Applied Ergonomics 74 (2019) 85–96

Contents lists available at ScienceDirect

Applied Ergonomics

journal homepage: www.elsevier.com/locate/apergo

Extra-laboratorial usability tests: An empirical comparison of remote and classical field testing with lab testing

Juergen Sauer*, Andreas Sonderegger, Klaus Heyden, Jasmin Biller, Julia Klotz, Andreas Uebelbacher

Department of Psychology, University of Fribourg, Fribourg, Switzerland

ARTICLE INFO

Keywords:
Usability test
Remote testing
Field testing
Synchronous testing
Asynchronous testing

ABSTRACT

The present article examined the effects of using different extra-laboratorial testing procedures in usability testing. Three experiments were conducted using different artefacts (website, computer-simulated mobile phone, fully operational smartphone) to compare different methodological approaches in field testing (synchronous and asynchronous remote testing, classical field testing) to lab-based testing under different operational conditions (dual task demands, poor product usability). Typical outcome variables of usability testing were measured, including task completion time, click rate, perceived usability and workload. Overall, the results showed no differences between field and lab-based testing under favourable operational conditions. However, under difficult operational conditions (i.e. dual task demands, poor product usability) differences between field and lab-based testing emerged (corresponding to small and medium effect sizes). The findings showed a complex pattern of effects, suggesting that there was no general advantage of one testing procedure over another.

1. Introduction

1.1. Factors of influence in usability testing

Usability testing is a widely used method in the evaluation of consumer products. While its utility is not generally questioned, there are concerns that various factors may have an undue influence on the outcomes of usability testing (Bevan and Macleod, 1994). This refers to factors that are related to the properties of the product being tested, to the characteristics of the user, to the tasks being selected or to the testing environment being chosen (Lewis, 2006).

A prominent question in usability testing has been concerned with choosing the best location for conducting the test. This has generally centred on the question of whether lab or field testing would be the better option (e.g. Kjeldskov and Stage, 2004). This discussion may be considered part of the more general debate in ergonomics and psychology about the pros and cons of experimental research in the lab and in the field (e.g. Anderson et al., 1999; Dipboye and Flanagan, 1979). In the context of usability testing, the costs incurring from tests are also an important issue (e.g. Kaikkonen et al., 2005), with field testing generally being more costly than lab testing. Choosing the most appropriate testing method also needs to consider the influence of factors such as task demands and product properties. For instance, users in lab-based testing environments using a single task scenario were found to be more responsive to product information than if the typical usage scenario involves the simultaneous completion of more than one task (Sauer and Sonderegger, 2011).

These issues are also addressed in the Four-Factor Framework of Contextual Fidelity (Sauer et al., 2010), which proposes a more formal model of the factors identified by Lewis (2006). It proposes four chief factors that are expected to influence the outcomes of usability testing: user characteristics, product properties, task scenarios and testing environment (see Fig. 1). The last factor is the focus of the present study. The factor testing environment relates to physical features (e.g. size of laboratory) but also to social ones (e.g. observer presence). Both may influence the test outcomes, sometimes in rather complex ways. For example, the direction of the impact of observer presence on user performance may depend on the difficulty of the task scenario (i.e. observer presence increases performance for simple tasks and decreases it for difficult ones), as it would be predicted by social facilitation theory (Zajonc, 1965). Such observer effects may also be caused by the presence of experimenters, which might lead to a 'social desirability bias' on the part of the test participant by behaving and responding in a way the experimenter would appreciate.

* Corresponding author. Department of Psychology, University of Fribourg, Rue de Faucigny 2, 1700, Fribourg, Switzerland.
E-mail address: juergen.sauer@unifr.ch (J. Sauer).

https://doi.org/10.1016/j.apergo.2018.08.011
Received 16 February 2017; Received in revised form 19 April 2018; Accepted 11 August 2018
Available online 18 August 2018
0003-6870/

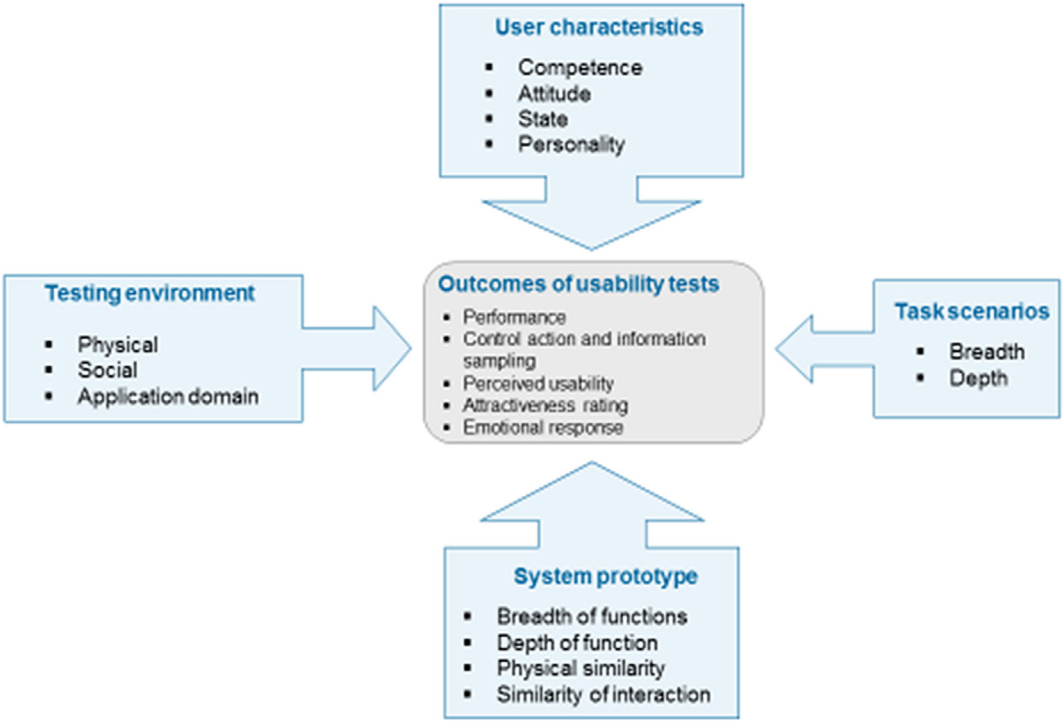

Fig. 1. Four-factor framework of contextual fidelity (adapted from Sauer et al., 2010).

1.2. Methodological issues in usability testing

1.2.1. Testing outside the lab

In usability testing, one may distinguish between different extra-laboratorial approaches to testing: synchronous remote testing, asynchronous remote testing, and classical field testing. They differ from lab-based testing in different ways (see Table 1). For example, remote usability testing may be considered as a particular form of field testing because the test administrator and the user are in different locations during the product evaluation process (Dumas and Fox, 2012). A remote test may be conducted while test administrator and user can communicate via technical means in real time (e.g. telephone, chat) or with a time gap (e.g. email). These two forms are referred to as synchronous and asynchronous remote testing, respectively (e.g. Alghamdi et al., 2013). The testing methods listed in Table 1 differ from each other on a number of criteria, including experimental control, presence of test administrator, presence of environmental distractors, and chosen location. The rating of the criteria was made by the authors of the present article to provide a coarse assessment of the differences between testing methods. The ratings were based on the typical set-up in each testing method. The ratings have revealed considerable differences between the three forms of extra-laboratorial testing on most criteria (e.g. experimental control). However, the results also showed similarities between the lab setting and certain forms of extra-laboratorial testing (e.g. presence of test administrator). In summary, the advantages of lab-based testing are high levels of control over a range of factors such as variations in noise or interruptions during task completion. The different extra-laboratorial testing methods have their advantages in the form of higher levels of ecological validity. Participants in extra-laboratorial testing may complete the tasks in environments, in which they are typically done (e.g. public domain, at home). Furthermore, the two remote testing methods are characterised by lower experimenter presence, which can represent an advantage in certain set-ups (e.g. reduced tendency to fulfil social expectations of the experimenter).

1.2.2. Remote testing

Remote usability testing provides a number of advantages over lab-based testing (e.g. Albert et al., 2010). From a practitioner's point of view, the advantages of remote testing over traditional lab-based testing refer to budgetary savings (e.g. travelling expenses for test participants, renting lab space, access to larger sample size) but also to methodological benefits (e.g. remote testing allows for the evaluation of a system with culturally diverse users). A further advantage would be that testing would be possible in a familiar and less artificial environment (e.g. if home-based testing is required). However, the downside of remote testing is reduced experimental control (e.g. users may be distracted during task completion).

A number of studies have examined the effects of remote usability testing (e.g. Andrzejczak and Liu, 2010; Madathil and Greenstein, 2011). Some work found differences between remote usability testing and lab-based testing, for example, in the form of longer task completion times (e.g. Andrzejczak and Liu, 2010; McFadden et al., 2002; Thompson et al., 2004) and lower task completion rates in remote testing (e.g. Ames and Brush, 2003). However, other work did not find any differences between testing methods (e.g. Madathil and Greenstein,

Table 1
Similarity of lab-based testing to different forms of extra-laboratorial testing.

	Experimental control	Presence of test administrator	Control of environmental distractors	Chosen location
Synchronous remote testing	low	medium	medium	user-selected
Asynchronous remote testing	very low	weak	low	user-selected
Classical field testing	medium	strong	very low	experimenter-selected
Lab testing	high	strong	high	experimenter-selected

2011; Tullis et al., 2002). Most of the work that examined remote testing only collected quantitative data. Only one study measured qualitative outcomes (e.g. the number and severity of usability problems) in addition to quantitative data (Andreasen et al., 2007). This work reports fewer usability problems detected in asynchronous usability testing than in synchronous usability testing. Task completion time was much longer in the asynchronous remote condition than in the lab or the synchronous remote condition, which was attributed to breaks and interruptions by the authors of the study.

1.2.3. Classical field testing

Few studies compared the conventional approach of lab-based usability testing with field testing. Comparing the operation of an electronic mobile guide in the lab and the field, the study found little difference with regard to the usability problems identified (Kjeldskov et al., 2005). In another comparative study, a usability test of a mobile phone was carried out in a lab and a field setting (Kjeldskov and Stage, 2004). While operating the mobile phone, in the field, test participants walked through a pedestrian zone while in the lab, they walked at varying speed on a constantly changing course that had been set up before. For the chief outcome measures, the findings showed no significant differences between testing methods (i.e. performance, perceived workload, and the number of usability problems reported). Another study also aimed to examine possible advantages of field testing for identifying usability problems (Kaikkonen et al., 2005). A mobile phone was tested on a series of 10 typical tasks, but no major differences emerged between the two testing approaches with regard to the usability problems reported. It may be important to note that in the field condition, when participants had to complete a cognitively demanding task, they were often found to withdraw to a quiet spot. This may be indicative of the influence of task demands and environmental distractors in usability testing. Other work compared testing the usability of a mobile phone in the lab and in the field (Duh et al., 2006). When using the phone on a busy commuter train (i.e. field setting), it resulted in more usability problems being identified than in the lab, users reporting higher levels of dissatisfaction and requiring longer to complete the tasks. Very similar results were found in a study comparing lab- and field-based testing of a mobile barcode scanner (Nielsen et al., 2006).

Overall, the research findings for classical field testing and remote testing showed an inconsistent pattern of effects. The review of the literature also revealed that the number and type of outcome variables differed considerably across studies. It is not clear why alternative methods to lab-based usability testing provided such an inconsistent pattern of findings, including positive, negative and nil effects on performance compared to lab-based testing. The inconsistent pattern might be related to differences in task difficulty across studies, as predicted by social facilitation theory, but they may also be due to a general lack of experimental control in all remote testing methods.

1.3. Product properties and task demands

1.3.1. Product properties

The properties of the product to be tested are an important factor in usability testing as proposed by the Four-Factor Framework (Sauer et al., 2010). For example, aesthetics as a prominent factor of influence represents such a product property, which was found to affect the outcomes of usability testing (e.g. Tractinsky et al., 2000). Other examples of product properties are usability and price, which both may be considered indicators of product quality. Usability may be considered a direct indicator and price an indirect indicator of product quality. While the link between usability and product quality appears to be rather straightforward, the link between price and product quality may need explaining. Price is considered an indirect indicator because if users know little about the quality of a product (e.g., at the time of purchasing it), product quality may be primarily judged based on product price (e.g. Kirchler et al., 2010).

A number of studies tested the impact of system usability on different outcome measures. Manipulations in system usability were found to affect user performance (Andreasen et al., 2007) and perceived usability (e.g. Sonderegger et al., 2012). Furthermore, differences in system usability influenced perceived aesthetics (e.g.Tuch et al., 2012; Ben-Bassat et al., 2006), which is surprising because they are conceptually unrelated to system usability.

Price as another important product property seems to be a particular important facet of product value (Cockton, 2004). However, the results of empirical work suggest that there is no uniform way of how price affects user reactions in usability testing, with individual and cultural price sensitivity playing an important role (Sonderegger and Sauer, 2013). This work revealed that some user groups perceived a high-price product to be more usable and to be associated with positive emotions while other user groups perceived exactly the opposite pattern. These unexpected findings may be related to the concept of price fairness since an unfair price can trigger negative emotions (Xia et al., 2004). These (unexpected and complex) findings suggest that price plays an important role but that the way in which it influences outcome measures of a usability test may be difficult to predict.

1.3.2. Task demands

The questions surrounding the effects of dual task completion represents a major research issue in the psychological and ergonomic research literature (e.g. Matthews et al., 2000; Heuer, 1996; Pashler, 1994). The findings generally suggest that dual task completion might lead to performance decrements but whether such a decrement is observed at all and to what extent may depend on many factors (e.g. task difficulty, resource capacity). This issue is relevant in usability testing because the use of many artefacts requires the completion of dual tasks (e.g. using mobile phone while walking in a crowded place).

In the field of usability testing, a direct comparison between single and dual task conditions was only made in a very small number of studies. Kjeldskov and Stage (2004) compared the operation of a mobile phone under task scenarios of different complexity levels. In the single task condition (i.e. participants were seated at a table) subjective workload was lower but more usability problems were recorded than in the more demanding dual task condition (i.e. participants walked while operating the mobile phone). Interestingly, most of the additional usability problems reported in single task condition were considered to be of little importance. This may suggest that too low task demands may lead to a too strong focus of test participants on minor usability problems.

1.4. The present studies

The studies presented aimed to examine a series of methodological factors that are related to the testing environment used in usability tests. This included the following factors: field-based, synchronous and asynchronous remote testing. They were all compared to conventional lab-based testing. The impact of using such testing procedures was examined in combination with pertinent factors of usability testing related to the properties of the product and the task. More specifically, product quality (e.g. usability, price) and task demands (single vs. dual task completion) were examined. The testing approach used in our work corresponds to summative usability testing rather than formative testing since we take quantitative measurements of several pertinent outcome variables (performance, perceived usability, etc.).

In all three experiments, typical tasks had to be accomplished for each device (e.g., finding pieces of information). The samples recruited for each experiment comprised university students who represented a comparatively homogeneous group of users. In each empirical study, several experimental conditions (ranging from 3 to 8) were compared by evaluating their effect on a set of primary outcome factors such as performance, perceived usability and workload. In particular, the

Fig. 2. Starting page of the hospital website'.

distinction between performance and perceived usability is important since the former concept corresponds to an objective usability measure whereas the latter is considered a subjective usability measure (see Hornbæk, 2006; Thielsch et al., 2015). Further variables were measured to respond to the specific needs of each study (e.g. emotion, usability problems reported). This corresponds to a broad definition of usability as, for example, suggested by the theoretical framework presented in Fig. 1. In each experiment, an extra-laboratorial methodological approach was compared to conventional lab-based usability testing.

We generally assumed that in a lab-based testing environment performance would be better than in the different forms of field testing due to fewer distractions, allowing for a stronger task focus. We also expected that the strong presence of a test administrator (e.g. in lab-based testing) would lead to higher usability ratings because users may feel more inclined to meet social expectations.

2. Study I 'remote testing'

2.1. Goals of study

The review of the research literature revealed no studies that made a direct comparison between synchronous and asynchronous remote testing. The first study therefore aimed to make such a direct comparison between two remote testing methods and conventional lab-based usability testing (as a control group) to determine whether typical outcome variables of usability tests would be affected (e.g. performance, perceived usability, workload, emotion, usability problems).

Due to the effects of social expectations (resulting from the strong presence of the test administrator), we predicted that users would give higher usability ratings and report fewer usability problems in the lab environment than in synchronous and asynchronous testing, with the latter showing the lowest ratings and the highest number of usability problems. Furthermore, we expected that users would show better performance in the lab than in synchronous and asynchronous testing because of lower levels of distraction and a social facilitation effect due to stronger administrator presence.

2.2. Method

2.2.1. Participants

The sample of this experiment consisted of 60 students from the University of Fribourg (63% female), aged from 18 to 38 years (M = 22.9). Participants were not paid but they received course credits in return for their participation. An analysis of the sample characteristics with regard to age and internet experience revealed no differences across experimental conditions (for both variables: $F < 1$; $df = 2, 54$; $p > .70$).

2.2.2. Experimental design

A one-factorial between-subjects design was used, manipulating the testing environment at three levels: synchronous remote testing, asynchronous remote testing and conventional lab-based testing.

2.2.3. Measures and instruments

In this experiment, the following measurements were made: task completion rate (%), task completion time (s), interaction efficiency index (required number of clicks divided by number of clicks made), perceived usability (PSSUQ; Lee and Koubek, 2010), subjective workload (NASA-TLX; Hart and Staveland, 1988), positive and negative affect (PANAS; Watson et al., 1988), and number and severity of usability

Table 2
Effects of testing method.

	Lab-based testing		Synchronous remote testing		Asynchronous remote testing	
	M	(SD)	M	(SD)	M	(SD)
Performance						
Task completion time (s)	114	(34)	136	(33)	133	(42)
Task completion rate (%)	41.4	(15.1)	45.9	(21.9)	42.1	(18.1)
Interaction efficiency index (0–1)	.61	(.17)	.62	(.20)	.59	(.20)
Subjective usability						
Perceived usability (1–7)	4.23	(1.04)	4.05	(1.20)	4.08	(.97)
No of problems reported	4.71	(2.29)	5.70	(2.11)	4.95	(1.55)
Severity of problems reported (1–3)	2.33	(.44)	2.10	(.36)	2.37	(.45)
Subjective workload						
NASA-TLX (1–20)	10.70	(2.29)	9.86	(2.89)	9.84	(2.93)
Affect						
Positive affect: post-test (1–5)	2.39	(.60)	2.59	(.63)	2.59	(.64)
Negative affect: post-test (1–5)	1.42	(.29)	1.51	(.52)	1.38	(.28)
Perception of testing situation						
Feeling observed (1–100)	55.2	(27.9)	38.4	(33.4)	41.7	(37.0)
Feeling disturbed (1–100)	31.0	(28.8)	24.3	(27.3)	31.5	(32.3)
Violation of privacy (1–100)	15.0	(20.0)	16.1	(23.8)	28.6	(28.5)

problems reported. In addition to these commonly applied measures in usability testing, we used three purpose-built items to capture the user's perception of the testing situation. Using visual analogue scales (1–100), the items were worded as follows: (a) To what extent did you feel being observed during task completion (not at all – a great deal)? (b) To what extent did you feel disturbed by being observed (not at all – a great deal)? (c) To what extent did you feel that your privacy was violated during task completion (not at all – a great deal)?

2.2.4. Materials

The study used the website of a regional hospital in Switzerland (see Fig. 2), which had been online for some time. The contents of the website did not change during the course of the experiment.

2.2.5. Tasks

Participants were asked to complete six tasks on the website. In each task, participants had to find information on the website (e.g. contact details, information on medical treatments, fees for sports participation) and enter the required piece of information into a text field. Participants were requested not to use the integrated search function of the website, as this would have made the task too simple. The maximum time for task completion was set to 3 min. The task was aborted automatically if not completed within the set time.

2.2.6. Procedure

Participants were mainly recruited by an email sent out to all students of the University of Fribourg. After agreeing to take part in the study, participants were assigned to one of the three experimental conditions and informed about their respective testing procedure (lab-based testing, asynchronous remote testing or synchronous remote testing).

In the lab-based condition, participants were welcomed by the experimenter who then explained the purpose of the experiment. Participants were informed that the current website of a hospital was to be tested for usability in order to receive some feedback on how to improve the website. It was checked whether participants had visited this particular website before (none of them had). Participants were then asked to conduct the test following online instructions. During the course of the experiment, the experimenter was sitting next to the participant, observing him/her using the webpage. The experiment began with a short tutorial on how to report usability problems that may be observed. This was followed by pre-test measurements of positive and negative affect (PANAS). To start each task, participants clicked on a link that opened the website of the hospital in a separate window. After completing a task, participants were asked to report all usability problems identified by entering them into a text field. Afterwards, they evaluated the severity of each problem reported. This was followed by the completion of the PANAS to measure positive and negative affect after the test. Finally, the following instruments were administered: PSSUQ for measuring perceived usability, NASA-TLX for measuring subjective workload, and the three items for measuring the perception of the testing situation. During task completion, an on-site data logger recorded the participant's screen.

In the synchronous remote testing condition, participants were invited by email to join the TeamViewer session, which allowed the experimenter to see and record the participant's screen during the course of the experiment. The experimenter used the chat function to communicate with the participant (e.g. to send the link to the online questionnaire). The instructions given to the participant via the chat function were identical in content to the ones given orally in the lab condition.

In the asynchronous condition, participants received an email containing all instructions (including the link to the online questionnaires). This also included an instruction of how to download a tool that allowed on-site recording of the computer screen during the experiment. The recording was subsequently sent to the experimenter for further analysis. All other elements of the procedure were identical to the other two conditions.

2.2.7. Data analysis

A one-way analysis of covariance was carried out to analyse the main effect of testing method. The covariates used were age, sex and internet experience. This was because previous research showed that such variables could influence the outcomes of usability tests (e.g. Sonderegger et al., 2016). If needed, post-hoc LSD tests were used to determine possible differences between single experimental conditions.

The data for positive and negative affect were analysed using analysis of covariance. Prior to the testing session, a baseline measurement of each type of affective state was taken, which was then used as a covariate. This procedure takes into account the considerable variance in affective states prior to the usability test but avoids the problems associated with the use of gain scores (e.g. Dimitrov and Rumrill, 2003).

2.3. Results

Performance. The data for the three performance measures are presented in Table 2. It showed that there was no effect of testing method on any of the three measures. Analysis of variance confirmed that there

was no significant difference between means for task completion rate ($F = 0.23$; $df = 2, 54$; $p = .793$; $\eta^2_{partial} = .009$), task completion time ($F = 2.22$; $df = 2, 52$; $p = .119$; $\eta^2_{partial} = .079$) and interaction efficiency ($F = 0.34$; $df = 2, 52$; $p = .711$; $\eta^2_{partial} = .013$).

Subjective usability. Table 2 presents the results of the subjective evaluation of usability. The analysis of the PSSUQ ratings revealed no difference between conditions ($F = 0.23$; $df = 2, 54$; $p = .793$; $\eta^2_{partial} = .009$). The qualitative analysis of usability problems reported showed an average of about five problems identified per participant. There was no difference between conditions with regard to the number of problems being reported ($F = 0.81$; $df = 2, 54$; $p = .450$; $\eta^2_{partial} = .029$). When examining the severity ratings of the usability problem (which showed rather high ratings), no effect of testing method was found ($F = 2.05$; $df = 2, 54$; $p = .139$; $\eta^2_{partial} = .070$).

Perceived workload. The levels of subjective workload measured by NASA-TLX are shown in Table 2. The results showed no difference between experimental conditions, which was confirmed by analysis of variance ($F = 0.74$; $df = 2, 54$; $p = .482$; $\eta^2_{partial} = .027$).

Affect. For positive and negative affect, no differences between testing methods were found (see Table 2). Analysis of variance confirmed the absence of such an effect for positive affect ($F = 0.01$; $df = 2, 53$; $p = .986$; $\eta^2_{partial} = .001$) as well as for negative affect ($F = 1.38$; $df = 2, 53$; $p = .261$; $\eta^2_{partial} = .049$). The covariate 'baseline measurement' was related with the post-experimental measurement. It showed a significant effect for post-experimental positive affect ($F = 45.40$; $df = 1, 53$; $p = .000$; $\eta^2_{partial} = .461$) and for post-experimental negative affect ($F = 4.76$; $df = 1, 53$; $p = .034$; $\eta^2_{partial} = .082$).

Perception of testing situation. For the three items (see Table 2), visual inspection of the means seems to suggest that there was an effect of testing method. However, analyses of variance showed that for none of the measures a significant difference between conditions was found ('feeling observed': $F = 1.70$; $df = 2, 54$; $p = .193$; $\eta^2_{partial} = .059$; 'feeling disturbed': $F = 0.05$; $df = 2, 54$; $p = .947$; $\eta^2_{partial} = .002$; 'violation of privacy': $F = 1.46$; $df = 2, 54$; $p = .242$; $\eta^2_{partial} = .051$). Interestingly, the ratings differed considerably between items. Feeling observed had the highest rating ($M = 45.1$), followed by 'feeling disturbed' ($M = 28.9$), and 'violation of privacy' ($M = 19.9$).

2.4. Discussion

In contrast to our expectations, we did not find any effects of testing method, with different forms of remote testing providing neither an advantage nor a disadvantage. This finding adds further support to the small body of research that did not find any effects of remote testing methods on test outcomes (e.g. Madathil and Greenstein, 2011), suggesting that it does not matter which of the three methodological approaches will be chosen. To gain a better understanding of why no differences were found, the analysis of the data on user perceptions of the testing situation (e.g. feeling of being observed) may provide some hints. The factors related to the perception of the testing situation may potentially have a considerable influence on the testing outcomes. However, in the present case user perception did not differ as a function of testing method. There was no main effect. Several reasons may account for this. Since in each condition there was some kind of observation (i.e. test administrator was present during lab testing, screen recordings were made in remote testing that could be seen by the administrator), users may not have experienced a strong difference between human and computer-based 'social control'. Overall, this kind of social control was not perceived negatively, as shown by the ratings recorded for all three items that measured how participants perceived the testing situation. Although participants felt observed to some degree (medium score), they did not feel disturbed by the observation (low score) or experienced a violation of privacy (very low score). The more critical the item appeared to be, the lower the score was. These ratings may have also been generally low because of ample prior experience of being observed and their behaviour being recorded, as this has increasingly become a part of many people's lives (e.g., smartphone recordings, closed-circuit TV cameras and video conferencing).

It is important to note that in the present study, the implication may not only be relevant to summative usability testing but also to formative testing. Two measures typically used in formative usability testing provided converging evidence that the use of any of the three methodological approaches would not result in noteworthy effects on outcome measures. This is remarkable since the present study adopted a very broad approach to measuring the outcomes of usability testing (including usability problems reported by users and a wide range of subjective ratings), which is different from the majority of studies found in the research literature.

While the findings of the present study provided no evidence for any advantages of certain methodological approaches, it remains to be determined whether they can be generalised to other artefacts. Furthermore, it needs to be seen whether the presence of moderating variables (e.g. poor product usability) would provide a different pattern of findings.

3. Study II 'remote testing and product properties'

3.1. Goals of study

There were no effects of the synchronous and asynchronous remote testing approaches in the first study. In the second study, we aimed to compare asynchronous remote testing (i.e. the form of remote testing that is most different from lab testing with regard to the exchange between test administrator and user) to lab-based testing. The experimental set-up was slightly modified to see whether the effects from the first study could be replicated under different circumstances. First, a different artefact was used (smartphone rather than a website). Second, the artefact appeared to be in an early stage of the product development process rather than in the final stage (computer-simulated smartphone rather than an operational website). Third, the artefact was tested by using different levels of product quality (e.g. smartphone with poor usability rather than a website with good usability). In order to achieve this, we employed product quality as a subsidiary independent factor in the form of product usability (i.e. a direct indicator of product quality) and product price (i.e. an indirect indicator of product quality).

We hypothesised that users would show better performance in the lab environment than during asynchronous remote testing because of lower levels of distraction and a social facilitation effect due to stronger test administrator presence. We also predicted that due to social expectations users would give higher usability ratings in the laboratory than during asynchronous remote testing. We also expected the effects of testing method (described in the first hypothesis) to be more pronounced when usability was high than when it was low.

3.2. Method

3.2.1. Participants

Eighty student participants took part in the study (70.0% were female). Their ages ranged from 18 to 39 years ($M = 23.1$, $SD = 4.11$). All students were from the University of Fribourg, reading a range of subjects including psychology, law, education and geography. In return for their participation, they were given cinema vouchers as compensation.

An analysis of the sample characteristics across experimental conditions with regard to age and smartphone expertise revealed few differences. We found only a difference in age between lab and remote conditions ($M_{lab} = 24.1$, $SD_{lab} = 3.66$; $M_{remote} = 22.1$, $SD_{remote} = 4.34$; $F = 4.69$; $df = 1, 72$; $p < .05$) but not for any of the other conditions. No difference was found for smartphone expertise.

3.2.2. Design

A 2×2 x 2 factorial between-subjects design has been

implemented. The main independent factor *testing method* was manipulated at two levels: lab-based testing versus (asynchronous) remote testing. The subsidiary independent factor 'product quality' was divided into two subfactors: product usability (low and high, as a direct indicator of product quality) and product price (low and high, as an indirect indicator of product quality).

Product usability was manipulated by adapting the respective wording of the different icon functions of the interface as well as the menu structure of the navigation. In the high usability prototype, icons were precisely labelled according to their function (e.g. *costs* being used as a label for the function providing information about communication expenses) and navigation paths were short and direct (e.g. costs → running/current costs → total expenses). In the low usability condition, labels were rather generic (e.g. *services* instead of *costs*) and the menu structure was complicated (e.g. services → cost information → invoices → current costs → total expenses).

Price as the second independent variable was manipulated by offering the products with different pricing information. In the low-price condition, the price of the appliance was CHF 99 (about € 91) and, in the high-price condition, it was CHF 699 (about € 643).

3.2.3. Measures and instruments

In this experiment, the following measurements were taken: Task completion time (s), number of clicks, perceived usability (PSSUQ), and subjective workload (NASA-TLX). All these measurements were also taken in the first study, where they were described in detail.

3.2.4. Materials

A computer-simulated prototype of a smartphone, developed by Hamborg et al. (2014), was used in the study (see Fig. 3). This prototype was purpose-built for research and was not modelled on an existing smartphone. It had the advantage that familiarity with a specific smartphone did not provide any benefits in the experiment. Of the smartphone icons shown on the screen, three were fully activated so that a multi-level human-artefact dialogue could develop. If the user clicked on an icon that was not activated, an error message emerged.

3.2.5. Tasks

Participants were asked to complete the following three tasks: (a) to enter the contact details of a friend into the address book, (b) to find out about the telephone charges incurred, and (c) to select a new ringing tone. All of the tasks represented typical tasks of smartphone users.

3.2.6. Procedure

Participants were mainly recruited by an email that was sent out to all students of the University of Fribourg. When agreeing to take part in the study, participants were assigned to one of the eight experimental conditions.

In the lab-based condition, participants were welcomed by the experimenter who then explained the purpose of the experiment. Participants were informed that the development of a new smartphone (which still was at an early stage of the product development cycle) was to be supported by future users by taking part in a usability test and providing feedback about the test results to the product development team. The experimenter told the participant that the smartphone would be launched in the 'low-budget' or luxury segment of the market. Participants were then asked to use the computer-simulated smartphone to complete three tasks (see above). This was followed by the completion of PSSUQ to measure usability and NASA-TLX to measure workload.

In the remote testing condition, participants logged onto the 'Unipark' server (following a link that was sent to them by email). They were then given exactly the same information in writing as participants in the laboratory condition received from the experimenter.

3.2.7. Data analysis

A three-way analysis of variance was carried out. We analysed the main effect of testing method as the primary independent variable to see whether asynchronous remote testing differed from traditional lab testing. Since we were also interested in whether this effect was modified by product usability and price (as two indicators of product quality), we also analysed the interaction of these two factors with testing method. A significant interaction was followed up by post-hoc LSD-tests to determine which cell means were different from each other. However, we will not report any main effects of the two indicators of product quality or any interaction between the two since these factors of product usability have not been the focus of the study.

Two manipulation checks were carried out. We tested whether the manipulation of product usability actually led to the expected different user perceptions of usability. As expected, users rated the 'high

Fig. 3. Main menu (left) and an example of a sub-menu (right) of smartphone.

Table 3
Effects of testing method (the means marked with the superscripts[a] and[b] were found to be significantly different from each other).

	Lab-based testing		Remote testing	
	M	(SD)	M	(SD)
Task completion time (s)	85.5	(65.2)	58.3	(28.6)
Good usability	38.5	(8.8)	32.7	(6.0)
Poor usability	132.3	(63.5)[a]	84.0	(16.1)[b]
Efficiency (number of clicks)	33.3	(19.2)	26.7	(8.6)
Good usability	19.7	(3.8)	19.2	(4.3)
Poor usability	46.9	(8.7)[a]	34.2	(3.9)[b]
Perceived usability (1–7)	5.22	(0.87)[a]	4.82	(1.16)[b]
Good usability	5.80	(0.40)	5.37	(0.97)
Poor usability	4.64	(0.84)	4.26	(1.07)
Perceived workload (1–20)	5.02	(2.24)	5.0	(2.33)
Good usability	3.90	(1.51)	4.24	(2.47)
Poor usability	6.13	(2.32)	5.72	(1.99)

usability'-appliance as more usable than users operating the 'poor usability'-appliance ($M_{high} = 5.59$, $SD_{high} = 0.97$; $M_{low} = 4.45$, $SD_{low} = 0.76$; $t = 5.79$, $df = 78$, $p < .001$; Cohen's $d = 1.308$). We also tested whether different product prices given to the user actually led to difference in price perceptions. As expected, the analysis showed that price perception between the two groups differed significantly ($M_{low} = 2.0$, $SD_{low} = 1.26$; $M_{high} = 4.72$, $SD_{high} = 0.55$; $t = 12.5$, $df = 78$, $p < .001$; Cohen's $d = 2.798$). The results showed that both manipulations were successfully implemented.

3.3. Results

Task completion time. Analysis of variance revealed that task completion time was significantly higher in the lab than during the remote testing condition (see Table 3; $F = 13.0$; $df = 1, 72$; $p < .001$; $\eta^2_{partial} = .152$). However, more important was the significant interaction between testing method and usability ($F = 8.01$; $df = 1, 72$; $p = .006$; $\eta^2_{partial} = .100$). It shows that there was little difference between testing methods when usability was good (post-hoc LSD-test: $p = .585$). In contrast, when usability was poor, users needed more time in the lab than during remote testing (post-hoc LSD-test: $p < .001$). None of the interactions involving price were significant (all $F < 1$). It was also remarkable that the standard deviation was much higher in the lab condition with poor usability than in the three other conditions, with 6 out of 20 users showing task completion times of 2.5–5 min.

Efficiency. The number of clicks made was used as an efficiency indicator and provided exactly the same pattern of results as task completion time (see Table 3). Analysis of variance revealed a significantly higher click frequency in the lab than during the remote testing condition ($F = 8.34$; $df = 1, 72$; $p < .005$; $\eta^2_{partial} = .104$). However, more important was the interaction between testing method and usability ($F = 7.24$; $df = 1, 72$; $p = .009$; $\eta^2_{partial} = .091$). Again, there was no difference between lab and remote testing when usability was good (post-hoc LSD test: $p = .889$). However, when usability was poor, users in the lab needed more clicks than during the remote testing condition (post-hoc LSD test: $p < .001$). None of the interactions involving price were significant (all $F < 1$).

Perceived usability (PSSUQ). The analysis showed that usability was rated higher when users were in the lab involving the presence of a test administrator than during remote testing (see Table 3). This difference was statistically significant ($F = 4.24$; $df = 1, 72$; $p = .043$; $\eta^2_{partial} = .055$). No significant interaction was observed (all $F < 1$).

Workload (NASA-TLX). Ratings of workload did not differ between the two testing methods examined (see Table 3). The analysis of variance did not reveal any significant effects for any of the four tests that had been carried out (all $F < 1$).

3.4. Discussion

The study revealed two important findings: the complex interaction of product properties and testing method, and the influence of social expectations in the laboratory setting. Both findings may be related to social processes.

The results showed that there was no difference in performance between remote and laboratory conditions providing that product usability was good. This corresponded to the findings of the previous study, where the artefact also enjoyed good usability. Conversely, when product usability was poor, users showed worse performance in the laboratory than during remote testing. The interaction between testing method and product usability was unexpected and defied any straightforward explanation. It is conceivable that a social inhibition effect (e.g., Zajonc, 1965; Geen, 1991) may be at the root of this finding. When faced with poor product usability, the task was difficult to complete by users in either testing method (requiring longer completion times and more clicks). The social inhibition effect predicts that performance on more difficult tasks is impaired during the presence of others (here: test administrator). This may have resulted in the additional performance deterioration in the laboratory condition under the presence of a test administrator. This issue is related to the question of whether there should be observers present during usability testing (e.g. members of the product development team). Research has shown that the presence of observers in usability tests may have an effect on psychophysiological parameters such as decreased heart rate variability (Sonderegger and Sauer, 2009), which is commonly considered an indicator of strain (Kaikkonen et al., 2005). The difficulty of a task largely determines whether social inhibition or social facilitation takes effect in the presence of others.

The results for perceived usability appear to be somewhat at odds with those for performance. Usability ratings were higher when users were tested in a traditional lab environment than in asynchronous remote testing whereas no such effect of lab testing was found for performance. On the contrary, participants faced more difficulties in using the device (i.e. they showed poorer performance) during the presence of a test administrator when product usability was poor. The interesting effect of testing method on perceived usability may also be related to social expectations. Users are sometimes hesitant about giving negative feedback directly to the test administrator in which the quality of a technical device is questioned since they fear that such criticism may lead to disappointment among members of the product development team who expended considerable effort into the design of the device. Due to the physical presence of the test administrator in the lab condition, social pressure and expectations of this kind may gain in influence compared to remote testing conditions involving no direct contact between user and test administrator. This confirmed the hypothesis of the present study but the finding was inconsistent with the results of the first study. These inconsistent findings may be related to the level of test administrator presence in the two studies. Test administrator presence was higher in the remote conditions of the first study because subjects would contact the experimenter (via video link or email), whereas there was no such possibility in the remote condition of study II. The issue of experimenter presence will be revisited in the general discussion.

In contrast to the influence of product usability as a direct indicator of product quality, price as an indirect indicator of product quality did not show such an effect. This may be because unlike poor usability, low price has no direct impact on task difficulty. The impact of price may be strongest if the user has not yet had much direct experience with the product. For example, during product purchase the user relies on price for quality assessment because there is little other information available to assess usability. With increasing duration of actual user-artefact interaction, any such influence is expected to wane. The present study indicated that even a very short interaction period was sufficient to reduce its influence as an indicator of quality. Product usability and price represent two aspects of an important contextual influencing

factor in usability testing (i.e. product quality). Whereas product price showed no effects as a function of testing method (i.e. lab vs. remote testing), such an effect was found for product usability. A reason for this difference in the pattern of effects between the two product-inherent factors might be that usability (but not product price) is linked to the difficulty participants encounter when completing a specific task. Since task characteristics are also considered as an influencing factor in the Four Factor Framework of Contextual Fidelity, it was addressed in our third study.

4. Study III 'classical field testing and dual task demands'

4.1. Goals of study

The second study showed that differences between remote and lab-based testing only emerged under difficult operational conditions, which were characterised by increased task difficulty. In the third study, we aimed to evaluate the effects of classical field testing (involving direct contact between administrator and user). The field testing scenario consisted of a rich social environment in the form of a public space (i.e. cafeteria with visual and auditory distractors) and was compared to a conventional lab-based environment. In addition, we manipulated task demands by using the dual task paradigm. Dual task scenarios represent difficult operational conditions in a similar way as poor usability does. As the previous study showed that task difficulty can mediate the influence of testing method, we aimed to examine whether similar effects will be observed under dual task demands.

We predicted a main effect of testing method with performance being better in the laboratory than in the field due to fewer distractions in the former. We also expected that this difference between testing methods would be larger under dual task completion than during single task completion. Furthermore, during field testing we expected ratings of workload to be higher and ratings of usability to be lower than in the laboratory. This was because the difficulties faced in the more challenging testing method (requiring more cognitive resources and leading to performance decrements) would be partly attributed to the poor usability of the device.

4.2. Method

4.2.1. Participants

Sixty-two students (55.9% female) from the University of Fribourg participated in the study. They were aged between 18 and 36 years (M = 22.8). Participants were offered course credits in return for their participation. An analysis of the sample characteristics with regard to age revealed no differences across experimental conditions (all $F > 4.69$; $df = 1, 55$; $p > .20$).

4.2.2. Design

A 2 × 2 between-subjects design was employed in the experiment. The independent variable *testing method* was varied at two levels: classical field setting vs. lab setting. The independent variable *task demand* was also varied at two levels: single task demands vs. dual task demands.

4.2.3. Measures and instruments

In this experiment, the following measurements were taken: Task completion time (s), index of interaction efficiency (i.e. minimum number of user inputs needed to complete the task divided by the number of user inputs actually made), perceived usability (PSSUQ), perceived workload (NASA-TLX), and affect (PSSUQ). For details about the measurements and instruments used, the reader is referred to section 2.2.1.

4.2.4. Materials

The device used in the study was an Apple iPhone 3G. This smartphone was running with the operating system MAC iOS x with a screen resolution of 320 × 480 pixels. The participants used for the task an application, called "London CityScouter" (see Fig. 4). This was an electronic city guide, allowing its user to find restaurants, cafes and other venues for entertainment in London. All other functions of the smartphone were disabled during the testing period.

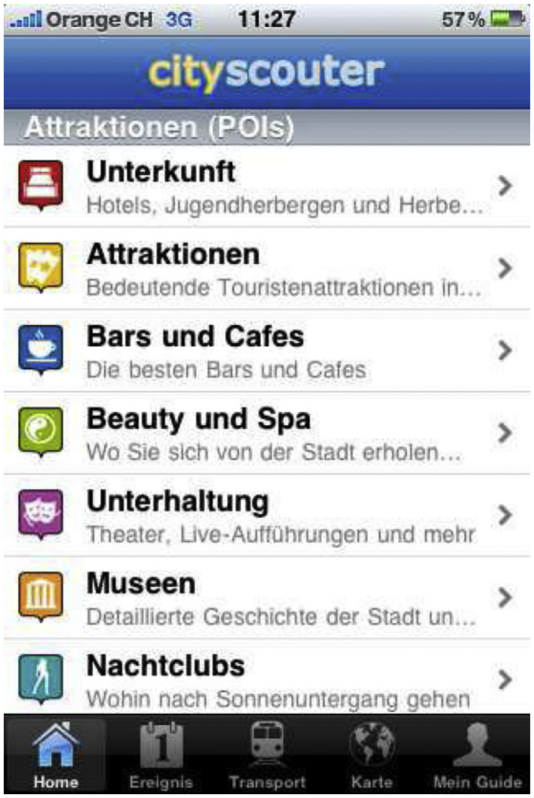

Fig. 4. Starting screen of application 'London CityScouter'.

4.2.5. Tasks

Participants had to complete four tasks that were embedded in a short scenario such as the following: "You would like to go out in London tonight. A friend recommended to you a club called 'Fridge'. You would like to find out what the entrance fee is. Please put down how much it is to get in".

The following four tasks were to be completed with the application 'London City Scouter': (a) Find out the entrance fee charged by the club 'Fridge'! (b) Find out the opening hours of the Churchill Museum! (c) Find out the address and phone number of a restaurant with a name ending ' … de Portugal'! (d) Find out how to get to a French restaurant specialising in regional cuisine from the Burgundy!

4.2.6. Procedure

The testing session took either place in a usability laboratory or in a university cafeteria. In the university cafeteria, students were always tested in the same location. Since noise levels varied in the cafeteria at different times of day, we measured ambient noise by using a sound level meter. This was to check whether noise had an influence on performance.

In the secondary task conditions, participants were asked to complete an additional task involving simple mental arithmetic (e.g. 2 + 7 = ?). This secondary task was transmitted by using the chat function of skype. Participants were told that in the scenario presented they would help a younger cousin with her homework.

4.2.7. Data analysis

A two-way analysis of variance was carried out. The main effect of testing method was analysed as the primary independent variable to determine whether classical field testing differed from traditional lab testing. Since it was also of interest whether this effect was modified by dual task demands, we also analysed the interaction of these two factors. A significant interaction was followed up by post-hoc LSD-tests to determine which cell means were different from each other. Since the factor task demand was not of interest in itself, we will not report any main effects of this factor. The data for positive and negative affective state were analysed by means of analysis of covariance, using a baseline measurement of each type of affective state (i.e. taken prior to the testing session) as a covariate.

The measurement of sound pressure levels revealed that ambient noise was higher in the field than in the lab (M_{field} = 70 dB; M_{lab} = 38 dB). A correlational analysis showed that noise was not related to any of the performance measures. Therefore, noise was not used as a covariate.

4.3. Results

Task completion time. The data in Table 4 showed that time needed to complete the task was unaffected by the testing method (F = 0.50; df = 1, 55; p = .48; $\eta^2_{partial}$ = .009). Surprisingly, there was a significant cross-over interaction between testing method and task demands, with the highest score being recorded in the condition 'lab/single task' and the lowest one for 'field/single task' (F = 6.40; df = 1, 55; p = .014; $\eta^2_{partial}$ = .104). Fisher's LSD-tests showed that the condition 'field/single task' was significantly different from the conditions 'lab/single task' and 'field/dual task' (both p < .05).

Task completion rate. A similar pattern was observed for task completion rate (see Table 4). Again, it showed no effect of testing method (F = 0.27; df = 1, 55; p = .60; $\eta^2_{partial}$ = .004). A significant cross-over interaction between testing method and task demands was found, with the worst performance being recorded in the condition 'field/dual task' and the best one for 'lab/dual task' (F = 5.72; df = 1,55; p = .02; $\eta^2_{partial}$ = .094). A post-doc LSD-test confirmed the two conditions to be significantly different from one another (p < .05) but none of the others.

Interaction efficiency. The interaction efficiency index (i.e. minimum number of user inputs needed to complete the task divided by the number of user inputs actually made) examined the number of clicks needed to complete the task (see Table 4). Analysis of variance revealed no difference between testing locations (F = 0.72; df = 1, 55; p = .39; $\eta^2_{partial}$ = .012) and there was no interaction (F = 1.05; df = 1, 55; p = .31; $\eta^2_{partial}$ = .018).

Perceived usability. User ratings of usability are presented in Table 4. While they did not differ much between testing methods overall (F = 1.15; df = 1, 58; p = .28; $\eta^2_{partial}$ = .019), analysis of variance revealed a significant interaction (F = 4.04; df = 1, 58; p = .048; $\eta^2_{partial}$ = .065). The interaction reflected the same pattern observed for task completion rate, with the conditions 'lab/dual task' showing the highest usability rating while the lowest was recorded for 'field/dual task'. Fisher's LSD-test confirmed that these two conditions were significantly different from one another (p < .05) but none of the others.

Subjective workload. Ratings of workload showed no difference between experimental conditions (see Table 4). This was confirmed by analysis of variance for the main effect (F = 0.17; df = 1,58; p = .67; $\eta^2_{partial}$ = .003) as well as for the interaction (F = 3.51; df = 1, 58; p = .066; $\eta^2_{partial}$ = .057).

Affective state. For this measure, an analysis of covariance was carried out, with measurements of positive and negative affect prior to the test serving as the respective covariates. The affective state data presented in Table 4 showed little difference between experimental conditions. This was confirmed by analysis of covariance for both components of affect for the main effect as well as the interaction (all F < 1; df = 1, 57; p > .35).

Secondary task performance. For the dual task conditions, we compared the completion rate of the secondary task in the two testing methods (see Table 4). The analysis revealed a marginally significant effect of testing method, with slightly better performance being recorded in the lab condition (F = 3.15; df = 1, 28; p = .086; $\eta^2_{partial}$ = .101).

Table 4
Effects of testing method (the means marked with the superscripts[a] and[b] were found to be significantly different from each other).

	Lab-based testing		Field-based testing	
	M	(SD)	M	(SD)
Performance				
Task completion time (s)	181.3	(54.3)	172.4	(53.7)
Single task	195.5	(38.5)[b]	151.5	(28.7)[a]
Dual task	167.2	(64.8)	191.8	(64.6)[b]
Task completion rate (%)	67.2	(23.7)	63.7	(24.6)
Single task	60.0	(24.6)	71.4	(16.5)
Dual task	74.4	(21.2)[a]	56.6	(29.1)[b]
Interaction efficiency index (0–1)	0.52	(0.14)	0.49	(0.16)
Perceived usability (1–7)	4.94	(0.93)	4.66	(1.11)
Single task	4.66	(0.91)	4.91	(1.15)
Dual task	5.22	(0.90)[a]	4.43	(1.05)[b]
Subjective workload (1–20)	9.95	(2.29)	10.22	(2.94)
Positive affect: post-test (1–5)	3.32	(0.51)	3.09	(0.91)
Negative affect: post-test (1–5)	1.33	(0.33)	1.42	(0.44)
Secondary task: task completion rate (%)	86.7	(11.0)	77.8	(16.1)

4.4. Discussion

There was no main effect of testing method. This was largely in line with the results from the two other studies. However, there was an important interaction between testing method and task demands for performance. While the interaction was not surprising in itself, the type of interaction was. It was a cross-over interaction between task environment and testing method, which was different from the interaction predicted by the hypothesis.

The unexpected interaction contained two surprising elements. First, task completion time showed a poorer performance score for the single task in the lab than for the single task in the field. Second, in the field, task completion time was poorer under dual task conditions than under single task conditions (as one would expect) but in the lab, no such pattern was found. If anything, task completion time was better in the lab under the additional load of the secondary task than when only a single task was completed. Interactions of a similar kind were also observed for the two other outcome measures: task completion rate and perceived usability. For both measures, the least favourable scores (i.e. low completion rate and low usability ratings) emerged in the condition 'field/dual task' (as one would expect) whereas, surprisingly, the most favourable scores were observed in the condition 'lab/dual task'. These findings may be surprising since they cannot be explained within the framework of resource theory (e.g. Wickens and Hollands, 2000), which is commonly used to describe the effects of multiple task performance. It is difficult to explain an improvement of performance when an additional task is given. This is because the completion of the additional task requires extra resources and typically leads to performance decrements. When the cognitive resources available are currently not fully used (i.e. task is data limited; Norman and Bobrow, 1975), performance may be unimpaired but should not improve.

A better explanation of the results may be achieved when social facilitation theory (Zajonc, 1965) and the role of arousal (Stennett, 1957) are used to complement resource theory. In the present study, the additional cognitive load of the secondary task was merely of minor magnitude (involving only very simple mental arithmetic). It led to a

performance decrement in the rather noisy cafeteria with visual and auditory distractions (e.g. people moving around, noise levels of about 70 dB) but did not do so in the less distracting lab environment (e.g. no movements, noise levels of about 38 dB). It even led to performance improvements in the lab under dual task demands. This was because the additional secondary task might have increased arousal to a more optimal level. For example, arousal at a medium level may be best to achieve optimal performance (Stennett, 1957).

Differences in demand across experimental conditions were also reflected in usability ratings. These ratings suggest that the smartphone has not been perceived as being equally suitable for all operational conditions it was used in (i.e. different locations, varying task demands). The most negative ratings were found for dual task demands in the field. This represents the most realistic operational condition (and it is at the same time the most difficult one), for which usability demands are usually highest. The variation of usability ratings across experimental conditions reiterates the need to provide realistic simulations of future usage scenarios.

5. General discussion

Overall, the findings of the three experiments showed no simple effects of testing method on the outcomes of usability testing, which would have allowed us to conclude that there was a general advantage of one method over the other. Instead, a rather complex pattern emerged that showed an influence of testing method when this variable was combined with other factors such as poor usability and dual task demands. This complex pattern may have been caused by the interplay between factors from social expectations, resource theory (e.g. Wickens and Hollands, 2000) and social facilitation theory (Zajonc, 1965).

There are two main implications of the experimental work. First, task demands seem to play an important role influencing the outcomes of usability testing in the different locations. Task demands may be increased by poor product usability (e.g. study II) or by assigning an additional task to the user (e.g. study III). Second, the direction in which such an increase in task demands takes effect may be difficult to predict. As the empirical data showed in the present work, too low demands as a result of a moderately challenging single task coupled with a very quiet, non-distracting testing location may also lead to decreased performance.

The findings suggest that the presence of others (influencing the user's physical arousal and the social expectations they may experience) may interact with task difficulty. As the present work indicated, the effects of social facilitation theory observed were rather inconsistent in their nature (i.e. study I showed neither an effect of social facilitation nor of social inhibition, indications for social inhibition were found in study II and for social facilitation in study III). Social facilitation theory postulates that it is contingent upon the difficulty of a task whether an effect of social facilitation or social inhibition occurs (Zajonc, 1965). The inconsistent pattern found in the present work may be due to such differences in task difficulty. A comparative analysis of task difficulty across studies would have been of great value since it would have allowed us to carry out a formal evaluation of its effect in the context of social facilitation theory. However, given that different settings, devices and tasks were used in the three studies, it is difficult to conduct such a comparative analysis.

An important factor in social facilitation theory is the influence exerted by the presence of others. Therefore, test administrator presence may have influenced the outcomes of usability testing in two ways. First, administrator presence may increase the physiological arousal of test participants due to the physical or virtual presence of others. Second, administrator presence may be associated with a need to conform to certain social expectations, which test participants may perceive as social pressure. Due to the differences between experimental conditions with regard to the presence of the administrator (ranging from physical presence through virtual presence to no presence), arousal levels and social expectations may have differed between studies. The presence of the administrator was high in the three lab conditions and in the field test of study III because of the physical presence (e.g. sitting at neighbouring table). It was considered moderate in the synchronous remote testing of study I because of the direct exchange of written information between test administrator and participant. It was judged slightly lower in the asynchronous remote testing conditions of study I because the test administrator sent emails to participants and screen-recording software was used. Finally, administrator presence was considered lowest in the remote condition of study II because there was no exchange between test administrator and participant.

There are a number of limitations of this study. First, a direct comparative analysis between the three studies could not be carried out because tasks and technical devices differed between studies. On the one hand, this represents a disadvantage because results may be influenced by task-specific and device-specific effects. On the other hand, it has the benefits of having used a broader range of tasks, usage environments and devices, which may make it easier to generalise the findings. Second, we acknowledge that larger sample sizes would be needed in future work to demonstrate the nil effect of testing methods that we found in the present work. Third, the studies presented in this article cover a broad range of potential influencing factors of interest for the usability evaluation in different contexts (e.g. lab vs. field). However, it cannot be excluded that further factors (e.g. product aesthetics) might mediate the influence of different extra-laboratorial testing procedures in usability testing. In this series of studies, we concentrated on operational conditions. However, in future studies it might be very interesting to focus on contextual factors such as product aesthetics or user characteristics (e.g. cultural background, age etc.).

Most interesting for practitioners may be the question of whether low-cost testing methods are as valid as high-cost methods. Put differently, would the use of low-cost methods come at the cost of obtaining a less accurate picture of product usability? More precisely, it entails the question of whether lab testing provides similar results as field testing (as the more costly method) and whether remote testing provides similar results as lab testing (as the more costly method). In the present studies, the results did not show any main effects of testing method. Based on these findings, one might conclude that it is not critical for practitioners to choose a particular testing method. However, our results have shown that the influence of the testing method is moderated by other contextual factors of the usability test such as task difficulty and experimenter presence. This indicates that the usability test itself is prone to influences of such contextual factors (Sonderegger, 2010) and it is therefore advisable to test as close as possible to real usage situations.

The main implication for future work is that multifactorial designs should be used in future usability research to gain a better understanding of the complex relationship between the many factors that influences the outcomes of usability testing methods. Some of these factors are well researched such as aesthetics (e.g. Ben-Bassat et al., 2006) whereas others are less well researched (e.g. use of secondary or loading task, cultural background) and would benefit from future work. The four-factor framework outlined in the introduction may provide some guidance to that end.

Acknowledgements

The experimental work was carried out as part of a research grant awarded by the Swiss National Science Foundation (No. 100014_140359), whose generous financial support is gratefully acknowledged. We are also grateful to Thomas Supersaxo for his help in collecting the data.

References

Albert, W., Tullis, T., Tedesco, D., 2010. Beyond the Usability Lab: Conducting Large-scele Online User Experience Studies. Morgan Kaufman, Burlington, MA.

Alghamdi, A.S., Al-Badi, A.H., Alroobaea, R., Mayhew, P.J., 2013. A comparative study of synchronous and asynchronous remote usability testing methods. Int. Rev. Basic Appl. Sci. 1 (3), 61–97.

Ames, M., Brush, A.J., 2003. Final report on remote usability studies.CRA-W Distributed Mentor Project. Retrieved September 9, 2005 from. www.cra.org/Activities/craw/dmp/awards/2003/Ames/report.html.

Anderson, C.A., Lindsay, J.J., Bushman, B.J., 1999. Research in the psychological laboratory truth or triviality? Curr. Dir. Psychol. Sci. 8 (1), 3–9.

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

Andrzejczak, C., Liu, D., 2010. The effect of testing location on usability testing performance, participant stress levels, and subjective testing experience. J. Syst. Software 83 (7), 1258–1266.

Ben-Bassat, T., Meyer, J., Tractinsky, N., 2006. Economic and subjective measures of the perceived value of aesthetics and usability. ACM Trans. Comput. Hum. Interact. 13, 210–234.

Bevan, N., Macleod, M., 1994. Usability measurement in context. Behav. Inf. Technol. 13, 132–145.

Cockton, G., 2004. From quality in use to value in the world. In: Proceedings of the 2004 Conference on Human Factors in Computing Systems, CHI, April 24 - 29 2004. ACM, New York, NY, USA, pp. 1287–1290.

Dimitrov, D.M., Rumrill Jr., P.D., 2003. Pretest-posttest designs and measurement of change. Work 20 (2), 159–165.

Dipboye, R.L., Flanagan, M.F., 1979. Research settings in industrial and organizational psychology: are findings in the field more generalizable than in the laboratory? Am. Psychol. 34 (2), 141–150.

Duh, H.B.L., Tan, G.C., Chen, V.H.H., 2006. Usability evaluation for mobile device: a comparison of laboratory and field tests. In: Proceedings of the 8th Conference on Human-computer Interaction with mobile Devices and Services. ACM, pp. 181–186.

Dumas, J.S., Fox, J.E., 2012. 2012 In: Jacko, J.A. (Ed.), The Human-computer Interaction Handbook – Fundamentals, Evolving Technologies and Emerging Applications. CRC Press, Boca Raton, FL, pp. 1221–1242 3. Aufl.

Geen, R.G., 1991. Social motivation. Annu. Rev. Psychol. 42 (1), 377–399.

Hamborg, K.C., Hülsmann, J., Kaspar, K., 2014. The interplay between usability and aesthetics: more evidence for the “what is usable is beautiful” notion. Advances in Human-Computer Interaction 2014, 15.

Hart, S.G., Staveland, L.E., 1988. Development of NASA-TLX (task load index): results of empirical and theoretical research. Adv. Psychol. 52, 139–183.

Heuer, H., 1996. Dual-task performance. Handbook of perception and action 3, 113–153.

Hornbæk, K., 2006. Current practice in measuring usability: challenges to usability studies and research. Int. J. Hum. Comput. Stud. 64 (2), 79–102.

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

Kirchler, E., Fischer, F., Hölzl, E., 2010. Price and its relation to objective and subjective product quality: evidence from the austrian market. J. Consum. Pol. 33 (3), 275–286.

Kjeldskov, J., Stage, J., 2004. New techniques for usability evaluation of mobile systems. Int. J. Hum. Comput. Stud. 60 (5–6), 599–620.

Kjeldskov, J., Graham, C., Pedell, S., Vetere, F., Howard, S., Balbo, S., Davies, J., 2005. Evaluating the usability of a mobile guide: the influence of location, participants and resources. Behavior & information Technology 24, 51–65.

Lee, S., Koubek, R.J., 2010. Understanding user preferences based on usability and aesthetics before and after actual use. Interact. Comput. 22 (6), 530–543.

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

Madathil, C.K., Greenstein, J.S., 2011. Synchronous remote usability testing: a new approach facilitated by virtual worlds. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM, pp. 2225–2234.

Matthews, G., Davies, D.R., Westerman, S.J., Stammers, R.B., 2000. Human Performance: Cognition, Stress, and Individual Differences. Taylor & Francis, Philadelphia, PA.

McFadden, E., Hager, D.R., Elie, C.J., Blackwell, J.M., 2002. Remote usability evaluation: overview and case studies. Int. J. Hum. Comput. Interact. 14 (3–4), 489–502.

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

Norman, D.A., Bobrow, D.G., 1975. On data-limited and resource-limited processes. Cognit. Psychol. 7 (1), 44–64.

Pashler, H., 1994. Dual-task interference in simple tasks: data and theory. Psychol. Bull. 116 (2), 220.

Sauer, J., Sonderegger, A., 2011. Methodological issues in product evaluation: the influence of testing environment and task scenario. Appl. Ergon. 42, 487–494.

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

Sonderegger, A., 2010. Influencing Factors in Usability Tests: Testing Situation, Product Prototype and Test User. PhD thesis. University of Fribourg, Switzerland).

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

Sonderegger, A., Sauer, J., 2013. The influence of socio-cultural background and product value in usability testing. Appl. Ergon. 44 (3), 341–349.

Sonderegger, A., Zbinden, G., Uebelbacher, A., Sauer, J., 2012. The influence of product aesthetics and usability over the course of time: a longitudinal field experiment. Ergonomics 55 (7), 713–730.

Sonderegger, A., Schmutz, S., Sauer, J., 2016. The influence of age in usability testing. Appl. Ergon. 52, 291–300.

Stennett, R.G., 1957. The relationship of performance level to level of arousal. J. Exp. Psychol. 54 (1), 54–61.

Thielsch, M.T., Engel, R., Hirschfeld, G., 2015. Expected usability is not a valid indicator of experienced usability. PeerJ Computer Science 1, e19.

Thompson, K.E., Rozanski, E.P., Haake, A.R., 2004. October). Here, there, anywhere: remote usability testing that works. In: Proceedings of the 5th Conference on Information Technology Education. ACM, pp. 132–137.

Tractinsky, N., Shoval-Katz, A., Ikar, D., 2000. What is beautiful is usable. Interact. Comput. 13, 127–145.

Tuch, A.N., Roth, S.P., Hornbæk, K., Opwis, K., Bargas-Avila, J.A., 2012. Is beautiful really usable? Toward understanding the relation between usability, aesthetics, and affect in HCI. Comput. Hum. Behav. 28 (5), 1596–1607.

Tullis, T., Fleischman, S., McNulty, M., Cianchette, C., Bergel, M., 2002, July. An empirical comparison of lab and remote usability testing of web sites. In: Usability Professionals Association Conference.

Watson, D., Clark, L.A., Tellegen, A., 1988. Development and validation of brief measures of positive and negative affect: the PANAS scales. J. Pers. Soc. Psychol. 54 (6), 1063–1070.

Wickens, C.D., Hollands, J.G., 2000. Engineering Psychology and Human Performance. Pearson Prentice Hall, New Jersey.

Xia, L., Monroe, K.B., Cox, J.L., 2004. The price is unfair! A conceptual framework of price fairness perceptions. J. Market. 68, 1–15.

Zajonc, B., 1965. Social Facilitation. Science. New Series 149 (3681), 269–274.