition of the term usability (see chapter 2): whether a product is usable or not depends on the user operating it. This implies that the same product (e.g. a car with a manual gearshift) can be usable for one user group (experienced stick shift drivers) but very difficult to use for another (experienced drivers of cars with automatic transmission). This example illustrates how influential user characteristics can be on the outcomes of usability tests. For that reason, one study of this thesis addresses the influence of user characteristics in usability tests. In detail, the four studies examine the following aspects of the 4FFCF.

The first study entitled “The Influence of Laboratory Setup in Usability Tests: Effects on User Performance, Subjective Ratings and Physiological Measures” (named henceforth the *lab-*

setup study) is about the influence specific aspects of the testing environment of usability tests. In usability practice, it is a common that one or more observers may be present in a usability test. The number of present observers may differ considerably depending on the aim and scope of the usability test. There have been concerns that the presence of other people during usability testing represents a source of stress for test participants, but no attempt has yet been made to evaluate empirically the impact of the testing situation on the outcomes of a usability test. Therefore, an experiment was conducted in which the number of observers was varied as a between-subjects variable.

The following two studies address another factor of the four factor framework of contextual fidelity, namely the system prototype. The study entitled "The influence of design aesthetics in usability testing: effects on user performance and perceived usability" (herein after referred to as *design-aesthetics study*) is focussed on the influence of design aesthetics of prototypes in usability testing. The influence of design aesthetics in usability tests has already been addressed in previous research. These studies mainly reported a positive correlation between design aesthetics and measures of perceived usability. While this relation already has been confirmed empirically, it is less clear how prototype aesthetics is linked with other measures recorded in usability tests (e.g. performance measures). The design-aesthetics study addresses this issue by comparing two functionally identical computer prototypes of a mobile phone of which one was designed very appealingly whereas the other was designed very unappealingly.

Despite the broad acceptance of prototypes in usability practice, only little is known about the consequences of their use in usability tests compared to the use of not simplified and abstracted real products. The study with the title "The influence of prototype fidelity and aesthetics of design in usability tests: effects on user behaviour, subjective evaluation and emotion" (herein after referred to as *prototype-fidelity study*) compares the results of usability tests conducted by means of a paper or computer prototype of two different mobile phones (a more and a less appealing one) with the results of a usability test conducted with the real products.

The fourth study presented in this thesis is entitled "The Influence of Cultural Background and Product Value in Usability Testing" (named henceforth the *culture study*) and addresses the factors user characteristics and system prototype of the four factor framework of contextual fidelity. This study examines the influence of the user's cultural background by comparing test

participants from two different regions, Switzerland and East Germany. Furthermore, product value as a characteristic of the system prototype was varied by manipulating the price of the product.

Table 1: Overview of the four studies presented in this thesis

Independent variables	Corresponding factor of the 4FFCF	Dependent variables	Participants	Reference
Observer presence - No observer - One observer - Three observers Task difficulty - High - Low	Testing environment	- User performance - Perceived usability - Perceived attractiveness - User emotions - Physiological data (HRV)	- Students - N = 60 - 74% female - Mean age = 23.4	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.
Prototype aesthetics - Appealing - Unappealing	System prototype	- User performance - Perceived usability - Perceived attractiveness	- Adolescents - N = 60 - 52% female - Mean age = 14.2	Sonderegger, A. & Sauer J. (2010). The influence of design aesthetics in usability testing: effects on user performance and perceived usability. *Applied Ergonomics, 41*, 403–410.
Type of prototype - Paper prototype - Computer prototype - Real appliance Prototype aesthetics - Appealing - Unappealing	System prototype System prototype	- User performance - Perceived usability - Perceived attractiveness - User emotions	- Students - N = 60 - 42% female - Mean age = 23.8	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*, 670-677.
Cultural background - Switzerland - East Germany Product value - High - Low	User characteristics System prototype	- User performance - Perceived usability - User emotions	- General public - N = 64 - 38% female - Mean age = 31.6	Sonderegger, A. & Sauer J. (under review*). The Influence of Cultural Background and Product Value in Usability Testing. *Behaviour & Information Technology.*

* Editor invited the authors to revise the manuscript and resubmit it
4FFCF = Four Factor Framework of Contextual Fidelity

5 The Influence of Laboratory Setup in Usability Tests: Effects on User Performance, Subjective Ratings and Physiological Measures

Abstract

This article examines the influences of situational factors on user behaviour in usability tests. Sixty participants carried out two tasks on a computer-simulated prototype of a mobile phone. Employing a 3 x 2 mixed experimental design, laboratory setup was varied as a between-subjects variable (presence of facilitator and two non-interactive observers, presence of facilitator or no person present) while task difficulty was manipulated as a within-subjects variable (low vs. high). Performance data, subjective measures, and physiological parameters (e.g. heart rate variability) were taken. The results showed that the presence of non-interactive observers during a usability test led to a physiological stress response, decreased performance on some measures and affected the emotional state of test participants. The presence of a facilitator (i.e. a participating observer) also influenced the emotional state of the test participant. Practitioners involved in usability testing need to be aware of undue influences of observers, in particular, if the observers are non-interactive.

Keywords: usability test; social facilitation; heart rate variability; usability lab, laboratory setup

Reprinted from *Ergonomics, 52* (11), Sonderegger, A. and Sauer, J., The Influence of Laboratory Setup in Usability Tests: Effects on User Performance, Subjective Ratings and Physiological Measures, 1350 - 1361, Copyright (2009), with permission from Taylor & Francis.

Introduction

This study is concerned with the impact that observers in usability tests may have on the test outcomes. Usability tests are a widely used method in product development to identify usability problems, with a view to maximize the usability of the final product (Lewis, 2006). To identify usability problems, a prototype of the product is tested with future users who perform a range of typical tasks in a usability laboratory, which represents an artificial testing environment that models the context of future product usage. The testing environment can vary with regard to a number of features, such as the technical equipment being used, size of the facilities, and the number of persons being present during the test. In addition to the test facilitator who guides the test participant through the test and is therefore considered a participating observer, one or several non-interactive observers (e.g., members of the product design team) may attend the session to monitor the testing process. In practice, the laboratory setup can vary quite considerably (Rubin, 1994). Although there have been concerns that the presence of other people during usability testing represents a source of stress (Shrier, 1992; Salzman & Rivers, 1994; Patel & Loring, 2001), no attempt has yet been made to evaluate the impact of the testing situation on the outcomes of a usability test in a controlled study.

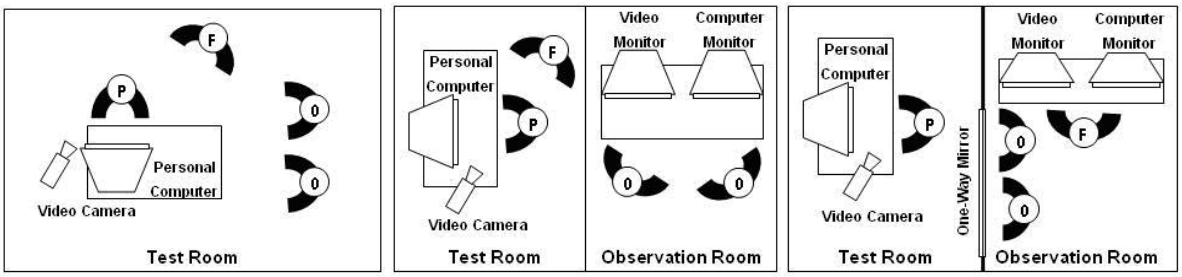
Setup of usability laboratories

The setup of usability laboratories can range from a simple low-cost laboratory to a rather sophisticated testing environment. Rubin (1994) distinguishes between three different testing configurations: single-room setup, electronic observation room setup, and classic testing laboratory setup (see figure2a-c). All setups have in common that the user is placed in front of the product to be tested, for software evaluation typically facing a computer while a video camera is available to record the testing procedure. However, the setups differ with regard to the number of people that are in the same room as the test participant.

The single-room setup (see figure 2a) represents the common minimum standard for a usability test. It consists of a single room where the test facilitator and the non-interactive observers are present to observe the participant's behaviour directly. Participating as well as non-interactive observers are usually positioned behind the test participant to minimize distraction. In the electronic observation room setup (see figure 2b), the test facilitator is still in the same room as the test participants while the non-interactive observers are placed in a separate room, allowing

them to observe the testing procedure on a closed circuit TV screen. In the classic testing laboratory setup (see figure 2c), the participant is alone in the testing room while the facilitator and the non-interactive observers are in the observation room, from which they can monitor the testing procedure through closed circuit television or/and a one-way mirror.

There are various factors to take into account when selecting a particular laboratory setup. These have been widely discussed in the usability literature (for an overview see Rubin, 1994). However, most recommendations in the usability literature about the advantages and disadvantages of different setups are based on practitioners' experience rather than scientific research. Therefore, there is a need for a more controlled examination of the multiple effects of the setups referred to above. This should include a range of measures that assess the effects of different setups at several levels: physiological response, performance and subjective evaluation. This corresponds to the three levels of workload assessment used in the work domain (Wickens & Hollands, 2000).



(a) (b) (c)

Figure 2: Setup of usability laboratories (P: Participant; F: Facilitator; O: Observer): (a) Single-room setup, (b) Electronic observation room setup, (c) Classic testing laboratory setup.

Multi-level analysis of test outcomes

Psychophysiological response.

Any testing situation may result in a change in psychophysiological parameters due to the arousal that is typically associated with the evaluation of a person (Kirschbaum, Pirke & Hellhammer, 1993). The presence of observers is expected to increase user arousal even further, as can be predicted by the theory of social facilitation (Geen, 1991). Arousal may be primarily observed in physiological parameters such as heart rate and heart rate variability (HRV). While heart rate is influenced by the physical effort expended during task completion (Boucsein & Backs, 2000),

HRV is considered to be a good indicator for mental stress and negatively toned affect (Kettunen & Keltikangas-Järvinen, 2001). Of the different frequency bands that can be derived from spectral analyses (high: 0.15- 0.4 Hz; low: 0.04 - 0.15 Hz; very low: 0.003 - 0.04 Hz; Task Force, 1996), two of them appear to be highly relevant for measuring mental and physical stress responses. The high frequency (HF) band of HRV is considered to be a suitable indicator (similar to heart rate) of the physical demands of task completion (Berntson & Cacioppo, 2004). The low frequency (LF) band is generally considered to be a suitable measure for mental demands (Boucsein & Backs, 2000). However, Nickel and Nachreiner (2003) have argued that the LF band indicates general activation rather than task-specific mental demands. Social stressors (e.g., observers being present during a usability test) may have such an activating influence since some work has demonstrated that social stress (induced by an observer while the participant completed a memory task) led to a decrease of HRV in the LF band (Pruyn, Aasman & Wyers, 1985). In addition to the LF band, the Task Force of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology proposes the computation of the LF/HF power ratio to increase the reliability of physiological measures reflecting psychological phenomena (Task Force, 1996). It is acknowledged that there has been some controversy in the literature about the sensitivity and diagnosticity of HRV and how the different types of stressors are related to HRV on the different frequency bands (e.g., Berntson & Cacioppo, 2004; Nickel & Nachreiner, 2003). Despite the on-going debate, the research presented above provides some justification for using the LF frequency band and the LF/HF ratio as indictors of stress. For the purpose of this study, it is assumed that a decrease in either of the two measures represents an increase in individual stress levels (Nickel & Nachreiner, 2003; Task Force, 1996). While there is a considerable body of scientific work on the use of psychophysiological data to determine operator stress in a work context, until now there has been no research that examined the effects of observers on physiological responses of test participants in consumer product usability tests, perhaps due to the difficulties associated with data collection and analysis.

Performance.

An important measure in any usability test is the performance shown by the test participant, which has been typically measured with regard to effectiveness and efficiency (Jordan, 1998a). Effectiveness refers to the extent to which a task goal or task steps are successfully achieved with the product (e.g., percentage of users that complete a task) while efficiency is a straight

productivity measure which is concerned with the level of resources deployed to achieve a task goal (e.g., task completion time, number of user inputs). These measures proved to be useful, in particular, for summative usability evaluations (i.e. comparative evaluation of product against another product or a reference standard).

User performance in the different laboratory setups may be moderated by the kind of task given. Social facilitation theory predicts differential effects as a function of task difficulty because of the role of arousal (Geen, 1991). Social facilitation theory postulates that the optimal level of arousal for performing a given task is inversely related to the difficulty of that task. On easy tasks, increased arousal is expected to lead to enhanced performance whereas on complex tasks, increased arousal results in impaired performance (Guerin, 1986).

Subjective evaluation.

Perceived usability. As the collection of performance data, the measurement of user satisfaction represents a standard procedure in usability tests, usually in the form of perceived usability (Jordan, 1998a). The collection of subjective usability in addition to objective usability data based on performance is of high importance since the two types of usability data may not always be in accord (e.g., Jordan, 1998a; see also Wickens & Hollands, 2000 in the context of work). A wide range of standardised instruments is available that can be employed for measuring perceived usability and its facets (for an overview, see Lewis, 2006). Criteria for selecting one of the instruments are clearly degree of specificity (generic vs. highly specific to a certain product), length (ranges from 10-item SUS to 71-item QUIS) and type of facets covered (e.g., ISO standard). Most of the instruments have acceptable psychometric properties and are therefore applicable from a methodological point of view.

Emotion. While the measurement of perceived usability of a product has a long tradition in usability testing, more recently the evaluation of emotional responses to products has gained increasing attention in product design (Marcus, 2003). Emotions are important in usability tests because they may have an influence on action regulation such as information seeking and user judgments (Dörner & Stäudel, 1990; Forgas & George, 2001). For example, it was found that the affective dimension of a product has a stronger influence on consumer decision-making than cognitive components (Shiv & Fedorikhin, 1999). This may be because emotions represent a

more immediate reaction to an object than a pure cognitive evaluation (Khalid, 2006). The reliable and valid measurement of emotions during and after product usage is also important because emotions are not only influenced by product features but also by situational factors such as laboratory setup. It is therefore vital to separate different sources of influences (product features, testing procedure, etc.) because the primary question of interest in a usability test concerns the emotions that are triggered by the product features rather than circumstantial factors (cf. Seva, Duh & Helander, 2007).

Attractiveness. Product features that trigger off emotions may refer to attractive and innovative functions or to the aesthetic appeal of the product. For example, work has shown that user emotions were more positively affected by the operation of an attractive product than by a less appealing one (Sauer & Sonderegger, in press). Furthermore, there is evidence for a positive relationship between product aesthetics and perceived usability of a product (e.g., Tractinsky, Katz & Ikar, 2000). This suggests that product aesthetics is an important aspect in a usability test. While there is some research on the effects on aesthetics on various outcome variables, much less is known about factors that influence attractiveness ratings.

The present study

Although there have been indications that the setup of usability tests has an influence on test participants (cf. Schrier, 1992), this aspect has not been given much consideration in usability practice and research. In particular, no controlled study has yet attempted to measure the effects of this factor. Against this background, the main research question aims to examine the extent to which the presence of observers influences the test results, employing the multi-level analysis of test outcomes. To answer this question, usability tests were conducted in three different laboratory settings using a computer-based prototype of a mobile phone. The laboratory settings corresponded to the settings outlined in figure 2. During the usability test, participants completed typical tasks of mobile phone users.

With the first level of analysis being concerned with the psychophysiological response, instantaneous heart rate was measured during the usability test, allowing for the calculation of HRV. It was hypothesized that *with an increasing number of observers in a usability test, the power on the LF* band *as well as the LF/HF ratio decreases*. It was expected that all three conditions were significantly different from each other. This assumption was based on the

research evidence that the presence of observers represents a social stressor that evokes a change in psychophysiological parameters (e.g., Pruyn et al., 1985). We are aware that stress responses differ as a function of gender (e.g., Stroud, Salovey & Epel, 2002). Therefore, we assigned an equal number of males and females to each experimental condition.

At the second level of analysis, performance was measured on four dependent variables (e.g., task completion time, interaction efficiency). It was hypothesized that *an increasing number of observers in a usability test will lead to performance decrements on difficult tasks but to performance increments on easy tasks*. The predicted interaction between 'lab setup' and 'task difficulty' was based on the assumption of the theory of social facilitation (Geen, 1991).

At the third level of analysis, subjective user responses to the testing situation were measured. It was hypothesized that an increasing number of observers in a usability test will lead to an increased intensity of negative user emotions and a decreased intensity of positive user emotions. It was expected that all three conditions were significantly different from each other. This is due to the social stress induced by the presence of observers in an evaluation context, which has been found to be linked with negative affect (Lazarus, 1993).

In addition to these dependent variables, we also measured perceived usability, attractiveness and heart rate (though they were not referred to in any of the hypotheses) to explore their relationship with the manipulated independent variables.

Method

Participants

The sample of this study consisted of 60 students (74% female) of the University of Fribourg, aged between 18 and 31 years (M= 23.4, SD= 3.1). Participants were not paid for their participation.

Experimental design

In a 3 x 2 mixed design, test situation was used as a between-subjects variable, being varied at three levels: According to the different setups of usability laboratories described in the introduction, the usability tests were conducted either in the single-room setup (in the following referred to as multi-observer setup), the classic testing laboratory setup (i.e. single-observer

setup) or the electronic observation room setup (i.e. no-observer setup). As a within-subjects variable, task difficulty was varied at two levels: low and high.

Measures and instruments

Heart rate data.

The heart rate of the participants was continuously recorded during the whole experiment. To measure the effect usability test situations have on participants, the heart rate and HRV during the tests were compared with a heart rate and HRV baseline taken prior to task completion while the participant was relaxing. According to recommendations of the Task Force (1996), for each phase a minimum recording of 5 minutes was used for the analysis, excluding the first and last two minutes of an activity. For the relaxation phase, the period from min 2 - 7 (out of a total measurement period of 10 min) was included in the data analysis while for the testing phase, the period from min 2 – 7 was employed (out of a total measurement period of 10-15 min). The changes in HRV between testing phase and relaxation phase were calculated and used for data analysis. Since a minimum recording time of 5 minutes was required for the calculation of the HRV data, an analysis of the physiological data on task-level was not possible.

User performance.

Four measures of user performance were recorded: (a) *Task completion rate* refers to the percentage of participants that were able to complete the task within five minutes. (b) *Task completion time* indicated the time needed to complete the task successfully. (c) The *interaction efficiency index* measured the ratio of minimum number of user inputs required divided by actual number of user inputs. (d) The *number of error messages* referred to the number of times participants left the optimal dialogue path by more than two clicks (in which case the error message “wrong path, please go back” was displayed).

Subjective evaluation.

Perceived usability. To measure the user’s satisfaction with the system usability, the Post Study System Usability Questionnaire (PSSUQ; Lewis, 1995) was translated into German and employed in this study. The PSSUQ was chosen over alternative instruments (e.g. SUMI, SUS) because it was especially developed for usability tests in laboratory settings. On a seven-point Likert scale (1 = strongly agree; 7 = strongly disagree) users rated 16 items (example item: “I

could effectively complete the tasks and scenarios using this system). The overall internal consistency of the questionnaire (Cronbach's $\alpha > .90$) is high.

Emotions. To measure the two independent dimensions of mood (positive and negative affect), the German version of the "Positive and Negative Affect Schedule" (PANAS, Watson, Clark & Tellegen, 1988) was employed. The German-language questionnaire enjoys good psychometric properties (Cronbach's $\alpha = .84$; Krohne, Egloff, Kohlmann & Tausch, 1996). The instrument consists of 20 adjectives describing different affective states (e.g. active, interested, excited, strong). The intensity of each affect is rated on a five-point Likert scale (very slightly or not at all, a little, moderately, quite a bit, extremely).

Attractiveness. The attractiveness rating of the mobile phone was made on a one-item five-point Likert scale, with the item being phrased: "The design of the mobile phone is appealing" (scale: agree, partly agree, neither agree nor disagree, partly disagree, disagree).

Materials

Heart rate monitor and video camera.
The heart rate was recorded continuously throughout the experiment with a Polar S810i™ heart rate monitor. A video camera (Panasonic™ NV-MS5EG) was positioned next to the user's work space.

Computer prototype.
Based on a SonyEricsson™ SE W800i mobile phone, a computer simulation of the dialogue structure was developed using html and JavaScript. The interaction data was recorded by a PHP-script. The simulation was running on an Apache™ server (XAMPP) installed on a Toshiba Portege™ M200 TabletPC equipped with a touch screen. This specific screen enabled the user to interact directly with the computer prototype instead of having to use a mouse. This ensured that a similar kind of interface is used for the computer prototype compared to the real product. The computer prototype allowed the user to carry out a range of tasks in a similar way as with the real product. The dialogue structure was modelled in full depth for the task-relevant menu items. For the functions that were irrelevant for task completion, only the two top levels of the dialogue structure were modelled in the simulation. If the user selected a menu item that was not modelled in the menu structure (i.e. more than two clicks away from the optimal dialogue path), an error

message was displayed ("Wrong path, please go back"). It is acknowledged that displaying this error message indicates to the test participant that the technical is not yet fully operational. Furthermore, it represented some support to the participant by pointing out deviations from the optimal dialogue path. In total, 124 different menu configurations were modelled in the prototype.

User Tasks.

For the usability test, two user tasks were chosen. The first task ("text message") was to send a prepared text message to another phone user. This represents a task frequently carried out by users and was considered to be of low difficulty. The second task ("phone number suppression") was to suppress one's own phone number when making a call. This was a low-frequency task that required a higher number of clicks to be completed (15 clicks) compared to the first (9 clicks) and was therefore considered to be more difficult. To prevent participants from accidentally discovering the solution for the easy task during completion of the difficult task, the order of task completion was fixed, with the easy task always being presented first.

Procedure

The study was conducted in a usability laboratory at the University of Fribourg. Each participant was randomly assigned to one of the three experimental conditions. The two experimenters welcomed the participant and explained that the purpose of the experiment was to determine the usability of a computer-simulated prototype of a mobile phone. To measure the heart rate, the electrode of the Polar T61™ transmitter was moistened and attached on the participant's chest and the Polar S810i™ heart rate monitor system was fastened at the participant's wrist. Subsequently, the first experimenter guided the participant to a relaxation room where he/she was asked to remain seated for 10 minutes in a comfortable armchair listening to relaxing music. During that time period, a 5-min recording of physiological data was made, which later served as a baseline for a comparison of the changes in HRV in the usability test.

After 10 minutes, the participant was guided to the usability laboratory where the second experimenter (here: test facilitator) explained the steps in the testing procedure. First, the participant completed a short warm-up task (unrelated to the use of a mobile phone) to become familiar with the touch screen. The participant began completing the experimental tasks about 5 minutes after he/she had been seated, which provided sufficient time for physiological adaptation

following the physical movement from the relaxation room to the usability lab (the two rooms were situated adjacent to each other). In all three laboratory setups, the entire testing procedure was videotaped. In the one-observer setup, the test facilitator (i.e. second experimenter) was present but did not provide any assistance to the participant when help was requested during task completion. In this case, the facilitator deflected the question and asked participants to proceed with the task as well as they could. In the multiple-observer setup, a test facilitator and two non-interactive observers taking notes were present. Again, the facilitator did not provide any assistance to the participant during task completion. The two non-interactive observers (both male, aged 25 and 63 yrs) were introduced to the participant as two designers of a company involved in the development and evaluation of the mobile phone to be tested. In the no-observer setup, the test facilitator left the room as the testing procedure began and the test participant was alone in the laboratory. There was no one-way mirror in the laboratory. The display of the user was mirrored through a VNC server-software to a computer in a separate room. This allowed the experimenter to monitor the testing procedure without the test participant becoming aware of it. After the two tasks had been completed, the mood of the participant was measured with the PANAS. This was followed by the presentation of the PSSUQ and the attractiveness scale. At the end of the experiment, the participant had the opportunity to give feedback to the second experimenter about the prototype and the testing procedure.

Analysis of heart rate data and statistical data

The recorded heart rate data were controlled for eliminating artefacts (as proposed by Berntson & Stowell, 1998), using the Polar Precision Performance™ Software for automatic and Microsoft Excel™ for manual artefact correction. The data were further processed using the HRV-analysis software (V1.1), developed by the Biosignal Analysis and Medical Imaging Group from the University of Kupio in Finland (Niskanen, Tarvainen, Ranta-Aho & Karjalainen, 2004). Using the Fast Fourier Transformation Method, HRV was calculated in the LF band (0.04 - 0.15 Hz) and the HF band (0.15 - 0.4 Hz).

For physiological measures and subjective user ratings, a one-factorial analysis of variance (ANOVA) was carried out, followed by a priori multiple planned pair comparisons (one-tailed). For performance measures, a two-factorial ANOVA was conducted, with task difficulty being the second independent variable. Again, one-tailed planned pair comparisons

were carried to test for significant differences between cell means. For explorative post-hoc comparisons, the Tukey HSD method was applied if appropriate.

Results

Physiological measures

Heart rate variability.

Considered to be a sensitive indicator of participant stress, HRV in the LF-band was compared to the baseline levels (i.e. during relaxation phase). A decrease in power in the LF band is assumed to indicate an increase in participant's stress level and vice versa. The results showed a decrease of power in the LF band in the two test setups with observers, whereas in the no-observer setup the power in the LF band increased (see table 2). An overall difference among the laboratory setups was found ($F = 3.23$; $df = 2, 57$; $p < .05$). Planned contrasts revealed significant differences between multi-observer and no-observer setup ($t = 2.48$; $df = 38$; $p < .01$) and between multi-observer and single-observer setup ($t = 1.74$; $df = 38$; $p < .05$). These findings indicate increased stress levels for test participants in the presence of non-interactive observers. The comparison between single-observer setup and no-observer setup was not significant ($t < 1$). In contrast to the data for the LF band, changes in the HF band did not differ significantly between the laboratory setups ($F < 1$; see table 2).

As for the HRV in the LF band, the LF/HF ratio represents an indicator of participants stress, with a decrease in ratios representing an increases in stress levels compared to the baseline measurement (see table 2). The analysis revealed that the changes in the LF/HF ratio differed significantly between the laboratory setups ($F = 3.41$; $df = 2, 57$; $p < .05$). Planned contrasts showed a significant difference between the decrease of LF/HF ratio in the multi-observer setup and the increase in the no-observer setup, indicating higher stress levels in the setup condition with non-interactive observers being present ($t = 2.6$; $df = 57$; $p < .05$). No significant difference was found among the other conditions ($t < 1$).

Heart rate.

Analogous to the analysis of HRV data, for the heart rate the difference between the baseline measure and the beginning of the testing phase (2 – 4 min into the task) was calculated. The main effect of laboratory setup on heart rate was significant ($F = 4.01$; $df = 2, 57$; $p < .05$). The mean heart rate showed an overall increase from the relaxation phase ($M = 73.9$ bpm) to the testing

phase (M = 80.4 bpm). However, the size of the increase was much higher in the presence of observers (see table 2). Planned pair contrasts showed that in the multi-observer setup, heart rate showed a significantly higher increase compared to the baseline than in the no-observer setup ($t = 1.71$; $df = 38$; $p < .05$). The contrasts between the other conditions were not significant.

To test whether psychophysiological changes occurred during the course of the testing phase, a post-hoc analysis was carried out, comparing the heart rate at the beginning and at the end of task completion by calculating the mean value during two 2-min periods (2 - 4 min into the task vs. final 2 minutes of task completion). The results showed a significant reduction in heart rate over the course of the testing phase (from 80.4 bpm to 74.7 bpm; $F = 43.4$; $df = 1, 58$; $p < .01$). There was no significant difference among the groups with regard to the magnitude of the reduction of HR during the testing phase (no-observer: - 3.2 bpm; one-observer: - 6.9 bpm; multi-observer - 7.0 bpm; $F = 2.2$; $df = 2, 57$; $p > .05$), suggesting a general calming-down effect of the participants during the testing phase.

For HRV, a time-on-task effect could not be examined since the task completion time was not sufficiently long for conducting data analysis. It would have required two data collection periods of a minimum duration of 5 min each (Jorna, 1992; Task Force, 1996).

Table 2: Changes in physiological parameters (testing phase compared to baseline in relaxation phase) as a function of laboratory setup.

	multi-observer setup M (SD)	single-observer setup M (SD)	no-observer setup M (SD)
LF[a] power (ms^2)	-149.4[d] (534.1)	-50.2 (306.1)	+177.4 (371.6)
HF[b] power (ms^2)	-332.4 (660.9)	-120.1 (322.8)	-195.0 (546.5)
LF/HF ratio	-0.7 (2.7)	+0.5 (2.5)	+1.4 (2.2)
Heart rate (bpm[c])	+9.5 (8.0)	+6.3 (5.2)	+3.7 (4.5)

[a]LF: low frequency
[b]HF: high frequency
[c]bpm: beats per minute
[d] Negative values denote a decrease in that parameter.

User performance

Task completion rate.

The data of the measure of effectiveness are presented in table 3. The data showed no significant difference among conditions of laboratory setup ($F = 2.01$; $df = 2, 57$; $p > .05$). Furthermore, there was no significant interaction of test situation and task difficulty on task completion rate (F

= 2.01; df = 2, 57; p > .05). The main effect of task difficulty on task completion rate was significant (F = 37.9; df = 2, 57; p < .001), with users showing higher effectiveness in the easy task than in the difficult one. Because all test users completed the easy task (100% task completion rate), planned contrasts were only calculated for the difficult one. These comparisons revealed that subjects were most effective in the single-observer setup. Test users in this condition were significantly more effective than those in the multi-observer setup (t = 1.97; df = 38; p < .05). The other comparisons were not significant.

Task completion time.

The data of task completion time are presented in table 3. The analysis revealed a main effect of the test situation on this measure (F = 3.42; df = 2, 57; p < .05), with users requiring more time in the multi-observer setup than in the other two setups. However, no significant interaction of laboratory setup and task difficulty on task completion time was found (F < 1). This was in contrast to the predictions of social facilitation theory. Planned comparisons revealed that for the easy task, participants needed significantly more time in the multi-observer setup than in the single-observer setup (t = 2.68; df = 38; p < .01) and in the no-observer setup (t = 2.33; df = 38; p < .05). For the difficult task, no such differences among lab setups were found (all planned comparisons: p >.05). As expected, a main effect of task difficulty emerged, with the completion of the difficult task taking significantly longer than the easy task (F = 202.2; df = 1, 58; p < .001).

Interaction efficiency index.

Considering the impact of laboratory setup and task difficulty on the efficiency of user interaction (minimum number of clicks required / actual number of clicks), no significant main effect of laboratory setup (F < 1) as well as no significant interaction of laboratory setup with task difficulty (F < 1) was found (see table 3). The main effect of task difficulty was significant (F = 68.1; df = 1, 58; p < .001), revealing a higher interaction efficiency for the easy task than for the difficult task. In addition to the analysis of counting user inputs, a separate analysis measured the number of error messages displayed to the participant (i.e. being two clicks off the optimal dialogue path). Since the analysis of that error parameter showed a very similar pattern of results like the efficiency index, detailed results are not reported here.

Table 3: Measures of user behaviour as a function of laboratory setup and task difficulty (TD: task difficulty).

	Multi-observer setup M (SD)	Single-observer setup M (SD)	No-observer setup M (SD)	Overall M (SD)
Task completion rate (%)	72.5 (0.26)	87.5 (0.22)	82.5 (0.24)	
Low TD	100	100	100	100
High TD	45 (0.51)	75 (0.44)	65 (0.49)	62 (0.49)
Task completion time (s)	160 (36.4)	125.9 (48.5)	136 (41.3)	
Low TD	77.8 (53.7)	44.3 (15.5)	48.6 (16.4)	56.9 (36.3)
High TD	242.1 (76.8)	207.4 (90.4)	223.4 (76.0)	224.3 (81.2)
Interaction efficiency (optimal number of clicks / actual number of clicks)	0.45 (0.13)	0.54 (0.17)	0.51 (0.12)	
Low TD	0.75 (0.33)	0.86 (0.22)	.86 (0.25)	0.82 (0.27)
High TD	0.23 (0.1)	0.3 (0.21)	.24 (0.14)	0.26 (0.16)

Subjective user ratings

Emotions.

At a descriptive level, the data analysis revealed that negative affect was overall quite low and positive affect was slightly above midpoint on the 5-point scale (see table 4). The inferential statistical analysis ($F = 4.39$; $df = 2, 57$; $p < .05$) showed an influence of laboratory setup on positive affect. Participants in the no-observer setup showed higher positive affect than participants in the two other conditions (multi-observer setup: $t = 2.37$; $df = 38$; $p < .01$; single-observer setup: $t = 2.73$; $df = 38$; $p < .005$). For negative affect, visual inspection of the data showed a similar effect but the statistical analysis did not confirm a significant effect of laboratory setup ($F = 2.5$; $df = 2, 57$; ns).

Perceived usability.

The data of the PSSUQ-questionnaire are presented in Table 4. Regarding the influence of the laboratory setup on the subjective usability evaluation, no differences can be reported for the overall evaluation of usability ($F < 1$). A separate analysis for each of the three subscales showed the same pattern.

Attractiveness.

Table 4 contains the data of participants' appraisal of the aesthetic appeal of the mobile phone. The calculated ANOVA showed no significant effect of laboratory setup on the attractiveness rating of the tested mobile phone ($F < 1$).

Table 4: User ratings of emotions, usability, and attractiveness.

	Multi-observer setup M (SD)	Single-observer setup M (SD)	No-observer setup M (SD)
Positive affect (1-5)	2.8 (0.5)	2.7 (0.58)	3.2 (0.37)
Negative affect (1-5)	1.7 (0.67)	1.5 (0.56)	1.3 (0.36)
Usability rating (1-7)	4.3 (0.97)	4.3 (0.99)	4.7 (0.96)
Attractiveness (1-5)	2.6 (0.94)	2.6 (0.68)	2.4 (0.88)

Discussion

The main goal of the present study was to determine how laboratory setups commonly used in usability evaluation practice influence outcomes of usability tests. The main results showed that the presence of observers during a usability test had an effect on physiological measures, performance and emotion. However, no effects were recorded for perceived usability and attractiveness.

The results showed that the presence of a facilitator and non-interactive observers in the laboratory led to psychophysiological changes in test participants, which became mainly evident in the form of decreased HRV. This finding was supported by subjective participant reports in the debriefing session, which revealed that the presence of others had been experienced as a social stressor. In particular, the multi-observer condition was regarded as very stressful, with about half of the participants explicitly referring to the two non-interactive observers as a source of stress. This hints at possible differential effects of facilitators and non-interactive observers on test participants. The data from the present study indicated that non-interactive observers may be perceived as potentially more threatening since they did not communicate with the test participants. This may have raised concerns about their exact role, resulting in an increased fear of evaluation among test participants (cf. Hembree, 1988).

The changes induced by the presence of observers in physiological parameters were paralleled by decrements in various performance measures. Although the pattern of decrement was slightly inconsistent across task parameters (e.g, observer presence impaired performance on the easy task for task completion time and efficiency index and on the difficult task for task completion rate), we did not observe in a single parameter that presence of observers (non-interactive or facilitator) led to performance improvements. This is indicative of the adverse

effects of observer presence on performance in usability tests and, at the same time, it rejects the hypothesis based on social facilitation theory (i.e. observer presence would lead to improvements for easy tasks). Both tasks were novel to the participants and both were problem-solving tasks (i.e. current state and target state were known but the procedure to change from one to the other needed to be identified). To demonstrate the effects of social facilitation theory, it needs perhaps a more extreme difference in task difficulty, for example, a well practiced task or a simpler task type (e.g., perceptual-motor task). Either of the demands is difficult to meet in usability testing since these tasks are typically problem-solving tasks and are often unpractised (because they are embedded in a novel interface and dialogue structure). We may assume a general negative effect of observer presence, though positive benefits for individual test participants may be possible.

The results of the present study indicate that situational factors such as the setup of the usability test laboratory can have an influence on the participant's emotional state. While the overall level of negative emotions experienced during the usability test was rather low, there was nevertheless a significant effect of the presence of others (facilitator as well as non-interactive observers). Test participants under observation rated their emotional state significantly more negatively than those who were alone during the usability test. Since the user's emotional state can also be influenced by properties of the consumer product (Marcus, 2003), it is important to separate these respective influences, in particular as the product-induced emotions are considered a central outcome of product design while emotions induced by the test environment are to be regarded as an undesirable side-effect. Therefore, it is important to make efforts to ensure that the user's affective state is only influenced by product properties and not by situational features such as lab setup.

In contrast to measures of performance and emotion, the setup of the usability test laboratory did not influence the subjective appraisal of a product's usability. Although there were no hypotheses put forward that predicted a relationship of this kind (i.e. the variables were measured on an exploratory basis), it is of some interest that no such relationship was found. This corresponds to the results of a meta-analysis of Nielsen and Levy (1994), which revealed that subjective usability ratings were influenced by product characteristics but not by situational factors. Similarly, attractiveness ratings were not influenced by situational factors in the present study. Product aesthetics and the user's response to it are clearly an important factor in usability testing since there has been evidence that aesthetics influences perceived product usability (Tractinsky et al., 2000). Since the relationship between usability and aesthetics is not yet fully

understood, negative evidence of this kind is also helpful to discount the influence of situational factors on attractiveness ratings.

Of interest is also the question to what extent any of the observed effects would remain stable with increasing duration of the usability test. While temporal stability was not included as a research question in the experimental design, it was still worth examining this issue since some of the collected data could be used for that purpose. A calming-down effect was found in heart rate for all three laboratory setups. Participant reports in the debriefing session corroborated this finding in users they felt less affected by the testing situation as the usability test progressed. At the same time, about half of the participants in the multi-observer condition stated that they had perceived the presence of the non-interactive observers as a constant source of stress with little habituation taking place. The data did not provide conclusive evidence about the size of the calming-down effect (which the study never set out to examine but was included as a post-hoc analysis). Despite the degree of uncertainty associated with this issue (partly due to the impossibility to determine HRV), it appears to be safe to argue for an extension of the calming-down period by giving the test participants a warm-up task (which would not be part of the usability test). Furthermore, as it is currently not clear to what extent the effects of the presence of non-interactive observers will diminish after a certain time period, non-interactive observers (being placed in the same room like the test participants) should only be employed with caution.

Using physiological measures in the present study corresponded to the demands put forward by several researchers who argued that physiological scanning technologies should be integrated more strongly into ergonomic research (e.g. Hancock, Weaver & Parasuraman, 2002; Wastell & Newman, 1996; Wilson & Sasse, 2000). While previous lab-based experiments have shown that cognitive stressors (such as mental arithmetic tasks, reaction time tasks or the Stroop interference task) resulted in an increase of HRV in the LF band and a decrease in the HF band (Berntson & Cacioppo, 2004; Jorna, 1992), the results of the present study indicated that the presence of observers as a social stressor influences HRV in the LF band in the opposite direction as the cognitive stressors. No difference between stressors was found for the HF band. These results reiterate the need for a greater differentiation between stressors since they may have even opposite effects on different HRV bands. This is in line with the argument put forward by Berntson and Cacioppo (2004) in which they state that “it is clear that no single pattern of autonomic adjustments and associated changes in HRV will apply universally across distinct

stressors" (p. 59). These results indicate that physiological reactions to mental workload and social stressors may be different (Jorna, 1992; Pruyn et al., 1985).

The present study has a number of implications for usability practice as well as for future research. First, there is a need to examine the difference between participating and non-interactive observers. The one-observer setup showed the same results as the no-observer setup for performance (visual inspection indicated even better results for the former on all performance parameters), which suggests the possibility that a facilitator who has established a good rapport with the test participant may represent a source of support with performance-enhancing effects. Second, the study raises the question to what extent product-related effects can be separated from other influences on the different test outcomes (e.g. environmental effects due to poor setup of usability test). Since the reason for testing is to examine the effects of user-product interaction, additional environmental effects such as lab setup that impinge upon the test results clearly represent undesirable side-effects that need to be minimized. In the present study, users were able to make a clear distinction between the product (considered to be usable) and the test environment (considered inadequate if observers are present), resulting in a product evaluation (i.e. subjective usability measures) that was unaffected by the test environment. However, performance (i.e. objective usability measures) and the user's emotional state were both affected by the test environment, demonstrating the influence of such interfering variables in usability tests. Third, it is currently unclear whether the effects of the presence of non-interactive observers will disappear after sufficient exposure. Therefore, for the time being it appears advisable to refrain from placing non-interactive observers in the same room like the test participants. This may favour the use of remote usability testing as new product evaluation method which has gained in importance in usability practice over recent years (Dray & Siegel, 2004). Fourth, there was evidence for the sensitivity of HRV parameters to pick up variations in user stress, providing support for the utility of these measures. Despite these encouraging results, there may be concerns about the current suitability of HRV as an appropriate measure for the standard usability test given the considerable resource requirements and the need for substantial analyst expertise. In spite of these concerns, it appears to be promising to pursue these research activities since with technical advancements in measurement technology and in data analysis tools, the process of using HRV in usability tests is likely to become much simpler in the future.

6 The influence of design aesthetics in usability testing: effects on user performance and perceived usability

Abstract

This article examined the effects of product aesthetics on several outcome variables in usability tests. Employing a computer simulation of a mobile phone, 60 adolescents (14 –17 yrs) were asked to complete a number of typical tasks of mobile phone users. Two functionally identical mobile phones were manipulated with regard to their visual appearance (highly appealing vs not appealing) to determine the influence of appearance on perceived usability, performance measures and perceived attractiveness. The results showed that participants using the highly appealing phone rated their appliance as being more usable than participants operating the unappealing model. Furthermore, the visual appearance of the phone had a positive effect on performance, leading to reduced task completion times for the attractive model. The study discusses the implications for the use of adolescents in ergonomic research.

Keywords: usability test; aesthetics; adolescent; performance; mobile phone

Reprinted from *Applied Ergonomics, 41*, Sonderegger, A. and Sauer, J., The influence of design aesthetics in usability testing: effects on user performance and perceived usability, 403–410, Copyright (2010), with permission from Elsevier.

Introduction

Design aesthetics

Research in consumer ergonomics has indicated that product usability may not be the only major determinant of user satisfaction but that other design features also play an important role (Tractinsky et al., 2000; Norman, 2004b). Over recent years, this has led to a continual shift in consumer ergonomics, moving from a functional view of usability issues (with a focus on improving efficiency and effectiveness of product usage) towards an experiential perspective, which takes into consideration the whole user experience (Forlizzi & Battarbee, 2004; Brave & Nass, 2008). User experience comprises the entire set of effects elicited by the use of a product, including aesthetic experience, experience of meaning, and emotional experience (Desmet & Hekkert, 2007). This suggests that aesthetics may play an important role in product and systems design.

The issue of aesthetics enjoys a long historic tradition in the research literature, with psychologists and philosophers having carried out theoretical and empirical work in that field. This topic has been the subject of discussions by ancient Greek philosophers such as Plato (beautiful objects incorporate proportion, harmony, and unity among their parts) and Aristotle (universal elements of beauty are order, symmetry, and definiteness). In the domain of psychology, issues of aesthetics were first raised by Fechner (cited in Liu, 2003) whose aim was to discover the relationships between different design dimensions and perceived attractiveness through systematic manipulations of visual stimuli such as rectangles and ellipses. More recently, these ideas were taken up again to identify the features of stimuli (such as shape, colour, complexity, order, rhythm and prototypicality) that influence the attractiveness of an object (Liu, 2003; Hekkert & Leder, 2007).

In the research literature, the term design aesthetics is employed in two ways: it may refer to the *objective* features of a stimulus (e.g. colour of a product) or to the *subjective* reaction to the specific product features. To make a distinction between the two meanings, in the present study *aesthetics* refers to the objective design aspects of a product, including form, tone, colour, and texture (Postrel, 2003). Conversely, *attractiveness* refers to the individual's reaction to these product features and represents "the degree to which a person believes that the [product] is aesthetically pleasing to the eye" (van der Heijden, 2003; p. 544).

The response to aesthetic design is not only influenced by specific design factors (such as form or surface attributes) but may also be modified by characteristics of the individual, such as age, personality, cultural background or gender (Crilly et al., 2004). Because of its role in product marketing and consumer behaviour research (e.g. Meyers-Levy & Sternthal, 1991), gender may also be of particular interest in consumer ergonomics, though the evidence of the direction of the influence is far from being unequivocal. While some research has concluded that gender has little or no effect on aesthetic judgments (e.g. Lubner-Rupert & Winakor, 1985; Minshall et al., 1982; Morganosky & Postlewait, 1989, there is other work that did find differences (e.g. Holbrook & Corfman, 1984; Holbrook, 1986). However, since all of the work cited referred to non-interactive products such as clothes, it remains to be seen how gender moderates the effects of aesthetics in the context of operating interactive consumer products.

Usability testing

Given the role of aesthetics in product development, there is a need to examine the influence aesthetics have in usability testing. Usability testing is considered to be one of the most important and most widely used methods to evaluate product designs (Lewis, 2006). It aims to assess the usability of a product by simulating the user-product interaction under controlled conditions.

Usability is defined according to the International Standardisation Organisation as "the *effectiveness*, *efficiency* and *satisfaction* with which specified users can achieve specified goals in a particular environment" (ISO, 1998). *Effectiveness* refers to the extent to which a task goal is successfully achieved (e.g., proportion of users that are able to complete a given task). *Efficiency* refers to the amount of resources a user expends to reach a task goal. It can be measured by the deviation from the optimal user behaviour (e.g., task completion time, number of user actions to complete a task). Both effectiveness and efficiency represent different kinds of performance measures. *Satisfaction* can be considered as an attitude towards the product. It is a subjective measure that is typically collected in usability tests by means of questionnaires (e.g. Chin et al., 1988; Lewis, 1995; Kirakowski et al., 1998; Willumeit et al., 1996).

Design aesthetics and perceived usability

The influence of aesthetics on perceived usability has already been addressed in several studies. These studies reported a positive correlation between perceived attractiveness and perceived

usability for a range of products, such as computer-simulated cash machines (Kurosu & Kashimura, 1995; Tractinsky et al., 2000), websites (Hartmann et al., 2007; Schenkman & Jonsson, 2000) and computer software (Hassenzahl, 2004). While in these studies design aesthetics (attractive vs. unattractive) was not manipulated experimentally (and hence it cannot be excluded that perceived attractiveness and perceived usability were confounded), there are also studies in which an experimental manipulation of aesthetics was carried out. This includes the variation in colour settings of a webpage (Nakarada-Kordich & Lobb, 2005), the manipulation of the shape of an electronic phonebook-simulator (Ben-Bassat et al., 2006), the variation in the design of a webpage (following mathematical rules and two choices of colour settings; (Brady & Phillips, 2003), and the manipulation of the colour of casing and screen of a mobile phone (Sauer & Sonderegger, 2009). All these experiments confirmed that perceived usability was positively influenced by the aesthetics of the product. With regard to the psychological mechanisms behind this effect, the halo-effect has been put forward as a possible explanation. The halo effect describes the phenomenon that a specific, salient characteristic of a person or an object influences the apperception of other characteristics. This is analogous to the "what is beautiful is good"-stereotype, known from social psychology, that has been postulated to explain the phenomenon that physically attractive persons are considered to possess more positive personality traits than unattractive persons (Dion et al., 1972). Since attributes of physical beauty are obvious and accessible to others very early in the interaction between humans, they are assumed to colour later perceptions of other personal characteristics. Similarly, in usability testing the user's attitude towards a product is formed very rapidly (i.e. in about 50 ms) during user-product interaction (Lindgaard et al., 2006), which exemplifies the importance of the very first impression. Overall, there is ample evidence of the positive influence of aesthetics on perceived usability.

Design aesthetics and user performance

While the positive relation between aesthetics and perceived usability has been well demonstrated by empirical research, it is less clear how aesthetics is linked with objective measures of performance in usability tests. Only very few studies have examined the effect of aesthetics on performance measures, albeit with somewhat inconsistent findings. Two studies found evidence of performance decrements when using an aesthetically pleasing product. For example, test participants showed poorer performance using an appealing computer simulation of

an electronic phonebook (Ben-Bassat et al., 2006). Similar results were obtained in a study in which the aesthetics of a mobile phone was manipulated experimentally (Sauer & Sonderegger, 2009). However, two other studies found no effect of aesthetics on performance. Hartmann, Sutcliffe and de Angeli (2007) reported no correlation between perceived attractiveness and task completion time when comparing three different webpages. Thüring and Mahlke (2007) varied the design aesthetics of existing MP3-players, with the results showing no effects of aesthetics on task completion time and error rate.

One may envisage two different effects of aesthetics on performance measures: an "increased motivation"-effect (i.e. increments in performance) or a "prolongation of joyful experience"-effect (i.e. decrements in performance). For the "increased motivation" effect, one may speculate that technology that is aesthetically pleasing might put the user at ease (Lindgaard, 2007) or put the user "in flow" (Csíkszentmihályi, 1997), which both may result in increased performance (e.g. reduced task completion time). In contrast, the "prolongation of joyful experience"-hypothesis would predict decreased user performance because the user enjoys the beauty of the product and therefore concentrates less on the task to be completed. This may lead to longer task completion times due to extended observation times during user-system interaction. The empirical findings reported above provided cautious support for the "prolongation of joyful experience"-explanation while no support has yet been found for the "increased motivation"-effect.

The present study

The primary research question of this study addressed the influence of aesthetics on central outcome variables of usability testing, such as perceived usability and user performance. For this purpose, two functionally identical mobile phones were manipulated with regard to their visual appearance to make them either aesthetically appealing or unappealing. In all system features other than aesthetic appeal, the two appliances were identical. The mobile phone was chosen as a technical device because it has a stronger affective component than most other interactive consumer products (e.g., vacuum cleaner). This will give additional weight to design aesthetics. The present study was conducted with adolescents as an important group of mobile phone users (Milanese, 2005). In addition to the influence of aesthetics, we have examined the influence of gender as a secondary research question.

Based on the research literature reviewed, the following three hypotheses were formulated: (a) User performance will be better for the more aesthetically pleasing product than for the less pleasing one. (b) Perceived usability will be higher for the aesthetically more pleasing product than for the less pleasing one. (c) The difference in perceived usability between the two conditions will be less pronounced after the usability test than prior to it, due to the diminishing influence of aesthetic after the user had actual experience with the product. Because of the equivocal pattern for gender, no hypothesis was formulated for the effects of gender.

Method

Participants

The sample of this study consists of 60 participants (52% female). All of them were pupils doing their GCSEs (General Certificate of Secondary Education) at a secondary school in Thun (Switzerland), aged between 13 and 16 years (M = 14.2). Self reports showed that they were quite experienced mobile phone users, employing their mobile phone on average 8.7 times per day (SD = 10.6). Their self-rated expertise in operating a mobile phone was M = 65.0 on a 100 mm visual analogue scale. The ends of the scale were labelled “very little experience”, and “a great deal of experience”, with higher values indicating more experience. The two experimental groups did not differ in their self-rated expertise in mobile phone usage ($t < 1$) and in their stated frequency of daily phone usage ($t = 1.57$, $df = 55.7$, $p > .05$). There was no difference between male and female participants with regard to their perceived expertise in mobile phone usage ($t < 1$) and their reported frequency of usage ($t < 1$).

Experimental design

A 2 x 2 mixed design was employed in the experiment, with *aesthetics of design* as a between-subjects variable. Participants were randomly assigned to a group using a prototype of mobile phone with an appealing design or an unappealing one. To determine the effects of product usage experience, some measures were recorded repeatedly during the usability test. This within-participants variable was varied at two levels: prior to the product usage in the usability test and following the usability test.

The influence of gender was examined by using this variable as a covariate. The distribution of gender across conditions was unequal (e.g. 12 females used the unappealing phone

while 19 females operated the appealing phone), due to the particular distribution of gender in the participating school classes.

Measures and instruments

Perceived product attractiveness

The attractiveness of the appliance was measured before and after product usage. The measure (prior to usability test) served as a manipulation check. A one-item scale was used ("the design of the mobile phone is very appealing"), with a seven-point Likert scale (strongly agree, agree, partly agree, neither agree nor disagree, partly disagree, disagree, strongly disagree) as a response format. A single-item scale was chosen, to ensure that participant motivation was maintained throughout the testing session. Since the main goal of the study was to attain an overall assessment, the use of a 1-item measure is justifiable if the item is unambiguous and captures the main concept (Wanous et al., 1997). This type of scales has been employed in previous usability studies (e.g. Tractinsky et al., 2000).

Perceived usability

Similar to the evaluation of the attractiveness of the prototype, test participants were asked to assess the usability of the mobile phone before and after product usage on a one-item scale ("The mobile phone seems to be very usable"). Again, a seven-point Likert scale was used (strongly agree, agree, partly agree, neither agree nor disagree, partly disagree, disagree, strongly disagree).

As a more detailed measure of the system usability comprising several subscales, a German translation of the Post System Study Usability Questionnaire (PSSUQ) (Lewis, 1995) was employed after product usage. This instrument has been widely applied for usability testing in laboratory settings. The questionnaire was slightly modified by removing four items that were irrelevant for the intended application area. The remaining items are presented in table 5. To improve comprehensibility, items were adapted to the appliance it was used for (e.g., "system" was replaced by "mobile phone"). Users rated the items on the same seven-point Likert scale as the single-item scale above. The PSSUQ comprised the following three subscales: system usefulness, information quality and interface quality. The overall internal consistency of the questionnaire as well as the internal consistency of the subscales was found to be satisfactory (see table 5).

Table 5: Adapted version of Post Study System Usability Questionnaire (PSSUQ) (Lewis, 1995)

Post Study System Usability Questionnaire *(Cronbach's α=.88)*
Subscale "system usefulness" (Cronbach's α=.91)
Overall, I am satisfied with how easy it is to use this mobile phone.
It was simple to use this mobile phone.
I was able to complete the tasks and scenarios quickly using this mobile phone.
I felt comfortable using this mobile phone.
It was easy to learn to use this mobile phone.
I believe I could become productive quickly using this mobile phone.
Subscale "information quality" (Cronbach's α=.68)
The mobile phone gave error messages that clearly told me how to fix problems.
Whenever I made a mistake using the mobile phone, I could recover easily and quickly.
The information provided by this mobile phone was clear.
It was easy to find the information I needed.
The information was effective in helping me complete the tasks and scenarios.
The organisation of information on the mobile phone's display was clear.
Subscale "interface quality" (Cronbach's α=.87)
The interface of this mobile phone was pleasant.
I liked using the interface of this mobile phone.
Overall satisfaction
Overall, I am satisfied with this mobile phone.

User performance

Three measures of user performance were recorded. *Task completion time* referred to the time needed to accomplish the task. *Interaction efficiency* is a composite parameter, dividing the optimal number of user manipulations by the actual number of user inputs. Lastly, the *number of error messages* that have been displayed when the user chose a wrong navigation option was recorded.

Materials

Two functionally identical computer prototypes of a mobile phone were used in this study. One version was aesthetically appealing, the other one not so (see figure 3). It is useful to note that

users only employed the navigation buttons in the top section of the interface for task completion (i.e. they did not need to use the numeric keys). The buttons in the top section were of the same size for both appliances. The functionality of the two appliances was exactly the same. This was because the overlaid event triggers (in the form of invisible push buttons) were exactly of the same size for both appliances while only the form (but not the size) of the visible shell differed slightly between appliances. To control for objective usability differences between the two appliances, we calculated the average time per click. The results of the analysis showed that there was no difference between the appealing ($M = 3.0$; $SD = 0.8$) and the unappealing design ($M = 3.1$; $SD = 0.8$) ($F < 1$). This suggests that task difficulty for the two appliances was the same.

The two designs were developed, using graphic design software (Photoshop). The designs were based on previous research, which identified a number of factors that determine object attractiveness, such as colour, texture, symmetry, and clarity (Ngo et al., 2003; Postrel, 2003).

The dialogue structure of the mobile phone was based on the functionality of a SonyEricsson™ SE W800i. Compared to the original appliance, the functionality of the prototype was limited. Only for the task-relevant menu items, the dialogue structure was modelled in full depth. For functions that were irrelevant for task completion, only the two top levels of the dialogue structure were represented. An error message was displayed ("wrong path, please go back") when a user selected a function that was not simulated in the prototype (i.e. more than two clicks away from the optimal dialogue path). The computer simulation of the dialogue structure was developed using Powerpoint™. Both computer simulations (appealing and unappealing) were installed on a Toshiba Portege™ M200 TabletPC. For the interaction with the prototype, a computer mouse (Logitech Pilot Optical) was used.

Pilot study

In a pilot study, different design alternatives for the prototype of the mobile phone were compared. 10 participants (aged between 14 and 16 yrs) evaluated the attractiveness of these designs (one aesthetically appealing and two aesthetically unappealing ones). The participants were recruited from the same population as the sample of participants of the main study. The two aesthetically unappealing designs differed in form and colour setting compared to the appealing one. Both unappealing prototypes consisted of a disharmonious facia of different colours (blue, yellow, pink, and grey). The buttons were either purple or looked like if they were made out of

wood. On a scale raging from 1 to 10, ratings of perceived attractiveness differed remarkably between the appealing and the two unappealing prototypes: (a) M = 8.1 (SD = .88); (b) M = 2.0 (SD = 1.33); (c) M = 2.6 (SD = 1.50). The prototype with the highest score (fig.1a) and the one with the lowest score (fig.1b) in the attractiveness rating were selected for the main study.

Figure 3: Two prototypes employed in experiment: (a) aesthetically appealing design; (b) aesthetically unappealing design

User tasks

For the usability evaluation, two tasks had to be completed by test users. These tasks were chosen because they represent typical activities in mobile phone usage. The first task ("text message") involved sending a prepared text message to another phone user. This task could be completed with a minimum number of 9 clicks. In the second task ("phone number suppression"), test users had to change the mobile phone settings in such a way that one's own phone number is

suppressed when making a call. To complete this task, a minimum number of 16 clicks were necessary.

Procedure

The study was conducted in a computer lab of the school. Participants were recruited from different classes on a voluntary basis and within each class, participants were randomly assigned to one of the experimental conditions. Any difference in age or ability between experimental groups is expected to be balanced by the procedure of randomly allocating participants. Participation in the study took about 20 minutes. All participants were tested individually. After being welcomed by the experimenter, participants were informed that they would take part in a usability test and would have to operate a computer-simulated prototype of a mobile phone. Prior to operating the prototype, participants were asked to rate their previous experience with mobile phones and to rate attractiveness and usability of the mobile phone on the two single-item scales. Then, participants completed the two experimental tasks. Immediately after task completion, the two single-item scales and the PSSUQ were presented. The experiment was concluded with a debriefing session, in which the participant was given the opportunity to give further feedback about the prototype and the testing procedure.

Statistical analysis

To examine the impact of design aesthetics and product usage on subjective evaluations of attractiveness and usability, a two-factorial analysis of variance was used. For the analysis of the performance data a one factorial analysis of variance was carried out. In both cases, the influence of gender was examined by entering this factor as a covariate.

Results

Perceived product attractiveness

The data of the attractiveness evaluation of the two prototypes before and after product usage is presented in figure 4. Representing a manipulation check, the data confirmed that the aesthetically appealing prototype was rated significantly more attractive than the unappealing prototype ($M_{appealing} = 5.3$ vs. $M_{unappealing} = 3.15$; $F = 39.8$; $df = 1, 58$; $p < .001$). Furthermore, an

interaction between prototype and product usage was found ($F = 4.7$; $df = 1, 57$; $p < .05$), showing an increase in the perceived attractiveness rating of the aesthetically appealing prototype after product usage whereas the attractiveness-rating of the unappealing prototype decreased after product usage. The main effect of product usage (before vs. after) was not significant ($F = 1.5$; $df = 1, 57$; $p > .05$). The covariate *gender* was not related to the perceived product attractiveness, neither before nor after user-product interaction (all $F < 1$).

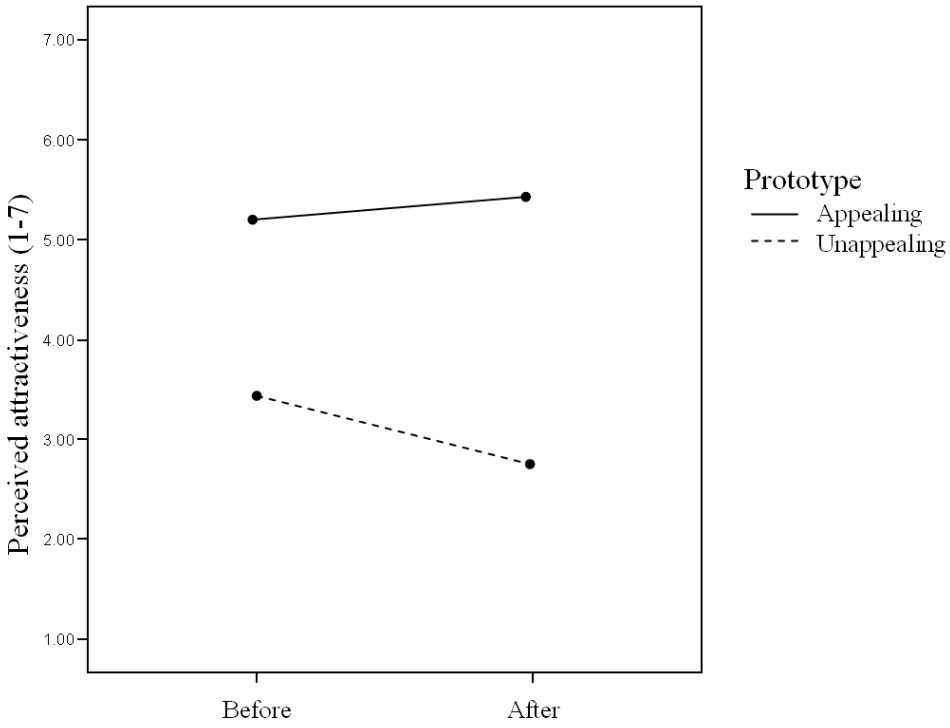


Figure 4: User ratings of perceived attractiveness (1-7) of the prototype before and after product usage as a function of design aesthetics

Perceived usability

Perceived usability was measured prior to task completion and after task completion by the 1-item scale as well as with the PSSUQ after task completion. The ratings on the 1-item scale differed significantly between the two prototypes ($M_{appealing} = 6.14$ vs. $M_{unappealing} = 5.32$; c.f.

figure 5). The appealing prototype was rated more usable than the unappealing one ($F = 9.8$; $df = 1, 57$; $p < .01$). The actual use of the prototype did not influence the user's usability rating ($F < 1$) and also the interaction between prototype and product usage was not significant ($F < 1$). Gender was not related to the perceived usability, neither before nor after user-product interaction (all $F < 1$).

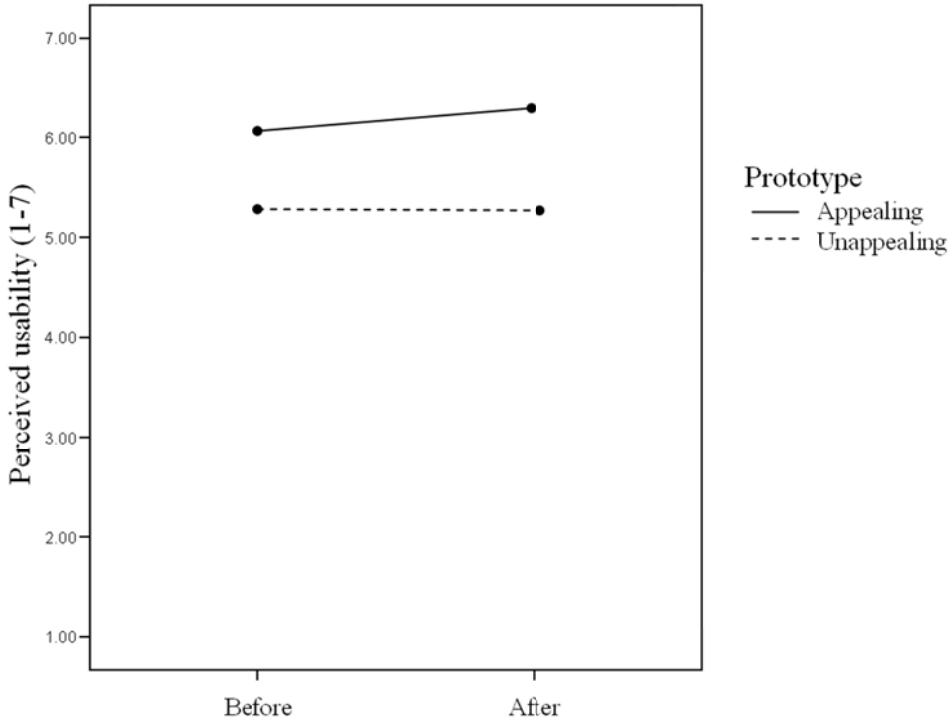


Figure 5: User ratings of perceived usability (1-7) on the one-item scale before and after product usage as a function of design aesthetics

The perceived usability ratings after product usage on the PSSUQ are similar to the ratings on the 1-item scale (c.f. table 6). The analysis revealed that overall ratings were significantly higher for the appealing prototype than for the unappealing one ($F = 20.8$; $df = 1, 57$; $p < .001$). A separate analysis of the three subscales confirmed the same pattern for system usefulness ($F = 13.6$; $df =$

1, 57; $p < .001$), information quality ($F = 7.2$; $df = 1, 57$; $p < .01$) and interface quality ($F = 14.5$; $df = 1, 57$; $p < .001$). Gender showed no relationship with perceived usability (all $F < 1$).

Table 6: Perceived usability (1-7) on the Post System Study Usability Questionnaire as a function of design aesthetics

	Appealing prototype M (SD)	Unappealing prototype M (SD)
Overall scale (item 1-15)	6.13 (0.48)	5.19 (0.91)
System usefulness (item 1-6)	6.29 (0.62)	5.32 (1.22)
Information quality (item 7-12)	6.27 (0.56)	5.66 (0.87)
Interface quality (item 13-14)	5.30 (1.38)	3.60 (1.69)

User performance

Task completion time. The analysis of the data of task completion time revealed a significant difference for the two designs (c.f. Table 7). It showed that participants using the appealing prototype needed less time to complete their tasks than the participants using the unappealing prototype ($F = 8.9$; $df = 1, 57$; $p < .01$). The covariate *gender* was not related to task completion time ($F = 1.9$; $df = 1, 57$; $p > .05$).

Interaction efficiency. Similar to the findings of task completion time, the analysis of the data on interaction efficiency (optimal click number divided by actual number of clicks; c.f. table 7) indicated a significant effect of design aesthetics ($F = 8.8$; $df = 1, 57$; $p < .01$). Participants using the appealing prototype needed fewer clicks to complete their tasks than participants using the unappealing one. Gender was not related to interaction efficiency (all $F < 1$).

Errors. The analysis of errors that occurred during task completion (c.f. table 7) revealed that participants using the attractive prototype committed significantly fewer errors than the participants using the unappealing prototype ($F = 12.0$; $df = 1, 57$; $p < .001$). This shows that all three performance measures indicate better performance when operating an appealing prototype. Gender was related to the error rate ($F < 5.1$; $df = 1, 57$; $p < .05$), indicating that female participants committed more errors than male ones.

Table 7: Measures of user performance as a function of design aesthetics and gender

	Appealing prototype M (SD)	Unappealing prototype M (SD)	Overall M (SD)
Task completion time (s)	147.7 (58.0)	198.5 (83.6)	173.1 (76.2)
female	*147.2 (57.8)*	*230.1 (102.4)*	*179.3 (86.8)*
male	*148.5 (64.0)*	*177.4 (62.9)*	*166.4 (63.8)*
Interaction efficiency index (%)	59 (17)	46 (19)	53 (19)
female	*58 (16)*	*42 (20)*	*52 (19)*
male	*61(19)*	*48 (18)*	*53 (19)*
Number of errors (per trial)	2.0 (2.0)	4.8 (5.0)	3.4 (4.0)
female	*2.2 (2.1)*	*7.0 (6.4)*	*4.1 (4.8)*
male	*1.5 (1.6)*	*3.3 (3.1)*	*2.7 (2.8)*

Discussion

The findings showed that perceived usability was higher for appealing products than for unappealing ones, even though there was no difference between the two appliances in the objective quality of usability. This pattern was observed for the one-item scale as well as for the more elaborate instrument PSSUQ on all its subscales. These results provide further confirmation of the positive influence of aesthetics on perceived usability observed in previous work (Nakarada-Kordich & Lobb, 2005; Ben-Bassat et al., 2006; Brady & Phillips, 2003; De Angeli et al., 2006). This tendency which was consistently observed across different adult user populations was also applicable in the case of adolescent users. Furthermore, it is noteworthy that the actual completion of the experimental tasks did not change perceived usability as one would expect. If there was an influence of aesthetics prior to using the appliance, one would expect this influence to decrease in size as the user becomes more familiar with the appliance. However, the ratings remained stable (if anything, the difference widened rather than narrowed as visual inspection of the data suggests). This stability in ratings observed before and after the usability test was also observed in an experiment with adult users, employing a similar experimental set-up (Sauer & Sonderegger, 2009). This suggests that the observed effects are consistent across age groups.

Furthermore, similar to social psychology, where the "what is beautiful is good"-stereotype seems to represent a cross-cultural phenomenon (Chen et al., 1997), the cross-cultural quality of the effect also appears to apply to judgements of perceived attractiveness on technical artefacts since similar findings obtained with the Swiss sample in the current study were reported from studies conducted in Japan (Kurosu & Kashimura, 1995), Israel (Tractinsky et al., 2000), and Germany (Thüring & Mahlke, 2007).

While the effects on perceived usability were in line with previous work, the influence of aesthetics on user performance was in contrast to previous findings. The present study provided support for an "increased motivation"-effect, with users showing better performance with the appealing prototype. Previous work, however, found support for the "prolongation of joyful experience"-effect, with users taking more time to complete a data entry task (Ben-Bassat et al., 2006) and to operate a mobile phone (Sauer & Sonderegger, 2009) when using the more appealing version of the technical artefact. These differences may be due to inherent domain characteristics (leisure vs work context). One may assume that the "increased motivation" effect would be more likely to occur in a work context while the "prolongation of joyful experience"-explanation would be more likely to be observed in a leisure context. As the present study was carried out in a school setting (which most pupils would not consider a leisure-oriented environment), a stronger performance-orientation may have ensued from this, resulting in a higher motivation to complete the tasks as fast as possible. Interestingly, this effect was opposite to the one observed in a previous study (Sauer & Sonderegger, 2009), which used a similar experimental set-up with a mobile phone being operated but in a leisure-oriented context. In such a context, the focus may be less on performance but more on fun and enjoyment which supports the mechanism of the "prolongation of enjoyable experience" effect.

While the results clearly demonstrated that the manipulation check was successful (since the two mobile phones were rated very differently with regard to their perceived attractiveness), more interesting was the observation that the difference in perceived attractiveness between high and low aesthetics widened after the usability test. This observation may be interpreted by referring to the attitude polarization effect (Lord et al., 1979). The initial attitude (which is formed very early during user-product interaction; (Lindgaard et al., 2006)) may have become more extreme due to biased information assimilation (MacCoun, 1998). The occurrence of attitude polarization among adolescents was also demonstrated in the context of reasoning about religious affiliations (Klaczynski & Gordon, 1996). It showed that adolescents' reasoning was

systematically biased to protect and promote pre-existing beliefs. Overall, the present findings may suggest that usability has little influence on perceived attractiveness. Otherwise, one would have expected some narrowing of the difference, as users gained increasing experience with the usability of the product (which was identical for both conditions). This speculative explanation needs to be empirically tested by manipulating product usability and determine its effects on perceived attractiveness, with particular consideration to be given to the long-term effects over repeated practice trials.

Overall, gender had little effect on outcome variables. This is in line with the bulk of the literature (albeit a small number of studies did find an effect), suggesting that the influence of aesthetics is not only observed across cultures and age groups but also across gender. Although an effect was recorded on a single measure (i.e. suggesting that female users committed more errors with the unappealing prototype than male users), it was difficult to interpret and, given its small effect size, it should not be taken as evidence for a general consideration of gender as a crucial variable that moderates the influence of aesthetics in usability evaluations. Although the unequal distribution of gender reduced the power of the covariance analysis, even with a more balanced distribution, it is unlikely that the effect of gender would have been significant, given the size of the effects.

It is important to note that the results in this study are based on a sample of adolescent test users. The use of adolescents as a separate user group seems to be increasingly relevant, given their growing financial freedom of manœuvre (e.g. Shim, 1996), their influence in family decision-making (e.g. Beatty & Talpade, 1994; Foxman et al., 1989) and their role as future (adult) customers with whom it is important to establish an early brand relationship (Khadir, 2007). Against this background, it is justifiable and increasingly necessary to carry out research with adolescent users. The current study also provided first hints about possible differences in the effects of aesthetics compared to adults, though we do not know whether these were due to differences in user groups (i.e. adults vs. adolescents) or in usage context (work vs. leisure). We would therefore caution against a generalisation of the findings of the present study to other user groups.

Some limitations with regard to the interpretation of the results are acknowledged. While the effects of aesthetics may be due to the mechanisms discussed above, alternative explanations are also possible. The attractive phone might have been perceived as a conventional phone that can be purchased in the shops (and is fully usable) whereas the unattractive one might have been

perceived as having a rather unusual design (which is not yet fully developed). Therefore, the more conventional product might have been evaluated more positively, resulting in a confounding effect of familiarity and aesthetics. On the basis of the available data, it is not possible to control for such a confounding effect. However, participant feedback after task completion did not indicate a difference in the perception of prototypicality between the two prototypes. Furthermore, to minimise a possible effect of familiarity, users of a recent SonyEricssonTM mobile phone (upon which our prototypes were based) were not allowed to take part in the study. As the present study and previous work have demonstrated, there seems to be increasing evidence for the influence of aesthetics beyond subjective parameters such as perceived usability. Indeed, aesthetics may influence performance, with empirical evidence having been observed for both effects ("prolongation of joyful experience"-effect vs. "increased motivation"-effect). This suggests the need for experiments to address the following issues in future research. First, direct comparisons between adolescents and adults should be made. We may predict that adolescents might attach even more importance to the mobile phone's aesthetics (resulting in a stronger effect of aesthetics on usability test outcomes) since they are often prone to extreme self-focus and are excessively concerned with what their peers think of them (Magrab, 2005). Second, different usage contexts such as the domestic and work domain should be compared. We may predict a stronger influence of aesthetics in the domestic (and leisure) domain than in the work domain. Third, it would be of interest to determine to what extent the influence of aesthetics is moderated by the prestige value associated with the product. The prestige value of a mobile phones may be considered high (Dedeoglu, 2004), compared to other products such as a vacuum cleaner or an electric fire. As a concluding remark, we would like to point out that the usage of interactive consumer products should not be considered separately from the purchase decision. As the user is often the buyer of a product, ergonomic issues become more strongly interwoven with issues pertaining to consumer psychology.

7 The influence of prototype fidelity and aesthetics of design in usability tests: effects on user behaviour, subjective evaluation and emotion

Abstract

An empirical study examined the impact of prototype fidelity on user behaviour, subjective user evaluation and emotion. The independent factors of prototype fidelity (paper prototype, computer prototype, fully operational appliance) and aesthetics of design (high vs. moderate) were varied in a between-subjects design. The 60 participants of the experiment were asked to complete two typical tasks of mobile phone usage: sending a text message and suppressing a phone number. Both performance data and a number of subjective measures were recorded. The results suggested that task completion time may be overestimated when a computer prototype is being used. Furthermore, users appeared to compensate for deficiencies in aesthetic design by overrating the aesthetic qualities of reduced fidelity prototypes. Finally, user emotions were more positively affected by the operation of the more attractive mobile phone than by the less appealing one.

Keywords: usability test; prototype fidelity; aesthetics; mobile phone

Reprinted from *Applied Ergonomics, 40*, Sauer, J. and Sonderegger, A., The influence of prototype fidelity and aesthetics of design in usability tests: effects on user behaviour, subjective evaluation and emotion, 670-677, Copyright (2009), with permission from Elsevier.

Introduction

Prototype fidelity

Product designers are typically faced with the problem that human behaviour in operating a system needs to be predicted although the system has not yet been fully developed. The system may only be available in a rudimentary form, which falls well short of a fully operational prototype. This may range from specifications (descriptions based on requirement analysis) through cardboard mock-ups to virtual prototypes.

The question of which prototype is to be used for usability testing is strongly influenced by a number of constraints that are present in industrial design processes, notably time pressure and budgetary limitations. This usually calls for the use of low-fidelity prototypes (e.g., paper prototype) because they are cheaper and faster to build. Although prototypes of various forms are widely used in industry, there is little comparative research on the utility of prototypes at different fidelity levels. A review of the research literature has revealed a total of 9 studies in which comparative evaluations of different prototypes were carried out (Sefelin et al., 2003; Virzi et al., 1996; Säde et al., 1998; Nielsen, 1990; Catani and Biers, 1998; Walker et al., 2002; Wiklund et al., 1992; Hall, 1999; Sauer et al., 2008). The majority of studies concluded that the reduced fidelity prototypes provided equivalent results to fully operational products. Only three studies (Nielsen, 1990; Hall, 1999; Sauer et al., 2008) reported some benefits of higher fidelity prototypes over lower fidelity prototypes.

The decision of selecting a prototype for human factors testing entails a dilemma. On the one hand, a prototype of too high fidelity is very time-consuming and expensive to build, hence valuable resources are wasted. On the other hand, the findings obtained with a prototype of too low fidelity may not be valid. This requires the careful consideration of what level of fidelity would be best to opt for. The concept of prototype fidelity is quite broad in scope, encompassing a number of different dimensions upon which a prototype can differ from the reference product. Virzi et al. (1996) have suggested a classification system that distinguishes between four dimensions of fidelity: degree of functionality, similarity of interaction, breadth of features, and aesthetic refinement.

Degree of functionality is concerned with the level of detail to which a particular function has been modelled. For example, the user-product dialogue for taking a picture with a mobile phone can be modelled in its entirety or in a reduced form. *Interactivity* refers to the type of

interface (i.e. controls and displays) with which the prototype is modelled. For example, on a computer-based simulation of a telephone, one may use a touch screen to enter a phone number directly with the fingers (higher fidelity) or use a mouse to do the same on a conventional screen (lower fidelity). *Breadth of functions* refers to the extent to which all functions of the target product are modelled in the prototype (e.g., 4 out of 5 displays and 3 out of 4 control elements of the real system are represented in the prototype). *Aesthetic refinement* refers to the extent to which there are similarities between the prototype and the target product with regard to physical properties, such as shape, size, colour, texture and material. This dimension has also been referred to as the 'look' of the prototype (e.g., Snyder, 2003). The model of Virzi et al. (1996) clearly indicates that a prototype can differ from the reference product in many different aspects. Overall, the model of Virzi et al. may represent a useful framework for designers to guide the prototype development process.

Usability testing

In order to assess the utility of prototypes, usability tests are often used since they allow for user-product interaction to be measured under controlled conditions. The ISO Standard of usability (ISO 9241-11) refers to the three main aspects of usability: effectiveness, efficiency, and user satisfaction. Effectiveness and efficiency may be considered objective measures since they examine actual user behaviour while user satisfaction refers to subjective measures that take into account the user's opinion and feelings.

User behaviour

Effectiveness refers to the extent to which a task goal is successfully achieved with the product (Jordan, 1998a). This may be measured by the proportion of users that can actually complete a given task. In addition to rate of task completion, effectiveness may also be measured by the quality of the output (e.g., taste of a cup of coffee brewed with a coffee maker). *Efficiency* refers to the amount of resources expended to accomplish a task goal (Jordan, 1998a). Typical measures of efficiency are deviations from the critical path (e.g., number of superfluous clicks on a menu during task completion), error rates (e.g., number of wrong commands), and time on task (e.g., time needed to accomplish the task).

All these measures may be taken during usability tests. However, knowledge about the influence of different levels of prototype fidelity on these outcome measures is limited. Most of the studies cited in the literature review above focused on usability problems alone, with a

smaller number of studies also measuring user satisfaction (e.g., Catani and Biers, 1998; Wiklund et al., 1992). The review of the studies also suggests that empirical research has concentrated very much on effectiveness measures, with efficiency issues being somewhat neglected. The focus on usability errors may have contributed to a largely positive evaluation of prototypes of lower fidelity in usability tests, which might not be entirely justified. It remains to be empirically tested whether this positive evaluation can still be maintained when a wider range of measures of user behaviour is examined.

Subjective user evaluations and emotions

In addition to objective data, data on user satisfaction are often collected during usability tests by means of standardised questionnaires and semi-structured interviews. The questionnaires range from rather short instruments (e.g., 10-item Software Usability Scale of Brooke, 1996) to very elaborate instruments that measure different facets of user satisfaction (e.g., Questionnaire for User Interaction Satisfaction containing 71 questions; Chin et al., 1988). These questionnaires have been typically employed on fully operational products so that it remains to be seen to what degree reduced fidelity prototypes provide valid data to estimate user satisfaction with the real product.

While user satisfaction has been a notion in usability testing for some time, more recently consumer product design has also become concerned with concepts such as joy, pleasure and fun (Norman, 2004a; Jordan, 1998b; Jordan, 2000). While the concept of satisfaction may be considered an attitude towards the product (i.e. like the concept of job satisfaction in a work context; e.g., Schleicher, Watt and Greguras, 2004), joy, pleasure and fun (which appear to be used largely synonymously in the usability literature) represent emotions, which, in contrast, have a clear focus on the internal state of the user. Emotions are increasingly considered to be an important issue in consumer product design, as a rising number of publications have paid testimony to (e.g., Helander and Khalid, 2006; Norman, 2004a; Brave and Nass, 2003). For example, there is evidence that the emotional response to a product is more influential than cognitive components in determining consumer decision-making (Shiv and Fedorikhin, 1999). Emotions are also of particular interest because they represent a faster and more immediate reaction to an object than a pure cognitive evaluation (Khalid, 2006).

Concerning the effects of prototype fidelity, it is of particular interest to what extent emotions associated with product utilisation can be predicted from low- and medium fidelity

prototypes. In order to assess the user's emotional response, product developers typically use prototypes of higher fidelity for this purpose (e.g., 3D mock-up), which are characterised by considerable aesthetic refinement. This is due to concerns that lower fidelity prototypes (e.g., involving only a rough sketch of the design) would not elicit the same emotional response. If a prediction of the emotional response was possible on the basis of a prototype with reduced fidelity, it would allow designers to measure the impact of a product on user emotions at an earlier stage in the design process rather than having to wait until an aesthetically refined prototype can be made available.

Closely related to emotions is the aesthetic appeal of a product. There are a number of concepts in the research literature that refer to the exterior properties of a product and the user's response to it, such as aesthetics, appearance, attractiveness and beauty (e.g., Hekkert, Snelders and van Wieringen, 2003; Chang, Lai and Chang, 2007; Hassenzahl, 2004). However, these concepts are not employed consistently across research communities and research fields. For example, with regard to the concept of aesthetics, Lavie and Tractinsky (2004) have distinguished between the factors classical and expressive aesthetics while Hekkert et al. (2003) have identified novelty and typicality as factors. Other work considers the term aesthetics as the user's response to the appearance of the product (Crilly et al., 2004). In the present article, we will use the term aesthetic of design to refer to the visual appearance of a product (i.e. independent variable) whereas the users' response to these product properties is referred to as attractiveness (i.e. dependent variable).

Aesthetics of product design has long been considered an important issue in the field of industrial design (e.g., Yamamoto and Lambert, 1994). However, in the field of ergonomics, only more recently there have been calls for a stronger consideration of aesthetics as a pertinent factor in system design in addition of safety, usability and comfort (e.g., Liu, 2003). While aesthetics has also been linked to consumer decision-making, its influence may not be limited to that field since it may also affect the perceived usability of products. For example, research has indicated that aesthetic products are perceived as being more usable than less appealing ones (Tractinsky, 1997). This finding suggests that the influence of aesthetics is not limited to the product's appeal to the user but also affects usability ratings and, possibly, the way the product is being used.

The present study

The review of the literature revealed that there is only little work that examined the effects of prototypes fidelity on efficiency measures, user satisfaction, emotions and attractiveness. The limited work available mainly focussed on effectiveness measures (e.g., number of users that were able to complete the task). Against this background, the main research question examines the extent to which data obtained in usability tests with prototypes of reduced fidelity allow the prediction of user responses (i.e. observed behaviour and subjective evaluations) to the fully operational system. This was investigated by comparing paper and computer-simulated prototypes with fully operational products. A subsidiary research question was concerned with the aesthetic appeal of the design and to what extent it may modify the relationship between prototype fidelity and user responses.

The mobile phone was used as a model product. This appliance was regarded as particularly suitable for the purpose of this study because it is not only functionality and usability that are important for this product group. A mobile phone may be considered a lifestyle product to which a certain prestige value is attached, which may trigger off stronger emotional reactions during user-product interaction than a conventional product. The measures taken in this study covered the main outcome variables of a usability test. This included various performance measures as well as subjective measures ranging from usability ratings to emotional states.

Based on the research reviewed, the following research assumptions were formulated:

(a) User performance would be higher for the fully operational product than the two reduced fidelity prototypes (task completion time and efficiency of operation).

(b) The difference in user behaviour and subjective usability ratings between the fully operational product and reduced fidelity prototypes would be larger for the paper prototype than for the computer-based prototype since the latter is more similar to the fully operational product.

(c) An aesthetically more appealing appliance would create more positive emotions and would receive higher usability ratings than a less appealing product.

(d) For the fully operational product, the effects of design aesthetics on emotion and subjective usability would be more pronounced than for the reduced fidelity prototypes (i.e. interaction *fidelity level* x *appliance usability*). This is because a less appealing aesthetic design would be

more tolerable to users on an unfinished prototype than on a fully operational product with a finalised design.

Method

Participants

Sixty participants (58.3 % male, 41.7 % female) took part in the study, aged between 19 and 41 yrs (M = 23.8 yrs). They were students of the University of Fribourg and all of them were regular users of a mobile phone. A strict selection criterion was that participants should not have been familiar with the particular mobile phone they were going to use in the study. Participants were not paid for their participation.

Some of the participants had however experience with other models of the same brand they used in the experiment. In total, 23 participants were found to have such previous experience. However, post-hoc tests comparing participants with and without previous brand experience showed no difference for any of the dependent variables (all $t < 1$), suggesting no significant influence of this factor.

Experimental design

A 3 x 2 between-subjects design was employed in the study. The main independent variable *prototype fidelity* was varied at three levels: paper prototype, computer-based prototype, and fully operational appliance. A second independent variable *aesthetics of design* was manipulated at two levels: highly appealing vs. moderately appealing (see section 2.4.1). Each participant was randomly assigned to one of the six experimental conditions.

Measures and instruments

User behaviour

Two measures of user behaviour were recorded: *Task completion time* (s) referred to the time needed to accomplish the task. *Interaction efficiency* was a composite parameter, dividing the optimal number of user inputs by actual number of user inputs.

Subjective usability evaluation

The German-language questionnaire *Multimetrix*^{S} (Ollermann, 2001) was employed to measure usability ratings of the user. This instrument was largely based on the design principles suggested

by the ISO Standard (ISO-9241-11). The questionnaire was slightly modified by removing items that were irrelevant for the intended application area (e.g., the subscales "media quality" and "suitability of individualisation" were removed since they were considered not to be applicable). This reduced the number of items from 86 to 58. The statements had to be rated on a 5-point Likert scale (agree, partly agree, neither agree nor disagree, partly disagree, disagree). If the item was not applicable, the user was given the choice to tick the appropriate category. The psychometric properties of the Multimetrix are sufficient, with Cronbach's alpha ranging from .63 to .89 for the different scales (Willumeit et al., 1995). The subscales of the instrument were as follows:

- Suitability for the task ("The system forces me to carry out unnecessary actions")
- Conformity with user expectations ("Messages of the software always appear at the same place")
- Information and information structure ("The software contains all relevant information")
- Suitability for learning ("The functions of the software can be easily learnt")
- Self descriptiveness ("I can use the software straight away without the help from others")
- Controllability ("I feel that I have control over the software at any time")
- Error tolerance ("Correcting errors involves little effort")
- User acceptance (example item translated from German: "The software is overloaded with graphical design features")

Emotions and attractiveness

Learning affect monitor (LAM). This is a 32-item questionnaire developed by Reicherts et al. (2005) to capture emotions experienced in daily life. It was slightly adapted to make it suitable for the purpose of the present study. Only a subset of 10 items was employed and analysed, excluding those items that were considered to be less relevant for user-product interaction. The items had a 9-point Likert scale ranging from "not at all" to "very much". The selection of items was based on examining the emotions covered by PrEmo (Desmet, 2003). PrEmo is an instrument that aims to measure emotions relevant to consumer product evaluation by using a cartoon character that depicts each emotion during a short animation. The selected items were identical or very similar to the set of 14 emotions measured by PrEmo (excluding 4 emotions for which no equivalent emotion had been found in the LAM instrument). The remaining 10 items

referred to the following emotions: irritation, boredom, disappointment, delight, enthusiasm, surprise, contentment, disgust, anger, and happiness.

Attractiveness. The attractiveness of the product was measured on a one-item 5-point Likert scale, with the item being phrased: "The design of the mobile phone is appealing" (agree, partly agree, neither agree nor disagree, partly disagree, disagree). The item, translated from German, was self-developed and intended to capture very broadly the user's response to the aesthetic design of the product.

Materials

Mobile phones (high fidelity prototype)

Two mobile phones (SE W800i from Sony Ericsson and M V3690 from Motorola) were selected for the study (see figure 6a). The SE W800i (launched onto the market in the year 2005) was considered to be an aesthetically appealing appliance whereas the M V3690 (launched in 1999) was chosen as a model for a moderately appealing appliance. The selection of the two mobile was based on expert judgement, involving the two authors and two other raters who independently rated a total of 15 telephones for aesthetic appeal. The two appliances with the most extreme ratings at either end were selected for the experiment. The manipulation check in the experiment was successful since it was later confirmed by the participants who rated the two telephones very differently for their attractiveness (see section 3.3)..

(a) (b) (c)

Figure 6: Prototypes of mobile phone: (a) high fidelity, (b) medium fidelity, (c) low fidelity

Touch-screen computer (medium fidelity prototype)

For the medium fidelity condition, computer-based simulations of the dialogue structure of each mobile phone were developed by using Microsoft PowerPoint (see figure 6b). Each prototype allowed the user to interact with the mobile phone and to carry out the same tasks as with the real

product. For the purpose of the study, only the top two levels of the dialogue structure were fully developed for all functions rather than providing an emulation of the complete functionality of each mobile phone. The dialogue structure was only modelled in full depth for the task-relevant menu items. If the user left the optimal dialogue path by more than two levels of the menu structure, an error message was displayed ("Wrong path, please go back"). To obtain the sketchy appearance often found for prototypes employed in usability tests, both icons and text were drawn by hand on a graphic tablet using an electronic pen. The simulation was run on an IBM ThinkPad x41 Tablet PC with a touch screen, which enabled the user to interact directly with the prototype instead of having to use a mouse. This ensured that a similar kind of interface (i.e. high interactivity) is used for the prototype compared to the real product (cf. McCurdy, Connors, Pyrzak, Kanefsky and Vera, 2006).

Paper prototype (low fidelity prototype)

The paper prototype consisted of a collection of cards (sized 90mm x 180mm for Sony Ericsson and 80mm x 210mm for Motorola) upon which all configurations were printed that were modelled by the computer simulation (see figure 6c). These were basically exact replications of the different screen shots. The cards were kept in an indexed cardboard box and presented to the user by the experimenter during the usability test. The user performed the task by pointing the finger to one of the buttons on the paper prototype. Based on the user's selection, the experimenter presented the card reflecting the change in display content initiated by the action.

User tasks

For the usability test, two user tasks were chosen. The first task ("text message") was to send an already prepared text message to another phone user. This represents a task frequently carried out by a typical user. The second task ("phone number suppression") was to suppress one's own phone number when making a call. This task is a low-frequency task compared to the first and was therefore considered to be slightly more difficult.

The tasks differed slightly with regard to the number of commands entered by the user to complete the task successfully. For the "text message" task, this was 8 inputs for the mobile phone from Sony Ericsson and 13 inputs for the Motorola phone. For the "phone number suppression" task, the optimal way of completing the task consisted of 14 inputs (Sony Ericsson)

and 8 inputs (Motorola). The tasks were always presented in the same order, beginning with the "text message" task and followed by the "phone number suppression" task.

Procedure

The study was conducted in a usability laboratory at the University of Fribourg. After welcoming the participant and providing a briefing about the purpose of the experiment, a biographical questionnaire was administered, followed by the LAM questionnaire to obtain a baseline measure of the participant's emotional state. Participants were randomly assigned to one of the experimental conditions (if a participant had already gained some experience with that particular mobile phone, the participant was removed from that experimental condition). The next activity of participants was the completion of the two experimental tasks (see 2.5). During the entire testing procedure, an experimenter was present and took notes. Immediately after the two tasks had been completed, the emotions of the participant were measured again with the LAM questionnaire. This was followed by the presentation of the one-item attractiveness scale and the Multimetrix questionnaire. Finally, the participant was given the opportunity to provide feedback to the experimenter about the prototype and the testing procedure.

Results

User behaviour

Task completion time. This measure was not taken for the paper prototype since it would not have been an adequate reflection of user performance. The measure would have been confounded with the response time of the human "playing" the computer. Overall, the data showed a strong between-participant variation with regard to this performance measure (e.g., task completion times ranged from 46s to 498s). The analysis revealed a main effect of prototype fidelity (see table 8), with users requiring significantly more time when using the computer-based prototype than the fully operational appliance ($F = 9.72$; $df = 1,36$; $p <. 005$). This main effect was modulated by a significant cross-over interaction between fidelity and appliance aesthetics ($F = 6.59$; $df = 1,36$; $p <. 05$). While in the condition "computer prototype / highly aesthetic design", the completion time was longest, the same model showed the fastest task completion times when the fully operational appliance was used ($F = 6.59$; $df = 1,36$; $p <. 05$; LSD-tests: $p <. 05$). No significant main effect of appliance aesthetics was found ($F = 2.31$; $df = 1,36$; $p >. 05$).

Interaction efficiency. The results of the efficiency of user-product interaction (i.e. optimal number of user commands divided by actual number of user commands) are presented in table 8. Due to a failure of the data logging facility of the computer prototype, the number of user-system interactions was not accurately counted so that no data were available for this experimental condition. For the remaining conditions, no differences between cells were found. This was confirmed by a two-factorial ANOVA, which showed no effect of prototype fidelity ($F<1$), none of design aesthetics ($F = 2.51$; $df = 1,36$; $p > .05$), and no interaction between the two factors ($F<1$). In the medium fidelity condition, the experimenter made the interesting observation that many users clicked several times directly on the display of the mobile phone presented on the computer touch-screen rather than the buttons until users realised that only the computer had a touch-screen but not the simulated mobile phone. This type of error related to prototype interactivity was not observed under the paper prototype condition.

Table 8: Measures of user behaviour as a function of prototype fidelity and aesthetic design of appliance (N/A: not available)

	Paper prototype (low fidelity)	Computer-based prototype (medium fidelity)	Fully operational appliance (high fidelity)	Overall
Task completion time (s)	N/A	268.8	157.2	
Highly aesthetic design	N/A	342.1	138.4	240.2
Moderately aesthetic design	N/A	195.6	175.9	185.7
Interaction efficiency index	.58	N/A	.58	
Highly aesthetic design	.61	N/A	.63	.62
Moderately aesthetic design	.55	N/A	.53	.54

Subjective usability evaluation

A multivariate analysis of variance (MANOVA) was carried out to test for overall effects of the independent variables on 8 rating scales of the Multimetrix. The MANOVA showed an overall effect for appliance aesthetics ($F = 12.4$; $df = 7, 48$; $p <. 001$) but not for prototype fidelity ($F < 1$) and no interaction was found ($F = 1.05$; $df = 14, 96$; $p >. 05$). Separate analyses on each scale revealed that the highly appealing design (i.e. SE W800i) was given higher usability ratings than

the moderately appealing design on all 8 scales (all scales were strongly correlated with each other, suggesting that users did not distinguish much between them) as well as on the overall scale. All effects were highly significant, as the data in table 9 demonstrate. The absence of an effect of prototype fidelity suggests that fidelity does not influence the perceived usability of a product.

Table 9: User ratings on overall scale and each subscale (1-5) of Multimetrix ^{S} (***p<.001)

	Highly aesthetic design	Moderately aesthetic design	Results of analysis of variance
Overall scale	3.81	2.88	F (1,54) = 77.3***
Suitability for the task	3.75	2.77	F (1,54) = 46.3***
Conformity with user expectations	4.00	3.41	F (1,54) = 21.7***
Information and information structure	3.73	2.37	F (1,54) = 73.8***
Suitability for learning	3.92	2.87	F (1,54) = 27.6***
Self descriptiveness	3.71	2.37	F (1,54) = 80.7***
Controllability	4.14	3.15	F (1,54) = 33.2***
Error tolerance	3.23	2.74	F (1,54) = 14.7***
Acceptance	4.04	3.36	F (1,54) = 22.9***

Emotions and attractiveness

Emotions. A MANOVA was carried out on the 10 LAM items. The analysis revealed no effect for fidelity level ($F < 1$) but an effect for design aesthetics was observed ($F = 2.86$; $df = 10, 45$; $p <. 01$). No interaction between the two factors was recorded ($F < 1$). In table 10 the means of participant ratings at t_0 (i.e. prior to usability test) and t_1 (i.e. after usability test) are presented as a function of design aesthetics. Separate univariate analysis of variance on single items revealed significant effects for 5 items. The strongest effect was found for ‘delighted’, followed by ‘disappointed’, ‘happy, ‘irritated’, and ‘angry’. No significant differences were found for the five other emotions. The data in table 10 also indicated that the emotion “surprised” showed a very strong increase from t_0 to t_1 for both appliances ($F = 16.1$; $df = 1, 59$; $p <. 001$).

Table 10: Mean ratings of emotions at t_0 (prior to usability test) and t_1 (after usability test) as a function of appliance usability on a 9-point Likert scale; significant differences as a function of design aesthetics are indicated by stars (*p<.05, ***p<.001)

	Highly aesthetic design			Moderately aesthetic design		
	t_0 (SD)	t_1 (SD)	Difference ($t_1 - t_0$)	t_0 (SD)	t_1 (SD)	Difference ($t_1 - t_0$)
Irritated *	2.80 (1.9)	2.33 (1.4)	-.47	2.53 (1.5)	3.27 (2.0)	+.73
Bored	2.37 (1.7)	2.3 (1.6)	-.07	3.0 (1.8)	2.56 (1.5)	.-44
Disappointed *	2.33 (2.0)	1.9 (1.3)	-.43	1.70 (1.1)	2.3 (1.7)	+.83
Delighted ***	5.80 (1.5)	6.43 (1.6)	+.63	5.73 (1.6)	5.1 (1.7)	-.63
Enthusiastic	5.2 (2.0)	5.0 (2.2)	-.20	4.57 (1.8)	4.27 (1.8)	-.30
Surprised	2.53 (1.5)	3.47 (2.1)	+.94	2.27 (1.5)	3.46 (2.1)	+1.19
Contented	6.3 (1.9)	5.93 (2.1)	-.37	6.07 (1.6)	5.4 (1.8)	-.67
Disgusted	1.57 (1.5)	1.43 (1.2)	-.14	1.27 (0.8)	1.3 (0.7)	+.03
Angry *	2.0 (1.7)	1.63 (1.4)	-.36	1.86 (1.2)	2.1 (1.6)	+.23
Happy *	5.63 (2.1)	6.13 (2.1)	+.50	5.76 (1.7)	5.23 (1.7)	-.53

Attractiveness. The ratings of the attractiveness scale are presented in table 11. As expected, the analysis revealed a strong main effect of aesthetics, with the highly aesthetic appliance being given higher ratings ($F = 25.3$; $df = 1,53$; $p < .001$). This demonstrated that the experimental manipulation had been successful. More interesting was the strong interaction between prototype fidelity and design aesthetics ($F = 4.6$; $df = 2,53$; $p < .05$), with the moderately aesthetic design of the fully operational appliance having a significantly lower rating than all the other conditions (LSD-test: $p <. 005$). No significant difference was found between the two paper prototypes and the two computer prototypes (both LSD-tests: $p >. 05$). Finally, a main effect of prototype fidelity was found ($F = 3.4$; $df = 2,53$; $p < .05$), which was only due to the low score of the real appliance with the moderately aesthetic design.

Table 11: User ratings of attractiveness of product (1-5) as a function of prototype fidelity and design aesthetics

	Paper prototype (low fidelity)	Computer-based prototype (medium fidelity)	Fully operational appliance (high fidelity)	Overall
Overall	3.3	3.6	2.9	
Highly aesthetic design	3.7	4.0	4.0	3.9
Moderately aesthetic design	3.0	3.3	1.8	2.7

Discussion

The central question of this article concerned the utility of prototypes that are of lesser fidelity than the reference system during usability tests. The main results showed that task completion time may be overestimated when a computer-based simulation is used. Furthermore, the effects of fidelity levels on attractiveness ratings appeared to be stronger for less appealing products than for attractive ones. It also emerged that objective performance parameters collected during the usability test and subjective usability ratings were not associated. Finally, the results showed no evidence for fidelity level affecting emotions or subjective user evaluation.

The results showed that task completion time was higher for the computer-based simulation than when a fully operational product was being used. This effect was observed for both mobile phones, though they differed with regard to the strength with which this effect occurred. The increased task completion time under the computer condition was partly caused by prototype-specific errors being made by users that resulted from differences in the interactivity of prototype (cf. model of Virzi et al., 1996). With the computer prototype, ineffective clicks were made by participants because they erroneously extended the interactivity of the device from the computer screen (direct manipulation was possible) to the display of the simulated mobile phone (direct manipulation was not possible). However, it was only the representation of the mobile phone's controls on the touch screen that were interactive. Although the touch screen permits a more natural interaction of the user with the mobile phone than a conventional screen (for which the user needs to use a mouse), this advantage may be accompanied, as observed in the present case, by unanticipated side-effects in the form of negative transfer,

For the attractiveness rating of the appliances, an interesting interaction between prototype fidelity and design aesthetics was observed. While there was no difference in ratings across different fidelity levels for the highly aesthetic mobile phone, the moderately aesthetic phone was rated lower on attractiveness for the original appliance than for the reduced fidelity prototypes. The fact that the two reduced fidelity prototypes had similar ratings like the original appliance for the highly aesthetic design is in itself a somewhat surprising result. This suggests that some compensatory activity on the part of the user took place since neither the paper prototype nor the computer-based prototype was aesthetically refined (e.g., lacking colour and shape of the reference appliance). Users may have mentally anticipated of what the real appliance might look like and employed this mental picture as a basis for their rating. For the moderately

aesthetic phone design, users may have engaged in a similar process in that they extrapolated the appearance of the computer and paper prototypes to the real appliance (indeed, there were no significant differences between the two computer-based prototypes and the two paper prototypes across phone types). Since the computer-based and paper prototypes were judged to be more attractive than the real appliance, it can be speculated that under the reduced fidelity conditions users created a mental model of the real appliance representing a much more attractive design than the real appliance actually enjoyed. This may suggest a kind of "deficiency compensation"-effect. As this interaction between prototype fidelity and design aesthetics was not predicted, it needs to be treated with some caution but, if confirmed in subsequent studies, it would have implications for the use of reduced fidelity prototypes for the purpose of attractiveness judgements.

The results showed no association between objective performance parameters and subjective usability evaluation. While there was a clear preference of users for the more aesthetic appliance because of higher attractiveness ratings and higher perceived usability, this was not paralleled by better objective usability of that appliance. This suggests that perceived usability may be more strongly associated with attractiveness ratings than objectively measured usability parameters. This result is in support of the findings of Tractinsky (1997), who proposed that the beauty of design would positively affect perceived usability. While in Tractinsky's study no user-product interaction took place (with the usability rating of users being based on the mere look of the product), the present study provided similar evidence even for the case when user-product interaction occurred. If this finding was to be found consistently, it would imply that the beauty of a product was such an important aspect that it would also need to be considered by designers and engineering psychologist when designing for usability.

The changes in emotions during the usability test (i.e. from t_0 to t_1) were quite substantial, suggesting that user-product interaction constitutes a significant emotional experience. The intensity of the emotional experience may have been increased by two factors. First, the usability testing procedure that included the presence of an experimenter may have intensified the emotions recorded because of the increased arousal induced by the presence of others, as suggested by social facilitation theory (Cottrell et al., 1968). Second, it may be that at t_0 emotions were measured but at t_1 measurements of sentiments were taken. Sentiments refer to the user's feelings towards the appliance rather than reporting their internal state (Brave and Nass, 2003). These may have been evoked during product utilisation, resulting in a considerable change in

user ratings. At t_0 users reported their general internal emotional state while at t_1 their self-reported state was closely linked to the directly preceding experience with product utilisation. This may explain the considerable changes across measurement points. Similar to the findings for attractiveness ratings, there was no evidence for a different emotional reaction being triggered off by reduced fidelity prototypes compared to the real appliance. The same was observed for subjective usability evaluation (i.e. even prototypes of lower fidelity seemed to be useful to assess subjective usability). Users may have achieved this by creating a mental model of the real appliance (under paper and computer prototype conditions) upon which their judgements are based.

The use of reduced fidelity prototypes raises the broader issue of validity of usability testing. Concerns have been expressed about the validity of usability tests, given the remarkable inconsistencies in test outcomes that were observed across tests (e.g., Lewis, 2006). While it is generally agreed that usability testing improves the usability of products (as opposed to not conducting any usability test), the validity of the test could increased if we had a better understanding of the factors that influence validity. Of the many forms of validity, ecological and predictive validity may be of particular interest. In order to improve the ecological validity of a usability test (i.e. the extent to which behaviour in a test situation can be generalised to a natural setting), the influence of the wider testing environment needs to be considered (e.g., Brehmer and Dörner, 1993). This refers in particular to the physical and social aspects of test environment (e.g., lab set-up, presence of observers). For this purpose, a model (called the Four-Factor Framework of Contextual Fidelity) has been proposed, which explicitly refers to these factors (Sauer, Seibel and Ruettinger, under review). Predictive validity coefficients of paper and computer prototypes may also be determined in future studies, using a similar approach as in personnel selection where the validity of different selection methods has been determined. Test participants would first complete a set of tasks with a reduced fidelity prototype and subsequently (after a time interval) with a real product. Lastly, we would like to point out a methodological weakness of this study. This refers to the exhibition of the mobile phone's brand name in the high fidelity condition. The brand name was left uncovered to produce a more natural testing situation but it cannot be excluded that this may have influenced emotion and attractiveness ratings.

Finally, there is a need to carry out more research into the effects of prototype fidelity and design aesthetics to examine whether the findings of the present study can be replicated with modified design characteristics and also with different interactive consumer products. For

example, it would be important to see whether the interaction found for attractiveness ratings can be replicated if the reduced fidelity prototypes had been aesthetically more refined instead of presenting a rough sketch. The question of which prototype should be used would not only be relevant in the context of usability testing but also when designers present prototypes of the work that was commissioned by their clients. In this situation, the issue of aesthetics is also of great importance since they may influence the client's decisions. Overall, the findings suggest that prototypes of reduced fidelity may be suitable for modelling the reference system. From the findings of the present work, it appears that in order to design a highly usable product, an appealing design would be one of the *necessary* product features. This would suggest that the issue of aesthetics should be closer to the heart of the ergonomic design process than perhaps previously thought.

8 The Influence of Cultural Background and Product Value in Usability Testing

Abstract

This article examines the influence of cultural background and product value on different outcomes of usability tests. A quasi-experimental study was conducted in two different countries, Switzerland and East Germany, which differed with regard to their wellbeing-orientation. Product value (high vs. low) was varied by manipulating the price of the product. Sixty four test participants were asked to carry out five typical user tasks, measuring performance, perceived usability, and emotion. The results showed that in a wellbeing-oriented culture, high-value products were rated higher in usability than low-value products whereas in a less wellbeing-oriented culture, high- value products were evaluated lower in usability than low-value products. A similar interaction effect of culture and product value was observed for user emotion. Implications are that the outcomes of usability testing do not allow for a simple transfer across cultures and that the mediating influence of perceived product value needs to be taken into consideration.

Keywords: usability testing; culture; product value; coffee machine

Because of copyright restrictions, this chapter cannot be published in this thesis. Please contact me by email (andreas.sonderegger{at}unifr.ch) to obtain a copy of the manuscript.

9 General discussion

9.1 Overview of findings

The primary purpose of this thesis was to evaluate the influence of typical factors of reduced contextual fidelity in usability tests on their outcomes. To this end, aspects of the product prototype (type of prototype, prototype aesthetics, and product value), of the testing situation (observer presence) and of the test participants (cultural background) were experimentally manipulated in four separate studies. The findings of the studies are summarized in table 12. The different performance measures (task completion rate, error rate, interaction efficiency and task completion time) were pooled together since they were always influenced in the same direction (although very often the influence did not show to be significant on every measure of user performance).

9.2 Integration of findings

The summary of the findings (see table 12) indicates that contextual factors in usability tests affect a wide range of usability measures. Interestingly, user performance was affected in all studies. Measures of user performance are important for the usability of future products because design adaptations and decisions about specific design alternatives are generally based on the evaluation of user performance (Rubin & Chisnell, 2008). Perceived usability is a further measure of capital importance in usability evaluation. A review of current practice in measuring usability revealed that a vast majority of studies published in ergonomic literature included measures of perceived usability (Hornbæk, 2006). The fields of application for such measures are very wide. They are instrumental in the gathering of information about users' expectations, users' appreciation of the product and users' preferences, etc. and hence influence decisions in the whole product development process (Kuniavsky, 2003). Since the presented findings point out that measures of perceived usability may be considerably biased by contextual factors, this may affect the usability of the future product. Additionally, less classical measures recorded in usability tests such as user emotions and physiological strain were influenced by contextual factors. Most contextual factors impinged on user emotions, which is a measure of increasing importance in product design (cf. section 2.4.3). Again, design decision (and hence product usability) based on information about

Table 12: Summary of effects of the four studies (↓ indicates a decrease on that measure and ↑ an increase)

		User performance	**Perceived usability**	**Perceived Attractiveness**	**User emotions**	**Physiological strain**
Lab-setup study	Increasing observer presence	↓	n.s.	n.s.	↓	↑
	Increasing task difficulty	↓	n/m	n/m	n/m	n/m
Design-aesthetics study	Decreasing prototype aesthetics	↓	↓	↓	n/m	n/m
Prototype-fidelity study	Reduced prototype fidelity	↓	n.s.	↓[a]	n.s.	n/m
	Decreasing prototype aesthetics	n.s.	↓	↓[a]	↓	n/m
Culture study	Increasing well-being orientation	↑	↑↓[b]	n/m	↑↓[c]	n/m
	Increasing product value	n.s.	↑↓[b]	n/m	↑↓[c]	n/m

Notes: n.s. = non-significant; n/m = not measured

a) This effect was observed for the real product but not for the reduced fidelity prototypes (i.e. interaction type of prototype x product aesthetics)

b) perceived usability of high-value products was rated high in a culture with high well-being orientation and low in a culture with low well-being orientation (i.e. interaction well-being orientation x product value)

c) Usage of high-value products increases user emotion in a culture with high well-being orientation and decreases emotions in a culture with low well-being orientation (i.e. interaction well-being orientation x product value)

user emotions may be affected by contextual factors. Their effect on user emotions may however even have a further impact on product usability: emotions are known to affect human behaviour, memory, attitudes, decision making, etc (e.g., see Berkowitz, 2000). These are all aspects that play an important role in usability testing. Affected emotions might hence influence other measures recorded in usability tests such as user performance and perceived usability (see section 9.4 below for a discussion on relations between different outcome measures in usability tests). Measurement of physiological data, notably as an indicator of a user's workload while using the product, is rather a novel approach in usability testing (Lin, et al., 2005). Results presented in this thesis indicate that contextual factors such as laboratory setup affect such measures and hence may impinge on product usability. This indicates the importance of the influence of contextual factors for usability testing and product usability. For a thorough understanding of the influence of the contextual factors examined in this thesis, the results of the presented studies are elaborated in the following sections.

With regard to aspects of the system prototype, the results indicate that the effects on the different outcomes of usability tests are not entirely congruent. However, a certain pattern can be discerned, especially with regard to the factors of prototype aesthetics and type of prototype. For both factors, a reduction in the fidelity level led to a decrease in different usability measures (see below for an analysis of their influence on specific outcomes of usability tests). Product value as a further aspect of prototype fidelity is somewhat a special case since in the culture study, the scope of value was not confined to high and low levels of fidelity but rather to two different levels of low fidelity - one above and one below market level. This needs to be taken into consideration for the analysis of the findings. However, as for prototype aesthetics and type of prototype also for product value it can be concluded that the level of fidelity influences outcomes of usability tests. In this case though, the influence depends on the cultural background of the test participant (e.g. overestimation of product usability for high value products in a culture with high well-being orientation and low value products in a culture with low well-being orientation).

With regard to the influence of aspects of the system prototype on measures of user performance**,** results indicate that user performance was influenced by prototype aesthetics and type of prototype. The influence of type of prototype on user performance was caused by differences in the interactivity of the prototype compared to the real physical product. Participants working with the computer prototype made more ineffective clicks than participants operating of the real product because they extended the interactivity of the touch screen from the

computer simulating the mobile phone to the display of the mobile phone. While this effect was explained by prototype-related errors being made by users, the nature of the influence of prototype aesthetics on performance measures is less apparent. The results in relation to this effect differ in the two studies in which design aesthetics was manipulated. Interestingly, design aesthetics only showed an influence on performance measures in the design-aesthetics study but not in the prototype-fidelity study. This distinction might be due to differing influences of the situational context in the two studies. In the design-aesthetics study, test participants might have felt as being in a work-oriented context, which is assumed to lead to an increase in performance due to an increased-motivation effect. The situational context of the prototype-fidelity study might on the other hand have been somewhat ambiguous. A comparison of the two studies' procedures and test settings may serve as indicator for this assumption. In the design-aesthetics study, the usability test was conducted with pupils during their lessons in the school's computer lab, which represented for the pupils a work context rather than a leisure context (which would argue for the increased-motivation effect). On the other hand, the prototype-fidelity study was conducted in a laboratory of the university with university students. The students participated in the usability test in their leisure time, however the laboratory at the university might also represent a work context. This ambiguity in the situational context might be a reason for the missing influence of prototype aesthetics on performance measures in the prototype-fidelity study. This interpretation is however somewhat speculative and needs to be confirmed empirically in future research on the influence of contextual factors in usability tests.

The analysis of the effects of the system prototype on measures of perceived usability reveals that type of prototype is the only aspect in this thesis that showed no influence on perceived usability. These measures seem hence not to be affected by the use of paper or computer prototypes. In contrast to this, measures of perceived usability were influenced by aspects of prototype aesthetics as well as product value. The reported effects of prototype aesthetics on perceived usability corroborate the findings of previous work (e.g. Tractinsky, Shoval-Katz & Ikar, 2000; Hassenzahl, 2004) and are further indicators for the existence of a halo effect of product attractiveness on perceived usability ratings. A similar effect on perceived usability was found for product value. However, this effect differs with regard to its influencing direction, depending on the cultural background of the test participants. Participants from Switzerland rated the perceived usability of the high priced product more positively than the low priced one. For participants from East Germany, the converse pattern was observed. As for the

influence of aesthetics on usability measures, this result can be explained by the halo effect. This indicates that different aspects of the product prototype may influence the apperception of other characteristics of the product. In addition to the well known stereotype ‘‘what is beautiful is good’’, known from social psychology (Dion, Bersheid & Walster, 1972), the results of the culture study indicate that there might be also a stereotype “it is good when the price is right” in usability testing. This might be considered as argument for an increased integration of marketing aspects into product development (as proposed e.g. by McClelland & Brigham, 1990; Benini, Batista & Zuffo, 2005), since the presented results indicate that a product evaluation may also depend on the attributed product value or product price.

The pattern of the influential aspects of the prototype on user emotions in usability tests is very similar to the effects observed on perceived usability. As for perceived usability, measures of user emotions were more positive when a highly aesthetic prototype was used. Also with regard to the influence of product value, measures of user emotions showed the same pattern of influence as measures of perceived usability: a high product value was associated with more positive emotions for Swiss users and more negative emotions for users from East Germany. In addition, type of prototype had neither an influence on measures of perceived usability nor on user emotions. This evident pattern indicates that measures of user emotions and perceived usability are somehow linked to each other. However, until today only little is known about the link between those two usability outcomes (for a further discussion of the link between these measures see section 9.4).

Overall, the presented results indicate that the design of prototypes may be an important source of error in usability testing. This should be considered as a complement to general assumptions made in the literature on usability testing and prototyping. For example, Snyder (2003) states that “the evidence suggests that paper prototyping is as valid a technique for uncovering problems in an interface as usability testing the real thing” (p. 289). Similar assumptions with regard to low fidelity prototypes are also made by Virzi, Sokolov and Karis (1996) and Catani and Biers (1998). Since usability problems have not been measured explicitly in the studies presented in this thesis, it is not possible to evaluate these assumptions. However, with regard to an approach that considers a broader range of measures in usability testing (and especially with regard to summative usability evaluation) the results indicate that the use of reduced fidelity prototypes may play a vital role for the validity of usability tests as a method of evaluation.

Findings of the lab-setup study indicate that in addition to characteristics of the product prototype aspects of the testing environment may also influence usability test outcomes. This empirically confirms assumptions that have been made by different authors about the potential source of stress for test participants represented by the testing environment (e.g. Schrier, 1992; Salzman & Rivers, 1994; Patel & Loring, 2001). However, the understanding of the influence of aspects of the testing environment in usability tests is still very limited. Findings obtained from other fields of psychological research indicate that the presence of a camera in psychological experiments affects arousal level and performance patterns of participants (Cohen, 1979; Kelsey et al., 2000). Being monitored by (imaginary) observers behind a one-way mirror can diminish subjects' available processing resources and lead to decreased performance (Seta, Seta, Donaldson & Wang, 1988). Electronic performance monitoring in the workplace also proved to be linked with increased levels of stress and decreased levels of productivity and work quality (Aiello & Kolb, 1995). Such findings may serve as indicators for the potential influence of further aspects of the testing environment on outcomes of usability tests. However, in usability practice it is not always possible to change the testing environment (e.g. the presence of a camera is needed for a retrospective evaluation of the test run). Therefore it is important to acquire specific information about the influence of different aspects of the testing environment on usability test outcomes. Specific knowledge about the effect of such environmental factors would help to better understand and interpret the results of laboratory based usability tests. Since the context of usage represents an integral part of the ISO definition of usability (usability is defined as 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; cf. chapter 2.1), an explicit knowledge about the influence of the testing set-up is critical to be able to make assumptions about the validity of the evaluation method.

As for the testing environment, *user characteristics* are also part of the ISO definition of usability, which implies that the term can be defined only with regard to a specific user group. The recruitment of participants representing the future end users is therefore essential for the validity of the usability evaluation. Even though this issue has been discussed already in usability literature (e.g. by Rubin, 1994; Snyder, 2003; Nielsen, 1993), empirical evidence of the influence of user characteristics on usability test outcomes are rather limited. The findings of this study clearly indicate that even by comparing two relatively similar user groups, it was possible to find considerable culture-based differences in the influence of product value on usability evaluation.

This suggests that culture-specific usability testing is of paramount importance in usability practice (cf. DeAngeli & Kyriakoullis, 2006; Simon, 2003).

9.3 Effect sizes of influencing factors

The studies presented in this thesis have shown that reduced fidelity of contextual factors may influence the outcomes of usability tests. With regard to this influence, it is of great interest for usability research and usability practice to be aware of how strong these effects really are. Effect sizes however are hardly ever reported in usability literature. As a remedial measure, an attempt has been made in this thesis. The effect sizes of the different influences of contextual factors on outcome measures of usability tests are presented in table 13. Effect sizes (Cohen's *d*) were calculated according to Faul, Erdfelder, Lang & Buchner (2007) on the basis of planned comparisons of the different group means. In table 13, only the largest effect sizes are reported. For example, in the culture study, the effect sizes varied considerably with regard to the influence of well-being orientation on different measures of user performance. The effect size of the influence of well-being orientation on task completion rate (d = .49) was lower than on of task completion time (d = .80) and number of user interactions (d = .80). In table 13, only the largest effect size (d = .80) is reported. With regard to measures of perceived usability containing different sub-factors (e.g. PSSUQ with the sub-factors system usefulness, information quality and interface quality), only the effect size of the overall measure was considered, even if the effect size on a sub-factor was larger (e.g., the effect size of prototype aesthetics on information quality was d = 2.37, nonetheless the effect size of d = 1.76 for the overall measure is reported). According to Cohen (1992), small, medium, and large effect sizes are d = . 20, d = . 50, and d = . 80. Overall, the effect sizes presented in table 13 are considerable. With regard to all outcome measures, almost all effects are large, indicating that the differences between the experimental groups are substantial. This can be considered as an indicator for the importance of the influence of contextual factors on usability test outcomes.

Although all the effect sizes reported in this thesis are considerable, it might be interesting to compare the influence of contextual factors with the influence usability issues may have on results of usability tests. The comparison of effects of contextual factors with effects of usability issues would help to appraise the impact contextual factors have on usability test outcomes. Unfortunately, only very few studies in usability literature varied specific contextual factors in

Table 13: Comparison of effect sizes (Cohen's *d*) in the four studies and a comparison study (Thüring & Mahlke, 2007)

		User performance	**Perceived usability**	**Perceived Attractiveness**	**User emotions**	**Physiological strain**
Lab-setup study	Increasing observer presence	.80	n.s.	n.s.	.76	.85
	Increasing task difficulty	2.66	n/m	n/m	n/m	n/m
Design-aesthetics study	Decreasing prototype aesthetics	.74	1.29	1.51	n/m	n/m
Prototype-fidelity study	Reduced prototype fidelity	.91	n.s.	1.82	n.s.	n/m
	Decreasing prototype aesthetics	n.s.	1.76	1.31	.92	n/m
Culture study	Increasing well-being orientation	.80		n/m		n/m
	Increasing product value	n.s.	.72	n/m	.95	n/m

Notes: n.s. = non-significant; n/m = not measured;

combination with different levels of inherent usability. One example for such a study was published by Thüring and Mahlke (2007). Regrettably, the authors did not report effect sizes. Therefore, I calculated in a subsequent analysis effect sizes of their published data. The analysis reveals large effects of inherent usability on measures of user behavior (d = 1.01), perceived usability (d = 1.73) and user emotions (d = 1.33) whereas the effects of aesthetics were only small or even negligible (d = .08 for user performance, d = .26 for perceived usability, d = .37 for perceived attractiveness, and d = .41 for user emotions). The small effect sizes of prototype aesthetics could be considered as consequence of the stimulus material used in the study of Thüring and Mahlke (2007). Prototype aesthetics was varied by manipulating different design dimensions (e.g. symmetry, color combination, and shape) of a computer prototype of a digital audio player. This manipulation is very similar to the manipulation of visual aesthetics in the design-aesthetics study, where also aspects of symmetry, color and shape have been manipulated (cf. chapter 6). However, the difference between the two designs used by Thüring and Mahlke (2007) might be considered as being less substantial compared to the designs used in the design-aesthetics study (cf. figure 7).

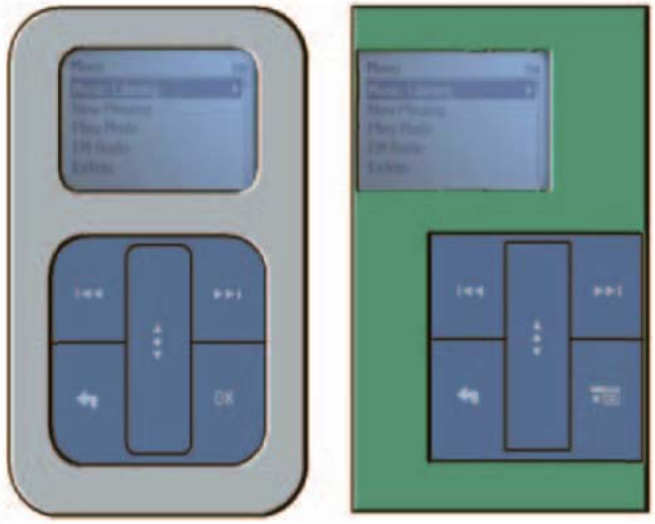


Figure 7: Variations of visual aesthetics of prototypes used in Thüring & Mahlke (2007; p. 260)

9.4 Relations between the different outcome measures in usability tests

The summary of the reported results indicates that the influence of contextual factors in usability tests on objective measures of user behavior is not equivalent to subjective measures of perceived usability (cf. table 12). With the exception of the design-aesthetic study, contextual factors were either influencing measures of user performance or on measures of perceived usability but not both together. In the lab-setup study, the presence of observers had an influence on task completion time (representing a measure of user performance) whereas perceived usability ratings were unaffected. In the prototype-fidelity study, the type of prototype affected task completion time whereas the perceived usability ratings were not affected by this factor. In the same study, design aesthetics influenced the perceived usability ratings but did not influence objective efficiency measures. A similar pattern of results can be found in the culture study: culture had an effect on user performance, but not on perceived usability ratings. On the other hand, the interaction of product price and culture had an effect on perceived usability ratings but not on measures of user performance. The link between (objective) measures of performance and perceived usability ratings has been already addressed in usability research before. A meta-analysis accents a positive association between objective measures of user behavior and perceived usability ratings (Nielsen & Levy, 1994). The correlation of 0.46 indicates that measures of user behavior explain for 21% of the variance of users' subjective evaluations. Several other factors may account for the remaining variance in the subjective evaluation scores. Nielsen and Levy (1994) mention graphic design quality as a further possible influencing factor. The results of the two studies on design aesthetics presented in this thesis support this assumption: the perceived usability rating was in both studies influenced by the design aesthetics of the product. More recent meta-analyses (Fokjaer, Hertztum & Hornbaek, 2000; Hornbaek & Law, 2007) however indicate that the average correlations among subjective and objective usability ratings are very moderate (e.g. $r = .20$ for effectiveness and user satisfaction). Based on these findings, Hornbaek & Law (2007) propose that effectiveness, efficiency and satisfaction should be considered as independent aspects of usability. Fokjaer et al. (2000) assume that the correlations among the different usability measures depend on the application domain, the user's experience, the task complexity and the use context in a complex way. Such factors were already discussed with regard to the inconsistent effects of prototype aesthetics on performance measures in the design-aesthetic study and the prototype-fidelity study (cf. section 9.2) and might explain

why only in the design-aesthetics study, measures of performance and perceived usability were jointly influenced. It is still not clear however, based on which processes these contextual factors influence the different usability measures. Furthermore, the association between subjective and objective usability measures is not yet understood comprehensively. Therefore, further research is needed scrutinizing the relations between these two types of measures as well as the underlying processes.

An interesting relationship appears to exist as well between the measures of user emotions and perceived usability. Both measures were affected by experimental conditions in a very similar manner: when participants reported positive emotions, they also rated perceived product usability higher. A similar positive association of user emotions with perceived usability has already been revealed in previous work (e.g., Tractinsky et al., 2000), however without discussion of a possible rationale for such a correlation. A possible explanation of this link might be found in social psychological research. Measures of perceived usability may be considered as summary evaluations of a specific object. Summary evaluations of people or objects are described in social psychology as attitudes (Petty, Wegener & Fabrigar, 1997; Eagly & Chaiken, 1993) and can be separated into an evaluative component, an affective component, and a belief component (Olson & Zanna, 1993; Petty et al., 1997; for an overview of the empirical justification of this tripartite conceptualization of attitude see Weiss, 2002). Social psychological research has shown that attitudes are influenced by affective, cognitive and behavioral antecedents (Olson & Zanna, 1993; Weiss, 2002). Based on such findings, it can be assumed that user emotions as affective consequence of product usage are influencing the overall evaluation of product usage (notably perceived usability). However, the precise nature of that influence, whether moderator or mediator, is still unclear. In contrast to prototype fidelity, observer presence did not generate a similar pattern of effects on user emotions and measures of perceived usability. Whereas user emotions were more negative when observers were present in the usability laboratory, this presence had no influence on measures of perceived usability. This indicates that the connection between user emotions and perceived usability might not be comprehensive but depend on the source of emotional change. If the source is some product-inherent factor such as prototype aesthetics or product value, then it follows that emotions of test users are influencing their usability rating. Whereas measures of perceived usability are not influenced by user emotions when the emotional change is caused by aspects of the testing

environment. This assumption however is highly speculative and needs to be addressed in future research.

9.5 Implications for the 4FFCF

The different studies presented in this thesis have shown that certain contextual factors such as the presence of observers or the design aesthetics of the prototype may cause an issue for the validity of usability test results. These results indicate the necessity of the model for research on such issues of validity in usability testing. Based on the 4FFCF, further studies should be planned this time, aiming at analyzing the impact of other influencing factors on the outcomes of usability tests. For example, with regard to the influence of aspects of the testing environment in usability tests, the presence of observers is only one of many possible influencing factors. Nielsen (1993) for example points out that the presence of a one-way mirror in the testing laboratory might be stressful for the participants. Schrier (1992) mentions the lab atmosphere, the presence of video cameras or the mere fact of using a new and unknown product as factors that might induce stress in test participants. Yet, an empirical evidence of such assumptions is still missing. By analyzing influences of further contextual factors, more information about the quality of the usability data can be collected. However, such an approach has some drawbacks. The 4FFCF neither provides an explicit indication of the direction of expected effects nor propositions about interaction effects of different influencing factors. Furthermore, it does not provide any theoretical explanation for the anticipated influences. It would therefore be helpful to have a model at hand that describes and illustrates the associations of different factors of the 4FFCF with measures typically collected in usability tests, additionally explaining the underlying processes. Figure 8 illustrates an attempt to give consideration to the first two critiques mentioned above. As for the 4FFCF, it defines four contextual factors influencing the main outcome variables in usability tests. Each of the factors can have either a direct effect on the different outcome variables or a moderating or mediating influence on the influence of another factor. Furthermore, assumed connections between different usability measures (cf. section 9.4) are indicated. The influences identified in the present thesis are indicated in figure 8 by a minus or a plus sign, depending on the direction of the factor's influence on the outcome measure (moderation or mediation effects are indicated by an asterisk).

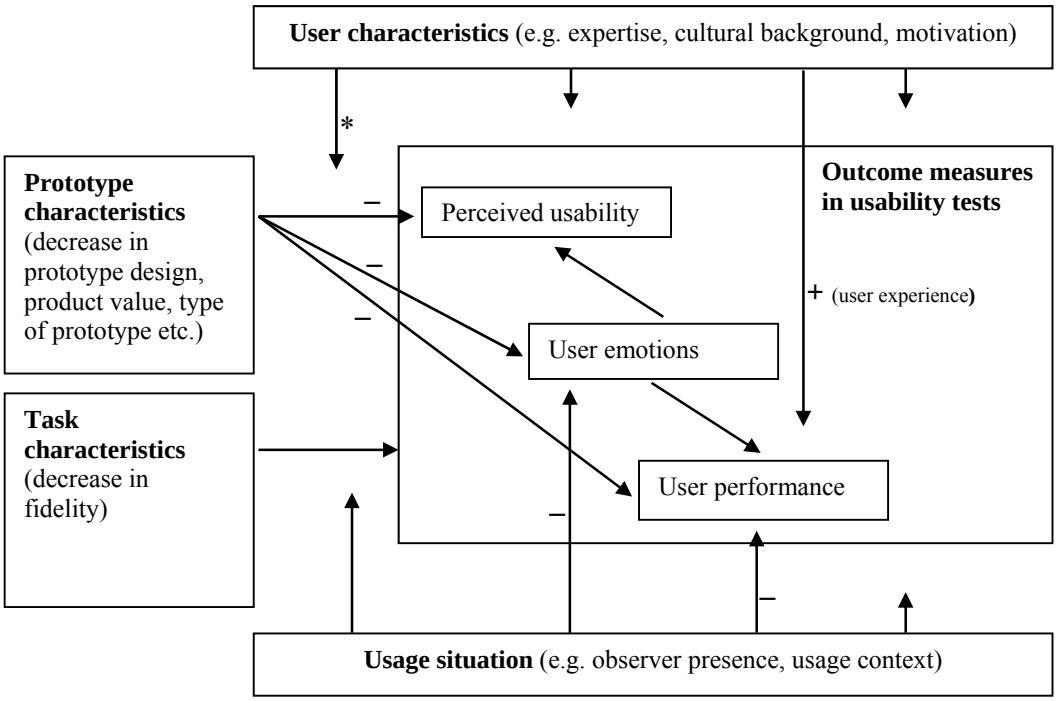


Figure 8: Extended Four Factor Framework of Contextual Fidelity (– indicating a negative influence, + indicating a positive influence, * indicating a moderating influence)

9.6 Implications for further research

Relations between different outcome measures. The analysis of the relation between different outcome measures of usability tests revealed some interesting patterns and raised various new questions. Why for example is the correlation between performance measures and perceived usability so low? Although this missing link has been addressed several times in ergonomic literature already (e.g. by Nielsen & Levy, 1994 and Fokjaer et al., 2000), no satisfactory explanation for such empirical findings has been provided so far. Is there just a weak link because factors other than efficiency and effectiveness of task completion are influencing the subjective evaluation of a product's usability (such as e.g. design aesthetics)? Or might the reason behind the low correlation be found in problems operationalizing the two concepts? Interestingly, the scope of instruments used in usability tests is usually only vaguely defined by their authors. Very often, usability questionnaires are meant to measure product usability in some subjective way (compared to "objective" performance measures). For most of these questionnaires however,

subjective usability is defined referring to the notion of user satisfaction (e.g. QUIS, PSSUQ; cf. section 2.4.2). User satisfaction and perceived usability however might be considered as different concepts. This indicates that it may often not be clear what usability questionnaires explicitly tend to measure. Regrettably, research on validity of instruments used in usability testing is very limited. A detailed analysis of the most popular and widely applied usability questionnaires is hence a very important agenda item for future research – perhaps this would also contribute to answer the question of why the correlations between objective and subjective measures of usability are so low.

In addition to that, the nature of the link between measures of user emotions and perceived usability should be investigated in future research in ergonomics. A better comprehension of the positive correlation of user emotions with perceived usability is essential for a comprehensive understanding of processes and mechanisms that influence behaviour and judgement of users in usability tests. With regard to such a research question, it would be interesting furthermore to consider different contextual factors of usability tests as potential moderating factors of the link between user emotions and perceived usability. The results presented above indicate a difference in the connection between user emotions and perceived usability with regard to source of emotional change (product-inherent factors vs. factors of the testing environment). It was assumed that measures of emotion and perceived usability are correlated only with regard to influencing factors of the product prototype but not of the testing environment. This assumption as well requires empirical verification.

Importance of the influence of contextual factors in usability tests. As mentioned above, it is rather difficult to make a statement about the importance of the observed effects of reduced fidelity of the different contextual factors that have been assessed in the presented studies. This is because in all the studies, only single units of contextual factors have been experimentally varied, whereas the primary object of interest in usability tests - the usability of the product - has not been modified. It would hence be interesting if future research could incorporate a variation of the inherent usability of a product, for example by manipulating the system reaction time, the menu structure or the readability of the displayed information in addition to the manipulation of the fidelity of different contextual factors such as prototype aesthetics and laboratory setup. In doing so, it would be of interest to differentiate between a variety of levels of usability (e.g. perfect usability, small usability problems, and severe usability problems) so as to be able to give

evidence of the influence of contextual factors compared to usability issues in usability tests. Another approach towards the same objective would be the accomplishment of a meta-analysis including all the published studies that varied both aspects of product usability and fidelity level of contextual factors. Given the small number of published studies that have manipulated fidelity of contextual factors and product usability however, it might be difficult to meet the statistical requirements for conducting a meta-analysis (e.g. Hunter & Schmidt, 2004).

Need for multifactorial research designs. The significant interaction effects reported in the culture-study and in the prototype-fidelity study indicate the need to consider more than one factor of the 4FFCF in future studies. For example with regard to the influence of prototype aesthetics on performance measures, it might be interesting for future research to consider the influence of different aspects of the usage context (leisure vs. work) in a multifactorial study design. This might provide further information about the rationale for the different effects of prototype aesthetics on user performance. Furthermore, in order to be able to appraise the effect of contextual factors on usability test outcomes, future research designs should include different levels of inherent usability in addition to the contextual factors of interest.

9.7 Implications for usability practice

The implications of the presented findings for usability practitioners are summarized in table 14. The table indicates possible consequences of reduced fidelity of contextual factors on different usability measures. The table could be read as follows (as example illustrated for the lab-setup study): The presence of observers in the laboratory may lead to an underestimation of product usability because test participants proved to be less efficient in a testing environment with observer presence. Furthermore, measures of user emotions risk being too low in a testing environment with present observers. On the other hand, observer presence seems to have no influence on measures of perceived usability in usability tests.

In summary, usability practitioners should consider different aspects of contextual factors when conducting a usability test. The important impact of reduced fidelity of *product prototypes* on measures of user performance, perceived usability and user emotions implies for usability practitioners that they should pay attention to the design of the prototype used in the usability

Table 14: Consequences of factors of reduced fidelity on usability measures: a table for practitioners

	User performance	**Perceived usability**	**Perceived Attractiveness**	**User emotions**	**Physiological strain**
Influence of observer presence	Usability may be under-estimated	✓	N/R	Emotions may be assessed too negatively	Strain overestimated
Influence of unaesthetic prototypes	Usability may be under-estimated	Usability may be under-estimated	Attractiveness may be over-estimated	Emotions may be assessed too negatively	?
Influence of reduced fidelity prototypes (e.g. paper or computer)	Usability may be under-estimated	✓	N/R	✓	?
Influence of culture	?	?	?	?	?
Influence of product value	✓	too negative if the price is "not right"	N/R	too negative if the price is "not right"	?

Note: ✓ = factor does not influence the measure; N/R = factor is considered not to be relevant for the outcome measure; ? = not yet possible to give evidence based on the available data.

test. Since it is not yet clear how strongly usability measures are influenced by fidelity issues of the product prototype (compared to issues of the inherent usability of the product), it might be recommended for to use prototypes that match the design of the final product as accurate as possible. When using reduced fidelity prototypes, practitioners should bear in mind that the recorded usability measures might turn out to be too high or too low, compared to measures obtained with the real product (see table 14 for a detailed overview of over and underestimation of the usability measures). For example, performance of participants using the computer prototype may be reduced compared to participants using the real product. Especially with regard to the growing availability and the increased use of inexpensive prototyping tools based on computer simulations (e.g. Engelberg & Seffah, 2002), usability professionals should attend to this issue because it can impact participants' performance, lead to a underestimation of usability, which might, ultimately, influence the quality of usability data (and finally lead to a poorly designed product). Furthermore, the influence of aesthetics and product value on outcomes of usability tests indicates the significant role product characteristics beyond usability play in usability tests. This suggests that such aspects should more often be considered in usability practice. Especially when the considerable prestige value associated with a product is high (which is the case e.g. for mobile phones, Dedeoglu, 2004), it might be recommended to integrate product characteristics such as price and value in usability into the tests. Doing so would help to contribute to the claim for an increased integration of issues of marketing and sale in the product development process (see e.g. Benini, Batista & Zuffo, 2005; McClelland & Brigham, 1990).

With regard to the influence of aspects of the *testing environment* in usability tests, the findings presented in this thesis indicate that usability practitioners should be aware of the possible interfering influence of the testing environment. This implies that influencing factors such as the presence of observers should be avoided as far as this is possible. If not, usability practitioners should be aware that the measures of user performance might turn out to be lower, user emotions more negative and measure of physiological strain higher than they would turn out in a real usage scenario.

The cultural background of test participants proved to be of particular importance in usability tests. However, it is not possible to make a specific assumption about the direction of the influence of cultural background on usability measures based on the presented findings. The results however indicate the difficulty of generalizing test results obtained in a specific user

population on a user groups with a different cultural background and hence provide clear evidence of the need of culture-specific usability testing.

9.8 Conclusion

In summary, the presented findings indicate that contextual factors have a considerable influence on results of usability tests. With regard to the importance of this evaluation method in product development, this may considerably affect usability of newly developed products. In other words, the still substantial number of products on the market being difficult to use (cf. Rubin & Chisnell, 2008) may either be due to the fact that their usability has not been evaluated or that contextual factors may have affected the outcomes of the evaluation of their usability.

10 References

Aiello, J. R. and Kolb, K. J. (1995). Electronic performance monitoring and social context: impact on productivity and stress. *The Journal of applied psychology*, *80*(3), 339-353.

Ashkanasy, N. M., Ashton-James, C. E. and Jordan, P. J. (2004). Performance impacts of appraisal and coping with stress in workplace settings: The role of affect and emotional intelligence. *Research in occupational stress and wellbeing*, *3*, 1–43.

Ashton-James, C. E. and Ashkanasy, N. M. (2008). What lies beneath? A process analysis of affective events theory. In: N.M. Ashkanasy, W.J. Zerbe and C.E.J. Härtel (Eds.), *The effect of affect in organizational settings* (pp. 23–46). Bingley: Emerald.

Atyeo, M., Sidhu, C., Coyle, G. and Robinson, S. (1996). Working with marketing. In *Conference companion on Human factors in computing systems: common ground* (pp. 313-314). Vancouver, Canada: ACM.

Bailey R. (1993). Performance vs. Preference. *Proceedings of the HFES 1993*, 282-286.

Bauersfeld, K. and Halgren, S. (1996). "You‘ve got three days!" Case Studies in Field Techniques for the Time-Challenged. In D. Wixon and J. Ramey (Eds.), *Field Methods Casebook for Software Design* (pp.177-195). New York, NY, USA: John Wiley & Sons.

Beatty, S.E. and Talpade, S. (1994). Adolescent Influence in Family Decision Making: A Replication with Extension. *J. Cons. Res., 21*, 332-341.

Benini, M. J., Batista, L. L. and Zuffo, M. K. (2005). When marketing meets usability: the consumer behavior in heuristic evaluation for web. In *Proceedings of the 2005 Latin American conference on Human-computer interaction* (pp. 307 - 312). Cuernavaca, Mexico: ACM.

Bennett, J.L. (1972). The user interface in interactive systems. *Annual Review of Information Science and Technology, 7*, 159–196.

Bennett, J.L. (1979). The commercial impact of usability in interactive systems. In: B.Shackel, (Ed.), *Man–Computer Communication*, Infotech State of the Art Report, vol. 2. Infotech International, Maidenhead.

Bennett, J.L. (1984). Managing to meet usability requirements. In: Bennett, J.L., Case, D., Sandelin, J., Smith, M. (Eds.), *Visual Display Terminals: Usability Issues and Health*

Concerns (pp. 161–184). Englewood Cliffs: Prentice-Hall.

Berkowitz, L. (2000). *Causes and Consequences of Feelings*. Cambridge: University Press.

Berntson, G.G. and Cacioppo, J.T. (2004). Heart rate variability: Stress and psychiatric conditions. *In:* M. Malik and A.J. Camm (eds.), *Dynamic Electrocardiography (pp. 57-64).* New York: Blackwell.

Berntson, G.G. and Stowell, J.R. (1998). ECG artifacts and heart period variability: Don't miss a beat! *Psychophysiology,* 35 (1), 127-132.

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

Bevan, N., Kiriakovsky, J. and Maissel, J. (1991). What is Usability? In *Proceedings of the 4th International Conference on HCI*, Stuttgart, September 1991.

Boucsein, W. and Backs, R.W. (2000). Engineering psychophysiology as a discipline: Historical and theoretical aspects. *In:* R.W. Backs and W. Boucsein (eds.), *Engineering Psychophysiology Issues and Applications (pp. 3-30).* Mahwah, N.J.: Lawrence Erlbaum.

Bradley, M.M. and Lang, P.J. (1994). Measuring emotion: the self-assessment manikin and the semantic differential. *Journal of behavior therapy and experimental psychiatry*, *25*(1), 49-59.

Brady, L. and Phillips, C. (2003). Aesthetics and usability: A look at color and balance. *Usability News, 5*, 1-4.

Brave, S. and Nass, C. (2003). Emotion in human-computer interaction. *In*: J. Jacko and A. Sears (Eds.), *The human-computer interaction handbook: Fundamentals, evolving technologies, and emerging applications (pp. 81–96).* Mahwah: Lawrence Erlbaum Associates.

Brehmer, B. and Dörner, D. (1993). Experiments With Computer-Simulated Microworlds: Escaping Both the Narrow Straits of the Laboratory and the Deep Blue Sea of the Field Study. *Computers in Human Behavior*, 9, 171-184.

Brooke, J. (1996). SUS: A Quick and Dirty Usability Scale. In: P.W. Jordan, B. Thomas, B.A. Weerdmeester and I.L. McClelland (Eds.), *Usability Evaluation in Industry* (pp. 189–194). London: Taylor & Francis.

Buur, J. and Bagger, K. (1999). Replacing Usability Testing with User Dialogue.

Communications of the ACM, 42(5), 63-66.

Carroll, J. and Carrithers, C. (1984). Training wheels in a user interface. *Communications of the ACM, 27*(8), 800-806.

Catani, M. B. and Biers, D. W. (1998). Usability evaluation and prototype fidelity: Users and usability professionals. In *Proceedings of the Human Factors and Ergonomics Society 42nd Annual Meeting*, pp. 1331–1335. Santa Monica, CA: HFES.

Chang, H.C., Lai, H.H. and Chang, Y.M. (2007). A measurement scale for evaluating the attractiveness of a passenger car form aimed at young consumers. *International Journal of Industrial-Ergonomics*, 37 (1), 21-30.

Chen, N., Shaffer, D. and Wu, C. (1997). On physical attractiveness stereotyping in Taiwan: A revised sociocultural perspective. *J. Soc. Psychol., 137,* 117-124.

Chin, J. P., Diehl, V. A. and Norman, K. L. (1988). Development of an instrument measuring user satisfaction of the human-computer interface. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 213-218). Washington, D.C., United States: ACM.

Cohen, J. (1992). A power primer. *Psychological bulletin*, *112*(1), 155-159.

Cohen, J. L. (1979). Social facilitation. Increased evaluation apprehension through permanency of record. *Motivation and Emotion*, *3*(1), 19-33.

Cottrell, N.B., Wack, D.L., Sekerak, G.J. and Rittle, R.M. (1968). Social facilitation of dominant responses by the presence of an audience and the mere presence of others. *Journal of Personality and Social Psychology*, 9, 245-250.

Crilly, N., Moultrie, J. and Clarkson, P. J. (2004). Seeing things: consumer response to the visual domain in product design. *Design Studies*, 25 (6), 547-577.

Csíkszentmihályi, M. (1997). Finding Flow: The Psychology of Engagement with Everyday Life. New York: Basic Books.

DeAngeli, A. D. and Kyriakoullis, L. (2006). Globalisation vs. localisation in e-commerce: cultural-aware interaction design. In *Proceedings of the working conference on Advanced visual interfaces AVI '06* (pp. 250-253). Venezia, Italy: ACM.

Dedeoglu, A. O. (2004). The Symbolic Use of Mobile Telephone Among Turkish Consumers. *Journal of Euromarketing*, *13*(2/3), 143-162.

Desmet, P. M. A. (2003). Measuring emotion: Development and application of an instrument to

measure emotional responses to products. In M. Blythe, C. Overbeeke, A. F. Monk and P. C. Wright (Eds.), *Funology: From Usability to Enjoyment* (pp. 111-124). Dordrecht: Kluwer.

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

Desmet, P.M.A., Hekkert, P. and Jacobs, J. (2000). When a car makes you smile: Development and application of an instrument to measure product emotions. In S. J. Hoch and R. J. Meyer (Eds), *Advances in Consumer Research, (Vol. 19, pp. 111-117).* Provo: Association for Consumer Research.

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* (pp. 26-30). Toronto, Ontario, Canada: ACM.

Dion, K., Berscheid, E. and Walster, E. (1972). What is beautiful is good. *Journal of personality and social psychology*, *24*(3), 285-290.

Dörner, D. and Stäudel, T. (1990). Emotion und Kognition [emotion and cognition]. *In:* K.H. Scherer (ed.), *Psychologie der Emotion (pp. 293-344).* Göttingen: Hogrefe.

Dray, S. and Siegel, D. (2004). Remote possibilities?: international usability testing at a distance. *Interactions,* 11 (2), 10-17.

Dumas, J.S. and Redish, J.C. (1999). *A Practical Guide to Usability Testing*. Exeter: Cromwell Press.

Eagly, A. H. and Chaiken, S. (1993). *The psychology of attitudes*. Fort Worth: Harcourt Brace Jovanovich College Publishers.

Earthy, J., Jones, B. S. and Bevan, N. (2001). The improvement of human-centred processes-facing the challenge and reaping the benefit of ISO 13407. *International Journal of Human Computer Studies*, *55*(4), 553–586.

Eason, K. D. (1984). Towards the experimental study of usability. *Behaviour & Information Technology*, *3*(2), 133-143.

Eason, K.D. (1981). A task-tool analysis of the manager–computer interaction. In: B. Shackel (Ed.), *Man–Computer Interaction*. Amsterdam: Sijthoff and Noordhoff.

Elmes, D. G., Kantowitz, B. H. and Roediger, H. L. (2003). *Research methods in psychology*

(7th edition). West Publishing Company.

Engelberg, D. and Seffah, A. (2002). A Framework for Rapid Mid-Fidelity Prototyping of Web Sites. In *IFIP World Computer Congress 2002*. Montreal, Canada: Kluwer Academic Publishers.

Englisch, J. (1993). *Ergonomie von Softwareprodukten*. Mannheim: BI Wissenschaftsverlag.

Farrington-Darby, T. and Wilson, J.R. (2005). The nature of expertise: A review. *Applied Ergonomics, 37,* 17–32.

Faul, F., Erdfelder, E., Lang, A. and Buchner, A. (2007). G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior research methods*, *39*(2), 175-91.

Forgas, J.P. and George, J.M. (2001). Affective influences on judgments and behavior in organizations: An information processing perspective. *Organizational Behavior and Human Decision Processes,* 86 (1), 3-34.

Forlizzi, J. and Battarbee, K. (2004). Understanding experience in interactive systems. *In: Proceedings of the 2004 conference on Designing Interactive Systems (DIS 04): processes, practices, methods, and techniques*. New York: ACM, 261-268.

Forlizzi, J. and Ford, S. (2000). The building blocks of experience: an early framework for interaction designers. In *Proceedings of the 3rd conference on Designing interactive systems: processes, practices, methods, and techniques* (pp. 419-423). New York City: ACM.

Foxman, E., Tansuhaj, P. and Ekstrom, K.M. (1989). Adolescents' Influence in Family Purchase Decisions: A Socialisation Perspective. *J. Bus. Res., 18*, 159-172.

Frijda, N. (1993). Moods, emotion episodes and emotions. In M. Lewis and J. M. Haviland (Eds.), *Handbook of emotions* (pp. 381-403). New York: Guilford Press.

Frøkjær, E., Hertzum, M. and Hornbæk, K. (2000). Measuring usability: are effectiveness, efficiency, and satisfaction really correlated? In *Proceedings of the ACM CHI 2000 Conference on Human Factors in Computing Systems* (pp. 345-352). The Hague, The Netherlands: ACM.

Gediga, G., Hamborg, K. C. and Duntsch, I. (1999). The IsoMetrics usability inventory: an operationalization of ISO9241-10 supporting summative and formative evaluation of software systems. *Behaviour & Information Technology*, *18*(3), 151-164.

Geen, R.G. (1991). Social motivation. *Annual Review of Psychology,* 42, 377-399.

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

Guzzo, R. A., Jackson, S. E. and Katzell, R. A. (1987). Meta-analysis analysis. *Research in Organizational Behavior*, *9*, 407-442.

Hall, R.R. (1999). Usability and product design: a case study. *In*: P. Jordan and W.S. Green, (eds.), *Human Factors in Product Design (pp. 85-91)*. London: Taylor & Francis.

Hancock, P.A., Weaver, J. L. and Parasuraman, R. (2002). Sans subjectivity – ergonomics is engineering. *Ergonomics*, 45 (14), 991-994.

Hartmann, J., De Angeli, A. and Sutcliffe, A. (2008). Framing the User Experience: Information Biases on Website Quality Judgement. *CHI 2008*, April 5–10, 2008, Florence, Italy.

Hartmann, J., Sutcliffe, A. and de Angeli, A.D. (2007). Investigating attractiveness in web user interfaces. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, San Jose, April 28 - May 3, pp. 387-396.

Hassenzahl, M. (2004). The Interplay of Beauty, Goodness, and Usability in Interactive Products. *Human-Computer Interaction*, *19*(4), 319-349.

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

Hassenzahl, M., Burmester, M. and Koller, F. (2003). AttrakDiff: Ein Fragebogen zur Messung wahrgenommener hedonischer und pragmatischer Qualität. In J. Ziegler and G. Szwillus (Eds.), *Mensch & Computer 2003. Interaktion in Bewegung* (pp. 187–196). Stuttgart: B. G. Teubner.

Hekkert, P. (2006). Design aesthetics: principles of pleasure in design. *Psychology Science, 48*(2), 157 – 172.

Hekkert, P. and Schifferstein, H.N.J. (2008). Introducing produc experience. In H.N.J. Schifferstein and P. Hekkert (Eds.), *Product experience* (pp.1-8). Amsterdam: Elsevier.

Hekkert, P., Snelders, H.M.J.J. and van Wieringen, P.C.W. (2003). Most advanced, yet acceptable: Typicality and novelty as joint predictors of aesthetic preference. *British Journal of Psychology*, 94, 111-124.

Helander, M.G. and Khalid, H.M. (2005). Affective and Pleasurable Design. *In*: G. Salvendy, (ed.), *Handbook of Human Factors and Ergonomics (pp. 543-572),* 3rd ed. New York:

Wiley Interscience.

Hembree, R. (1988). Correlates, Causes, Effects, and Treatment of Test Anxiety. *Review of Educational Research*, 58 (1), 47-77.

Holbrook, M.B. (1986). Aims, Concepts, and Methods for the Representation of Individual Differences in Esthetic Responses to Design Features. *J. Cons. Res., 13,* 337-347.

Holbrook, M.B. and Corfman, K.P. (1985). Quality and Value in the Consumption Experience: Phaedrus rides again. In: J. Jacoby, J. C. Olson (Eds.), *Perceived Quality: How Consumers View Stores and Merchandise (pp. 31-57).* Lexington, MA: Lexington Books.

Hornbæk, K. (2006). Current practice in measuring usability: Challenges to usability studies and research. *International Journal of Human-Computer Studies*, *64*(2), 79-102.

Hornbæk, K. and Law, E. L. (2007). Meta-analysis of correlations among usability measures. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 617-626). San Jose, California, USA.

Hunter, J.E. and Schmidt, F.L. (2004). Methods of meta-analysis : correcting error and bias in research findings. Thousand Oaks: Sage.

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

Izard, C.E. (1991). *The psychology of emotions*. New York: Plenum.

Jacobsen, N. E., Hertzum, M. and John, B. E. (1998). The Evaluator Effect in Usability Studies: Problem Detection and Severity Judgments. *Human Factors and Ergonomics Society Annual Meeting Proceedings, 42*, 1336-1340.

Jordan, P.W. (1998a). *An Introduction to Usability.* Lodon: Tylor & Francis.

Jordan, P.W. (1998b). Human factors for pleasure in product use. *Applied Ergonomics*, 29 (1), 25-33.

Jordan, P.W. (2000). *Designing Pleasurable Products*. New York, NY : Taylor & Francis.

Jordan, P.W. and Green, W.S. (2002). *Pleasure with products: beyond usability*. London : Taylor & Francis.

Jorna, P.G.A.M. (1992). Spectral analysis of heart rate and psychological state: A review of its validity as a workload index. *Biological Psychology,* 34 (2-3), 237-257.

Kaikonnen, A., Kekäläinen, A., Cankar, M., Kallio, T. and Kankainen, A. (2005). Usability

testing of mobile applications: A comparison between laboratory and field testing. *Journal of Usability Studies*, *1*, 4-16.

Keenan, S.L., Hartson, H.R., Kafura, D.G. and Schulman, R.S. (1999). The Usability Problem Taxonomy: A Framework for Classification and Analysis. *Empirical Software Engineering, 4*(1), 1382-3256.

Kelsey, R. M., Blascovich, J., Leitten, C. L., Schneider, T. R., Tomaka, J. and Wiens, S. (2000). Cardiovascular reactivity and adaptation to recurrent psychological stress: the moderating effects of evaluative observation. *Psychophysiology*, *37*(6), 748-756.

Keltner, D., Ellsworth, P. C. and Edwards, K. (1993). Beyond simple pessimism: effects of sadness and anger on social perception. *Journal of Personality and Social Psychology, 64*, 740–752.

Kessner, M., Wood, J., Dillon, R. F. and West, R. L. (2001). On the reliability of usability testing. In *CHI '01 extended abstracts on Human factors in computing systems* (pp. 97-98). Seattle, Washington: ACM.

Kettunen, J. and Keltikangas-Järvinen, L. (2001). Intraindividual analysis of instantaneous heart rate variability. *Psychophysiology,* 38 (4), 659–668.

Khadir, F. (2007). Marketing and its impact on vulnerable consumer groups like children, adolescents etc. In *Proceedings of the International Marketing Conference on Marketing & Society*, Kozhikode, April 8-10, pp. 433-441.

Khalid, H.M. (2006). Embracing diversity in user needs for affective design. *Applied Ergonomics,* 37 (4), 409-418.

Kindlund, E. and Sauro, J. A (2005). Method to Standardize Usability Metrics into a Single Score. *Proc. CHI 2005*, 401-409.

Kirakowski, J. (1996). The software usability measurement inventory: Background and usage. In P. W. Jordan, B. Thomas, B. A. Weerdmeester and I. L. McClelland (Eds.), Usability evaluation in industry (pp. 169-178). London: Taylor & Francis.

Kirakowski, J. and Corbett, M. (1993). Sumi: The software usability measurement inventory. British Journal of Educational Technology, 24(3), 210-212.

Kirakowski, J. and Dillon, A. (1988). *The computer user satisfaction inventory (CUSI): Manual and scoring key*. Cork, Ireland: Human Factors Research Group, University College of Cork.

Kirakowski, J., Claridge, N. and Whitehead, R. (1998). Human centered measures of success in web site design. In *Proceedings of the Fourth Conference on Human Factors and the Web, Basking Ridge, New Jersey (USA).*

Kirschbaum C., Pirke K.M. and Hellhammer D.H. (1993). The 'Trier Social Stress Test' – a tool for investigating psychobiological stress responses in a laboratory setting. *Neuropsychobiology,* 28 (1-2), 76-81.

Kjeldskov, J., Skov, M.B. and Stage, J. (2008). A longitudinal study of usability in health care: Does time heal? *International Journal of Medical Informatics*, *79* (6), 135-143.

Klaczynski, P.A. and Gordon, D.H. (1996). Self-Serving Influences on Adolescents' Evaluations of Belief-Relevant Evidenc., *J. Exper. Child Psychol., 62*, 317-339.

Krohne, H.W., Egloff, B., Kohlmann, C.W. and Tausch, A. (1996). Untersuchungen mit einer deutschen Version der "Positive and Negative Affect Schedule" (PANAS) [Studies with a German version of the positive and negative affect schedule]. *Diagnostica,* 42 (2), 139-156.

Kuniavsky, M. (2003). *Observing the user experience*. San Francisco: Morgan Kaufman.

Kurosu, M. and Kashimura, K. (1995). Apparent usability vs. inherent usability: experimental analysis on the determinants of the apparent usability. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, Denver, May 07 - 11, pp. 292-293.

Lavery, D., Cockton, G. and Atkinson, M. P. (1997). Comparison of evaluation methods using structured usability problem reports. *Behaviour & Information Technology*, *16*(4), 246. doi: 10.1080/014492997119824.

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

Lazar J., Meiselwitz, G. and Norcio, A. (2003). Novice user perception of error on the Web. U*niversal Access in the Information Society*, *3*(3), 202-208.

Lazarus, R.S. (1993). From psychological stress to the emotions: A history of changing outlooks. *Annual Review of Psychology,* 44, 1-21.

Lazarus, R.S. and Folkman, S. (1984). *Stress, appraisal and coping*. New York: Springer.

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. (2006). Usability testing. In: Salvendy, G. (Ed.), *Handbook of Human Factors and Ergonomics* (pp. 1275–1316). New York: John Wiley.

Lin, H.X., Choong, Y.-Y. and Salvendy, G. (1997). A Proposed Index of Usability: A Method for Comparing the Relative Usability of Different Software Systems. Behaviour and Information Technology, 16 (4/5), 267-278.

Lin, T., Omata, M., Hu, W. and Imamiya, A. (2005). Do physiological data relate to traditional usability indexes? In *Proceedings of the 17th Australia conference on Computer-Human Interaction: Citizens Online: Considerations for Today and the Future* (pp. 1-10). Canberra, Australia: Computer-Human Interaction Special Interest Group (CHISIG) of Australia.

Lindgaard, G. (2007). Aesthetics, visual appeal, usability, and user satisfaction: What do the user's eyes tell the user's brain. *Australian J. Emerging Tech. Society, 5,* 1-16.

Lindgaard, G., Dudek, G., Fernandes, G. and Brown, J. (2006). Attention web designers: You have 50 milliseconds to make a good first impression! *Behav. Info. Tech., 25,* 115-126.

Liu, Y. (2003). Engineering Aesthetics and Aesthetic Ergonomics: Theoretical Foundations and a Dual-Process Research Methodology. *Ergonomics*, 46, 1273-1292.

Lord, C.G., Ross, L. and Lepper, M.R. (1979). Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *J. Pers. Soc. Psychol., 37*, 2098-2109.

Lubner-Rupert, J. A. and Winakor, G. (1985). Male and Female Style Preference and Perceived Fashion Risk. *Family Cons. Sci. Res. J., 13,* 256-266.

MacCoun, R.J. (1998). Biases in the interpretation and use of research results. *Ann. Rev. Psychol., 49*, 259-287.

Magrab, P.R. (2005). The Adolescent Learner and the Aesthetic Experience: A Brief Overview. In *Proceedings of the Unesco Expert Panel Meeting*, Education through Art - Building Partnerships for Secondary Education, Newark, October 27, pp.7-11.

Mankoff, J., Dey, A. K., Hsieh, G., Kientz, J., Lederer, S. and Ames, M. (2003). Heuristic evaluation of ambient displays. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 169-176). Ft. Lauderdale, Florida, USA: ACM.

Mann, S. and Smith, L. (2006). Arriving at an agile framework for teaching software engineering. In *19th Annual Conference of the National Advisory Committee on*

Computing Qualifications (p. 183–190), Wellington, New Zealand.

Marcus, A. (2003). The emotion commotion. *Interactions,* 10 (6), 28-34.

Martin, H. and Gaver, B. (2000). Beyond the snapshot from speculation to prototypes in audiophotography. In *Proceedings of the conference on Designing interactive systems processes, practices, methods, and techniques - DIS '00* (pp. 55-65). New York, ACM Press.

Marty, P.F. and Twidale, M.B. (2005). Usability@90mph: Presenting and Evaluating a New, High-Speed Method for Demonstrating User Testing in Front of an Audience. *First Monday, 10*(7), 1-18.

McCarthy, J. and Wright, P. (2004). *Technology as Experience*. Cambridge: MIT Press.

McClelland, I. L. and Brigham, F. R. (1990). Marketing Ergonomics - how should ergonomics be packaged. *Ergonomics, 33*(5), 519-526.

McCurdy, M., Connors, C., Pyrzak, G., Kanefsky, B. and Vera, A. (2006). Breaking the fidelity barrier: an examination of our current characterization of prototypes and an example of a mixed-fidelity success. In *Proceedings of the SIGCHI conference on Human Factors in computing systems (pp. 1233-1242),* April 22 - 27, 2006, Montréal, Canada. New York: ACM.

Mehrabian, A. and Russel, J.A. (1977). Evidence for a three-factor theory of emotions. *Journal of Research in Personality*, *11*, 273-294.

Mentis, H. M. and Gay, G. K. (2003). User recalled occurrences of usability errors: Implications on the user experience. *Extended Abstracts of the Conference on Human Factors in Computing Systems, Ft. Lauderdale, Fl,* 736 –737.

Meyers-Levy, J. and Sternthal, B. (1991). Gender Differences in the Use of Message Cues and Judgments. *J. Marketing Res., 28*, 84-96.

Milanese, S. (2005). Adolescent ergonomics. In: P.D. Bust, P.T. McCabe (Eds), *Contemporary Ergonomics (pp. 327-330).* London: Taylor & Francis.

Miller, R.B. (1971). Human ease of use criteria and their tradeoffs. *IBM Report TR 00.2185*, 12 April. IBM Corporation, Poughkeepsie, NY.

Minshall, B., Winakor, G. and Swinney, J. L. (1982). Fashion Preferences of Males and Females, Risks Perceived, and Temporal Quality of Styles. *Family Cons. Sci. Res. J., 10,* 369-379.

Molich, R., Thomsen, A. D., Karyukina, B., Schmidt, L., Ede, M., Oel, W. V. and Arcuri, M. (1999). Comparative evaluation of usability tests. In *CHI '99 extended abstracts on Human factors in computing systems* (pp. 83-84). Pittsburgh, Pennsylvania: ACM.

Morris, W. N. (1989). *Mood: the frame of mind*. New York: Springer-Verlag.

Nakarada-Kordich, I. and Lobb, B. (2005). Effect of Perceived Attractiveness of Web Interface Design on Visual Search of Web Sites. In *Proceedings CHINZ '05*, July 6-8, Auckland, pp. 25-27.

Newman, M. W., Lin, J., Hong, J. I. and Landay, J. A. (2003). DENIM: an informal web site design tool inspired by observations of practice. *Hum.-Comput. Interact.*, *18*(3), 259-324.

Ngo, D.C.L., Teo, L.S. and Byrne, J.G. (2003). Modelling interface aesthetics*, Info. Sci., 152,* 25–46.

Nickel, P. and Nachreiner, F. (2003). Sensitivity and Diagnosticity of the 0.1 Hz component of Heart Rate Variability as an Indicator of Mental Workload. *Human Factors,* 45 (4), 575-590.

Nielsen, J. (1990). Paper versus Computer Implementations as Mockup Scenarios for Heuristic Evaluation. In *Proceedings of the IFIP TC13 Third Interational Conference on Human-Computer Interaction (pp. 315-320),* August 27-31, 1990, Cambridge, UK. Amsterdam: North-Holland,.

Nielsen, J. (1993). *Usability engineering*. Amsterdam: Morgan Kaufmann.

Nielsen, J. (1994). Usability laboratories. *Behaviour & Information Technology*, *13*(1), 3-8.

Nielsen, J. and Levy, J. (1994). Measuring usability: Preference vs. Performance. *Communication of the ACM,* 37 (4), 66-75.

Nielsen, J. and Molich, R. (1990). Heuristic evaluation of user interfaces. In *Proceedings of the SIGCHI conference on Human factors in computing systems: Empowering people* (pp. 249-256). Seattle, Washington, United States: ACM.

Nielsen, J., Clemmensen, T. and Yssing, C. (2002). Getting access to what goes on in people's heads? reflections on the think-aloud technique. *NordiCHI '02: Proceedings of the second Nordic conference on Human-computer interaction* (p. 101—110). New York, NY, USA: ACM Press.

Niskanen, J.-P., Tarvainen, M.P., Ranta-Aho, P.O. and Karjalainen, P.A. (2004). Software for

advanced HRV analysis. *Computer Methods and Programs in Biomedicine,* 76 (1), 73-81.

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

Norman, D.A. (2004b). Introduction to this special section on beauty, goodness, and usability. *Human-Computer-Interaction*, *19*, 311-318.

Ollermann, F. (2001). *Evaluation von Hypermedia-Anwendungen: Entwicklung und Validierung eines Instruments*. Unpublished Diploma thesis. Institute of Work and Organisational Psychology, University of Osnabrück, Germany.

Olson, J. M. and Zanna, M. P. (1993). Attitudes and attitude change. *Annual Review of Psychology, 44*, 117–154.

Öörni, K. (2003) What do we know about usability evaluation? – A critical view. *Proceedings of the Conference on Users in the Electronic Information Environments*, Espoo, Finland.

Partala, T. and Surakka, V. (2004). The effects of affective interventions in human-computer interaction. *Interacting with Computers*, *16*(2), 295-309.

Patel, M. and Loring, B. (2001). Handling awkward usability testing situations. *In: Proceedings of the Human Factors and Ergonomics Society annual meeting, 2, Santa Monica (USA),*1772-1776.

Petty, R. E., Wegener, D. T. and Fabrigar, L. R. (1997). Attitudes and attitude change. *Annual Review of Psychology, 48*, 609–647.

Postlewait, D.S. and Morganosky, M.A. (1989). Consumers' Evaluations of Apparel Form, Expression, and Aesthetic Quality. *Clothing Textiles Res. J., 7,* 11-15.

Postrel, V. (2003). *The substance of style: how the rise of aesthetic value is remarking commerce, culture, and consciousness*. Harper Collins, New York.

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

Rani, P., Sims, J., Brackin, R. and Sarkar, N. (2002). Online stress detection using psychophysiological signals for implicit human-robot cooperation. *Robotica*, *20*(06),

673–685.

Reicherts, M., Salamin, V., Maggiori, C. and Pauls, K. (2005). Psychometric characteristics of a computer-based monitoring system for emotion processing and affective states. In: *Proceedings of the 10th Spanish Conference on Biometrics, 25th–27th May 2005*, Oviedo.

Rowe, D. W., Sibert, J. and 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* (pp. 480-487). Los Angeles, California, United States: ACM Press/Addison-Wesley Publishing Co.

Rubin, J. (1994). *Handbook of usability testing: How to plan, design, and conduct effective tests*. New York: J. Wiley.

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

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

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

Salzman, M.C. and Rivers, S.D. (1994). Smoke and mirrors: Setting the stage for a successful usability test. *Behavior & Information Technology,* 13 (1), 9-16.

Sarodnick, F. and Brau, H. (2006). *Methoden der Usability Evaluation: wissenschaftliche Grundlagen und praktische Anwendung*. Bern: Huber.

Sauer J. and 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*, 670-677.

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

Sauer, J. and Sonderegger, A. (in press). Methodological issues in usability testing: the influence of testing environment and task. *Applied Ergonomics*.

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

Scheirer, J., Fernandez, R., Klein, J. and Picard, R. W. (2002). Frustrating the user on purpose: a

step toward building an affective computer. *Interacting with Computers*, *14*(2), 93-118.

Schenkman, B.N. and Jönsson, F.U. (2000). Aesthetics and preferences of web pages, *Behav.Info. Tech., 19* 367-377.

Schleicher, D.J., Watt, J.D. and Greguras,G.J. (2004). Reexamining the Job Satisfaction-Performance Relationship: The Complexity of Attitudes. *Journal of Applied Psychology*, 89 (1), 165-177.

Schrier, J.R. (1992). Reducing stress associated with participating in a usability test. *In: Proceedings of the Human Factors Society 36th Annual Meeting, Santa Monica (USA),* 1210-1214.

Scriven, M. (1967). The Methodology of Evaluation. In R. Tyler, R. Gagne and M. Scriven (Eds.), *Perspectives of Curriculum Evaluation (pp. 39-83).* Chicago: Rand McNally.

Sefelin, R., Tscheligi, M. and Giller, V. (2003). Paper prototyping – what is it good for? A comparison of paper-and computer-based prototyping. *In*: *Proceedings of CHI (778-779)*, April 5-10, 2003, Florida, USA. New York: ACM.

Seta, C. E., Seta, J. J., Donaldson, S. and Wang, M. A. (1988). The effects of evaluation on organizational processing. *Personality and Social Psychology Bulletin*, *14*(3), 604–609.

Seva, R.R., Duh, H.B-L. and Helander, M.G. (2007). The marketing implications of affective product design. *Applied Ergonomics*, 38 (6), 723-731.

Shackel, B. (1981). The concept of usability. In *Proceedings of IBM software and information usability symposium*, Poughkeepsie, NY, 15–18 September. pp. 1– 30.

Shackel, B. (1991). Usability – Context, framework, definition, design and evaluation. In B. Shackel and S.J. Richardson (Eds.), *Human Factors for Informatics Usability* (pp. 21-31). Cambridge: University Press.

Shedroff, N. (2001). *Experience design.* Indianapolis: New Riders Publishing.

Shim, S. (1996). Adolescent consumer decision-making styles: The consumer socialization perspective. *Psychology Marketing, 13,* 547-569.

Shiv, B. and Fedorikhin, A. (1999). Heart and mind in conflict: The interplay of affect and cognition in consumer decision making. *Journal of Consumer Research,* 26 (3), 278-292.

Simon, S. J. (2003). The Impact of Culture and Gender on Web Sites : An Empirical Study. *Data Base For Advances In Information Systems*, *32*(1), 18-36.

Snyder, C. (2003). *Paper Prototyping: The Fast and Easy Way to Design and Refine User Interfaces.* San Francisco: Morgan Kaufmann.

Somervell, J., Wahid, S. and McCrickard, D. S. (2003). Usability heuristics for large screen information exhibits. In *Proceedings of the Ninth IFIP TC13 International Conference on Human Computer Interaction* (pp. 904–907).

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.

Sonderegger, A. and Sauer J. (2010). The influence of design aesthetics in usability testing: effects on user performance and perceived usability. *Applied Ergonomics, 41*, 403–410.

Sonderegger, A. and Sauer J. (under review). The Influence of Cultural Background and Product Value in Usability Testing. *Behaviour & Information Technology*.

Stroud, L., Salovey, P. and Epel, E. (2002). Sex differences in stress responses: social rejection versus achievement stress. *Biological Psychiatry , 52* (4), 318 – 327.

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.

Thomas, B. (1996). Quick and dirty usability tests. In P. Jordan, B. Thomas, B. Weerdmeester and I. McClelland, *Usability evaluation in industry* (pp. 107-114). London: Taylor & Francis.

Thomas, J. C. and Kellogg, W. A. (1989). Minimizing ecological gaps in interface design. *IEEE SOFTWARE*, 78–86.

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

Tractinsky, N. (1997). Aesthetics and apparent usability: Empirically assessing cultural and methodological issues. In *Proceedings of the CHI (115-122),* March 22 - 27, 1997, Atlanta, USA. New York: ACM.

Tractinsky, N., Shoval-Katz, A. and Ikar, D. (2000). What is beautiful is usable. *Interacting with computers, 13*, 127–145.

Urokohara, H., Tanaka, K., Furuta, K., Honda, M. and Kurosu M. (2000). NEM: "Novice Expert ratio Method" A Usability Evaluation Method to Generate a New Performance

Measure. In *Extended Abstracts of CHI 2000*, pp 185-186, New York: ACM.

Väänänen-Vainio-Mattila, K., Roto, V. and Hassenzahl, M. (2008). Now let's do it in practice: user experience evaluation methods in product development. In *CHI '08 extended abstracts on Human factors in computing systems* (pp. 3961-3964). Florence, Italy: ACM.

van der Heijden, H. (2003). Factors influencing the usage of websites: the case of a generic portal in The Netherlands. *Information Management, 40,* 541–549.

Virzi, R., Sokolov, J. and Karis, D. (1996). Usability problem identification using both low- and high-fidelity prototypes. *Proceedings of the SIGCHI conference on Human factions in computings*, 236-243.

Vu, K.-P.L. and Proctor, R.W. (2006). Web site design and evaluation. In: Salvendy, G. (Ed.), Handbook of Human Factors and Ergonomics. John Wiley, New York, pp. 1317-1343.

Walker, M., Takayama, L. and 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),* September 29 - October 4, 2002, Baltimore, USA. Santa Monica: HFES.

Ward, R. D. and Marsden, P. H. (2003). Physiological responses to different WEB page designs. *International Journal of Human-Computer Studies*, *59*(1-2), 199-212.

Wastell, D. and Newman, M. (1996). Information system design, stress and organisational change in the ambulance services: A tale of two cities. *Accounting, Management and Information Technologies,* 6, 283-300.

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

Weiss, H.M. (2002). Deconstructing job satisfaction. *Human Resource Management Review, 12*, 173–194.

Wharton, C., Rieman, J., Lewis, C. and Polson, P. (1994). The cognitive walkthrough method: A practitioner's guide. *Usability inspection methods*, 105–140.

Wickens, C.D. and Hollands, J.G. (2000). *Engineering Psychology and Human Performance*. Upper Saddle River, N.J.: Prentice-Hall.

Wiklund, M., Thurrott, C. and Dumas, J. (1992). Does the fidelity of software prototypes affect the perception of usability? In *PROCEEDINGS OF THE HUMAN FACTORS SOCIETY*

36th ANNUAL MEETING (pp. 399-403). San Jose, CA, USA.

Willumeit, H., Gediga, G. and Hamborg, K.C. (1995). *Validation of the IsoMetrics usability inventory*. Unpublished research report, Institute of Work and Organisational Psychology, University of Osnabrück, Germany.

Willumeit, H., Gediga, G. and Hamborg, K.-C. (1996). Isometrics: Ein Verfahren zur formativen Evaluation von Software nach ISO 9241/10 [Isometrics: A tool for the formative software evaluation based on ISO 9241/10]. *Ergonomie & Informatik,* 27, 5-12.

Wilson, G.M. and Sasse, M.A. (2000). Investigating the Impact of Audio Degradations on Users: Subjective vs. Objective Assessment Methods. In *OZCHI 2000: Interfacing Reality in the New Millennium*, pp. 135 – 142 (Sydney, Australia).

Wright, P.C., McCarthy, J.C. and Meekison, L. (2003). Making sense of experience. In M. Blythe, C. Overbeeke, A. F. Monk and P. C. Wright (Eds.), *Funology: From Usability to Enjoyment* (pp. 43-54). Dordrecht: Kluwer.

Yamamoto, M. and Lambert, D.R. (1994). The Impact of Product Aesthetics on the Evaluation of Industrial Products. *Journal of Product Innovation Management*, *11* (4), 309-324.

Ziefele, M. (2002). The influence of user expertise and phone complexity on performance, ease of use and learnability of different mobile phones. *Behaviour & Information Technology, 21*(5), 303-311.