Published in "Applied Ergonomics 41 : 130-140, 2010"
which should be cited to refer to this work.

The influence of user expertise and prototype fidelity in usability tests

Jürgen Sauer [a,*], Katrin Seibel [b], Bruno Rüttinger [b]

[a] Department of Psychology, University of Fribourg, Rue de Faucigny 2, 1700 Fribourg, Switzerland
[b] Institute of Psychology, Darmstadt University of Technology, Alexanderstr. 10, 64283 Darmstadt, Germany

An empirical study examined the impact of user expertise and prototype fidelity on the outcomes of a usability test. User expertise (expert vs. novice) and prototype fidelity (paper prototype, 3D mock-up, and fully operational appliance) were manipulated as independent variables in a 2×3 between-subjects design. Employing a floor scrubber as a model product, 48 users carried out several cleaning tasks. Usability problems identified by participants were recorded. Furthermore, performance, system management strategies and perceived usability were measured. The results showed that experts reported more usability problems than novices but these were considered to be less severe than those reported by novices. Reduced fidelity prototypes were generally suitable to predict product usability of the real appliance. The implications for the running of usability tests are specific to the fidelity of the prototype.

1. Introduction

1.1. Background

Usability testing represents the most fundamental and important method to identify problems in user–product interaction (Nielsen, 1993). While it is generally agreed that usability tests improve product usability (e.g. Säde et al., 1998; Sefelin et al., 2003; Walker et al., 2002), it is less clear what needs to be done to maximise the effectiveness of a usability test (partly owing to the plethora of methods and approaches used by designers). The need for increasing the effectiveness of usability tests is demonstrated by evidence in the research literature, which found remarkable inconsistencies across usability tests with regard to the usability problems identified (Lewis, 2006). For example, in the study of Molich et al. (2004), nine usability laboratories carried out usability tests with the same product independently of each other. Out of a total of 310 usability problems identified, about 75% were reported by one team only and merely two problems were found by six or more usability labs. Other work has produced similar findings (e.g. Kessner et al., 2001), indicating little overlap in the usability problems identified across different usability testing teams. These studies raise concerns about the objectivity, reliability and validity of usability tests since their outcomes may differ considerably across tests, observers and methods. It is quite conceivable that the inconsistencies found were at least partly caused by uncontrolled, and not yet well understood, features of usability tests.

1.2. Four-factor framework of contextual fidelity

A theoretical framework is presented in this article, which aims to provide guidance to designers and researchers when conducting usability tests. This framework, termed the *Four-Factor Framework of Contextual Fidelity* may help identify causes of the inconsistencies in the outcomes of usability tests reported above. The attribute 'contextual' emphasises the wider context and the different aspects of fidelity that are to be considered in usability testing. The factors of the framework were derived from three main sources: (a) previous models that addressed the issue of fidelity in usability testing (see review below), (b) pertinent issues discussed in the usability literature (e.g. user competencies), and (c) issues that play a role in ergonomics beyond the usability literature (e.g. physical and social environment).

The framework draws upon various models that have addressed the issue of fidelity in usability testing (e.g. Virzi et al., 1996; Nilsson and Siponen, 2005) but extends these to aspects of fidelity that have not been previously examined in any detail. Typically, previous models have concentrated on the fidelity of the technical system (e.g. prototype fidelity) while notably neglecting issues such as user characteristics and the testing environment. For example, the model of Nilsson and Siponen (2005) distinguishes between three aspects of fidelity: implemented automaticity (i.e. degree to which a prototype can be operated by a user without the help of a test facilitator), perceived automaticity (i.e. the subjective user assessment of

* Corresponding author. Tel.: +41 26 3007622; fax: +41 26 3009712.
E-mail address: juergen.sauer@unifr.ch (J. Sauer).

automaticity level), and precision (i.e. the level of detail at which a prototype is modelled). A slightly different and somewhat broader understanding of the concept of fidelity is provided by Virzi et al. (1996) whose model proposes four dimensions: degree of functionality (i.e. the level of detail to which each function is modelled), similarity of interaction (i.e. the level of mapping between human–machine communication and the type of displays and controls), breadth of features (i.e. the number of features modelled in a prototype), and aesthetic refinement (i.e. the modelling of the product with regard to colours and shape). The broadest view of fidelity is adopted by a model of Elliot et al. (2004) since it also includes aspects of fidelity that go beyond prototype design, such as task characteristics (e.g. distributed team tasks) and operational requirements (e.g. mission goals). The review of the models further suggests that none of them explicitly considers the wider testing environment, in which human–machine interaction takes place. While the literature on usability testing has acknowledged to some extent the importance of the wider usage context (Nielsen, 1993; Snyder, 2003), the focus was on the system, with comparatively little guidance given to designers about what fidelity level is to be used for the other factors such as user characteristics and the testing environment.

The four-factor framework of contextual fidelity aims to adopt a view of the wider context in which usability testing takes place. The framework with its main factors and subordinate factors is presented in Fig. 1. Each factor has a number of subordinate factors that outline the issues to be taken into consideration when conducting a usability test. These factors refer to various aspects of the issue of fidelity. In a usability test, the usage context is typically modelled with a fidelity level that is lower than in the future usage situation, owing to various constraints. The fidelity of the testing situation may differ from the future usage situation on four dimensions. First, the participant in a usability test may be different from the future user (e.g. short-haired male engineers are used to test a new hair dryer). Second, a prototype is available that is not yet fully operational (e.g. hair dryer has only a power setting but the temperature controls have not been implemented yet). Third, the task given may not be representative or sufficiently complex (e.g. appliance is used to dry a wig rather than the user's own hair). Fourth, the testing environment may differ physically from the future usage context (e.g. hair drying takes place in a lab rather than in the user's home). These four factors make up the context of usability testing while the level of fidelity on each factor will influence user behaviour and user satisfaction during the test. Therefore, these factors may represent potential threats to the reliability and validity of the usability test. Reliability and validity are important notions in psychological testing and many of these principles also apply to usability testing. In psychological testing as well as in usability testing, reliability and validity are influenced by the objectivity with which the testing procedure is carried out, the test results are scored, and the findings are interpreted (i.e. striving for consistency across testing sessions). This suggests a need for stronger standardisation of the testing procedure (e.g. consistent instructions, similar selection criteria for test users) to improve its reliability and validity, an endeavour to which the Four-Factor Framework of Contextual Fidelity may be able to make a contribution.

Each of the four factors can be divided into subordinate factors that describe more precisely the issues that need to be considered by the designer to increase the validity of the testing procedure (see Fig. 1). For the factor *testing environment*, one may distinguish between physical features, social features, and the application domain. The *physical* testing environment refers to aspects such as the size of the laboratory, noise levels, and location, which may all influence user behaviour, as is known from work on physical stressors (McCoy and Evans, 2005). However, work related to usability testing showed somewhat inconsistent results. While a study comparing the influence of a lab-based testing environment with a field test showed overall little evidence for differences between testing environment (Kaikonnen et al., 2005), other work very tentatively suggested that the behaviour-shaping effects of an information label were stronger in the lab than in the field (Sauer and Ruettinger, 2004). The *social* testing environment refers to the presence of other humans during the usability test (e.g. product design team) and the effects this may have on test outcomes. Following social facilitation theory (Zajonc, 1965), the presence of observers may influence appliance operation in usability testing. First, there is evidence that the presence of observers in usability tests may have negative effects on physiological parameters and some aspects of performance (Sonderegger and Sauer, in press). Second, the outcomes of a usability test may be moderated by the domain in which the appliance is used, such as at work, in the domestic domain or in the public domain (e.g. "walk-up-and-use"-products). For example, using a phone at work may be more strongly dominated by performance-related goals than in a leisure context, in which the joyful experience of the user with the product is of greater importance.

For the factor *task scenario*, one may distinguish between the *breadth* and the *depth* of a task scenario. The *breadth* of a task scenario refers to the degree to which the complexity of the natural task environment is modelled in the usability test. For example, if the operation of a car stereo is tested in the form of a single task, the task scenario is characterised by lower breadth than when the operation of a car stereo is part of a multiple-task environment including car navigation. A study comparing mobile phone operation using task scenarios of different breadth revealed that under the single task condition (phone operation while seated at a table), test participants reported more usability problems and lower overall workload than in a dual task condition (phone operation while walking in a pedestrian zone) whereas no difference was found for performance measures (Kjeldskov and Stage, 2004). The *depth* of a task scenario refers to the level of detail with which a particular task is completed. For example, this relates to the question of whether a task like writing a letter with a word processor is complete (i.e. it includes all task elements) or comprises a selection of task elements (e.g. cutting and pasting, changing line spacing). The two other factors from the model, *user characteristics* and *system prototype*, were empirically examined in the present study and are therefore explored in more detail in the next sections.

The ultimate purpose of the four-factor framework of contextual fidelity (following empirical testing) is to make predictions about which outcome measure is influenced by which factors of the framework. Prior to making these predictions, the different factors of the framework need to be empirically tested to ascertain their respective influence on the different outcome measures. Based on these empirical tests, the framework may need to be modified by adding, redefining or deleting factors. It is also acknowledged that the factors are not independent of each other. For example, if a prototype of a certain fidelity level is chosen, this places some constraints on other dimensions. In the example of a usability test of a cleaning appliance, a paper prototype would permit the cleaning task to be modelled as a decision-making task (e.g. user would inform experimenter about what power controls setting is to be chosen), but would not allow a sensori-motor task to be carried out (e.g. a rotary knob had to be turned to select desired setting).

1.3. User characteristics

Choosing appropriate users for testing represents a difficult task for designers. Potential test participants that are readily available (e.g. colleagues, friends and relatives of the designer, and students)

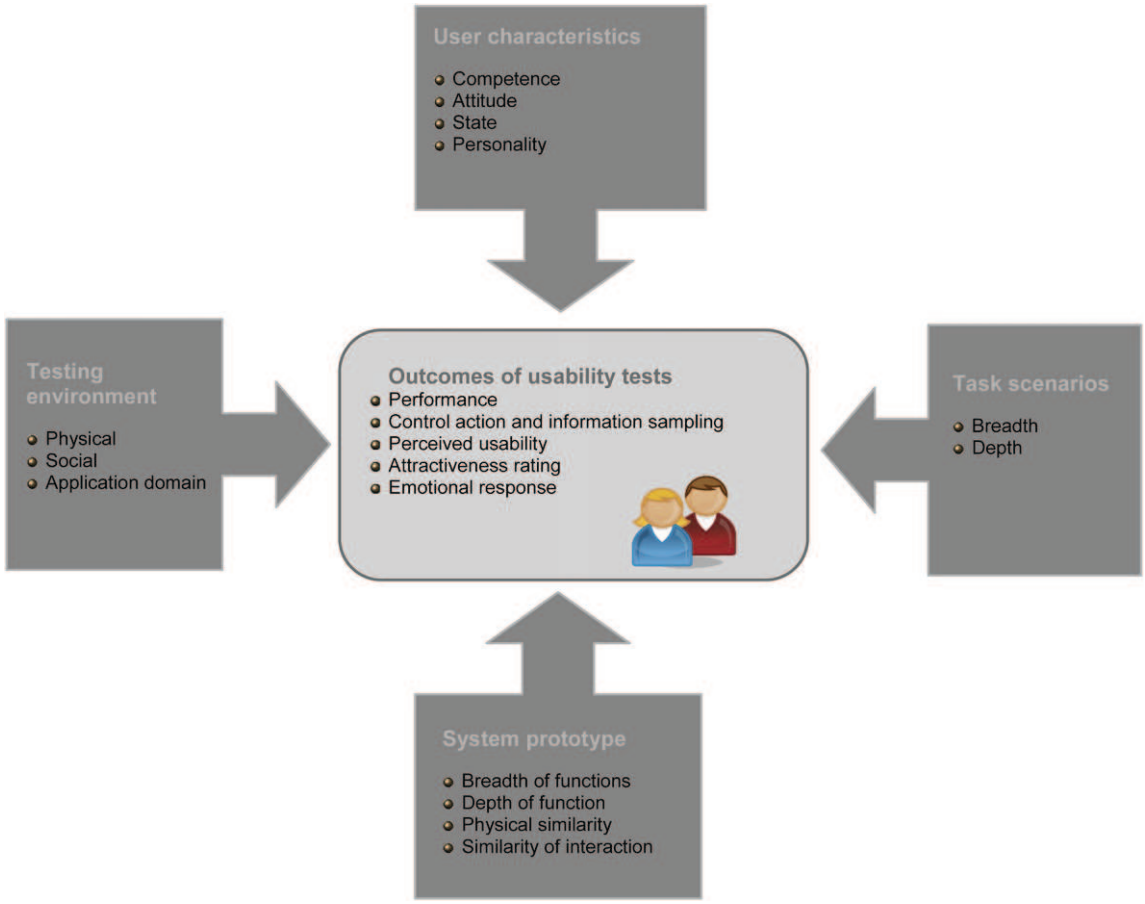

Fig. 1. Four-factor framework of contextual fidelity.

are often not representative of the future users of the product. The test participants may differ from future product users on a number of criteria: *User competence*, *user attitude*, *user state*, and *user personality*. These all contribute to the degree to which the usability test participants are good models of future product users.

For most products, the user attribute of the highest importance is *competence*, encompassing different aspects such as knowledge, skills and abilities. Expertise may be considered as a part of competence that refers to highly specific skills and knowledge of a person about a subject (here: floor scrubber operation). This determines whether a user is to be considered a *novice* or an *expert*, which represents an important dichotomous distinction in user selection. In the practitioners' literature, it is often recommended that users from both groups should be tested (Rubin, 1994; Snyder, 2003). However, the views on this are not unanimous; with Nielsen (1993) arguing that a novice should normally be used for usability testing and only under some circumstances a product needs to be tested on experts as well. To ensure that the lower end of the expertise continuum is also covered, the concept of the *least competent user* has been introduced (Rubin, 1994). Empirical work by Kjeldskov et al. (2005) showed that more usability problems were identified by novices than experts when operating an electronic patient record system. The question of whether experts or novices are better suited as test participants also depends on what specific aspects of usability are to be examined (e.g. learnability, efficiency). For example, novices are better suited than experts when learnability is measured than when the focus is on efficiency (which assumes that the user has already learnt the system; Nielsen, 1993). Although the level of expertise clearly represents a continuum (with novices and experts positioned at both ends), for reasons of simplicity and in conformity with the research literature, a dichotomous distinction of user expertise is adopted throughout this article.

The other user characteristics referred to above are also of importance in usability testing. Compared to user competence, their influence may however be limited to more specific circumstances. *User attitude* (e.g. environmental concern, openness towards technology) may influence user–product interaction. For example, if a product is to be designed for environmentally friendly use, environmentally concerned users may benefit more from enhanced system feedback on energy consumption as they are keen to reduce resource usage (Sauer and Rüttinger, 2004). *User state* may also need to be considered in application areas in which temporary conditions of the user have an influence on human–machine interaction. For example, the effectiveness with which an alarm clock is operated is influenced by the state of fatigue of the user, that is, the typical situation of a not yet fully awake user trying to operate an alarm clock in the dark needs to be modelled when testing different design options (Voûte et al., 1993). Lastly, *user personality* may influence the outcome of usability tests. For example, users scoring high on the personality factor conscientiousness may identify more usability problems because they approach the testing procedure more thoroughly.

1.4. Fidelity of prototype

The question of what kind of prototype is to be used is influenced by constraints that are inherent to the industrial design process, notably time pressure and budgetary limitations. This usually calls for the employment of reduced fidelity prototypes

(e.g. paper prototype, computer simulation, and mock-up) because they are cheaper, faster to build and more utilisable in earlier stages in the product development cycle than fully operational prototypes. However, there are concerns that this may be achieved at the cost of a less accurate picture of actual user behaviour.

There are a number of studies that have addressed the issue of prototype fidelity. Most studies concluded that the reduced fidelity prototypes provided equivalent results to fully operational products (Sefelin et al., 2003; Virzi et al., 1996; Catani and Biers, 1998; Wiklund et al., 1992; Walker et al., 2002). However, other work found differences in user behaviour as a function of prototype fidelity (Nielsen, 1990; Säde et al., 1998; Hall, 1999; Prümper et al., 1993; Sauer et al., 2008; Sauer and Sonderegger, 2009). This inconsistency in the research literature may be accounted for by several factors. First, there were differences with regard to the kind of prototype used and the target product being modelled (e.g. a paper prototype modelling 2D software package vs. a 2D computer simulation modelling a 3D product). Generally, reduced fidelity prototypes were more effective when the prototype and target product had the same number of dimensions. Second, some studies focused on the identification of usability errors while others examined efficiency measures (e.g. task completion time). Generally, reduced fidelity prototypes were more effective when usability errors were examined rather than efficiency measures. Overall, the number of studies published is not yet sufficient to allow a more precise analysis of this pattern to provide general recommendations about the utility of reduced fidelity prototypes.

1.5. The present study

The study examined two aspects of the framework of contextual fidelity. The first aspect was concerned with user competence by examining the respective role of expert and novice users in usability tests. The second aspect referred to the effects of using prototypes with different fidelity levels. This addressed the question of whether user behaviour with fully operational products can be accurately predicted from prototypes of lower fidelity.

1.5.1. User competence

Designers are faced with a number of questions when selecting participants for usability testing. A major question concerns the effect of using novices rather than experts for usability tests because the former are usually easier to recruit (unless it concerns a very widely used consumer product). It remains unclear to what extent and into which direction this will bias the results of the usability test. While the practitioners' literature recommends that users from both groups should be used (Rubin, 1994; Snyder, 2003), there is little research that has examined this question.

1.5.2. Prototype fidelity

Similarly, important decisions have to be made by the designer when selecting a prototype. It would be important for the designer to know how accurately user behaviour can be predicted from reduced fidelity prototypes. Furthermore, the effects of prototype fidelity may be moderated by user characteristics in that prototypes of reduced fidelity may have better predictive qualities when used with experts than novices. This might be due to their better mental representation of the task and the future product, which allows them to better predict forthcoming usability problems on the basis of a reduced fidelity prototype. Considering the widespread use of prototypes for usability tests, there is surprisingly little comparative research on the utility of prototypes at different fidelity levels, with the literature review above revealing only about a dozen studies that have addressed this issue. It is acknowledged that there may be some confounding between prototype and task. For example, the use of a paper prototype as opposed to a fully operational appliance may change the nature of the task. Some guidance of what kinds of task are particularly affected may be provided by task or resource models. For example, the multiple-resource model of Wickens and Hollands (2000) distinguishes between processing stages (perception, cognition, and responding), perceptual modalities (visual and auditory) and response modalities (manual and vocal). It is expected that the response stage is most strongly influenced by prototypes since control elements differ between prototypes of different fidelity. For example, turning a knob on a paper prototype does not have the same degree of resistance. We assumed that the perceptual stage would also be influenced by prototype fidelity, though more moderately. For example, the way information is presented may differ with regard to richness and dynamics (e.g. a dynamic display of an appliance can only be modelled in a static form with a paper prototype). The cognitive stage is not directly influenced by prototype fidelity but the cognitive processes are affected by preceding and subsequent phases (i.e. information input and output). In the present study, this problem has been addressed by focussing on tasks parameters that can be measured with all three prototypes (e.g. changing setting of a sliding control). In addition, some task parameters were used that were only applicable to the high-fidelity prototype (e.g. water consumption), which were then analysed only as a function of expertise.

The floor scrubber was chosen as a model product for the present study because it places higher demands on user skills than the average domestic appliance. This allows for a clear distinction to be made between expert and novice users. The floor scrubber comprises three primary functions: (a) navigation (i.e. speed, direction), (b) cleaning (i.e. mechanical cleaning by changing settings of brush pressure, chemical cleaning by controlling supply of cleaning solution), (c) maintenance and system monitoring (i.e. filling detergent into tank, checking battery status). This indicates that a range of skills is required for the user if the appliance is to be operated efficiently, including perceptual-motor skills (e.g. navigating the device across the floor) as well as process control skills (e.g. determining and monitoring the amount of cleaning solution used in cleaning process).

Based on the research reported by Kjeldskov et al. (2005), the recommendations offered in the practitioner literature, and our own deliberations, the following assumptions were put forward: (a) Novices would report more usability problems during testing than experts. (b) Experts would report more severe usability problems than novices. (c) Experts would show better performance than novices on the following dependent variables: achieved floor cleanness, task completion time, and water consumption. (d) More appropriate control settings would be made by experts than novices. (e) This difference in behaviour between experts and novices would be larger for the low- and medium-fidelity conditions than when the real appliance was operated. This prediction was made on the assumption that experts would have a better mental model of the technical system, which would allow them to extrapolate more successfully from the reduced fidelity prototype to the real appliance.

2. Method

2.1. Participants

Forty-eight participants were recruited for the study (31% female; mean age 42.7 yrs; age range: 18–72 yrs). It was attempted to recruit a sample of participants very similar to the population of real users with regard to education level and professional background. Most participants had basic school education (69% went to school for

9 years) while a smaller proportion of the sample achieved intermediate and advanced grades (17% went to school for 10 years and a further 14% for 13 years). This classification corresponds to the three chief school grades of the German education system. The study participants (novices and experts) worked in typical jobs that may include the operation of floor scrubbers (e.g. caretakers, professional cleaners, and retail shop assistants), though the job responsibility of novices did not involve the use of floor scrubbers.

Half of the participants recruited were expert users of floor scrubbers while the other half was novice users. More precisely, experts were defined as users who employ the appliance at least once a month (usage frequency per month: $M = 7.6$; accumulated usage duration per month: 16.7 hrs). Conversely, participants would be considered novices if they had never operated a floor scrubber before. The experimental groups of experts and novices were matched with regard to age, gender and education level. This matching procedure was to ensure that novices and experts were of sufficient similarity with regard to factors such as general cognitive ability.

2.2. Design

A 2×3 between-subjects design was used, with user expertise and prototype fidelity as independent variables. User expertise was varied at two levels: high (experts) vs. low (novices). Prototype fidelity was varied at three levels: 2D paper prototype (low-fidelity), 3D mock-up (medium-fidelity) and 3D fully operational appliance (high-fidelity).

2.3. Experimental measures

In this study, a considerable number of measures were taken, which can be grouped under the following headings.

2.3.1. Usability problems

In a semi-structured interview following the experimental session, participants were asked to report usability problems they had experienced with the prototype and, if possible, to provide suggestions for improvements. To gain an estimate of the importance of the reported usability problems, their severity was subsequently rated by a panel of 8 usability specialists. Four of these usability specialists were product designers and engineers of the manufacturing company that developed the appliance. The other 4 were university-based human factors specialists, with all of them being highly familiar with the appliance. The rating was made by each usability specialist on a 5-point scale, ranging from low to extreme severity.

2.3.2. Performance

This refers to several objective measures that were collected during completion of the experimental task, including task completion time (s), water consumption (L), and achieved cleanness (experimenter rating on a scale ranging from 1 to 4). These measures were only taken in the experimental condition using the real appliance.

2.3.3. Controls settings and system intervention

This is concerned with different forms of user behaviour, such as setting of controls (water flow rate, brush pressure), frequency of system interaction (brush pressure, water flow rate, menu, lifting or dropping suction beam and brushes) and the appliance of cleaning strategies.

2.3.4. Subjective user ratings

After the completion of the task scenario, users were asked to rate the perceived usability of the appliance with a product evaluation questionnaire. Comprising 41 items, the product evaluation questionnaire was specifically developed for assessing the technical features of the appliance (e.g. positioning of controls, turning circle). All items for subjective user ratings used a 6-point scale, labelled "very good" (6) and "very poorly" (1) at the end points. Furthermore, the aesthetic appeal of the appliance was measured by a one-item scale ("How aesthetically appealing is the appliance?"), again using a 6-point response scale.

2.4. Materials

2.4.1. Fully operational floor scrubber (high-fidelity prototype)

The high-fidelity prototype was a fully operational walk-behind floor scrubber (Kärcher BR 55/60 W Bp). The size of the appliance was about 1.47 m (length) $\times$ 0.67 m (breadth) $\times$ 1.16 m (height). A photograph of the floor scrubber is shown in Fig. 2a.

2.4.2. 3D mock-up (medium-fidelity prototype)

The medium-fidelity prototype was operationalised by a partially operational 3D mock-up of the above model with all navigational functions being fully available (e.g. speed). This was achieved by placing a PVC/cardboard mock-up over the real appliance (see Fig. 2b). The mock-up had duplicates of the all functions (e.g. water flow rate) but the functions were non-operational so that they had no effect on cleanness levels. Since the real appliance was completely covered with the mock-up, users had the impression that they were operating a not fully operational prototype.

2.4.3. Paper prototype (low-fidelity prototype)

The paper prototype was a 2D representation of the interface containing all controls in a simplified form with regard to the aesthetic refinement and tactile representation (see Fig. 2c). The paper prototype was modelled in cardboard (sized 300 mm $\times$ 300 mm), upon which all possible configurations were drawn. The controls were made of foam rubber that was fixed by a paper clip on the cardboard allowing their pushing or turning.

2.5. Procedure

The participants recruited were randomly assigned to one of the three prototype conditions. Upon arrival at the testing facilities, the participants were informed that they were to operate and evaluate a floor scrubber with a view that the feedback given would be used to improve the appliance. The testing facilities contained a tiled floor area of approximately 52 m^2, which represented a section of a corridor situated in the basement of a school building. On the designated floor area, there were four patches (sized approximately 0.5 m by 0.5 m each) being visibly soiled with a different substance each (flour, ointment, shoe polish, and sugar syrup). The four substances differed in terms of the efforts required to remove them. Due to the different prototypes being employed, the task completion varied slightly between conditions.

2.5.1. Task completion with high-fidelity prototype

In the high-fidelity condition, the task scenario corresponded to a typical cleaning activity involving floor scrubber operation. Participants were given a demonstration of the appliance, including the different displays and controls available. The participants then carried out some basic operations with the floor scrubber on a separate floor area (e.g. moving the appliance back and forth) to ensure that they were able to operate the different functions of the

Fig. 2. Prototypes of floor scrubber: (a) high-fidelity, (b) medium-fidelity, and (c) low-fidelity.

appliance. After having completed this short practice trial, the experimental task scenario began, with users being instructed to clean the prepared floor area until all four substances on the floor had been removed (see above). Furthermore, they were told that they should use the appliance in the same manner as they would do if they were at work. The experimenter was present during task completion to observe and record user behaviour, employing a protocol sheet and a clipboard.

2.5.2. Task completion with medium-fidelity prototype

Similar to the high-fidelity prototype, participants were given a demonstration of the 3D mock-up before beginning the task scenario, followed by a practice trial. Furthermore, it was pointed out to the participants that the appliance was still in a design stage, with some functions being not yet connected to the controls. During the experimental trial, participants manoeuvred the 3D mock-up across the prepared floor area, with the participant being able to set the controls but with no direct feedback being provided about the effects of the user's cleaning efforts. Instead, feedback was given in an indirect form by presenting pictures to the user with the likely impact of their action on the floor cleanness. The pictures were placed on the chassis of the floor scrubber by the experimenter whenever appropriate. The pictures were taken during extensive testing sessions prior to the experiment, using the real floor scrubber to document the effects of different control settings on floor cleanness. Again, as in the condition with the high-fidelity prototype, users were instructed to use the appliance in the same way as they would do at work. User behaviour was recorded by the experimenter observing the experimental trial.

2.5.3. Task completion with low-fidelity prototype

The participants employing the low-fidelity prototype to complete the task were sitting at a table, with the set-up corresponding to a typical usability test employing a paper prototype. Participants were explained the function of the different displays and controls available. Feedback on the impact of their control actions was given by using the same set of pictures employed with the medium-fidelity prototype. The soiled floor area employed in the two other conditions was also visible to users in this condition. The experimenter also adopted the role of a test facilitator during the usability test (i.e. changing the state of prototype based on the user's interaction with the system). The participants were able to complete control actions by sliding, pressing and turning the paper-made controls. Based on the user's selection, the experimenter presented the card reflecting the change in display content initiated by the action. As in the two other conditions, users were instructed to behave as if they operated the appliance in a work context.

In all three experimental conditions, users were interviewed by the experimenter after task completion. The user was asked to report any usability problem referring to the operation of the

appliance or any issue regarding design characteristics. This was followed by the completion of the product evaluation questionnaire and the aesthetics rating scale.

3. Results

3.1. Usability evaluation

3.1.1. Number of usability problems

In total, 266 usability problems were reported by users (this figure represents a simple count of each problem mentioned by a user). Generally, experts mentioned more problems than novices (157 vs. 109). This difference between experts and novices became even more pronounced when the usability problems were corrected for those being mentioned several times. This reduced the total number to 116 distinct problems, of which 56.0% were identified by experts, 8.6% by novices and 35.3% by members of both groups.

Table 1 presents the data for the mean number of usability problems identified by each user group (the type of usability problems identified is presented in Section 3.1.2). It shows that experts mentioned significantly more usability problems than novices ($F = 7.25$; df $= 1, 42$; $p < .01$). This difference appeared to be larger for the reduced fidelity prototypes than for the fully operational appliance. However, statistical tests did not confirm this interaction to be significant ($F < 1$). No main effect of prototype fidelity was found ($F < 1$).

3.1.2. Type of usability problems

All usability problems mentioned by users in the post-experimental interviews were assigned to a category system to gain a more holistic perspective on the kind of usability problems faced by users. The category system was developed by the experimenter and a second rater who was also familiar with the technical system (see Table 2). The categories include positioning and operation of controls (e.g. controls are not within easy reach), efficiency and functionality (e.g. suction beam is too small), inadequate functions (e.g. setting maximum speed in the menu is cumbersome), device navigation (e.g. turning circle is too small), intuitiveness and comprehensibility (e.g. scaling of brush pressure control violates population stereotype), and maintenance (e.g. emptying detergent tank is awkward). The allocation of usability problems was done independently by the two raters, with a satisfactory inter-rater reliability coefficient emerging (Cohen's $K = .74$).

The results showed that most usability problems reported concerned positioning and operation of controls ($M = 2.0$), followed by functionality and efficiency ($M = 1.7$). Usability problems from other categories were referred to considerably less frequently (see Table 2). There were a number of issues which experts were more concerned with than novices, such as functionality and efficiency, inadequate functions, maintenance, and safety and device protection. Conversely, there was one issue that novices reported more often than experts. This referred to the manoeuvring of the appliance, which represents an activity requiring considerable perceptual-motor skills.

Table 1
Mean number of usability problems reported by each user as a function of levels of expertise and prototype fidelity.

	Experts	Novices	Overall
Overall	6.5	4.5	
Low fidelity	6.5	4.4	5.4
Medium fidelity	7.3	4.5	5.9
High fidelity	5.9	4.8	5.3

Table 2
Mean number of usability problems identified by users in each category as a function of expertise and prototype fidelity.

	Paper prototype	3D mock-up	Fully operational appliance	Overall
Positioning and operation of controls	2.3	2.3	1.5	
Experts	1.9	2.9	1.5	2.1
Novices	2.6	1.8	1.8	2.0
Efficiency and functionality	1.3	1.6	2.2	
Experts	1.8	2.6	2.8	2.4
Novices	0.8	0.5	1.6	1.0
Inadequate functions	0.8	0.9	0.4	
Experts	1.4	0.8	0.5	0.9
Novices	0.3	1.0	0.4	0.5
Device navigation	0.1	0.7	0.4	
Experts	0	0.4	0.1	0.2
Novices	0.1	1.0	0.6	0.6
Intuitiveness and comprehensibility of interface	0.2	0.4	0.4	
Experts	0.1	0.5	0.4	0.3
Novices	0.3	0.3	0.3	0.3
Maintenance, set-up and shut-down procedures	0.5	0.1	0.2	
Experts	0.8	0.1	0.4	0.4
Novices	0.3	0	0	0.1
Others	0.1	0	0.1	
Experts	0.3	0	0.3	0.2
Novices	0	0	0	0

Differences between prototype conditions also emerged with regard to the mean number of usability problems in each category (see Table 2). Under the reduced prototype fidelity conditions, users were more concerned with the positioning and operation of controls and less with functionality and efficiency than when operating the real appliance. Furthermore, manoeuvring the appliance gained in relative importance under the 3D mock-up condition, compared to the two others.

3.1.3. Severity of usability problems

Since the number of usability problems mentioned may not necessarily be a good indicator of their contribution to better product usability, the severity of each problem reported by users was rated by the panel of usability specialists. As the data in Table 3 show, the usability problems reported by novices were considered to be more severe by the 8 usability specialists than those identified by experts ($F = 20.2$; df $= 1, 6$; $p < .001$). Furthermore, more severe usability problems were identified under high and low-fidelity than under medium-fidelity ($F = 14.1$; df $= 2, 12$; $p < .001$; post-hoc LSD-tests: $p < .05$). Finally, a significant interaction was observed between expertise and prototype fidelity ($F = 40.8$; df $= 2, 12$; $p < .001$). This was because novices were more effective in identifying the more serious usability problems under high and low-fidelity ($p < .05$) but not under medium-fidelity ($p > .05$).

Table 3
Severity of usability problems rated by human factors specialists (1: not severe at all; 5: very severe).

	Experts	Novices	
Overall	2.3	2.6	
Low fidelity	2.2	3.0	2.6
Medium fidelity	2.3	2.2	2.2
High fidelity	2.5	2.7	2.6

A further analysis carried out separately for university-based and industry-based raters revealed that the same pattern of effects was observed for each group of human factors specialists. However, there was a difference with regard to the severity of rating usability problems since the human factors specialists from the manufacturer considered the reported problems to be less severe than the university specialists ($M = 2.83$ vs. $M = 2.10$; $F = 10.8$; df = 1, 6; $p < .05$).

3.2. Performance

The data of the various performance measures are presented in Table 4. The performance data were only analysed for main effects of expertise in the high-fidelity condition since they were not collected for the other two prototypes.

The vast majority of measures did not indicate any differences between experts and novices as the data in Table 4 demonstrate. There was no significant difference between groups with regard to task completion time ($F = 2.26$; df = 1, 42; ns). Similarly, no difference was recorded for the distance covered by users during task completion ($F < 1$). Water consumption was the only parameter for which a marginally significant effect was observed in the high-fidelity condition. Experts consumed more water than novices during task completion ($F = 3.42$; df = 1, 42; $p = .086$). Interestingly, the increased water consumption of experts did not result in higher cleanness levels ($F < 1$). During the experimental trial, it was observed that novices tended to focus more strongly on the soiled patches (e.g. by starting off with these) than experts and, as a consequence, failed to clean the unsoiled floor area (50% of novices; 25% of experts). A Chi-square test showed that this difference just failed to be significant ($Chi^2 = 3.2$; df = 1; $p = .07$).

3.3. Setting of controls and frequency of system interaction

Whereas the collection of performance measures required the availability of a fully operational appliance, user–product interaction with regard to the setting of controls and interaction frequency could also be measured with reduced fidelity prototype. Critical controls for the floor scrubber are water outflow and brush pressure. Brush pressure was set by a control lever and water outflow by a rotary knob, with both having 6 discrete settings labelled from 1 to 6. Speed was selected by using a menu (operated by 6 push buttons and a rotary knob) that allowed for a setting to be chosen, ranging from 1 to 10. A continuously adjustable control lever was also available to increase and decrease speed very rapidly, with the upper limits being determined by the speed chosen in the menu. The data for these parameters are presented in Table 5.

3.3.1. Water flow rate

The results showed that users overestimated the amount of water needed when operating reduced fidelity prototypes ($F = 6.7$; df = 2, 42; $p < .01$). This overestimate appeared to be more pronounced for novices than experts, though the interaction was not significant ($F = 1.1$; df = 2, 42; ns). There was no main effect of expertise ($F = 1.7$; df = 1, 42; ns).

Table 4
Performance data as a function of levels of expertise for the fully operational prototype condition.

	Experts	Novices
Task completion time (min)	12.1	9.9
Distance covered (m)	114	108
Water consumption (L)	5.8	4.4
Achieved cleanness (1–4)	3.76	3.72

Table 5
Mean settings of controls as a function of levels of expertise and prototype fidelity.

	Experts	Novices	Overall
Water outflow (1–6)	3.8	4.0	3.9
Low fidelity	4.2	4.7	4.4
Medium fidelity	3.5	3.9	3.7
High fidelity	3.7	3.5	3.6
Brush pressure (1–6)	4.2	4.3	4.2
Low fidelity	4.3	4.8	4.5
Medium fidelity	4.3	4.5	4.4
High fidelity	4.0	3.5	3.8
Maximum speed (number of selections)	0.7	0.2	0.4
Low fidelity	0.5	0	0.3
Medium fidelity	0.4	0.3	0.3
High fidelity	1.3	0.3	0.8

3.3.2. Brush pressure

Similarly, there was a strong effect of prototype fidelity, with users choosing higher settings for brush pressure on the paper prototype and the mock-up than with the fully operational appliance ($F = 3.7$; df = 2, 42; $p < .05$). Again, this overestimation was more pronounced for novices than experts but failed to reach significance ($F = 1.6$; df = 2, 42; ns). The main effect of expertise was not significant ($F < 1$).

3.3.3. Maximum speed

The frequency of selecting maximum speed was also recorded since it represents an efficiency indicator showing that users can make use of the full range of functions offered by the machine. The analysis revealed that maximum speed was more often used by experts than novices ($F = 4.07$; df = 1, 42; $p < .05$) but was not affected by prototype fidelity ($F = 1.4$; df = 2, 42; ns). No interaction was observed ($F < 1$).

3.4. Subjective user rating

3.4.1. Usability

After the completion of the task scenario, users were asked to rate the perceived usability of the appliance with a product evaluation questionnaire. As the data in Table 6 show, no difference between experts and novices emerged with regard to the usability of the floor scrubber ($F < 1$). It is remarkable that the usability ratings of the appliance were not affected by prototype fidelity ($F < 1$). When users operated a prototype with reduced fidelity, they made similar judgements as they did for the real appliance. Separate analyses for each scale (e.g. position of displays and controls) showed broadly the same pattern.

3.4.2. Aesthetics

The analysis of the aesthetics ratings revealed that experts found the appliance less appealing than novices (see Table 6). This

Table 6
Subjective user ratings of usability and aesthetics as a function of levels of expertise and prototype fidelity.

	Experts	Novices	Overall
Perceived usability (1–6)	3.5	3.6	
Low fidelity	3.5	3.5	3.5
Medium fidelity	3.4	3.7	3.6
High fidelity	3.5	3.5	3.5
Aesthetics (1–6)	3.5	3.8	
Low fidelity	3.6	3.6	3.6
Medium fidelity	3.2	3.9	3.6
High fidelity	3.6	3.9	3.7

difference was significant ($F = 4.82$; df = 1, 42; $p < .05$). As already observed for the usability ratings, it is interesting that no effect of prototype fidelity was observed ($F < 1$), with the reduced fidelity prototypes not being differently rated than the real appliance. There was no significant interaction ($F = 1.61$; df = 2, 42; ns).

4. Discussion

This first aim of the study was to examine the respective roles of novices and experts in usability tests, employing prototypes of different fidelity levels. A main finding was that experts identified more usability problems than novices. The usability problems reported by novices were judged to be more severe than those identified by experts. All other dependent variables provided no strong evidence for the superiority of one group of users over the other. The second aim of the study was to examine the effects of prototype fidelity. It emerged that using reduced fidelity prototypes for determining user behaviour with real appliances may lead to a general overestimate of control settings since users employing a reduced fidelity prototype chose generally higher control settings than those using the real appliance.

With regard to the identification of usability problems, the present study provided evidence for specific advantages of each user group, depending on the primary goal of the usability test. If the primary goal is to gain an overview of all usability problems associated with the appliance, the consultation of experts may be advantageous because they provide a more complete listing of possible usability problems than novices. The finding of experts being more productive than novices appears to be in contrast to the results of work by Kjeldskov et al. (2005), which found that more usability problems were identified by novices than experts. This may however be due to methodological differences since, in contrast to the present study, Kjeldskov et al. used an outcome-based measurement of usability problems (i.e. a usability problem was recorded when user failed to complete a task). In their study, expert users were able to adopt compensatory strategies (permitting them to work around usability problems) so that fewer usability problems were recorded for this group than for novices who did not have these compensatory strategies available. In the present study, not only actual usability problems were reported by users but also potential ones, that is, those that did not occur in the present task scenarios but may occur in other ones. Experts reported more of those due to their higher level of expertise, which allowed them to anticipate usability problems that may occur in task scenarios they have previously experienced. For example, experts reported that the suction beam was too high for navigating the floor scrubber underneath some shelves and that the suction beam was too wide for navigating a narrow corridor. Neither of the two points represented a problem in the current task scenario but may well do so in others. This shows that experts give much stronger consideration to future usage scenarios than novices, resulting in a larger number of usability problems being reported by that user group. In order to tap into the considerable experience of expert users, it is advisable to include self-reported usability problems as a measure, in particular, if there is a wide range of possible task scenarios of which most cannot be covered in the usability test.

If the primary goal is to identify the most severe usability problems as quickly as possible, there seem to be some benefit of relying on novices rather than experts. This point has also been made by usability practitioners who have expressed a preference for novices over experts (Nielsen, 1993). Empirical research evaluating an electronic patient record system has also shown that novices identified usability problems of higher severity during task completion than experts (Kjeldskov et al., 2005). However, this advantage of novices may be due to the following reasons. A closer look at our data revealed that the usability problems most frequently reported by experts were efficiency and functionality issues, which were related to the size and shape of various system elements (e.g. disk brush is too small). These issues were not considered to be critical aspects of usability by the human factors specialists in their ratings. This was because these usability problems would not prevent the completion of a task but would affect usage efficiency only (e.g. by increasing task completion time). Usability problems that prevent task completion are clearly the more important ones since they also impinge on usage efficiency but not vice versa (i.e. a product may be considered to be highly inefficient by users although all user tasks can be successfully completed).

With regard to performance and controls settings, the expected superiority of experts over novices surfaced in two subtle forms. Firstly, it was observed that in comparison to experts, novices focussed more strongly on the soiled patches at the expense of the unsoiled floor areas. This may be due to novices not having sufficient experience to adopt a more holistic view of the task so that they concentrate on the most conspicuous problems, representing a form of encystment (cf. Dörner and Brehmer, 1993). Giving attention to the most salient aspect of the task environment is a typical pattern of human behaviour but it may be sub-optimal. This attentional narrowing becomes stronger under increasing task demands (Hockey, 1979), with task demands being clearly stronger for novices than for experts. Since the parameter of cleaning performance was mainly affected by the way users dealt with the four soiled patches, this provided relative benefits to the novice group (which focused very strongly on these) for cleaning performance since they neglected the remaining floor area (resulting in an overestimate of cleaning performance). Additionally, this neglect of the remaining floor area led to a reduction of task completion time for the novice group. This may explain why the expected effects of expertise on achieved cleanness and task completion time were not found. Secondly, the results revealed that experts consumed more water than novices. There is evidence from the manufacturer's engineering tests that using ample water increases the longevity of certain parts of the floor scrubber (e.g. disk brush). According to the manufacturer's testing data, the advantages of more generous water usage are not limited to reducing wear and tear, it also has positive effects on cleaning performance (in particular when combining this with an appropriate use of cleaning agents). Experts have managed the appliance to that end but, for methodological reasons, the benefits of such an approach could not be unequivocally demonstrated in the performance data.

The setting of controls during appliance operations represents an important parameter since it influences usage efficiency and, additionally, it can be measured across fidelity levels. This allows us to determine the extent to which a reduced fidelity prototype can be employed to make an accurate prediction of user behaviour with the real appliance. The findings suggested that reduced prototype fidelity led to an overestimate of control settings in the present study, that is, users selected a higher setting than they did with the real appliance. This overestimate of control settings for reduced fidelity prototypes was observed in other work, too (Sauer et al., 2008). Due to the lack of system and environmental feedback given by the reduced fidelity prototype (e.g. effectiveness with which the chosen setting cleans the floor area), users may find it difficult to set controls correctly. This lack of feedback does not support the use of closed-loop control (e.g. Wickens and Hollands, 2000) since no information is provided to the user about the appropriateness of the chosen setting. It therefore requires users to rely very much on their mental model of the appliance (i.e. their understanding of

how the system works) to predict the consequences of their actions. In the absence of feedback, users may prefer to err on the side of caution by choosing a too high control setting (rather than a too low one), which makes it more likely to achieve task goals. However, one may also envisage task environments in which users select too low settings in the absence of feedback (e.g., if it entails a risk of causing damage).

Interestingly, subjective usability ratings were unaffected by prototype fidelity. This is quite remarkable since certain aspects of product usability can only be judged when the user can actually operate the fully operational prototype. However, it appears that these aspects are not critical since users seem to be quite capable of extrapolating from a reduced fidelity prototype to the real appliance. This phenomenon bears some resemblance to the law of closure (e.g. Eysenck and Keane, 2005), which also suggests a compensatory process by human cognition. A similar effect was also recorded for the rating of the appliance's aesthetic appeal. For usability ratings as well as aesthetic assessment, it suggests that users carry out some kind of compensatory activity to make up for the lower level of detail and diminished information content provided by the reduced fidelity prototypes. This pattern has already been observed in previous work and was termed the 'deficiency compensation'-effect (Sauer and Sonderegger, 2009). Overall, this suggests that subjective usability ratings are not much influenced by the type of prototype used, thus allowing for a reasonable assessment of usability even on the basis of a low or medium-fidelity prototype.

Against the background of the Four-Factor Framework of Contextual Fidelity, the present study provides a first empirical test of how user competence in the form of expertise interacts with prototype fidelity. The results provide no evidence for a general superiority of one user group over the other in usability tests. The relative advantage of each user group seems to depend on the specific purpose of the usability test (e.g. identification of a maximum number of usability problems or identification of the most severe ones). However, taking the overall pattern of results into account, there was evidence that would justify a preference to be given to experts over novices. For example, problems associated with a reduced efficiency of product operation are best identified by expert users who are also better able to address issues that go beyond the once directly relevant in given task scenarios of the current usability test.

When presenting the four-factor framework in this article, we have already addressed general methodological limitations that concern the possible confounding between prototype fidelity and task. In the present study, the breadth of the task environment was reduced in the low-fidelity prototype. For example, appliance navigation as a motor skill activity was not part of the task scenario, which may have led to more cognitive resources being available for carrying out the task with the low-fidelity prototype. Furthermore, feedback quantity and quality was reduced for the low- and the medium-fidelity prototype (e.g. no resistance of control, no direct and immediate feedback of the selected water flow rate on cleaning result), which may have made it harder to achieve task-related goals. However, with regard to the identification of usability problems, the nature of the task was largely unaffected in the present study since it required a vocal response in all three conditions (i.e. usability problems were always reported in the same way).

Regarding prototype fidelity as an important factor of the framework, the findings of the present study and from other empirical work suggest that reduced fidelity prototypes (i.e. paper prototypes, mock-ups but also computer simulations) are suitable for usability tests if the following specific weaknesses are taken into account: (a) there is a tendency of users to overestimate the required setting of controls for paper prototypes and mock-ups but also for computer simulations in comparison with the real appliance; (b) there are limitations to the kind of outcome measures that can be taken when using reduced fidelity prototypes (notably for paper prototypes and computer simulations but to a lesser extent this also applies to mock-ups); (c) a 'deficiency compensation'-effect may be observed, with reduced fidelity prototypes being more positively rated than real appliances.

The four-factor framework permits to address the issue of reliability and validity in usability testing in a more structured manner by pinpointing potential threats. While it may be argued that the completion of a usability test with moderate reliability and validity is better than no usability test at all, there is a clear need to examine ways of improving reliability and validity of usability tests. Although some of the subordinate factors have already been examined in a number of studies (notably aesthetics and prototype fidelity), research is still required to determine their effects in combination with other factors. Two other factors may be of particular interest for future research in usability testing: breadth of task scenario and application domain. There is a need to choose a realistically broad and complex task scenario by modelling a multiple-task scenario rather than just focussing on a single task. The application domain of the product selected for the usability test is also important since it may moderate the influence of other factors (e.g. aesthetics may have a stronger influence in the domestic domain than at work). Overall, the present study provides a further empirical evaluation of the influence of factors, representing a further step towards a comprehensive evaluation of the framework.

Acknowledgements

We gratefully acknowledge the financial support of the German Research Foundation (DFG) for carrying out this work (Research grant: TFB55). We would also like to thank Dr H. Gehringer and Mr W. Callenius from Alfred Kärcher GmbH & Co. KG for providing us with technical support and advice.

References

Catani, M.B., Biers, D.W., 1998. Usability evaluation and prototype fidelity: users and usability professionals. In: Proceedings of the Human Factors Society 42nd Annual Meeting, pp. 1331–1335.

Dörner, D., Brehmer, B., 1993. Experiments with computer-simulated micro-worlds: escaping both the narrow straits of the laboratory and the deep blue sea of the field study. Computers in Human Behaviour 9, 171–184.

Elliot, L.R., Dalrymple, M.A., Schifflett, S.G., Miller, J.C., 2004. Scaling scenarios: development and application of C4ISR sustained operations research. In: Schiflett, S.G., Elliot, L.R., Salas, E., Coovert, M.D. (Eds.), Scaled Worlds: Development, Validation, and Applications. Ashgate Publishing Limited, Aldershot, pp. 119–133.

Eysenck, M.W., Keane, M.T., 2005. Cognitive Psychology. A Student's Handbook, fifth ed. Psychology Press, Hove.

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

Hockey, G.R.J., 1979. Stress and cognitive components of skilled performance. In: Hamilton, V., Warburton, D.M. (Eds.), Human Stress and Cognition. Wiley, Chichester, pp. 141–178.

Kaikonnen, 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, 4–16.

Kessner, M., Wood, J., Dillon, R.F., West, R.L., 2001. On the reliability of usability testing Conference on Human Factors in Computing Systems. In: CHI '01 Extended Abstracts on Human Factors in Computing Systems, pp. 97–98.

Kjeldskov, J., Skov, M.B., Stage, J., 2005. Does time heal?: a longitudinal study of usability. In: Proceedings of the 19th Conference of the Computer–Human Interaction Special Interest Group of Australia on Computer–Human Interaction: Citizens Online: Considerations for Today and the Future, pp. 1–10.

Kjeldskov, J., Stage, J., 2004. New techniques for usability evaluation of mobile systems. International Journal of Human–Computer Studies 60, 599–620.

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

McCoy, J.M., Evans, G.W., 2005. Physical work environment. In: Barling, J., Kelloway, E.K., Frone, M.R. (Eds.), Handbook of Work Stress. Sage, London, pp. 219–245.
Molich, R., Ede, M.R., Kaasgaard, K., Karyukin, B., 2004. Comparative usability evaluation. Behaviour & Information Technology 23, 65–74.
Nielsen, J., 1990. Paper versus computer implementations as mockup scenarios for heuristic evaluation. In: Proceedings of the IFIP TC13 Third International Conference on Human–Computer Interaction, pp. 315–320.
Nielsen, J., 1993. Usability Engineering. Academic Press, Boston.
Nilsson, J., Siponen, J., 2005. Challenging the HCI concept of fidelity by positioning Ozlab prototypes. In: Proceedings of the Fourteenth International Conference on Information Systems Development, pp. 349–360.
Prümper, J., Heinbokel, T., Küting, H.J., 1993. Virtuelle Prototypen als Werkzeuge zur benutzerzentrierten Produktentwicklung. Anwendung einer handlungstheoretischen Fehlertaxonomie auf reale und simulierte Oberflächen von Waschmaschinen. Zeitschrift für Arbeitswissenschaft 4, 160–167.
Rubin, J., 1994. Handbook of Usability Testing: How to Plan, Design, and Conduct Effective Tests. John Wiley, New York.
Säde, S., Niemenen, M., Riihiaho, S., 1998. Testing usability with 3D paper prototypes – case Halton system. Applied Ergonomics 29, 67–73.
Sauer, J., Franke, H., Rüettinger, B., 2008. Designing interactive consumer products: utility of low-fidelity prototypes and effectiveness of enhanced control labelling. Applied Ergonomics 39, 71–85.
Sauer, J., Rüettinger, B., 2004. Environmental conservation in the domestic domain: the influence of technical design features and person-based factors. Ergonomics 47, 1053–1072.
Sauer, J., Sonderegger, A., 2009. The influence of prototype fidelity and aesthetics of design in usability tests: effects on user behaviour, subjective evaluation and emotion. Applied Ergonomics 40, 670–677.
Sefelin, R., Tscheligi, M., Giller, V., 2003. Paper prototyping – what is it good for? A comparison of paper- and computer-based prototyping. In: Proceedings of CHI, pp. 778–779.
Snyder, C., 2003. Paper Prototyping: The Fast and Easy Way to Design and Refine User Interfaces. Morgan Kaufmann, San Francisco.
Sonderegger, A., Sauer, J. The influence of laboratory set-up in usability tests: effects on user performance, subjective ratings and physiological measures. Ergonomics, in press.
Voûte, C.C.C., Kanis, H., Marinissen, A.H., 1993. User involved redesign of a clock-radio. Unpublished Manuscript, Delft University of Technology, Department of Product and Systems Ergonomics.
Virzi, R.A., Sokolov, J.L., Karis, D., 1996. Usability problem identification using both low- and high-fidelity prototypes. In: Conference Proceedings on Human Factors in Computing Systems: CHI 96, pp. 236–243.
Walker, M., Takayama, L., Landay, J., 2002. High-fidelity or low-fidelity, paper or computer medium? In: Proceedings of the Human Factors and Ergonomics Society 46th Annual Meeting: HFES2002, pp. 661–665.
Wickens, C.D., Hollands, J.G., 2000. Engineering Psychology and Human Performance. Prentice Hall, New Jersey.
Wiklund, M., Thurrot, C., Dumas, J., 1992. Does the fidelity of software prototypes affect the perception of usability? In: Proceedings of the Human Factors and Ergonomics Society 36th Annual Meeting: HFES1992, pp. 399–403.
Zajonc, R.B., 1965. Social facilitation. Science 149, 269–274.