doi:10.1093/iwc/iws002

Effects of Perceived Prototype Fidelity in Usability Testing under Different Conditions of Observer Presence

A. UEBELBACHER*, A. SONDEREGGER AND J. SAUER

Department of Psychology, University of Fribourg, Rue de Faucigny 2, CH-1700 Fribourg, Switzerland
**Corresponding author: andreas@uebelbacher.ch*

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

RESEARCH HIGHLIGHTS

- Performance, emotions and psychophysiology are not affected by stage of system development.
- Observer presence in usability testing cause higher stress levels among participants.
- Observers impair participants' performance more in final than in early system testing.
- A single-item usability measure is more strongly affected by test instructions than a multiple-item scale.

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

Editorial Board Member: Kaisa Väänänen-Vainio-Mattila

Received 24 May 2011; Revised 18 July 2012; Accepted 26 September 2012

1. INTRODUCTION

The importance of usability as a goal in product development is increasingly acknowledged and the benefits of usability testing as a core method in this endeavour are hardly controversial (Lewis, 2006). One of the method's advantages is its flexibility, which allows an application at various stages of product development. At early stages, usability testing typically uses prototypes of the system to conduct a formative evaluation, identifying specific usability improvements (Gediga *et al.*, 2002). At later stages, the final product is available for tests with users and a summative approach is often applied taking global measures of performance and making an overall assessment of usability [e.g. using standardized questionnaires, such as Software Usability Measurement Inventory, Post Study System Usability Questionnaire (PSSUQ)] (Tullis and Albert, 2008). The tested prototypes can be very different from the final product with respect to various dimensions of their fidelity, and so can the respective test outcomes (e.g. a specific interaction pattern can only be tested if the prototype offers the required richness of interaction) (McCurdy *et al.*, 2006; Virzi *et al.*, 1996). However, how prototype fidelity is perceived by participants in a usability test can be very different even when

testing the same system (e.g. it may depend on task instructions). Therefore, the question arises whether the perceived prototype fidelity might have an impact on test outcomes. For usability practitioners, conducting prototype tests at various stages of product development is a reality and it is of high importance to know how a prototype should be presented to the participants to avoid any undesired side effects.

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

1.1. Perceived prototype fidelity

In the development of interactive products, it is important to gain user feedbacks on variants of a system design as early as possible, to avoid the high costs of product changes after implementation of a system (Mantei and Teorey, 1988). Therefore, the usability practitioner often faces the situation that user data need to be collected even before a working system is available for testing. Prototype testing then becomes the method of choice, and surveys among usability practitioners prove that the method of iterative user testing with early prototypes is very common (Vredenburg *et al.*, 2002). The main requirements for prototypes are low cost of production and sufficient similarity to the final product to reach valid test outcomes. Therefore, a very important characteristic of prototypes is their 'fidelity' to the final product, including aesthetic refinement, similarity of interaction and breadth of functions (McCurdy *et al.*, 2006; Sauer *et al.*, 2010; Virzi *et al.*, 1996).

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

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

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

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

The comparison of these two forms of usability evaluation reveals that one important difference between the two is the stage in the product development at which the evaluation typically takes place (i.e. early for formative testing and late for summative testing). Whether a usability test is conducted early or late in the product development cycle may have different consequences if the test participants become aware of the product's development stage. In a test conducted early in the product development cycle (such as informative testing), the participants may be happy to give critical feedback about the weaknesses of the product because there is plenty of time to remedy any such shortcomings. In contrast, giving such open and critical

feedback in a usability test conducted late in the product development cycle (such as in summative testing) may have more dramatic consequences if the test outcomes require a delay in product launch (Karat, 1994). Therefore, the test participants may feel that the social desirability of achieving positive test results would be higher in a late test than in an early one.

1.2. Observer presence in usability testing

In addition to the perceptions of product development stages, there are further factors that may influence the test participants' perception of the testing situation. The set-up of the laboratory represents such a factor in usability testing, which may also affect user behaviour during the test. Very few studies have investigated the influence of observer presence as an important social factor during the testing process. For example, Sonderegger and Sauer (2009) compared three conditions of observer presence in user testing (facilitator present with two observers, only facilitator present, test subject working alone in the room) and found that the presence of two additional observers had a negative impact on participants' heart rate variability (HRV), performance and emotions. Harris *et al.* (2005) found some evidence for effects of observer presence, as their work showed higher error rates for complex tasks when a facilitator was present than when the participant was alone. Grubaugh *et al.* (2005) also found higher error rates in usability testing when the laboratory set-up was more intrusive in terms of monitoring equipment used.

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

1.3. The present study

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

To investigate whether a previously demonstrated effect in usability tests (i.e. impact of observer presence) would equally occur in early prototype testing as in final system testing, we implemented two experimental conditions, which proved in previous work by Sonderegger and Sauer (2009) to have an effect on test outcomes. In one condition, a test facilitator (introduced as a university researcher) and two non-interacting observers (introduced as representatives of the product developer) were present in the test room throughout the test session. While the facilitator explained the procedure and answered participants' questions during the instruction phase, during task completion no interaction between participant and facilitator was allowed. In the other condition, no observers and no facilitator were present in the test room.

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

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

2. METHOD

2.1. Participants

Eighty participants (70% female) took part in the experiment, aged between 17 and 65 years (age: mean 27.9; SD 10.2). They were recruited from the general public and among students, using a test participant pool from the Universities of Basel and Fribourg. They all had no prior interaction with the experimenter or the observers being present during the experiment. Prior to the experiment, it was checked that participants did not own the specific device to be used in the experiment and were excluded if this had been the case. They were paid 25 Swiss francs (approximately €20) for participation.

2.2. Design

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

2.3. Measures and instruments

2.3.1. Performance

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

2.3.2. Perceived usability

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

Additionally, we used a visual analogue scale (0–100; ranging from 'not at all' to 'very much') to measure an overall estimation of perceived usability ('The software is usable'). Single-item measures of usability proved to be valuable and reliable in previous research (Christophersen and Konradt, 2010).

2.3.3. Emotions

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

2.3.4. Task load

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

2.3.5. Psychophysiology

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

2.3.6. Additional measures

As a control variable, previous experience with the specific test device was assessed by means of a visual analogue scale (0–100; ranging from 'no experience at all' to 'a great deal of experience') to rule out an impact of different knowledge and skill levels with respect to the hardware and software in question. As a manipulation check, we introduced four visual analogue scales, questioning participants for their estimation of: (a) the test system in terms of developmental state, (b) the distance to market introduction, (c) whether they felt observed and (d) how much they felt disturbed by observation (see Section 2.7).

2.4. Materials

2.4.1. Test device and data recording hardware

As a mobile phone test device, a black Apple, Inc. iPhone™ 3G was used with 16 GB memory and a touch screen with

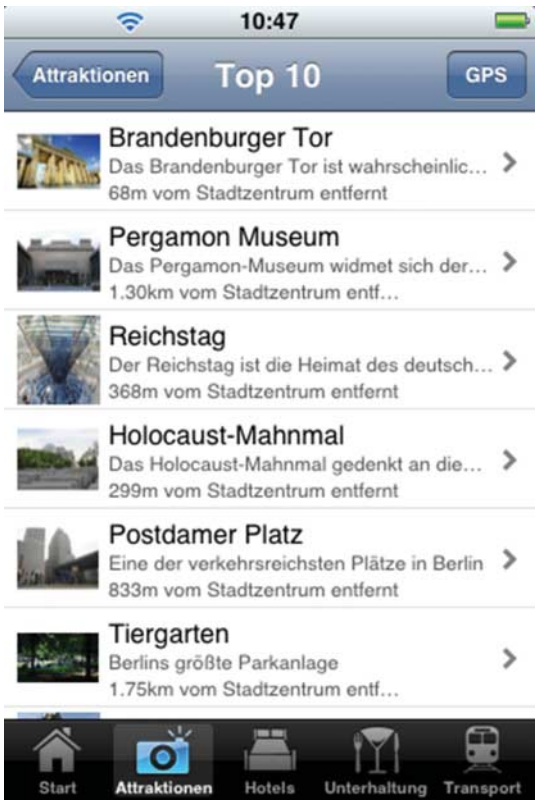

Figure 1. Interface of the city guide.

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

2.4.2. Software

The test device was running on iOS2.2™. Veency™ (v1.0.4) was installed, which allowed for displaying the mobile phone screen directly on a nearby laptop computer wirelessly. On the laptop, TightVNC™ (v1.3.10 for Windows XP™) was installed to support this connection.

As a test system, the cityscouter™ Berlin application (v2.01) for the Apple iPhone™ was used, which is an electronic travel guide for the city of Berlin (Figure 1). The software offers tourist information on Berlin city sights, restaurants, hotels and information on city transport. The application is fully menu driven so that no touchscreen keyboard usage was necessary. All the data necessary for the test tasks were available offline in the application.

As a screen capturing tool on the portable computer, the software CamStudio™ (v2.0) was used to record all mobile phone screen manipulation during the test sessions.

2.5. Tasks

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

2.6. Procedure

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

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

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

Afterwards, the test participant filled in the PANAS questionnaire to measure the emotional baseline and was then guided to the usability laboratory. In the observer-present

Table 1. Measures of user performance as a function of perceived developmental stage, observer presence and TD.

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

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

Then the participant was seated on an office chair at a table and was introduced to the test device, specifically the touch screen, the home button and the pinch-zoom functionality, and the tasks were described. In case of difficulties, the participant was invited to keep working on the tasks until a message would automatically be displayed on the mobile phone screen prompting to pass to the next task (which was manually triggered after 5 min). The tasks were described as all having a solution but showing a variation in difficulty, and the participants were given the opportunity to ask questions.

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

2.7. Manipulation check

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

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

For the manipulation of observer presence, the manipulation check also confirmed the desired effects. On a visual analogue scale, ranging from 'not felt observed at all' to 'felt strongly observed' (0–100), participants reported significantly stronger feelings of being observed (mean 36.5) when observers and test facilitator were present in the same room, compared with when they were working on their own (mean 26.7), $t(76) = 1.72$, $P < 0.05$. There was also an association of age with the level of feeling observed across all conditions, with older participants indicating to have felt less observed than younger ones ($r = -0.26$, $P < 0.05$). When asked about how much they felt disturbed by observation (0–100; ranging from 'not felt disturbed by observation at all' to 'felt strongly disturbed by observation'), they reported feeling significantly more disturbed by observation when observers and test facilitator were present in the room during task accomplishment (mean 21.6) compared with when working alone (mean 13.4), $t(68.5) = 1.72$, $P < 0.05$.

3. RESULTS

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

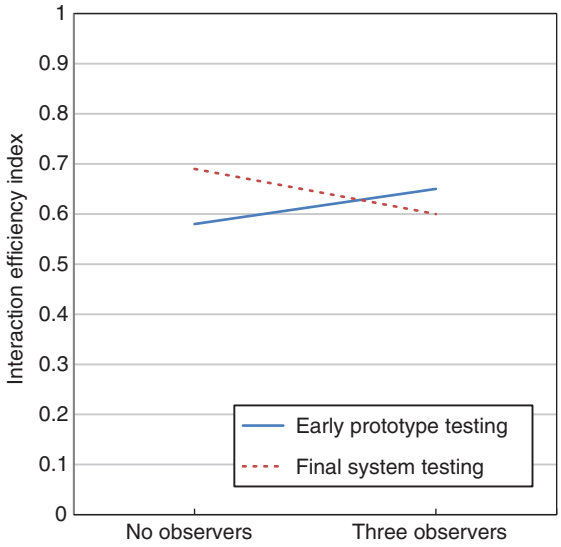

Figure 2. Efficiency of user–product interaction (minimum number of pages to be viewed/number of pages viewed) as a function of testing approach and observer presence.

3.1. User performance

3.1.1. Task completion rate

The performance data are presented in Table 1. The data analysis showed that neither the developmental stage of the system nor the observer presence had a significant impact on the task completion rate (both $F's < 1$). Furthermore, there was no interaction between the two factors ($F < 1$). As expected, task difficulty (TD) had a strong influence on the completion rate, with easy tasks showing a significantly higher completion rate than difficult tasks ($F(1.75) = 79.94$, $P < 0.001$).

3.1.2. Task completion time

As reported above for the task completion rate, there was no significant effect of experimental conditions on the task completion time (cf. Table 1). Neither the developmental stage of the system ($F < 1$) nor the observer presence ($F(1.78) = 1.07$, *ns*), had any significant impact on completion times and also the respective interaction proved non-significant ($F(1.78) = 1.01$, *ns*). As expected, TD determined the task completion time, with easy tasks being accomplished much faster than difficult tasks ($F(1.75) = 438.53$, $P < 0.001$).

3.1.3. Efficiency of user-product interaction

An analysis of the efficiency of task completion revealed overall high levels of efficiency, as indicated by the task efficiency index (minimum number of pages to be viewed/actual number of pages viewed during task completion) presented in Figure 2. As the data show, there were no main effects of developmental stage of the test system or of observer presence (both $F's < 1$). However, the corresponding interaction between the two experimental conditions was significant ($F(1.78) = 4.37$, $P < 0.05$). This was because more efficient performance occurred in the final system testing than in the early prototype condition, when no observers were present ($F(1.38) = 4.48$, $P < 0.05$). While with observers being present, participants in the two developmental stage conditions performed no different ($F < 1$). Finally, an expected main effect of TD occurred, with the participants performing significantly more efficiently on easy (mean 0.71) than on difficult tasks (mean 0.48) ($F(1.65) = 61.59$, $P < 0.001$). No further effects were recorded.

3.2. Subjective ratings

3.2.1. Perceived usability

The data for perceived usability are presented in Table 2. In contrast to our expectations, in the early prototype condition, participants rated the system's usability on the one-item scale significantly less positively (mean 60.5) than in the final system testing condition (mean 71.0) ($F(1.75) = 6.64$, $P < 0.05$). However, for the overall PSSUQ score, there was no such significant effect of developmental stage ($F(1.78) = 2.37$, *ns*).

Observer presence did not have an impact on perceived usability on the one-item scale or for the PSSUQ overall rating (both $F's < 1$). Overall, the correlation (r) between the usability one-item scale and the total PSSUQ score was 0.71 ($P < 0.001$). Interestingly, age showed a significant effect on PSSUQ overall score, with older participants rating the system more negatively ($F(1.74) = 9.78$, $P < 0.01$).

3.2.2. Task load

Task load data are presented in Table 2. There was no effect of the developmental stage ($F < 1$) or of the observer presence ($F(1.78) = 1.23$, *ns*) on experienced task load, and there was no interaction effect for the two conditions ($F < 1$). A separate analysis of the subscales of the NASA-TLX also provided no significant effects.

3.2.3. Emotions

Table 2 presents the data for participants' emotions. Developmental stage had no significant effect on the change of reported positive affect from before to after task completion ($F < 1$), and neither had observer presence ($F(1.78) = 3.06$, *ns*). There was no significant interaction ($F < 1$). Similarly, for negative affect, none of the effects were significant, showing the same pattern of results and are therefore not reported in detail. Interestingly, taking part in the experiment had both an impact on participants' positive and negative affect. Participants reported significantly more positive affect after the test (mean 3.21) than before (mean 2.88) ($F(1.78) = 25.66$, $P < 0.001$). For negative affect, there was also an increase from pre- to post-test measurement with participants reporting higher negative affect after (mean 1.34) than before task completion (mean 1.24) ($F(1.78) = 7.78$, $P < 0.01$).

Table 2. Measures of perceived usability, emotions and mental load as a function of perceived developmental stage and observer presence.

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

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

Table 3. Changes in physiological parameters from baseline to task completion phase, as a function of perceived developmental stage and observer presence.

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

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

3.3. Physiological measures

3.3.1. Heart rate

We compared participants' heart rate changes between resting phase and task completion phase. Overall, there was a significant increase in the heart rate from the resting phase (mean 73.3) to the task completion phase (mean 76.8), which was statistically highly significant ($F(1.72) = 26.62$, $P < 0.001$, see Table 3). This increase differed across observer presence conditions. When observers were present, there was a much stronger increase in the mean heart rate from the resting phase to the task completion phase (+5.37 bpm) than when participants were working on their own (+1.57 bpm) ($F(1.72) = 7.96$, $P < 0.01$). There was no main effect on the developmental stage of the test system ($F(1.72) = 1.08$, *ns*) and no interaction effect ($F(1.72) = 1.33$, *ns*).

3.3.2. Heart rate variability

Using HRV in the LF band (0.04–0.15 Hz) as a sensitive indicator for participants' stress response and mental effort, we compared baseline levels as measured during an initial resting phase with measurements during the task completion phase. To control for outliers, we excluded from the analysis eight participants with LF band values lying outside a range of ±2 standard deviations from the mean. Our analysis showed an impact of observer presence on these difference values (see Table 3). There was a decrease during the task completion phase for the power in the LF band when observers were present, indicating higher stress levels. In contrast, when working alone, participants showed a significant increase in power in the LF band during the task completion phase ($F(1.64) = 4.81$, $P < 0.05$). For the developmental stage of the system ($F < 1$) or the interaction between developmental stage and observer presence, there were no such effects ($F(1.64) = 2.12$, *ns*).

4. DISCUSSION

The primary research question of our study was to investigate the impact of perceived prototype fidelity on central outcome

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

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

While results for the PSSUQ generally did not find any evidence for differences between an early prototype and a final system in subjective evaluation, a different picture emerged for the one-item measure of usability. On this single item, participants judging a final system rated its usability more positively than when judging an early prototype. This was in contrast to our expectation that in the final system condition, users would be less tolerant and have higher expectations towards the finished system, resulting in lower ratings of usability. However, the results for the single-item measure are consistent with the assumption that expectations towards a more favourable evaluation of the test system were higher in the final system testing condition.

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

The results provided evidence that the presence of additional observers in a usability test setting caused higher stress levels among test participants. The participants with observers present during task accomplishment reported to have felt more disturbed by the observation, an effect which was mitigated by age, as older participants indicated that they generally felt less observed than younger ones. When observers were present, participants also showed changes at the physiological level, in the form of decreased HRV and increased heart rate when working on tasks. These results on the physiological level confirm findings of a previous study (Sonderegger and Sauer, 2009), providing further evidence for the HRV in the LF band to be a sensitive indicator for social stress in usability testing and even beyond. However, we could not confirm the previous study's (Sonderegger and Sauer, 2009) findings concerning a negative impact of the observer presence on the performance. Participants with observers present generally performed no worse than those working alone in a room. The question arises why we could not replicate the findings of impaired performance under observation, although we implemented the same conditions and used very similar performance indicators that should have been equally sensitive. One possible explanation for this difference in results between the two studies might be the gender of the observers. In contrast to the previous study, which used an all-male group of observers, we had one female observer. This may have mitigated the effect of observer presence by reducing the evaluative characteristics of the social situation, as suggested by research on gender differences (Leary *et al.*, 1994). This work suggests lower self-presentation concerns of participants in social situations when all interaction partners were female rather than male. Research on gender stereotypes also suggests that females are expected to show behaviours as taking care of the well-being of others in a social context, while men are expected to act task-oriented, even at the expense of the well-being of others (Eagly and Karau, 2002; Leon, 2005). This may have also contributed to a female observer not being perceived as socially threatening as a male observer.

There are implications of our findings for usability practitioners and researchers alike. First, usability test outcomes may be more robust against different instructions given to

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

ACKNOWLEDGEMENTS

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

FUNDING

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

REFERENCES

Bentley, T. (2000) Biasing Web Site User Evaluation: a Study. In Proc. Aust. Conf. Hum.–Comput. Interact. (OZCHI 2000), 4–8 December 2000, Sydney, pp. 130–134. IEEE Computer Society Press.

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

Bond, C.F. and Titus, L.J. (1983) Social facilitation: a meta-analysis of 241 studies. Psychol. Bull., 94, 265–292.

Christophersen, T. and Konradt, U. (2010) Reliability, validity, and sensitivity of a single-item measure of online store usability. Int. J. Hum.–Comput. Stud., 69, 269–280.

Eagly, A.H. and Karau, S.J. (2002) Role congruity theory of prejudice toward female leaders. Psychol. Rev., 109, 573–598.

Gediga, G., Hamborg, K.-C. and Düntsch, I. (2002) Evaluation of Software Systems. In Kent, A. and Williams, J.G. (eds.), Encyclopedia of Computer Science and Technology, vol. 45, pp. 127–153. Marcel Dekker, New York, NY.

Grubaugh, B., Thomas, S. and Weinberg, J. (2005) The Effects of the Testing Environment on User Performance in Software Usability Testing. In Hamza, M.H. (ed.), Proc. IASTED Int. Conf. Hum.–Comput. Interact., 14–16 November 2005, Phoenix, pp. 39–43. ACTA Press, Anaheim, CA.

Guerin, B. (1986) Mere presence effects in humans: a review. J. Exp. Soc. Psychol., 22, 38–77.

Guerin, B. (1993) Social Facilitation. Cambridge University Press, Cambridge.

Harris, E., Weinberg, J., Thomas, S. and Gaeslin, D. (2005) Effects of Social Facilitation and Electronic Monitoring on Usability Testing. In Proc. Usability Prof. Assoc. Conf., 27 June–1 July 2005, Quebec, published on CD.

Hart, S.G. and Staveland, L.E. (1988) Development of NASA-TLX (task load index): Results of Empirical and Theoretical Research. In Hancock, P.A. and Meshkati, N. (eds), Human Mental Workload, pp. 139–183. Elsevier, Amsterdam.

Hartmann, J., De Angeli, A. and Sutcliffe, A. (2008) Framing the User Experience: Information Biases on Website Quality Judgments. In Proc. CHI, 5–10 April 2008 Florence, pp. 855–864. ACM Press, New York, NY.

Hix, D. and Hartson, H.R. (1993) Developing User Interfaces: Ensuring Usability Through Product and Process. Wiley, New York.

Hix, D., Hartson, H.R. and Nielsen, J. (1994) A taxonomy for developing high impact formative usability evaluation methods. SIGCHI Bull., 26, 20–22.

Hornbaek, K. and Law, E.L.-C. (2007) Meta-analysis of Correlations among Usability Measures. In Proc. CHI, 28 April–3 May 2007, San Jose, pp. 617–626. ACM Press, New York, NY.

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

Izsó, L. and Láng, E. (2000) Heart period variability as mental effort monitor in human computer interaction. Behav. Inf. Technol., 19, 297–306.

Jokela, T. (2002) Making User-centred Design Common Sense: Striving for an Unambiguous and Communicative UCD Process Model. In Proc. NordiCHI, 19–23 October 2002, Århus, pp. 19–26. ACM Press, New York, NY.

Jokela, T., Koivumaa, J., Pirkola, J., Salminen, P. and Kantola, N. (2006) Methods for quantitative usability requirements: a case study on the development of the user interface of a mobile phone. Pers. Ubiquitous Comput., 10, 345–355.

Karat, C.-M. (1994) A Business Case Approach to Usability Cost Justification. In Bias, R.G. and Mayhew, D.J. (eds), Cost-justifying Usability, pp. 45–70. Academic Press, New York, NY.

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

Leary, M.R., Nezlek, J.B., Downs, D., Radford-Davenport, J., Martin, J. and McMullen, A. (1994) Self-presentation in everyday interactions: effects of target familiarity and gender composition. J. Pers. Soc. Psychol., 67, 664–673.

Leon, G.R. (2005) Men and women in space. Aviat. Space Environ. Med., 76, 84–88.

Lewis, J.R. (1995) IBM computer usability satisfaction questionnaires: psychometric evaluation and instructions for use. Int. J. Hum.–Comput. Interact., 7, 57–78.

Lewis, J.R. (2002) Psychometric evaluation of the PSSUQ using data from five years of usability studies. Int. J. Hum.–Comput. Interact., 14, 463–488.

Lewis, J.R. (2006) Usability Testing. In Salvendy, G. (ed.), Handbook of Human Factors and Ergonomics, pp. 1275–1316. Wiley, Hoboken, NJ.

Manstead, A.S.R. and Semin, G.R. (1980) Social facilitation effects: mere enhancement of dominant responses? Br. J. Soc. Clin. Psychol., 19, 119–136.

Mantei, M.M. and Teorey, T.J. (1988) Cost/benefit analysis for incorporating human factors in the software lifecycle. Commun. ACM, 31, 428–439.

Mayhew, D. (1999) The Usability Engineering Lifecycle: a Practitioner's Handbook for User Interface Design. Academic Press, San Diego.

McCurdy, M., Connors, C., Pyrzak, G., Kanefsky, B. and Vera, A. (2006) Breaking the Fidelity Barrier. In Proc. CHI, 22–27 April 2006, Montréal, pp. 1233–1242. ACM Press, New York, NY.

Nielsen, J. (1993) Usability Engineering. Academic Press, San Diego.

Pruyn, A., Aasman, J. and Wyers, B. (1985) Social Influences on Mental Processes and Cardiovascular Activity. In Orlebeke, J.F., Mulder, G. and van Doornen, L.F.P. (eds), Psychophysiology of Cardiovascular Control (Models, Methods, and Data), pp. 865–877. Plenum Press, New York, NY.

Raita, E. and Oulasvirta, A. (2011) Too good to be bad: favorable product expectations boost subjective usability ratings. Interact. Comput., 23, 363–371.

Redish, J., Bias, R., Bailey, R., Molich, R., Dumas, J. and Spool, J.M. (2002) Usability in Practice: Formative Usability Evaluations—Evolution and Revolution. In Terveen, L.G. and Wixon, D.R. (eds), Extended Abstracts of the 2002 Conf. Human Factors in Computing Systems, 20–25 April 2002, Minneapolis, pp. 885–890. ACM Press, New York, NY.

Rowe, D.W., Sibert, J. and Irwin, J. (1998) Heart Rate Variability: Indicator of User State as an Aid to Human–Computer Interaction. In Proc. CHI, 18–23 April 1998, Los Angeles, pp. 480–487. ACM Press, New York, NY.

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

Scriven, M. (1967) The Methodology of Evaluation. In Tyler, R., Gagne, R. and Scriven, M. (eds), Perspectives of Curriculum Evaluation, pp. 39–83. Rand McNally, Chicago, IL.

Seffah, A. and Habieb-Mammar, H. (2009) Usability engineering laboratories: limitations and challenges toward a unifying tools/practices environment. Behav. Inf. Technol., 28, 281–291.

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, 1350–1361.

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

Tarvainen, M.P. and Niskanen, J.-P. (2008) Kubios HRV Version 2.0 user's guide [online]. University of Kuopio. http://kubios.uku.fi/ (accessed 17 February 2010).

Tullis, T. and Albert, W. (2008) Measuring the User Experience: Collecting, Analyzing, and Presenting Usability Metrics. Morgan Kaufmann, Burlington, MA.

van den Haak, M., De Jong, M. and Schellens, P.J. (2003) Retrospective vs. concurrent think-aloud protocols: testing the usability of an online library catalogue. Behav. Inf. Technol., 22, 339–351.

Virzi, R.A., Sokolov, J.L. and Karis, D. (1996) Usability Problem Identification Using Both Low- and High-fidelity Prototypes. In Proc. CHI, 13–18 April 1996, Vancouver, pp. 236–243. ACM Press, New York, NY.

Vredenburg, K., Mao, J.-Y., Smith, P.W. and Carey, T. (2002) A Survey of User-centered Design Practice. In Proc. CHI, 20–25 April 2002, Minneapolis, pp. 471–478. ACM Press, New York, NY.

Watson, D., Clark, L.A. and Tellegen, A. (1988) Development and validation of brief measures of positive and negative affect: the PANAS scale. J Pers. Soc. Psychol., 54, 1063–1070.