The influence of laboratory set-up in usability tests: effects on user performance, subjective ratings and physiological measures

Andreas Sonderegger* and Juergen Sauer

Department of Psychology, University of Fribourg, 1700 Fribourg, Switzerland

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 × 2 mixed experimental design, laboratory set-up 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. The findings presented in this paper have implications for the practice of usability testing. They indicated a considerable influence of observers on test participants (physiology and emotions) and on the outcomes of usability tests (performance measures). This should be considered when selecting the set-up of a usability testing procedure.

Keywords: usability test; social facilitation; heart rate variability; usability laboratory; laboratory set-up

1. 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 maximise 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 set-up 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 (Schrier 1992, Salzman and Rivers 1994, Patel and 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.

1.1. *Set-up of usability laboratories*

The set-up 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 set-up; electronic observation room set-up; classic testing laboratory set-up (see Figure1a-c). All set-ups 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 set-ups differ with regard to the number of people who are in the same room as the test participant.

The single-room set-up (see Figure 1a) 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 minimise distraction. In the electronic observation room set-up (see Figure 1b), the test facilitator is still in the same room as the test participants while the non-interactive observers

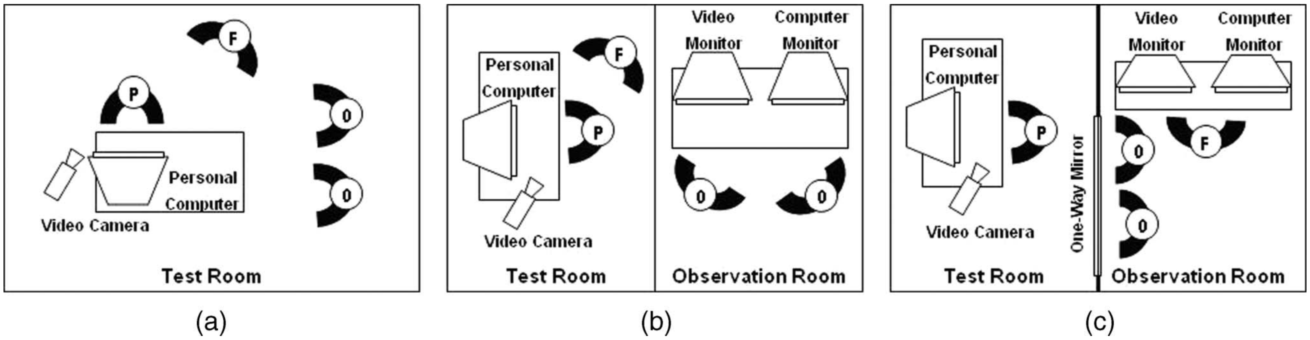

Figure 1. Set-up of usability laboratories (P = Participant; F = Facilitator; O = Observer): (a) Single-room set-up; (b) Classic testing laboratory set-up; (c) Electronic observation room set-up.

are placed in a separate room, allowing them to observe the testing procedure on a closed circuit television screen. In the classic testing laboratory set-up (see Figure 1c), 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 and/or a two-way mirror.

There are various factors to take into account when selecting a particular laboratory set-up. 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 set-ups 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 set-ups referred to above. This should include a range of measures that assess the effects of different set-ups at several levels: physiological response, performance and subjective evaluation. This corresponds to the three levels of workload assessment used in the work domain (Wickens and Hollands 2000).

1.2. Multi-level analysis of test outcomes

1.2.1. 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 *et al.* 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 and Backs 2000), HRV is considered to be a good indicator for mental stress and negatively toned affect (Kettunen and 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 of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology 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 and Cacioppo 2004). The low frequency (LF) band is generally considered to be a suitable measure for mental demands (Boucsein and 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 *et al.* 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 (1996) proposes the computation of the LF/HF power ratio to increase the reliability of physiological measures reflecting psychological phenomena. 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. Nickel and Nachreiner 2003, Berntson and Cacioppo 2004). Despite the ongoing 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 (Task Force of the European Society of

Cardiology and the North American Society of Pacing and Electrophysiology 1996, Nickel and Nachreiner 2003). 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.

1.2.2. 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 1998). 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 that 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 set-ups 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).

1.2.3. Subjective evaluation

1.2.3.1. 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 1998). 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 1998, see also Wickens and 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 (Brooke 1996) to 71-item Questionnaire for User Interface Satisfaction (QUIS; Chin, Diehl and Norman 1988) 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.

1.2.3.2. 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 judgements (Dörner and Stäudel 1990, Forgas and 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 and 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 set-up. 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 *et al.* 2007).

1.2.3.3. 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 and Sonderegger 2009). Furthermore, there is evidence for a positive relationship between product aesthetics and perceived usability of a product (e.g. Tractinsky *et al.* 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.

1.3. The present study

Although there have been indications that the set-up 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 1. 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 hypothesised 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). The present authors are aware that stress responses differ as a function of gender (e.g. Stroud *et al.* 2002). Therefore, an equal number of males and females were assigned 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 hypothesised 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 'laboratory set-up' 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 hypothesised 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, perceived usability, attractiveness and heart rate were also measured (although they were not referred to in any of the hypotheses) to explore their relationship with the manipulated independent variables.

2. Method

2.1. Participants

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

2.2. Experimental design

In a 3 × 2 mixed design, test situation was used as a between-subjects variable, being varied at three levels. According to the different set-ups of usability laboratories described in section 1, the usability tests were conducted either in the single-room set-up (in the following referred to as multi-observer set-up), the classic testing laboratory set-up (i.e. single-observer set-up) or the electronic observation room set-up (i.e. no-observer set-up). As a within-subjects variable, task difficulty was varied at two levels: low and high.

2.3. Measures and instruments

2.3.1. Heart rate data

The heart rate of the participants was continuously recorded during the whole experiment. To measure the effect that 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 of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology (1996), for each phase a minimum recording of 5 min was used for the analysis, excluding the first and last 2 min 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 min was required for the calculation of the HRV data, an analysis of the physiological data on task level was not possible.

2.3.2. User performance

Four measures of user performance were recorded: (a) task completion rate refers to the percentage of participants who were able to complete the task within 5 min; (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 that participants left the optimal dialogue path by more than two clicks (in which case the error message 'wrong path, please go back' was displayed).

2.3.3. *Subjective evaluation*

2.3.3.1. 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 (Kirakowski 1996), SUS) because it was especially developed for usability tests in laboratory settings. On a 7-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 $\alpha > 0.90$) is high.

2.3.3.2. 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 *et al.* 1988) was employed. The German-language questionnaire enjoys good psychometric properties (Cronbach's $\alpha = 0.84$; Krohne *et al.* 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 5-point Likert scale (very slightly or not at all, a little, moderately, quite a bit, extremely).

2.3.3.3. Attractiveness. The attractiveness rating of the mobile phone was made on a one-item 5-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).

2.4. *Materials*

2.4.1. *Heart rate monitor and video camera*

The heart rate was recorded continuously throughout the experiment with a Polar S810i^{TM} heart rate monitor (Polar S810i^{TM}, Kempele, Finland). A video camera (Panasonic^{TM} NV-MS5EG; Panasonic Corp., Kadoma, Japan) was positioned next to the user's workspace.

2.4.2. *Computer prototype*

Based on a SonyEricsson^{TM} SE W800i mobile phone (Sony Ericsson Mobile Communications, London, UK), a computer simulation of the dialogue structure was developed using html and JavaScript. The interaction data were recorded by a PHP-script. The simulation was run on an Apache^{TM} (Apache Software Foundation, Delaware, USA) server (XAMPP) installed on a Toshiba Portege^{TM} M200 TabletPC (Toshiba Corp., Tokyo, Japan) 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.

2.4.3. *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 (nine 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.

2.5. *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 heart rate, the electrode of the Polar T61^{TM} transmitter was moistened and attached to the participant's chest and the Polar S810i^{TM} 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 min 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 min, 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 min after he/she had been seated, which provided sufficient time for physiological adaptation following the physical movement from the relaxation room to the usability laboratory (the two rooms were situated adjacent to each other). In all three laboratory set-ups, the entire testing procedure was videotaped. In the one-observer set-up, 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 set-up, 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 years) 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 set-up, the test facilitator left the room as the testing procedure began and the test participant was alone in the laboratory. There was no two-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.

2.6. Analysis of heart rate data and statistical data

The recorded heart rate data were controlled for eliminating artefacts (as proposed by Berntson and Stowell 1998), using the Polar Precision Performance^{TM} software for automatic and Microsoft Excel^{TM} (Microsoft Corporation, Redmond, WA, USA) 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 *et al*. 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 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.

3. Results

3.1. Physiological measures

3.1.1. 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 set-ups with observers, whereas in the no-observer set-up the power in the LF band increased (see Table 1). An overall difference among the laboratory set-ups was found (F = 3.23; degrees of freedom (df) = 2, 57; $p < 0.05$). Planned contrasts revealed significant differences between multi-observer and no-observer set-up (t = 2.48; df = 38; $p < 0.01$) and between multi-observer and single-observer set-up (t = 1.74; df = 38; $p < 0.05$). These findings indicate increased stress levels for test participants in the presence of non-interactive observers. The comparison between single-observer set-up and no-observer set-up 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 set-ups (F < 1; see Table 1).

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 increase in stress levels compared to the baseline measurement (see Table 1). The analysis revealed that the changes in the LF/HF ratio differed significantly between the

Table 1. Changes in physiological parameters (testing phase compared to baseline in relaxation phase) as a function of laboratory set-up.

	Multi-observer set-up Mean* (SD)	Single-observer set-up Mean* (SD)	No-observer set-up Mean* (SD)
LF power (ms^2)	−149.4 (534.1)	−50.2 (306.1)	+177.4 (371.6)
HF 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)	+9.5 (8.0)	+6.3 (5.2)	+3.7 (4.5)

*Negative values denote a decrease in that parameter.
LF = low frequency; HF = high frequency; bpm = beats per min.

laboratory set-ups (F = 3.41; df = 2, 57; $p < 0.05$). Planned contrasts showed a significant difference between the decrease of LF/HF ratio in the multi-observer set-up and the increase in the no-observer set-up, indicating higher stress levels in the set-up condition with non-interactive observers being present (t = 2.6; df = 57; $p < 0.05$). No significant difference was found among the other conditions (t < 1).

3.1.2. *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 set-up on heart rate was significant (F = 4.01; df = 2, 57; $p < 0.05$). The mean heart rate showed an overall increase from the relaxation phase (mean 73.9 beats per min (bpm)) to the testing phase (mean 80.4 bpm). However, the size of the increase was much higher in the presence of observers (see Table 1). Planned pair contrasts showed that in the multi-observer set-up, heart rate showed a significantly higher increase compared to the baseline than in the no-observer set-up (t = 1.71; df = 38; $p < 0.05$). 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 min 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; $p < 0.01$). There was no significant difference among the groups with regard to the magnitude of the reduction of heart rate 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; $p > 0.05$), 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 of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology 1996).

3.2. *User performance*

3.2.1. *Task completion rate*

The data of the measure of effectiveness are presented in Table 2. The data showed no significant difference among conditions of laboratory set-up (F = 2.01; df = 2, 57; $p > 0.05$). Furthermore, there was no significant interaction of test situation and task difficulty on task completion rate (F = 2.01; df = 2, 57; $p > 0.05$). The main effect of task difficulty on task completion rate was significant (F = 37.9; df = 2, 57; $p < 0.001$), 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 set-up. Test users in this condition were significantly more effective than those in the multi-observer set-up (t = 1.97; df = 38; $p < 0.05$). The other comparisons were not significant.

3.2.2. *Task completion time*

The data of task completion time are presented in Table 2. The analysis revealed a main effect of the test situation on this measure (F = 3.42; df = 2, 57; $p < 0.05$), with users requiring more time in the multi-observer set-up than in the other two set-ups. However, no significant interaction of laboratory set-up and task difficulty on task completion time was found (F < 1). This was in contrast to the predictions of social facilitation theory. Planned comparisons

Table 2. Measures of user behaviour as a function of laboratory set-up and task difficulty (TD).

	Multi-observer set-up Mean (SD)	Single-observer set-up Mean (SD)	No-observer set-up Mean (SD)	Overall Mean (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)	0.86 (0.25)	0.82 (0.27)
High TD	0.23 (0.1)	0.3 (0.21)	0.24 (0.14)	0.26 (0.16)

revealed that, for the easy task, participants needed significantly more time in the multi-observer set-up than in the single-observer set-up (t = 2.68; df = 38; $p < 0.01$) and in the no-observer set-up (t = 2.33; df = 38; $p < 0.05$). For the difficult task, no such differences among laboratory set-ups were found (all planned comparisons: $p > 0.05$). 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; $p < 0.001$).

3.2.3. *Interaction efficiency index*

Considering the impact of laboratory set-up and task difficulty on the efficiency of user interaction (minimum number of clicks required/actual number of clicks), no significant main effect of laboratory set-up (F < 1) as well as no significant interaction of laboratory set-up with task difficulty (F < 1) was found (see Table 2). The main effect of task difficulty was significant (F = 68.1; df = 1, 58; $p < 0.001$), 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 to the efficiency index, detailed results are not reported here.

3.3. ***Subjective user ratings***

3.3.1. *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 3). The inferential statistical analysis (F = 4.39; df = 2, 57; $p < 0.05$) showed an influence of laboratory set-up on positive affect. Participants in the no-observer set-up showed higher positive affect than participants in the two other conditions (multi-observer set-up: t = 2.37; df = 38; $p < 0.01$; single-observer set-up: t = 2.73; df = 38; $p < 0.005$). For negative affect, visual inspection of the data showed a similar effect but the statistical analysis did not confirm a significant effect of laboratory set-up (F = 2.5; df = 2, 57; NS).

3.3.2. *Perceived usability*

The data of the PSSUQ are presented in Table 3. Regarding the influence of the laboratory set-up 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.

3.3.3. *Attractiveness*

Table 3 contains the data of participants' appraisal of the aesthetic appeal of the mobile phone. The calculated ANOVA showed no significant effect of laboratory set-up on the attractiveness rating of the tested mobile phone (F < 1).

4. **Discussion**

The main goal of the present study was to determine how laboratory set-ups 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

Table 3. User ratings of emotions, usability and attractiveness.

	Multi-observer set-up Mean (SD)	Single-observer set-up Mean (SD)	No-observer set-up Mean (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)

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). The researchers 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-practised 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). A general negative effect of observer presence may be assumed, although positive benefits for individual test participants may be possible.

The results of the present study indicate that situational factors, such as the set-up 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 laboratory set-up.

In contrast to measures of performance and emotion, the set-up 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.

Also of interest is the question as 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 set-ups. 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 as 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. Wastell and Newman 1996, Wilson and Sasse 2000, Hancock *et al.* 2002). While previous laboratory-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 (Jorna 1992, Berntson and Cacioppo 2004), 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 (Pruyn *et al.* 1985, Jorna 1992).

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 set-up showed the same results as the no-observer set-up 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 as to what extent product-related effects can be separated from other influences on the different test outcomes (e.g. environmental effects due to poor set-up of usability test). Since the reason for testing is to examine the effects of user–product interaction, additional environmental effects such as laboratory set-up that impinge upon the test results clearly represent undesirable side effects that need to be minimised. 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 as the test participants. This may favour the use of remote usability testing as a new product evaluation method, which has gained in importance in usability practice over recent years (Dray and 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.

Acknowledgement

We are very grateful to Nicole Tschuppert, Amadeus Petrig, Jules Zagyva and Eric Bourquard for their support in completing this study.

References

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*. New York: Blackwell, 57–64.

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

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*. Mahwah, NJ: Lawrence Erlbaum, 3–30.

Brooke, J., 1996. SUS: A quick and dirty usability scale. *In*: P. Jordan, B. Thomas, B. Weerdmeester, and I. McClelland, eds. *Usability evaluation in industry*. London: Taylor & Francis, 189–194.

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*, Washington, DC. New York: ACM Press, 213–218.

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

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

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.

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.

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

Hembree, R., 1988. Correlates, causes, effects, and treatment of test anxiety. *Review of Educational Research*, 58 (1), 47–77.

Jordan, P.W., 1998. *An introduction to 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.

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

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

Kirakowski, J., 1996. The software usability measurement inventory (SUMI): Background and usage. *In*: P. Jordan, B. Thomas, B. Weerdmeester, and I. McClelland, eds. *Usability evaluation in industry*. London: Taylor & Francis, 169–177.

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.

Krohne, H.W., *et al*., 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.

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

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

Lewis, J.R., 2006. Usability Testing. *In*: G. Salvendy, ed. *Handbook of human factors and ergonomics*, 3rd ed. Hoboken, NJ: Wiley, 1275–1316.

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

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

Niskanen, J.-P., *et al*., 2004. Software for advanced HRV analysis. *Computer Methods and Programs in Biomedicine*, 76 (1), 73–81.

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, CA: HFES, 1772–1776.

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)*. New York: Plenum Press, 865–877.

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

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.

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 (4), 670–677.

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, CA: HFES, 1210–1214.

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.

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.

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.

Tractinsky, N., Katz, A.S., and Ikar, D., 2000. What is beautiful is usable. *Interacting with Computers*, 13, 127–145.

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.

Wickens, C.D. and Hollands, J.G., 2000. *Engineering psychology and human performance*. Upper Saddle River, NJ: Prentice-Hall.

Wilson, G.M. and Sasse, A.M., 2000. Investigating the impact of audio degradations on users: Subjective vs. objective assessment methods. *In*: *Proceedings of OZCHI, 4*. Sydney, Australia, 135–142.