Int. J. Human-Computer Studies 72 (2014) 552–566

Contents lists available at ScienceDirect

Int. J. Human-Computer Studies

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

Using sketches and storyboards to assess impact of age difference in user experience ☆

Giorgio Brajnik *, Cristina Giachin

Dipartimento di Matematica e Informatica, Università di Udine, Via delle Scienze 206, 33100 Udine, Italy

ARTICLE INFO

Article history:
Received 20 March 2013
Received in revised form
8 October 2013
Accepted 5 December 2013
Communicated by S. Wiedenbeck.
Available online 16 December 2013

Keywords:
User experience
Usability
Ambient assisted living
Older adults
Experimental evaluation
Prototypes and storyboards

ABSTRACT

We compared two versions of a touch-screen digital thermostat using a framework encompassing several user experience (UX) characteristics, and here describe how the implementation of certain design factors (specialists, praises, tooltips and increased interactivity) was done on mixed-fidelity prototypes of the user interface. We illustrate how the experimental comparison, involving 20 university students and 20 older adults, revealed important differences in UX, including perceived ease of use, behavioral intentions, enjoyment, quality, satisfaction, trust and usability, measured mainly through established questionnaires.

Analysis revealed that using that kind of artifacts is a very cost effective way to elicit interesting and useful results; many UX variables are significantly affected by design factors and by age differences, as expected; effects of design factors go well beyond usability and therefore could not be caught by running an investigation focused only on usability.

Age difference matters: older adults do not respond to addition of specialists, praises and tooltips as younger users do. We argue that potential benefits of these design choices are outweigh by the increase in complexity of the user interface.

From a methodological viewpoint we suggest using a particular array of UX characteristics and metrics when testing mixed-fidelity prototypes. Not all the metrics that we adopted were equally useful, and in particular perceived usability, subjective mental effort, and emotions did not help us highlighting differences.

1. Introduction

Aspects of human computer interaction in general and usability in particular have been identified as being important in the context of technology usage. However, as Hassenzahl and Tractinsky pointed out, from a historic perspective, usability focused too much on instrumental aspects and task orientation (Hassenzahl, 2004; Tractinsky, 1997). Experiential traits are oftentimes neglected, especially in contexts such as the living environment, where they could have a big influence on design of devices and user interfaces.

Smart home technology, and in general ambient assisted living (AAL), is seen as a promising defense against the threats of the demographic change. The potential of AAL to increase autonomy, health and safety, to promote socialization and inclusiveness, to cope with disabilities that older adults face is enormous. And it is in this context that user experience (UX) aspects become particularly important because they act as precursors of acceptability and actual usage of devices and user interfaces (Venkatesh et al., 2003).

Yet, UX aspects that go beyond usability cover a gamut of properties and metrics that is extremely varied and to some extent ephemeral: many properties are contextual, depending on the individual, his or her mood, the particular situation in which interaction takes place, the particular perspective under which interaction is considered, the goals and motivations that drive interaction, and of course characteristics of the device/user interface (Law et al., 2009). This makes it very difficult to mesh them into software development processes. To make UX evaluations more feasible and cost-effective, it is important to embed these practices into early stages of user interface development. It is necessary therefore to understand what kinds of design artifacts can be used, what kinds of design factors to implement, what UX aspects can and should be measured, and how.

In this paper we describe how the implementation of certain design factors (specialists, praises, tooltips and increased interactivity) was done on mixed-fidelity prototypes of a touch-screen thermostat, and how an experimental comparison, involving 20

☆This paper has been recommended for acceptance by Effie Lai-Chong Law.

* Corresponding author. Tel.: +39 0432 558445; fax: +39 0432 558499.
E-mail addresses: brajnik@uniud.it (G. Brajnik), cristina.giachin@gmail.com (C. Giachin).
URL: http://www.dimi.uniud.it/giorgio (G. Brajnik).

1071-5819/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.ijhcs.2013.12.005

university students and 20 older adults, revealed important differences in several UX aspects, including perceived ease of use, behavioral intentions, enjoyment, quality, satisfaction, trust and usability.

The contribution of this paper consists of

(i) The identification of a combination of design factors, design artifacts, range of UX aspects and metrics that allows one to elicit important differences. Results we obtained are significant, reliable and valid, providing strong evidence that the particular combination works well.
(ii) The different impact that the design factors we considered have on younger *vs.* older users: what works for younger people has not the same effect on older ones. While younger users rate the user interface that implements those design choices as being simpler to use and more enjoyable, older adults are slowed down, and rate it as more complex.
(iii) We show that this kind of UX evaluation technique can be adopted during early phases of product development, and results can be used to make important design decisions that go beyond usability and visual layout. Because UX metrics we used are mainly based on questionnaires (the exception being usability metrics), we argue that the technique can be easily adopted by mainstream designers, at least as much as "guerrilla usability" techniques can.

For these reasons we believe that the way in which we characterized UX is useful for engineering high quality user interfaces. Practical implications of these decisions could include more effective use of a heating system, with more satisfaction, less effort, more energy savings, and increased attractiveness.

2. Related work

In this section we summarize the research work that was done in the areas that are mostly relevant to our project, namely user experience, older adults and ambient assisted living.

2.1. User experience

User experience (UX) is a complex notion. According to Law et al. (2009) it can be defined as "the entire set of affects that is elicited by the interaction between a user and a product, including the degree to which all our senses are gratified (aesthetic experience), the meanings we attach to the product (experience of meanings) and the feelings and emotions that are elicited (emotional experience)". And furthermore "UX is the consequence of a user's internal state (*e.g.*, predispositions, expectations, needs, motivations, and mood), the characteristics of the designed system (*e.g.*, complexity, purpose, usability, and functionality) and the context within which the interaction occurs (*e.g.*, organizational/social setting, meaningfulness of the activity, voluntariness of use)".

Therefore, one could view UX as an umbrella concept that goes beyond usability and accessibility, embracing a range of properties that deal with many psychological, physiological and social human phenomena. In particular, in addition to usability/accessibility, UX covers at least the following aspects:

Aesthetics: Aesthetics has since long been an important attribute of devices and user interfaces, affecting how people feel and behave with respect to other beings or things. It was claimed (Tractinsky, 1997; Tractinsky et al., 2000; Lavie and Tractinsky, 2004) that there is a dependency between aesthetics and perceived usability (dubbed "halo effect of aesthetics"); one could argue that because aesthetics can be appraised during a quick first impression (Lindgaard et al., 2006), it taints other properties as well, including perceived usability. Subsequent studies, however, challenged or qualified better such a link: Law and Hornbæk (2007) highlight that in several studies dealing with usability and aesthetics there are issues with definitions and with measurements; Tuch et al. (2012) found strong experimental evidence that aesthetics does not affect perceived usability, but instead bad usability lowers aesthetics and hedonic attributes ratings.

Emotions: In his book, Norman (2003) discusses the notion of "emotional design" which is based on an underlying model of affect in HCI. According to the model, three levels of our nervous system (visceral, behavioral, reflective) are tightly coupled and intertwined within rich feedback loops that allow us to appraise situations from an affective point of view (*i.e.*, assigning arousal and valence value). This influences the affective state of a person, which in turn affects how the person thinks and behaves, which further influences how the situation is appraised, priming therefore a complex feedback loop.

Perceived usability: Before using a device or user interface, a user estimates the level of usability of the device on the basis of his/her experience, capabilities, abilities and more superficial qualities like aesthetics. Notice that this is different from objectively assessed usability, normally based on user performance indexes such as task completion time, success level and rate, number of errors (Rubin and Chisnell, 2008). The notion of perceived usability was studied, among others, by Hassenzahl (2004) who dubbed that "pragmatic attributes", *i.e.* connected to users' need to achieve goals.

Hedonic attributes: Hassenzahl introduced also attributes which refer to pleasure-related properties, and more specifically to growth (*i.e.*, how stimulating, novel and challenging a device is), to identification (*i.e.*, addressing the need to express one's self through objects), and to evocation (*i.e.*, the ability of the device to evoke memories).

Cognitive load: Closely related to pragmatic attributes and usability is the cognitive effort felt by users when trying to achieve goals. It is related with the complexity of task and user interface (Michailidou et al., 2008), as well as with how attention ebbs and flows between external stimuli and internal trains of thoughts (Varakin et al., 2004). It is also affected by how interruptions are handled, *i.e.* how the user who is involved in a primary task is notified of a pending secondary task, how s/he can be supported in switching context between the two, and in resuming the primary task (McCrickard et al., 2003).

Interactivity: Sundar and Kim (2005) discuss this notion and illustrate the wide range of definitions that could be used to characterize it: interactivity could be seen from a process point of view (emphasizing the conversation that occurs between the user and the system), or from the perspective of the range of different components that the user interface offers (basically the number of different widgets that are available and the granularity of user actions), or

from the point of view of how users perceive the flow of information (one-way or two-way exchange of information, personalized messages from the system). Regardless of how it is conceptualized, it considerably impacts on how the user reacts and how his/her affective state changes. For example, more interactive political web sites lead to better qualities being attributed to political parties. Wensveen et al. (2004) discuss how the links between user actions and their effects can be made stronger, and suggest to consider six coupling dimensions (time, location, direction, dynamics, modality, and expressivity) along which to anchor appropriate feedback and feedforward information.

Social responses: Reeves and Nass (1996) discuss a number of factors that may affect the quality of the user–system interaction. Relevant factors include whether the system behaves in a polite manner (*e.g.*, by greeting the user upon login), whether it flatters the user, and whether it adopts a specialist role. In these cases the system is perceived as being more valid, more trustful, more friendly, and one's own work performed with that system is also seen in a better light.

Persuasion: There are several models of persuasiveness of user interface elements. Cialdini and Martin (2004) and Fogg (2003) provide different viewpoints on the basic principles of persuasion and on what techniques could be effectively be implemented in user interfaces. The Elaboration Likelihood Model developed for studying advertising (Cho, 1999) is interesting because it connects to user attention. It posits that ads may persuade by following either a central route, applicable when the user is motivated and able to perceptually and cognitively process the ad, or a peripheral one, when either motivation or ability is missing. In the former case it is the content of the ad that counts, while in the latter case peripheral and secondary aspects (like colors, sounds, aesthetics, humor) are determinant of the persuasiveness of the ad. Thus, for users that are tired or not focused on the task, one may assume that peripheral aspects might catch user's attention. Several other aspects and theories of advert design are relevant for UX; see a review in Brajnik and Gabrielli (2010).

Acceptability: The Technology Acceptance Model (TAM) (Davis, 1989) assumes that perceived usefulness and perceived ease of use are determinants of user acceptance. Later on other models subsumed it, by encompassing also other variables such as experience, behavioral intention, performance and effort expectancy, amongst others (Venkatesh et al., 2003). Acceptability is particularly important because it is a determinant of actual adoption and usage.

In Section 3 we describe the particular subset of UX attributes that were considered in the experiment.

2.2. Older adults

Older adults (age 65+) is the fastest growing segment of the population in many countries (United Nations, 2010). Unfortunately, regarding ICT many of them can be considered *digital immigrants*, in the sense that they need to learn a new culture, typical of *digital natives* (Fozard et al., 2009). While the potential for older adults of effectively using ICT is enormous, ranging from increased and prolonged autonomy to managing health, from communicating to being informed, from handling bureaucracy to playing games, with aging and especially being above the 65 threshold, a number of disabilities with different degrees ensue, with the consequence of reducing acceptability and motivation to use such technologies.

Many types of barriers may arise (Arch, 2008; Lunn et al., 2009; Zaphiris et al., 2007). The following is a brief taxonomy.

Perceptual barriers: They may ensue due to the natural decline of visual and auditory capabilities, with the consequence of making the tasks of perceiving information and of being aware of the external environment, more difficult, more error prone and slower.

Motor barriers: They may arise because of declines in the ability to finely control movement of body parts, especially hands and fingers. The consequence is again reduced effectiveness and efficiency, especially when interacting with virtual or mechanical widgets (like mice, hard or soft buttons, and tree views).

Cognitive barriers: They may be due to declines in mental abilities, regarding memory, attention span, information processing speed, and reading abilities. The consequence is a reduced ability to make sense of what may be perceived and of the orienting response, to learn new concepts or new procedures, to compare outcomes of different processes, to carry out complex and long tasks, and to cope with information overload.

Cultural barriers: They concern lack of experience in the domain, about devices, or about the language. One possible partial explanation is that older adults have grown up in an age where electromechanical devices were used, and the switch to digital ones requires relearning the basics.

Emotional barriers: They concern attitudes and motivations. In many cases researchers observed that older adults fear breaking the devices they use, they feel unconfident, they show apprehension when trying out new tasks, they prefer reading and reflecting before trying out things, they feel unease in developing new mental models, task errors due to slips or mistakes have a stronger psychological effect than in younger users, they do not like to be stigmatized by using "specialized" devices (Coughlin et al., 2007; Lunn et al., 2009; Leung et al., 2012).

In many studies these barriers interact. Lunn et al. (2009) list a number of potential accessibility barriers of the Web. These range from perception related ones (like low visual contrast, moving content, and text size being too small), to barriers related with operability of web user interfaces (buttons and links being too small, too close to each other, and cascading menus), to those related with understanding (text of links being too poor a description, usage of technical jargon like "close window", few orientation clues, inability to remember information and data, inconsistent usage of terms, and preference of text over icons). Sometimes these barriers interact making it difficult to decide on possible solutions: for example, the typical remedy for decreased visual acuity is to increase the text size of the user interface; however this reduces the amount of information that is shown, posing additional requirements to a possibly already overloaded

cognitive system; in many cases the benefits of increasing the text size are outweigh by the disadvantages of having to remember the parts that are no more visible.

In a usability study, Hollinworth and Hwang (2009) found that even in simple word processing tasks, older adults struggled because of a poor understanding of the conceptual model underlying the user interface; upon forgetting what was the correct sequence of actions, they had to place a greater reliance on visual cues to get to know what were the preconditions in order to perform a certain action; and interpretation of icons or words used as visual cues was problematic. Likewise, in an eye-tracking usability study of an iTV application, Obrist et al. (2007) found elderly users took longer to find the UI components to focus on, were less effective in searching for information, had more difficulties in interpreting navigation bars, and were less effective in learning how to interact than younger users.

Technologies that could more easily address some of these issues are multimodal user interfaces; for instance, Fairweather et al. (2007) discuss the usage of vocal interfaces (speech input and output), highlighting the fact that "speech is natural", compared with writing or typing; in some cases speech can eliminate the need to memorize associations between required actions and visual cues, helping to pursue a given sequence of actions in the correct order, etc. However there is also an additional cognitive load when speech-related tasks have to be added to things that users must perform. Speech interfaces also interfere with performance when older adults cannot stop talking after giving a vocal command, which happens relatively often. Finally, the paper discusses the paradox where experts in the performance of a task may be unable to perform it because a speech interface demands the use of particular terms, which a procedural expert might have forgotten or never known.

An ethnographic field study (Sayago and Blat, 2009) reported that the barriers that hindered older adults in using the web are not related to the web itself nor user interface technologies, experience with the web or educational level. Instead the main difficulties were in remembering task-related steps, understanding technical terms, and using pointing devices (specifically the mouse). Sayago and Blat emphasize that the drivers for using the web were mainly independence (the ability to use the technology/device by themselves) and inclusiveness (the aspiration to interact with technology/devices by using the same user interface that others use, so not to give the impression that they need special assistance). Other drivers for using the web were not to lag behind society, to remain close to family and friends, to enjoy the opportunity of learning that they did not have in their childhood.

Older users do not take advantage of the help that younger family members could provide because often such a help is given hastily, too concisely, using terms that are unfamiliar, without slow-paced practice, which leads to incomplete, fragmented and possibly incorrect learning results. Another consequence of this sort of help is diminishing confidence in their own abilities. The approach to learning described in Forbes et al. (2009) is based on informal learning practices, occurring mainly within a group of older adults that form a social network, enriched with helpers that focus on learning of practical topics, small chunks at a time, using repetition to consolidate knowledge.

Leung et al. (2012) illustrate and demonstrate the kind of difficulties faced by and preferences that older adults have regarding learning how to use mobile devices. They show that older adults withstand adoption of a trial-and-error approach to using devices and user interfaces; one plausible reason is that trial-and-error poses significant requirements on cognitive abilities (one has to remember the outcome of a trial when discovering and pursuing an alternative one). This is unfortunate because trial-and-error is the most effective approach for learning how to use existing devices and user interfaces. Low self-efficacy, which may ensue because of high error rate, difficult perception, and increased complexity of tasks, decreases one's motivation to learn and to keep using the device. Trust in the system is also likely to decrease. Another finding of that study is that older adults prefer written step-by-step help messages to other sorts of help material. A review of several studies shows that it remains unclear whether ICT (Information and Communication Technologies) use can improve cognition and well-being in older adults, even though positive results were found for older people with depression (van der Wardt et al., 2012).

It can be seen that many of these issues manifest themselves along one or more of the UX dimensions discussed above and go beyond mere usability and accessibility.

2.3. Ambient assisted living and digital thermostats

Nowhere else higher stakes for involving older adults with ICT can be found than in Ambient Assisted Living (AAL): activities supported by AAL fall into the three broad categories of comfort, autonomy enhancement and emergency assistance. Monitoring persons for health related purposes, supporting them in quickly getting in touch with somebody for emergency reasons or easing daily tasks are a few examples of what could be achieved with AAL technology. The expected benefits are dramatic improvements in safety, in health, in one's independence, inclusivity and self-esteem (Coughlin et al., 2007).

Several research projects have been carried out in the recent years; however, most of them only consider one or a few of the dimensions potentially relevant for AAL. The hypothesis that AAL is a multidimensional challenge is, for example, supported by the study reported by Coughlin et al. (2007), based on focus group which showed that older adults identify issues related to four large themes: technology design (addressing usability, reliability, functionality), ethics (privacy, trust, loss of dignity while being at home), affective reaction (including pros like independence, but risks of stigma, of devices being designed for the "oldest old" or the frailest ones), and market-related issues (cost, availability of devices, affordability, public policies). These four themes are tightly connected to acceptability aspects.

Usability issues and incorrect mental models play a particularly important role for heating-ventilation-air-conditioning (HVAC) systems. The survey by Meier (2011) shows that many people (regardless of their age) hold many misconceptions about their own HVAC (*e.g.*, that a thermostat works as a proportional controller, that it can be used as an on/off switch for turning on/off production of heat or cold, rather than by modulating the temperature). These greatly reduce the effectiveness of the system. In turn, this is going to negatively affect self-efficacy, confidence and frustration, especially because it impacts home comfort and energy consumption. In the end also one's perception of usefulness of the thermostat could be affected.

We see that also in the context of HVAC (and more generally of AAL), UX aspects should play a crucial role towards acceptance and actual use of the system by older adults, making it an ideal ground for experimenting how to assess UX properties, despite it being a kind of applications where more ludic aspects play a minor role than effectiveness in task completion.

3. Research goals

The goal of this research is threefold. First, in the context of AAL we aim at an experimental validation of the effectiveness of design factors that are expected to bear upon UX, usability and acceptability. Second, we want to see if age plays an important role in the

effectiveness of such factors. And third, if adoption of low-fidelity prototypes (such as sketches and storyboards) as a means for implementing potentially useful UX factors is a viable technique for eliciting such an information.

The system considered in the experiment is a HVAC controller and digital thermostat, called Otouch, manufactured and sold by two local companies.[1] The system consists of a touch screen (7″ × 4.5″, 700 × 480) that embeds a Linux system and is able to acquire data from sensors and to control pumps, valves and heating or air conditioning units that are located in different rooms of a building, such as a large house, a school or hotel. The touch screen behaves as a remote thermostat for individual devices, allowing the user to supervise temperature, humidity and ventilation in any room, to change setpoints, to define daily, weekly and yearly programs, as well as ad-hoc programs to be used in special occasions, like vacations.

The user interface which we manipulated was the outcome of a user-centered design process, where usability investigations and redesigns already occurred. Therefore, we decided to focus on design aspects that could bear upon UX aspects. The rationale was that, especially because such a system is expected to be used by individuals at home when they are alone and cannot ask for support by anybody else, acceptability could be improved by going beyond usability and accessibility. This should be even more important for older adults.

The factors we manipulated are concerned with ways in which help and guidance information can be provided, and more specifically we worked on social responses and in part on interactivity (as described in Section 2.1): we tested whether the addition of specialists to provide certain kinds of information could improve interaction and whether providing tooltips and praises at particular points in the interaction could improve the interaction.

During the experiment we collected information in order to be able to assess the effects of these factors. There is a great deal of methods that can be used to gather UX data, based on questionnaires, on qualitative investigation methods and on physiological variables (*e.g.*, skin conductance) (Vermeeren et al., 2010). In this experiment, through questionnaires, we captured pragmatic and hedonic attributes, emotional responses, mental effort, behavioral intentions and acceptability, quality assessments, both before and after use of the prototype system; in addition, usability metrics were used. We believe that these UX dimensions and these ways to measure them are well catered to investigations based on low-fidelity prototypes. We did not address aesthetics (low fidelity prototypes are certainly not well suited for that); we did not address other kinds of social responses (such as team membership) because in the context of a thermostat they appear to be awkward; interactivity viewed as an exchange of information is also not well suited for a thermostat. Persuasion and politeness could have been considered, but were not because we wanted to focus on a few factors; they will be the subject of additional work in the future.

As a consequence we tackled the following research questions, with respect to the particular way UX was operationalized (see Section 4.3):

1. Do the design factors we considered have any effect on usability and UX?
2. Can these effects be detected by deploying sketches and storyboards?
3. Is there an effect due to age?
4. Are there interactions with age, so that preferences and responses of older adults differ from those of younger users?
5. Can these interactions be detected using sketches and storyboards?
6. Which UX dimensions are more or less affected?
7. Can these differences be detected with sketches and storyboards?

The potential implications include (i) suggestions on how to design a user interface for a thermostat (but potentially other AAL devices), and especially the functionalities for providing help, so that they are well suited for digital savvy users and, perhaps through personalization, for older adults. More light could be shed on what key differences in user interfaces would make them suitable for either one of these segments. (ii) These design solutions could improve the ease with which older adults learn to use this kind of technology, and they could reduce the impact of emotional, cultural and cognitive barriers. (iii) These solutions could improve acceptability, and hence motivation to use, especially for older adults, in an area (AAL) that holds many potential advantages for them.

4. Experiment

4.1. Material

The experiment was based on two versions of the thermostat user interface. One version, dubbed Rosy Raisin (RR), was based on screenshots of the actual running version of the thermostat; the second one, Smokey Brown (SB), was based on a modified version of those screenshots. Changes included

1. The addition of specialists for delivering appropriate information. Each specialists was given a name (for example, “Maurizio” was the specialist dealing with tasks of programming the thermostat, “Renzo” was in charge with issues regarding regulating tasks – such as manually changing the setpoint in some room, “Edoardo” was in charge with energy consumption, and “Manuela” was a “receptionist” in charge of handling coordination between specialists). Each specialist was shown in some screens of the UI via a human figure together with its name, when appropriate messages were intended to be shown.
2. Help information in SB was mediated by Manuela and delivered by one of the other specialists. The same information was available in RR but upon clicking the appropriate help button (a question mark icon on the top of the screen). See Fig. 1.
3. Energy consumption data and help information in SB was delivered by specialists with buttons to scroll to the next/previous item, whereas the same information in RR was delivered simply through a menu of choices, each entry leading to a screen containing the details. This was done with the purpose of making SB slightly more interactive, in the sense of granularity of user actions. See Fig. 2.
4. During complex tasks, when the user interface offered many options and especially when during pilot tests it was discovered that users often failed, specialists for regulation and for programming delivered short tooltips, located in the footer of the screen. See Fig. 3.
5. Specialists praised users, 2 times for more complex tasks, once for simpler ones; see Fig. 4. Examples of praises are (i) “You just turned on the devices. Congratulations!”, which occurred upon successful completion of the task of turning on/off one or more devices (complex task). (ii) “You modified the scenario successfully. Very good, many others weren't able to do that!”, at the end of the task of changing some aspect of a stored scenario (complex task). (iii) “Bravo! You just created a new scenario without making mistakes.”, after creation of a scenario (complex task). (iv) “PIN was modified correctly!”, at the end of

[1] Eurapo SrL http://www.eurapo.it and Aragon Engineering SrL http://www.aragon.it.

Fig. 1. Screenshots of the user interface; version RR on the left and SB to the right. These screenshots show how help information is delivered in the two versions, and the specialist called Renzo.

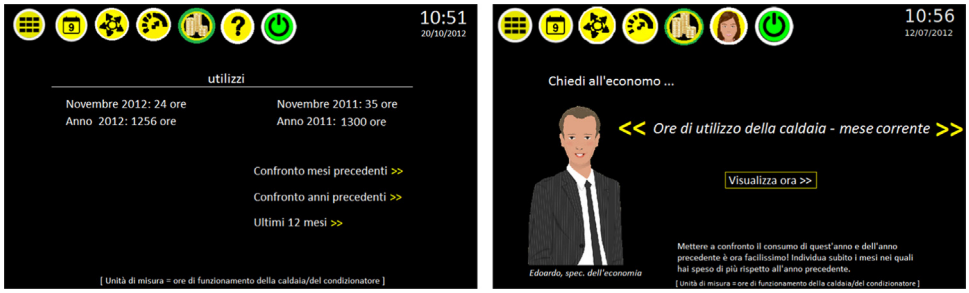

Fig. 2. These screenshots show how information about energy consumption are delivered. On the left (RR) via a menu with three options; on the right (SB) through the specialist Edoardo and scrolling buttons.

Fig. 3. Screenshots without tooltips (left, version RR) and with them (right, SB) delivered through the specialist Maurizio, in this case during the task of activating an already defined scenario.

Fig. 4. Screenshots with a praise delivered by Renzo and Manuela (in SB, right) and corresponding feedback screenshot in RR, without praises (left). The praise says "Comfort boost is active. Nature appreciates it! It's been a pleasure to collaborate with you.". And below that "Hi. I'm Manuela. Would you like to see energy consumption for this month? Shall we modify temperature for a number of rooms? Shall we add a new season program?".

resetting the PIN (simple task). (v) "Manual setting successfully defined.", at the end of choosing an ad-hoc setting for a room (simple task).

Praises are shown in interstitial screens, in many cases just after the final confirmation of a task but before reaching the screen that provides the feedback information. In the original version (RR) there are no such interstitial screens, and upon completing a task the subsequent screen is the one providing the feedback.

The two versions of the user interface were implemented as clickable storyboards in PDF[2] and shown on a laptop. Because we were not interested in eliciting data about affordance of controls, and because only a few of the items in the screen were clickable, the facilitator interacted directly with the laptop, following directions given by each of the participants.

[2] We used WireframeSketcher to develop them; http://wireframesketcher.com.

Task	Difficulty	Formulation
T1	easy	Change the setting of the room you are in to 21C°.
T2	easy	Turn off the heating system.
T3	medium	Add a new scenario that should start on Aug. 1, end on Aug. 20, and apply it to the family room, living room and bedroom.
T4	medium	Add a new season program, with specific temperatures for daily schedules between midnight and 8 AM, between 8 AM and noon, and between noon and midnight.
T5	difficult	Increase the temperature in each of the rooms by 3C° (use a shortcut – avoid "going" into each single room).
T6	difficult	Last March 2012, what was the system doing most of the time? Heating, air-conditioning or dehumidifying?

Fig. 5. List of tasks used in the second phase of the experiment (the original formulation is in Italian and provides also a 1-sentence description of a scenario).

Group	Version	T1	T2	T3	T4	T5	T6
A	RR	✓		✓		✓	
	SB		✓		✓		✓
B	RR		✓		✓		✓
	SB	✓		✓		✓	
C	RR		✓		✓	✓	
	SB	✓		✓			✓
D	RR	✓		✓			✓
	SB		✓		✓	✓	

Fig. 6. How different versions and tasks were allocated to four different groups of participants.

The experiment was structured in two phases. First, each participant was greeted and welcomed, asked to fill-in a short demographic questionnaire, and given a 5 min guided tour of the system, with the facilitator demonstrating how different activities could be carried out. The facilitator answered also any possible question regarding the user interface. At the end of this demonstration, the participant was given the pre-test questionnaire on paper (Appendix A). These two steps were repeated for the other version; filling in a second copy of the questionnaire completed the first phase. The order of the two versions was randomized.

In the second phase the participant was asked to perform six tasks, three for each version. Tasks were paired according to difficulty level (2 were easy ones, 2 difficult ones and 2 were in the middle – determined a-priori during pilot tests), and each participant carried out all of them; tasks are summarized in Fig. 5. To ensure counterbalancing learning, carryover and fatigue effects, participants were randomly classified into four groups according to table in Fig. 6; tasks were carried out in increasing order of difficulty.

No think-aloud was adopted. Participants were asked to perform three tasks with one version, then fill-in the previously used questionnaire (to make it explicit the fact that they were perhaps changing their opinion, compared to phase 1); they were then asked to fill-in a second questionnaire (Appendix B). After a quick break, the same procedure was repeated for three tasks on the other version.

Finally, a debriefing questionnaire was used (Appendix C). The entire session lasted less than a hour. At the end each participant was thanked and greeted. No compensation was given to participants.

4.2. Participants

Twenty university students (of different subjects) were recruited through personal contacts, and twenty older adults were recruited through the help of a local nonprofit organization. It was required that participants had a basic experience in using computers, derived from having attended specific classes and/or using computers at home or work. In both cases they were told that they would voluntarily contribute to a scientific experiment and that they could withdraw from the study at any moment, with no consequence what so ever.

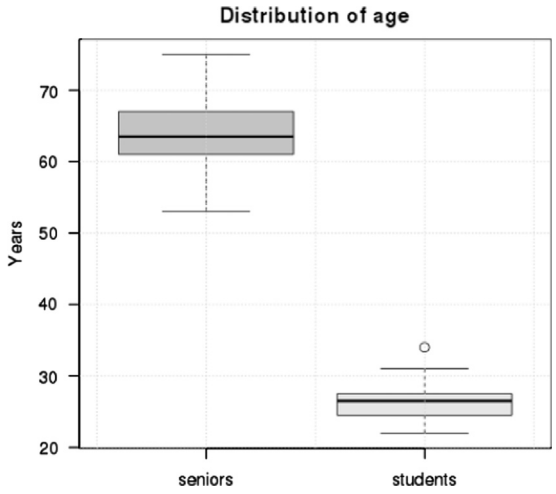

Fig. 7. Age distribution.

None did so. No student was involved in any class taught by the first author of this paper.

4.3. Independent and dependent variables

In order to tackle the research questions listed above (Section 3), the following independent variables were used:

User category: Which is related to participants age. *Students* were recruited from within the university, and their age ranged from 20 to 40; *seniors* were recruited from outside the university and their age ranged from 50 to 80. The two levels of the user category factor were determined both by how we recruited participants (within/outside the university) and their age; we chose 50 as a splitting threshold because it clearly separates participants (see Fig. 7).

Version: Of the user interface. One version (*RR*) was based on screenshots taken from the user interface of the actual system; the other one (*SB*) was based on screenshots that were changed so to implement the design factors discussed above. This is a within-subjects factor.

Phase: Some of the dependent variables refer to phase one (before use, *pre-test*), while others to phase two (after use, *post-test*). This is also a within-subjects factor.

Dependent variables included (see Appendix A for details):

Perceived ease of use: (PEOU), based on the scale reported by Davis and Venkatesh (1996) articulated on three 5-point Likert questions. This was used for both versions and in both phases. This variable belongs to the Perceived Usability aspect discussed in Section 2.1.

Behavioral intention: (BI), based on the scale used by Sundar and Kim (2005), two 5-point Likert questions. Also this variable was used for both versions and both phases. This variable belongs to Persuasion (Section 2.1).

Perceived usability: (PU), based on the scale reported by Davis and Venkatesh (1996) articulated on two 5-point Likert questions. This was used for both versions and in both phases. This variable belongs to Perceived Usability (Section 2.1).

Emotions: based on a subset of the scale discussed by Pekrun et al. (2004); for both versions, three

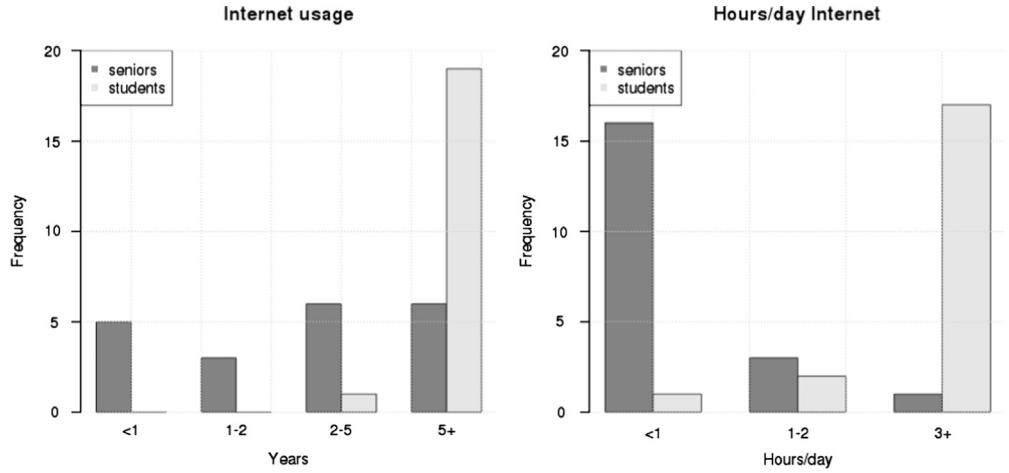

Fig. 8. Experience in using PCs and Internet.

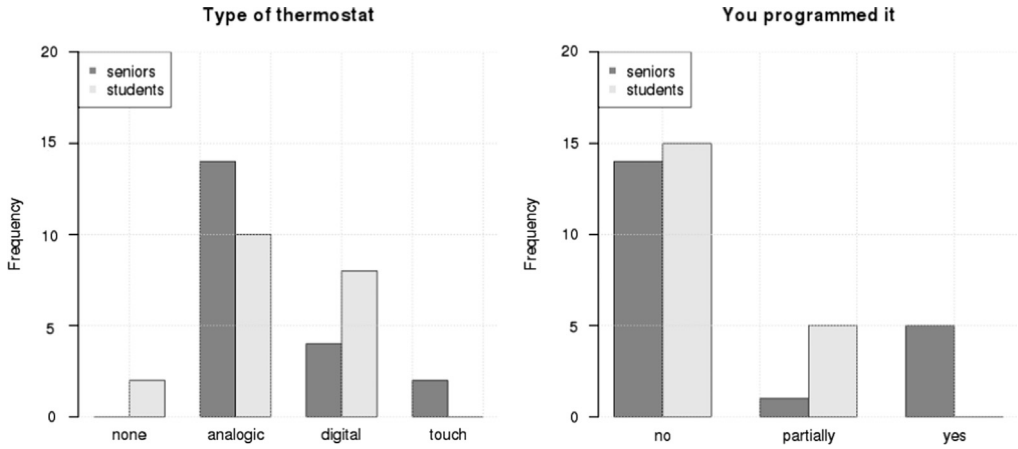

Fig. 9. Type of thermostat and who programs it.

5-point Likert questions were used in phase 1 and other three questions for phase 2.

Enjoyment: based on six adjectives (see Appendix A), used for both versions and in both phases. This variable belongs to Emotions (Section 2.1).

Subjective mental effort: based on SMEQ (Sauro and Dumas, 2009), for both versions and used only in the second phase. This variable belongs to Cognitive Load (Section 2.1).

Quality: based on a subset of the AttrakDiff questionnaire used by Hassenzahl (2004), covering pragmatic and hedonic attributes (23 adjectives in a semantic differential format). This was used for both versions and only on phase 2. This variable belongs to Perceived Usability, Hedonic Attributes and Acceptability (Section 2.1).

Overall satisfaction: based on the WAMMI questionnaire (Kirakowski et al., 1998), 18 5-point Likert questions, used for both versions and only on phase 2. This variable belongs to Hedonic Attributes and Acceptability (Section 2.1).

Trust: based on nine 5-point Likert questions derived from Lewis (1993), used for both versions and only on phase 2. This variable belongs to Social Responses (Section 2.1).

Usability: based on data collected during the second phase consisting of completion time, success level (binary), number of errors made (errors included clicking on an icon/button that does not get closer to the solution, or trying to use a functionality for an unrelated purpose), and number of undos/back buttons that were requested.

5. Results

5.1. Demographic data

Out of 20 students, 12 were female; out of 20 older adults, 6 were female. Mean age for students was 26.6 (sd=2.95), for older adults it was 63.8 (sd=5.05); see Fig. 7 for the distribution.

Experience of using computers and Internet varied across categories; by large, students had 5 or more years of experience and used them for more than 3 hours per day. Conversely, older adults were more equally distributed among those that use them since 5 or more years (6), 2–5 years (6), 1–2 years (3), less than 1 year (5). Most of them (16) use them for less than 1 h per day, three use it between 1 and 2 h per day. Fig. 8 shows a comparison. User category and years of experience are obviously associated ($\chi^2 = 18.33$, $p < 0.001$), as are category and hours per week ($\chi^2 = 27.66$, $p < 0.001$).

Fig. 9 shows the distribution of the kind of thermostats at home across our sample of participants, and who at home is in charge of programming them. We see a predominance of analogical thermostats, for both categories (14 for older adults and 10 for students); 4 older adults and 8 students use digital thermostats, of which

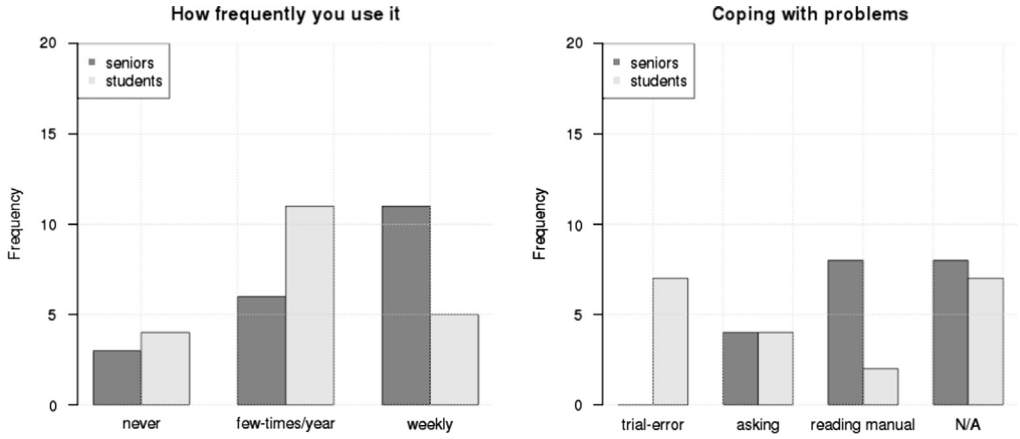

Fig. 10. Frequency of use and how to cope with usage problems.

	PEOU	Enjoyment	BI	PUs	Emotions (phase 1)
N items	3	6	2	2	3
Pre-test	0.88	0.84	0.83	0.72	0.74
Post-test	0.82	0.84	0.91	0.75	0.80
	Quality	Satisfaction	Trust		Emotions (phase 2)
N items	22	18	9		3
Post-test	0.94	0.91	0.90		0.66

Fig. 11. Cronbach α for the different scales that were used at the end of phases 1 and 2.

2 were touch-based thermostats for students. There is no significant association of category with type of thermostat. Regarding who programs it, 14 older adults and 15 students say they do not do it, 1 and 5 say they do it sometimes, 5 older adults and 0 students say they are in charge of that. These two aspects are significantly associated ($\chi^2 = 7.7,\ p = 0.022$).

Fig. 10 shows data regarding usage of thermostats. Four students never use it, compared to 3 older adults; 11 *vs.* 6 use it a few times a year; 5 *vs.* 11 use it weekly. No significant association found. Regarding how they cope with problems that may arise when using the thermostat, 7 students and 0 older adults follow a trial and error strategy; 4 and 4 ask somebody; 2 *vs.* 8 read the manual; the others did not respond. As expected these two aspects are significantly associated ($\chi^2 = 10.67,\ p = 0.014$).

5.2. Validity and reliability of data

No collected data point was invalid.

Regarding manipulation checks, ANOVA of completion time as a response variable shows a main effect of the difficulty level, a 3-valued factor (low/medium/high) ($F = 46.71, p < 0.001$) and no significant effect nor interaction with individual tasks. Similarly, for number of errors. There is a significant association between success level and task difficulty ($\chi^2 = 13.23, p = 0.0003$). Thus, our a-priori classification of tasks according to difficulty level appears to be valid; therefore the way in which praises and tooltips were distributed across different screens should be effective.

We found significant effects of the order with which tasks were executed on time, errors and undos; no association was found between order and success rate. This means that there might have been some habituation and learning effects taking place, even though they did not affect success rate. In any case, the randomization of tasks counterbalanced this.

Regarding reliability, Fig. 11 shows the Cronbach coefficients. Apart from the second scale for emotions, all α's are greater than 0.70 indicating a moderate-to-good reliability.

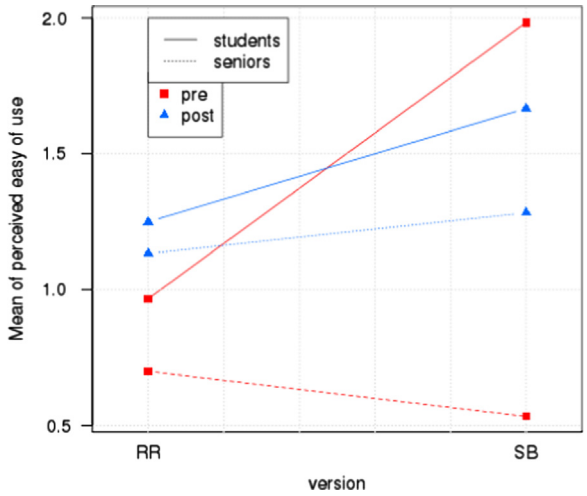

Fig. 12. Differences in means of PEOU, across versions, user categories and test phases. Notice the scale to the left, restricted to [0.5, 2.0].

Although PEOU is not normally distributed, application of the Bartlett tests for homogeneity of variances on PEOU with respect to version and category shows no significant differences, for either phase. Similarly for enjoyment, behavioral intentions, perceived usability, emotions, quality, satisfaction and trust. For these reasons, in the analysis reported below, for computing the ANOVA tables we used a mixed random effects linear model with `subject` as a grouping factor, more conservative than a repeated measures two/three way ANOVA.

5.3. Dependent variables

For the variable *perceived ease of use* (PEOU), we explored main effects and interactions of *category* (2 levels, `students` and `senior`), *version* (2 levels, `RR` and `SB`) and *phase* (2 levels, `pre` and `post`) on the response. We found significant main effects of version ($F = 6.5,\ p = 0.012$), of phase ($F = 4.30,\ p = 0.04$), a marginally significant main effect of category ($F = 4.06,\ p = 0.051$), significant interactions of version with category ($F = 6.83,\ p = 0.01$) and of category with phase ($F = 4.81,\ p = 0.03$). Version SB showed a smaller mean that RR ($M_{RR} - M_{SB} = -0.008$, in a scale between $[-3, 3]$; that is 0.1% of the total range – obviously negligible); performing the tasks has the effect of increasing PEOU (the mean changes by 0.59, or 9.9% of the scale); for students, PEOU decreased by 0.11 (1.9%); compared to RR, when using SB, students increase PEOU by 0.73 (12%) more than older adults; after performing the

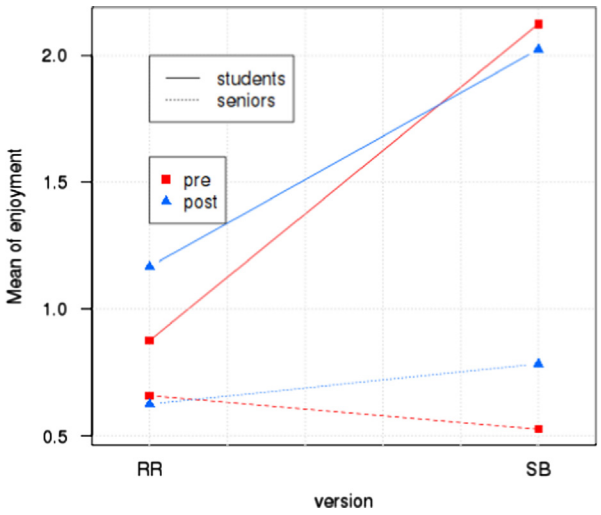

Fig. 13. Differences in means of enjoyment, across versions, user categories and test phases.

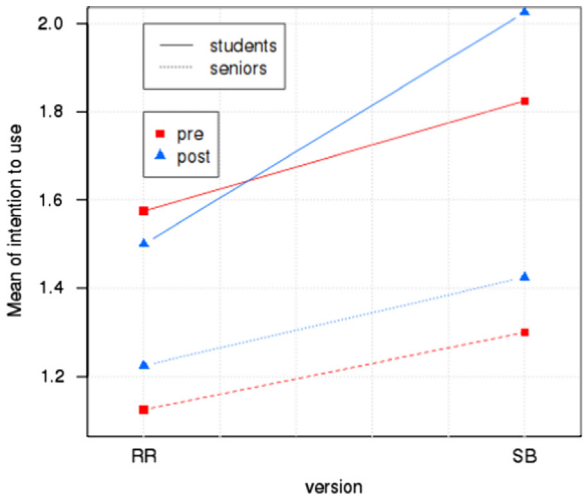

Fig. 14. Differences in means of behavioral intentions.

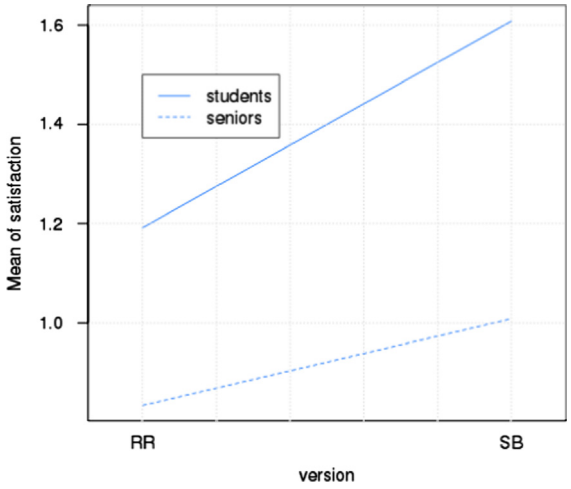

Fig. 15. Differences in means of satisfaction.

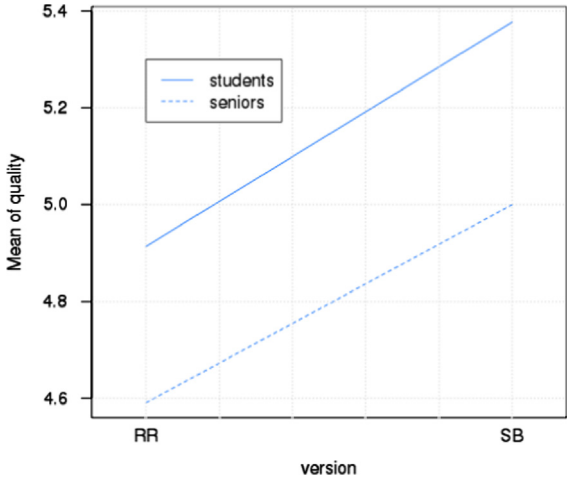

Fig. 16. Differences in means of quality.

tasks, for older adults PEOU shows a decrease of 0.61 (10.1%) compared to students. See Fig. 12.

Perceived usability is not significantly affected by version, category or phase, and no interaction is revealed.

Regarding enjoyment, we found a significant main effect of version ($F = 35.21,\ p < 0.001$), a significant main effect of category ($F = 10.62,\ p < 0.003$), and a significant interaction of version and category ($F = 33.58,\ p < 0.001$). SB increases enjoyment by 0.01 (0.2% – negligible); enjoyment for students increases by 0.38 (6.3%) compared to older adults; compared to RR, when using SB enjoyment for students increases even more, by 1.04 (17.3%), compared to older adults. No effect nor interactions were found for phase. See Fig. 13.

No significant main effect nor interactions among version, category or phase were found with emotions as a response variable.

Behavioral intentions are significantly affected only by version ($F = 8.52,\ p < 0.005$); no other main effects nor interactions were found. Compared to RR, for SB the mean increases by 0.28 (4.8%), from 1.36 to 1.64. See Fig. 14.

Satisfaction questions were asked only after the second phase, and therefore they were analyzed only as a response variable to version and category. There is only a main effect due to version ($F = 7.07,\ p < 0.012$); the mean increases by 0.30 (4.9%), from 1.01 for RR to 1.31 for SB. See Fig. 15.

Quality was also asked only after the second phase. Only a significant main effect of version was found ($F = 16.97,\ p < 0.0001$), with quality increasing from RR to SB by 0.44 (7.3%). See Fig. 16.

Trust was found to be only marginally affected by version ($F = 4.076,\ p = 0.05$), increasing from RR ($M = 1.22$, sd $= 0.99$) to SB ($M = 1.47$, sd $= 0.89$); a difference of 0.25 (4.2%).

No significant effect nor interactions were found for subjective mental effort.

One usability metric is completion time. There is a significant main effect of category on time ($F = 32.89,\ p < 0.0001$), while version has only a marginal effect ($F = 3.23,\ p = 0.07$); no interactions found. On average it took longer completing tasks to older adults ($M = 2.43,\ \text{sd} = 2.00$ min) than it took students ($M = 1.13,\ \text{sd} = 1.15$): a difference of 1.20 min. (a relative increase by 106.2%). Subjects were faster with RR ($M = 1.60,\ \text{sd} = 1.71$ min) than with SB ($M = 1.97,\ \text{sd} = 1.78$): SB slowed down subjects by 0.37 min (a relative increase of time by 23.1%).

Category has a significant main effect on errors ($F = 9.82,\ p = 0.003$); no other effects nor interactions were found. On average, there were more errors with older adults ($M =$

Table 1
Number of participants who preferred RR or SB.

Version	Older adults	Students	Sum
RR	6	5	11
SB	14	15	29
Sum	20	20	40

Table 2
Number of participants that gave a certain rank to each of the factor (older adults–students).

Rank	Tooltips	Praises	Specialists
Best = 1	6–4	0–2	4–**8**
2	**7–9**	1–0	**6**–4
3	5–7	2–2	**6**–2
4	2–0	**9**–4	2–2
Worst = 5	0–0	8–**12**	2–2
Sum	20–20	20–20	20–20

1.73, sd = 1.86) than with students ($M = 1.0$, sd = 1.44); an increase of 0.73 (in relative terms, 73%).

No significant main effect nor interactions on the number of undos were revealed.

While success rate and version are independent (χ^2 test), success rate and category are associated ($\chi^2 = 13.23$, $p < 0.001$): out of 120 tasks performed by students and by older adults, successful ones are 111 (92%) and 89 (74%), respectively.

5.4. Qualitative results

Table 1 shows how many participants preferred one or the other version of the prototype. There is no association between the two variables, but the table reveals a strong significant preference for SB (72.5%, $\chi^2 = 7.22$, $p = 0.007$).

In their comments participants said RR is minimalist (4 of them), clear (6), simple (8), clean (5), but also missing feedback (3) and missing help messages (11). They liked SB having specialists (10), help (10), rich communication with users (5), but they also indicated that there is a lot of information (8) and that it is impossible to disable help (4).

Table 2 shows the number of participants who ranked 1 (best) to 5 (worst) three design factors: tooltips, praises and specialists. There is no significant association between category and either of these rankings. The table reveals however that while tooltips and specialists were ranked high by the majority of the people, praises were ranked low by the majority.

For tooltips participants had only positive remarks: tooltips assist in carrying out tasks, they are easy to read, they are clear, they are necessary to continue operating the device, people feel more confident because of them.

Regarding praises, participants noticed that positive feedback reassures and encourages users, praises provide trust. On the other hand, praises are annoying and require closing windows, they did not help at all, people do not need that somebody tells them the task is completed, because a machine has no feelings then people do not need to be encouraged, it looks like making fool of the user.

Turning now to specialists, participants said: it is reassuring to know that somebody can help you, specialists helped carrying out tasks, they helped avoiding errors, without them somebody would not have completed tasks. On the other hand, participants found specialists to be distracting and invasive.

6. Discussion

Regarding whether design factors could have significant effects on usability and UX, results show that version has an effect on PEOU (significant albeit negligible); version interacts with category with respect to PEOU (students when using SB increase PEOU by 12% of the scale); version affects enjoyment (though negligibly); version interacts with category with respect to enjoyment too (students when using SB increase enjoyment by 17.5% compared to older adults); SB increases behavioral intentions (by 4.8%), overall satisfaction (by 4.9%), quality (by 7.3%), trust (marginally significant, with SB better by 4.2%), completion time (marginally significant, with RR faster by 23.1%). On the other hand, version did not affect perceived usability, emotions, subjective mental effort, errors, undos, success rate. For emotions, because of the relative low reliability, individual items were also analyzed, to no avail.

Participants clearly noticed the differences between RR and SB, expressing a strong preference for SB (72.3%) over RR, and preferring tooltips and specialists over praises.

One first conclusion is that design factors materialized using sketches and storyboards can be used to identify important UX (and usability) differences. The study by Newman and Landay (2000) reveals that several types of artifacts are used by web designers in early stages of the development process, including sketches, storyboards, site maps, wireframes, page mock-ups, and interactive prototypes. Consider also that the space of interactive prototypes and storyboards is quite structured. In fact, according to fidelity with respect to final products, *mixed-fidelity* prototypes can be classified along five orthogonal dimensions (McCurdy et al., 2006):

1. *visual refinement*: whose extremes are hand-drawn sketches *vs.* pixel accurate mock-ups;
2. *breadth of functionality*: corresponding to the number of implemented features;
3. *depth of functionality*: which is the level of detail of each feature;
4. *richness of interactivity*: defined in terms of the interactive elements captured and represented in the prototype; and
5. *richness of data model*: related to how representative of the actual domain the data manipulated by the prototype are.

The artifacts we used feature medium visual refinement fidelity (we did not pay too much attention to layout, colors, proportions, and graphical accuracy), medium breadth (only functionalities likely to be explored by chosen tasks were implemented), high depth (those functionalities were completely implemented), low interactivity and low data richness (as these were sketches in PDF without forms).

Not all types of investigations are equally well suited to the kind of artifacts that are developed. For example, in our case experimental *formal*, *laboratory* tests with *no think-aloud* worked well. This is motivated by the range of UX aspects that we wanted the test to be able to capture: unless one pays attention and isolate possible disturbance or confounding factors, it is very likely that results are biased or muddled.

We believe that the particular combination of design factors, of media used to implement them, and of variables used to measure UX effects, worked with synergy. We can see that variables that are mostly affected by these design factors are PEOU, enjoyment, quality, and satisfaction. Notice that some of the effects overlap with usability metrics (for example, completion time) but that some of the typical usability metrics were unaffected by these design factors (for example, success rate or errors). An implication of this is that running a usability investigation focusing on

measurable objective data only is likely to miss important differences bearing upon acceptability and actual usage.

Additionally, user preferences can be readily expressed by users and considered when designing the system. An implication is that sketches and storyboards allow users not only to detect design choices, but also to appraise them affectingly.

In this case, however, the experimental design was not suitable for separating the effects of individual design factors (for example, to contrast the effects of specialists against those of praises). There is no particular reason why this could not be done using the kind of artifacts and dependent variables that were used here. We could argue that, along with what was found by Reeves and Nass (1996), praises have a subconscious effect that goes beyond what people say. Thus it is possible that even though participants do not like praises, their behavior could be affected in a positive direction.

Therefore, overall, the design factors deployed in SB have a definite positive effect on measured UX aspects.

As expected age is important. It marginally affects PEOU (for older adults it is higher by 1.9%); it interacts with version when affecting PEOU (when moving from RR to SB, compared to older adults, students increase PEUO by 12%); it interacts with phase with respect to PEOU (performing the tasks leads to a decrease by 10.1% for older adults compared to students); it affects enjoyment (an increase for students by 6.3%); it interacts with version when affecting enjoyment (moving from RR to SB increases enjoyment for students by an additional 23.1%, compared to older adults); it increases completion time (older adults work 106.2% longer); older adults make 73% more errors; they are also less accurate in task execution by 18%. Conversely, we did not find age effects on perceived usability, emotions, behavioral intentions, satisfaction, quality, trust, subjective mental effort, and the number of undos. Because the experiment contrasted the behavior of older adults with that of younger people, differences are due to age and not other confounding factors, such as experimental protocol or the fact that participants did not directly operated on the sketches.

For older adults the extra features appearing on the screen do not improve PEOU (actually they appear worsening it). We can argue that the additional elements (human figures, buttons to scroll tips, more text to read) might lead them to perceive that version as being less easy to use. Actual usage of both prototypes for older adults also leads to a decrease in PEOU for RR. While in general PEOU increases by 9.9% after use, it looks like older adults before using a device overestimate its ease of use by 10.1% compared to what students do. Enjoyment is higher for students, who enjoy SB more than RR; this does not occur for older adults. And as expected, usability metrics reveal a worsening outcome for older adults.

Again, also in this case, we can see that implementing those design factors with sketches and storyboards is an effective technique for highlighting these differences. These outcomes are consistent with what was found previously in other studies concerning older adults: not relying on a trial-and-error strategy, preference for written step-by-step help descriptions, an increased cognitive workload, and a general decrease in usability. This can be considered one source of evidence of convergent validity of the evaluation method described in this paper.

We expected a larger number of undos for younger users, compared to older adults, which did not happen. This could be the result of the relatively simple tasks that they were asked to perform and/or of the relatively good usability already achieved in the user interface we considered. Thus, younger users did not need to adopt a trial-and-error strategy, while older adults refrained from adopting it despite the usability problems they faced.

We believe that trust was not significantly affected because of the limited exposure that participants had with the system and because of lack of real actual feedback: the prototype was not connected to any HVAC and therefore user tasks were unconnected to real effects.

Subjective mental effort was also not affected by age nor by version. Probably other methods for measuring attention could be adopted, like interference between a primary and secondary tasks, as used for example by Trafton et al. (2003). Even better, because more suited to the kind of artifacts that we used, could be the adoption of a secondary task based on randomly timed beeps that require users to dismiss them, as described by Reeves and Nass (1996).

In our experiment participants did not directly operate on sketches, but the facilitator animated them as appropriate. This could be a reason why perceived usability and emotions turned out to be insensitive to our independent variables. If participants were actually using an interactive prototype without an intermediary these dependent variables could show some significant effect. On the other hand, such an intermediation did not prevent other variables to show significant effects; it did not invalidate the experiment.

From a methodological point of view we argue that, when using clickable sketches and storyboards, UX aspects should be handled as follows:

- Perceived ease of use could be measured through a questionnaire (Section Appendix A).
- Behavioral intentions could be measured through a questionnaire (Section Appendix A).
- Perceived usability should not be measured, especially in light of the particular experimental protocol we used.
- Emotions could not be measured (at least not with the questionnaire we used, Section Appendix A); other means to detect these could be more effective, like skin conductance, pupillometry, blood pressure and flow, heart beat or even other types of questionnaires, like LEM-tool (Capota et al., 2007). Two undesired confounding factors could have played a negative role: intermediation of the facilitator could have held participants' emotions back; and medium-fidelity non-interactive sketches could have reduced intensity of emotions and increased variability of elicited emotions, reducing reliability of the questionnaires that were used.
- Enjoyment could be measured with a questionnaire (Section Appendix A).
- Subjective mental effort should not be measured with a questionnaire, but rather by introducing a secondary task requiring attention switch.
- Quality could be measured with a questionnaire (Section Appendix B).
- Overall satisfaction could be measured with a questionnaire (Section Appendix B).
- Trust could be measured with a questionnaire (Section Appendix B).

Practical implications of the results are that the ideal user interface should be customized. For younger users, the design factors we adopted lead to improvements along many variables, with the exception of completion time. Considering the low frequency with which people are expected to interact with a HVAC controller, however, completion time is not so important. The improvement of perceived ease of use (both before and after use), enjoyment, behavioral intentions, satisfaction, quality and trust, is even more important when tasks are rarely performed. This is because in version SB most of the design factors deal with delivering help information, on-the-fly explanations, and feedback, which are expected to impact casual users.

A different picture can be seen for older adults. The same factors that improve quality of interaction for younger users are likely to overload older adults. This is what probably happened in our case for specialists, additional written messages and additional buttons appearing on the screen. This was perceived both before and after using the prototype. With older adults, enjoyment was also not affected as for younger users. And, in any case, these factors apparently did not help improving actual usability. It is possible, however, that a redesign of these features, for example by distributing them in a larger number of screens with less choices to be made and perhaps with a better structured textual explanation, could make them more effective. It could also be the case that adopting another modality for delivering explanations and praises (like text-to-speech using different voices for the different specialists, or video-clips) achieves higher levels of effectiveness. Of course richer artifacts will be needed to assess that.

7. Conclusion

Set in the context of Ambient Assisted Living, this research is aimed at determining the effects that certain design factors (specialists, tooltips, praises, additional interactivity) have on a range of UX aspects, including perceived ease of use and usability, behavioral intentions, enjoyment, emotions, subjective mental effort, quality, satisfaction, trust and usability. We wanted to determine the extent to which age differences affect these aspects, and more importantly whether using design artifacts like sketches and storyboards could be used to elicit this information.

We performed an experiment comparing two versions of the user interface of a digital thermostat based on a touch-screen; in one version we implemented the design choices mentioned above. Twenty university students and 20 older adults were involved in the study.

Analysis revealed that using sketches and storyboards is an effective way to elicit interesting and useful results; many UX variables are significantly affected by design factors and by age differences, as expected; effects of design factors go well beyond usability and therefore could not be caught by running an investigation focused only on usability. Because what we found is consistent with results published in the literature, and because of the relatively good internal reliability of the subjective measures we adopted, we believe the results we got are valid and reliable.

Results show that even for an emotionally poor system such a HVAC controller, the use of sketches can be adopted to elicit UX responses that go well beyond usability data.

As mentioned above, age difference matters: older adults do not respond to addition of specialists, praises and tooltips as younger users do. We argue that potential benefits of these design choices are outweigh by the increase in complexity of the user interface.

From a methodological viewpoint we suggest using a particular array of UX aspects and metrics when testing mixed-fidelity prototypes like the ones we used. Not all the metrics that we adopted were equally useful, and in particular perceived usability, subjective mental effort, and emotions did not help us highlighting differences. We also argue that UX aspects should be tested after some usability investigations and redesigns have taken place, so that actual usability mishaps were cleared and would not influence UX aspects.

One limit of this study is that we did not consider collecting data regarding physiological responses from participants (such as skin conductance, eye pupils diameters, blood pressure and flow, and heart beat). Doing that would have provided us with a much richer picture of what subjective responses were, but at the cost of making the experiment more invasive and less ecologically valid. Other factors which we did not consider, but which could play an important role, include the “novelty effect”; to study this aspect, however, a longer longitudinal study is needed; likewise for directly measuring trust and acceptability. In particular, it is very likely that the prior attitude of participants with respect to technology affected the kinds of results that can be found; future experiments could try to isolate such a factor and determine its effects.

Second, some of the metrics we adopted featured a sub-optimal reliability (*i.e.*, emotions). Other questionnaires or elicitation methods could have provided better data. In addition, it would be interesting to compare effectiveness and UX coverage of the method we employed with other methods, such as those reviewed by Vermeeren et al. (2010).

Third, other kinds of artifacts, like including multimodality (voice output, videos) could also be considered. This could be considered in future studies, as well as studying the separate effect that the individual design factors have on UX aspects, giving designers more specific knowledge to enable them to make a better choice. Furthermore, different genres of applications more centered on ludic usage, could be more suitable for letting UX characteristics to emerge, even those that in our experiment did not play a role.

A further research opening is to consider also complexity measures, to see if and how much the design choices we made affect visual complexity, and how this could be explain the response we got from older adults.

Acknowledgments

We thank Eurapo SrL and Aragon Engineering SrL for their support in this project. We are also very grateful to the forty participants for their availability and willingness to help us. We thank also reviewers for suggestions on how to improve the paper.

Statistical processing was done with R (R Development Core Team, 2011).

Appendix A. Pre-test questionnaire

In this overall questionnaire questions were grouped by subscales as follows, which were filled-in in this order; we translated them into Italian.

Perceived ease of use. This questionnaire consists of 3 questions, formulated as a 5-point Likert scale (Davis and Venkatesh, 1996):

1. Interacting with RR/SB does not require a lot of my mental effort.
2. I find RR/SB easy to use.
3. I find it easy to get RR/SB to do what I want it to do.

Enjoyment. Formulated as a 5-point Likert scale:

1. RR/SB is friendly.
2. … pleasant.
3. … well mannered.
4. … boring.
5. … confusing.
6. … playful.

Intention to use. Formulated as a 5-point Likert scale (Sundar and Kim, 2005):

1. Assuming I had access to RR/SB, I intend to use it.
2. Given that I had access to RR/SB, I predict that I would use it.

Perceived usability. Formulated as a 5-point Likert scale, adapted from Davis and Venkatesh (1996):

1. Using RR/SB improves my performance in achieving a good temperature control.
2. Using RR/SB increases my productivity.

Emotions-pre. Formulated as a 5-point Likert scale, adapted from Pekrun et al. (2004):

1. I worry whether I will be able to complete tasks.
2. I am worried of making a fool of myself.
3. I feel embarrassed.

Appendix B. Post-test questionnaire

Perceived ease of use. Same as above.
Enjoyment. Same as above.
Intention to use. Same as above.
Perceived usability. Same as above.
Emotions-post. Formulated as a 5-point Likert scale, adapted from Pekrun et al. (2004):

1. I am proud of myself.
2. I am angry.
3. I feel nervous.

Subjective mental effort. One question based on Sauro and Dumas (2009).
Quality. Based on AttracDiff (Hassenzahl, 2004), semantic differential with 23 adjectives.
Satisfaction. Based on WAMMI (Kirakowski et al., 1998), 18 items, 5-point Likert scale.
Trust. Based on Lewis (1993), 5-point Likert scale:

1. Overall, I am satisfied with the ease of completing this task.
2. Overall, I am satisfied with the support information.
3. The information was effective in helping me complete the tasks and scenarios.
4. I could effectively complete the tasks and scenarios using the system.
5. I felt confident using the system.
6. I will recommend this product to a friend.
7. I think that the system is reliable.
8. I trust the system.

Appendix C. Final questionnaire

This questionnaire was for eliciting qualitative data, and included the following open questions:

1. Which version (RR or SB) do you prefer? Why?
2. What aspect of each system do you like the most?
3. Positive aspects of RR? of SB?
4. Negative aspects of RR? of SB?
5. For SB, do you remember some hints or suggestions given by the system? Do you remember who gave that suggestion?
6. Rank the following concepts (display of energy consumption data, specialists, tooltips, praises, scrollable manual).
7. Why did you rank … so high?
8. Why did you rank … so low?
9. Other suggestions?

References

Arch, A., 2008. Web accessibility for older users: a literature review. W3C, ⟨http://www.w3.org/TR/wai-age-literature⟩.

Brajnik, G., Gabrielli, S., 2010. A review of online advertising effects on the user experience. Int. J. Hum.-Comput. Interact. 26 (10), 971–997.

Capota, K., van Hout, M., van der Geest, T., August 2007. Measuring the emotional impact of websites: a study on combining a dimensional and discrete emotion approach in measuring visual appeal of university websites. In: Designing Pleasurable Products and Interfaces. ACM, Helsinki, Finland, pp. 135–147.

Cho, C., 1999. How Advertising Works on the www: Modified Elaboration Likelihood Model. Technical Report, University of Texas at Austin. ⟨ http://www.ciadvertising.org/studies/reports/info_process/jcira.html/⟩.

Cialdini, R., Martin, S., 2004. The science of compliance. NIMR Med. Rev. 3, 32–38.

Coughlin, J., D'Ambrosio, L., Reimer, B., Pratt, M., 2007. Older adult perceptions of smart home technologies: implications for research, policy and market innovations in healthcare. In: Proceedings of the 29th Annual Conference of Engineering in Medicine and Biology. Lyon, France, pp. 1810–1815.

Davis, F., 1989. Perceived usefulness, perceived ease of use and user acceptance of information technology. MIS Q. 13 (September (3)), 319–340.

Davis, F., Venkatesh, V., 1996. A critical assessment of potential measurement biases in the technology acceptance model: three experiments. Int. J. Hum.-Comput. Stud. 45, 19–45.

Fairweather, P., Basson, S., Hanson, V., 2007. Speech recognition and alternative interfaces for older users. Interactions (July–August) 26–29.

Fogg, B., 2003. Persuasive Technology. Morgan Kaufmann.

Forbes, P., Gibson, L., Hanson, V., Gregor, P., Newell, A.F., 2009. Dundee user centre—a space where older people and technology meet. In: Asset 2009. ACM Press, Pittsburgh, USA, pp. 231–232.

Fozard, J., Bouma, H., Franco, A., Bronswijk, J.v., 2009. Homo ludens: adult creativity and quality of life. Gerontechnology 8 (4), 187–196.

Hassenzahl, M., 2004. The interplay of beauty, goodness, and usability in interactive products. Hum.-Comput. Interact. 19 (December (4)), 319–349.

Hollinworth, N., Hwang, F., October 2009. Learning how older adults undertake computer tasks. In: ASSETS 2009. ACM, Pittsburgh, USA, pp. 245–246.

Kirakowski, J., Claridge, N., Whitehand, R., 1998. Human centered measures of success in web site design. In: Proceedings of the Fourth Conference on Human Factors & The Web. ⟨http://research.microsoft.com/en-us/um/people/marycz/hfweb98/kirakowski⟩.

Lavie, T., Tractinsky, N., 2004. Assessing dimensions of perceived visual aesthetics of web sites. Int. J. Hum.-Comput. Stud. 60, 269–298.

Law, E., Hornbæk, K., September 2007. User experience (UX) and usability measures: correlations and confusion. In: Law, E., Vermeeren, A., Hassenzahl, M., Blythe, M. (Eds.), Towards a UX Manifesto. HCI 2007 Workshop. Lancaster UK, pp. 49–56, ⟨http://www.cost294.org⟩.

Law, E., Roto, V., Hassenzahl, M., Vermeeren, A., Kort, J., April 2009. Understanding, scoping and defining user experience: a survey approach. In: CHI 2009. ACM, Boston, USA, pp. 719–728.

Leung, R., Tang, C., Haddad, S., Mcgrenere, J., Graf, P., Ingriany, V., 2012. How older adults learn to use mobile devices: survey and field investigations. ACM Trans. Access. Comput. (TACCESS) 4 (3), 11.

Lewis, J.R., 1993. IBM Computer Usability Satisfaction Questionnaires: Psychometric Evaluation and Instructions for Use. Technical Report 54.786, IBM. ⟨http://drjim.0catch.com/usabqtr.pdf⟩.

Lindgaard, G., Fernandes, G., Dudek, C., Brown, J., 2006. Attention web designers: you have 50 milliseconds to make a good first impression!. Behav. Inf. Technol. 25 (2), 115–126.

Lunn, D., Yesilada, Y., Harper, S., June 2009. Barriers Faced by Older Users on Static Web Pages—Criteria Used in the Barrier Walkthrough Method. Technical Report. HCW—SCWeb2 Technical Report WP1 D1, Human Centred Web Lab—School of Computer Science, University of Manchester, UK.

McCrickard, D.S., Czerwinski, M., Bartram, L., 2003. Introduction: design and evaluation of notification user interfaces. Int. J. Hum.-Comput. Stud. 58 (5), 509–514.

McCurdy, M., Connors, C., Pyrzak, G., Kanefsky, B., Vera, A., 2006. Breaking the fidelity barrier: an examination of our current characterization of prototypes and an example of a mixed-fidelity success. In: CHI 2006. ACM, ACM Press, New York, NY, pp. 1233–1242.

Meier, A., February 2011. Thermostat Interface and Usability: A Survey. Technical Report, Lawrence Berkeley National Laboratory, ⟨http://escholarship.org/uc/item/59j3s1gk⟩.

Michailidou, E., Harper, S., Bechhofer, S., 2008. Visual complexity and aesthetic perception of web pages. In: Proceedings of the 26th Annual ACM International Conference on Design of Communication. ACM, pp. 215–224.

Newman, M., Landay, J., 2000. Sitemaps, storyboards and specifications: a sketch of web site design practice. In: Designing Interactive Systems, DIS 2000. ACM Press, pp. 263–274.

Norman, D., 2003. Emotional Design: Why We Love (or Hate) Everyday Things. Basic Books.

Obrist, M., Bernhaupt, R., Beck, E., Tscheligi, M., 2007. Focusing on elderly: an iTV usability evaluation study with eye-tracking. In: Interactive TV: A Shared Experience. Springer, pp. 66–75.

Pekrun, R., Goetz, T., Perry, R.P., Kramer, K., Hochstadt, M., Molfenter, S., 2004. Beyond test anxiety: development and validation of the test emotions questionnaire (TEQ). Anxiety Stress Coping 17 (3), 287–316.

R Development Core Team, 2011. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, ISBN: 3-900051-07-0. ⟨http://www.R-project.org/⟩.

Reeves, B., Nass, C., 1996. The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places. Cambridge University Press.

Rubin, J., Chisnell, D., 2008. Handbook of Usability Testing, 2nd Edition Wiley.

Sauro, J., Dumas, J., 2009. Comparison of three one-question, post-task usability questionnaires. In: CHI 2009. ACM, ACM Press, Boston, MA, USA, pp. 1599–1608.

Sayago, S., Blat, J., 2009. About the relevance of accessibility barriers in the everyday interactions of older people with the web. In: W4A '09: Proceedings of the 2009 International Cross-Disciplinary Conference on Web Accessibility (W4A). ACM, New York, NY, USA, pp. 104–113.

Sundar, S.S., Kim, J., 2005. Interactivity and persuasion: Influencing attitudes with information and involvement. J. Interact. Advert. 5 (2), 6–29 ⟨http://jiad.org/article59⟩.

Tractinsky, N., 1997. Aesthetics and apparent usability: empirically assessing cultural and methodological issues. In: CHI '97: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM, New York, NY, USA, pp. 115–122.

Tractinsky, N., Katz, A., Ikar, D., 2000. What is beautiful is usable. Interact. Comput. 13 (2), 127–145.

Trafton, J.G., Altmann, E.M., Brock, D.P., Mintz, F.E., 2003. Preparing to resume an interrupted task: effects of prospective goal encoding and retrospective rehearsal. Int. J. Hum.-Comput. Stud. 58 (5), 583–603.

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.

United Nations, 2010. World Population Ageing 2009. Technical Report, United Nations. Department of Economic and Social Affairs. Population Division. ⟨http://www.un.org/esa/population/publications/WPA2009/WPA2009_WorkingPaper.pdf⟩.

vanderWardt, V., Bandelow, S., Hogervorst, E., 2012. The relationship between cognitive abilities, well-being and use of new technologies in older people. Gerontechnology 10 (4), 187–207.

Varakin, D.A., Levin, D.T., Fidler, R., 2004. Unseen and unaware: implications of recent research on failures of visual awareness for human-computer interface design. Hum.–Comput Interact. 19 (4), 389–422.

Venkatesh, V., Morris, M.G., Davis, G.B., Davis, F. D., 2003. User acceptance of information technology: toward a unified view. MIS Q, 425–478.

Vermeeren, A.P., Law, E. L., Roto, V., Obrist, M., Hoonhout, J., Väänänen-Vainio-Mattila, K., 2010. User experience evaluation methods: current state and development needs. In: Proceedings of the 6th Nordic Conference on Human–Computer Interaction: Extending Boundaries. ACM, pp. 521–530.

Wensveen, S., Djajadiningrat, J., Overbeeke, C., August 2004. Interaction frogger: a design framework to couple action to function through feedback and feedforward. In: Designing Interactive Systems 2004. ACM Press, pp. 177–184.

Zaphiris, P., Kurniawan, S., Ghiawadwala, M., 2007. A systematic approach to the development of research-based web design guidelines for older people. Univers. Access Inf. Soc. 6 (November (1)), 59–75 ⟨http://www.springerlink.com/content/087050g2771rj416/⟩.