



The Combined Walkthrough: Measuring Behavioral, Affective, and Cognitive Information in Usability Testing

Timo Partala^{1,2}

¹Tampere University of
Technology
Human-Centered Technology
Korkeakoulunkatu 6
FI-33720 Tampere, Finland
timo.partala@tut.fi

Riitta Kangaskorte²

²University of Oulu
Oulu Southern Institute
RFMedia Laboratory
Vierimaantie 5
FI-84100 Ylivieska, Finland
riitta.kangaskorte@oulu.fi

Abstract

This paper describes an experiment in studying users' behavior, emotions, and cognitive processes in single usability testing sessions using an experimental method called the combined walkthrough. The users' behavior was studied using task times and completion rates, and emotions were studied using bipolar scales for experienced valence and arousal. Cognition was studied after each task by revisiting detected usability problems together with the users and applying an interactive method based on cognitive walkthrough to each usability problem. An interactive media application was tested with 16 participants using these methods. The results of the experiment showed that the developed methods were efficient in identifying usability problems and measuring the different aspects of interaction, which enabled the researchers to obtain a more multifaceted view of the users' interaction with the system and the nature of the problems encountered.

Keywords

Usability, user experience, behavior, emotion, cognition, interactive media.

Introduction

Current usability evaluation methods can be divided into three categories: methods for usability testing, usability inspection, and inquiry. In usability testing, a product or service is evaluated by testing it on test users. Usability testing has been a central activity in the field of human-computer interaction (HCI) for almost two decades. It has had a substantial practical influence on the development of computing systems. Different usability professions now employ thousands of usability professionals worldwide. Some of the methods for usability inspection include heuristic evaluation (Nielsen, 1994) and the cognitive walkthrough method (Wharton, Rieman, Lewis, & Poson, 1994). Usability inquiry methods include, for example, questionnaires, surveys, and focus groups.

Usability cannot be directly measured (Nielsen & Levy, 2003), but it has been studied by measuring various different usability parameters and metrics. Nielsen (1994) presented the famous model consisting of five usability parameters: learnability, efficiency, memorability, error avoidance, and subjective satisfaction. Another well-known model, presented in ISO 9241-11: guidance for usability (1998), consists of the concepts of effectiveness, efficiency, and subjective satisfaction. In this model, effectiveness has been defined as the accuracy and completeness with which users accomplish their goals. Measures of effectiveness include, for example, quality of interaction outcome and error rates. Efficiency was defined as the relation between effectiveness and the resources used in achieving the task goals. Efficiency indicators include task completion times and learning times. Subjective satisfaction was defined as the user's comfort with and attitudes toward the use of the system. Satisfaction is typically measured using evaluation forms, for example, using questions from Software Usability Measurement Inventory (SUMI) (Kirakowski, 1996), which also includes one of the first attempts to include affective factors in a usability evaluation. An important problem in studying subjective satisfaction using questionnaires is the great number of different methods used. In his review of 180 studies of computing system usability, published in core human-computer interaction journals and proceedings, Hornbæk (2006) identified only 12 studies that had used standard questionnaires.

Many current usability evaluation methods concentrate on producing information related to one particular viewpoint only. For example, the cognitive walkthrough has become an important usability inspection method, but the information produced is limited to the cognitive challenges that a user interface might have. However, the need for producing information from different perspectives has already been acknowledged. For example, Frøkjær, Hertzum, and Hornbæk (2000) stressed that different usability criteria (e.g., measures of effectiveness, efficacy, and satisfaction) should be brought into usability evaluations and each criterion should be examined separately when the usability of a product is evaluated. Another important trend is broadening the measurement of subjective satisfaction so that the focus shifts toward the measurement of the users' experienced emotions. In an early study, Edwardson (1998) concluded that "It may indeed be far more useful to measure and understand customer happiness and customer anger as the primary exemplars of consumer experience rather than satisfaction" (p.11). According to Dillon (2001), affects cover elements related to attitudes, emotions, and mental events. The existence of this kind of phenomena has not been sufficiently taken into account in usability research. Besides measuring user satisfaction, it should be studied whether the user is frustrated, annoyed, insecure, or trusting. Focusing on the users' emotions shifts the focus from whether the users can use an application to whether they want to use it.

In the late 1990s the research field of affective computing started to emerge. Affective computing was defined as computing that relates to, arises from, or deliberately influences emotion or other affective phenomena (Picard, 1997). In this field, the main focus has been in the construction of systems with capabilities for recognizing affect. However, in this field physiology-based measures for evaluating affective interactions have also been developed (Partala & Surakka, 2003, 2004; Partala, Surakka, & Vanhala, 2006; Ward, 2005) utilizing, for example, the user's facial expressions or eye pupil size. During the current decade, more general methods for affective evaluation of HCI have started to emerge in the field of user experience. The field of user experience highlights the user's holistic experience that results from the use of technology. Hassenzahl, Platz, Burmester, and Lehner (2000) presented the hedonic quality scale. They defined hedonic quality as the pleasure-giving quality of a system.

Many psychometric assessments of users' responses to computing systems have focused on positive affective constructs (Hassenzahl & Tractinsky, 2006). This is consistent with the studies reported by Hornbæk (2006). In his review of 180 studies of computing system usability, Hornbæk identified 70 measures of specific attitudes that had been assessed by self-report. Of these, only 13 addressed explicitly negative emotional or physiological states. While positive affect has been found to have many kinds of positive consequences on cognition, for example, enabling more effective decision making (Isen, 2006), this kind of unidirectional measures provide only a partial view on the user's emotions. In fact, understanding the causes of negative experiences may be more important in order to further develop the tested system iteratively based on the test results.

Lately, emotions have been studied in interactive contexts, for example, by Mahlke and Thüring (2007) and Hazlett and Benedek (2007). These studies have indicated that variations in emotions—especially in terms of emotional valence ranging from positive to negative emotions—play an important role in interacting with technology. However, the new methods developed in this field have been largely independent of previous developments in the area of usability. Practical approaches for extending traditional usability testing methods with methods for studying experiential, especially affective aspects, have still been sparse.

There is substantial evidence from psychological research that affective experiences can be effectively organized using a dimensional model. Factor analyses have confirmed that three dimensions are enough to account for most of the variance in affective experiences. Currently, the most commonly used dimensional model of emotions, consisting of valence, arousal, and dominance, was presented by Bradley and Lang (1994). Of these three scales, the valence dimension ranges from negative to neutral to positive affect, while the arousal dimension ranges from very calm to neutral to highly aroused emotion. The dominance dimension varies from the sense of being controlled to neutral and to the sense of being in control (e.g., of a particular situation or an event). Valence and arousal are the most fundamental and commonly studied dimensions. The dominance dimension accounts for much less variation in semantic evaluation of emotions and it is consequently less often used in empirical studies. Bradley and Lang (1994) presented an easy-to-use pictorial method called Self-Assessment Manikin (SAM) for affective self-reports. Using their method, the participant chooses a picture that best represents his or her affective state on each scale. For example, the valence scale ranges from negative (an unhappy face) to neutral (a neutral face) to positive (a happy face) emotion. In addition to SAM, a non-pictorial version of their method has also been used successfully in many basic research experiments (e.g., Partala & Surakka, 2003). This method is based on the semantic differential method (Osgood, 1952) that connects scaled measurement of attitudes with the connotative meaning of words and has a long history of successful use in various fields. Using this method, the participants evaluate their emotional experiences on valence and arousal scales with emotional words used as anchors (e.g., on a 1-9 arousal scale: 1 = very calm, 5 = neutral, 9 = very highly aroused). These or similar methods have been used in user experience research, for example in Partala and Surakka (2004); Mahlke, Minge, and Thüring (2006); and Mahlke and Thüring (2007). However, the methods have not been largely used in the field of usability testing, but they could offer fast and reliable methods for affective self-reports.

Hornbæk (2006) suggested that the subjective measures of usability typically concern the users' perception of or attitudes toward the interface, the interaction, or the outcome. He also suggested the following important future challenges in the area of usability:

- the need to understand better the relation between objective and subjective measures of usability
- extending satisfaction measures beyond post-use questionnaires
- studying correlations between measures

This experiment addressed all three challenges pointed out by Hornbæk (2006). While we suggested that measurement of the users' subjective affective responses can be useful in evaluating traditional interaction (e.g., with graphical user interfaces), we also acknowledged that it would be most suitable for the evaluation of products, in which influencing the users' emotions is part of the intended user experience (e.g., interactive media products or entertainment applications). Based on their research, Sauro and Dumas (2009) suggested that post-task (cf. post-test) questions can be valuable additions to usability tests by providing

additional diagnostic information. In this experiment, we used post-task questions on a semantic differential scale for studying the participants' affective valence and arousal related to each task.

Some researchers, such as Matera, Costabile, Garzotto, and Paolini (2002), have suggested combining expert evaluation and usability testing. In this experiment, we presented an approach that also applies a version of a popular usability inspection method—the cognitive walkthrough—to usability testing. In the original cognitive walkthrough, the evaluator goes through a task phase by phase answering to a set of questions in each phase (e.g., "Will the user notice that the correct action is available?"). In this paper, we proposed an approach in which the evaluator observed the actual use of a system, detected potential problems in the system usage, and retrospectively asked the cognitive walkthrough questions directly to the participant for the detected potential usability problems.

In this paper, we described our experiences of using a method that combined measuring information about the participants' behavior, affects, and cognitive processes during human-computer interaction. For measuring behavior, traditional measures such as task times and task completion rates were used. For affective evaluations on valence and arousal, we used a non-pictorial version of Bradley and Lang's SAM method (Bradley & Lang, 1994). The previously mentioned interactive version of cognitive walkthrough was used for studying the participants' cognition. The ideas underlying the methods used in this experiment were first presented by Partala (2002). In this experiment, the methods were further developed and tested in practice for the first time.

Method

The following sections discuss the participants, equipment, materials and tasks, procedure, and data analysis used in this experiment.

Participants

Sixteen volunteer participants (seven females and nine males, mean age 40.3 years, range 30-51 years) participated in the experiment. The participants were unaware of the purpose of the experiment on arrival.

Equipment

The tests were run on a Fujitsu Siemens Amilo D7830 computer with a display resolution of 1024 x 768 pixels. The system was fast enough to run the software and play all the video and sound clips in real time. Volume was kept at a constant comfortable level throughout the experiment. The participants used a regular mouse to control the interactive media software and viewed the display from a distance of about 50 cm. The display of the computer and the participants' comments were recorded with a video camera for further analyses.

Materials and Tasks

In this experiment, the object of evaluation was *CD-Facta*, an interactive multimedia encyclopedia in Finnish, the native language of the participants. In this encyclopedia a particular information or media file could be found typically by navigating through 3-5 levels in the multimedia product hierarchy. Any information could be found using 3-6 different paths through the hierarchy. Consequently, there were different ways for the participant to accomplish the task goals. The main screen of the product had four different ways to start looking for information: themes, articles by topic, a map-based interface, and a multimedia gallery. In addition to textual articles, the encyclopedia contained audio and video samples, and the user interface contained pictures and icons.

The test participants were presented with seven information retrieval tasks. Each task required a different action chain through the interactive software (the tasks were designed so that completing previous tasks would not help in the subsequent tasks). The tasks were designed so that they were of approximately similar complexity and took on average about 2 minutes per task to complete. One task contained a highly positive audio element (an engaging sports commentary ending in a victory), one task contained a highly negative audio element (traditional old women's crying songs), one task contained a highly positive video element (an exciting portrayal of wild animals), and finally one task contained a highly negative video

element (air raid during a war). Completing the other three tasks did not necessarily involve any audio or video media elements. The tasks are presented in Table 1.

Table 1. The Experimental Tasks (Translated from Finnish)

Task	Audio/Video
In what year was the peace treaty of Uusikaupunki signed?	No
How many people reside in Jämsä?	No
In what year was Nordea bank Finland established?	No
Listen to the commentary of 4 x 400 m relay in a derby against Sweden in 1956. Who was the last runner for Finland?	Positive audio
Who were the main performers of crying songs in the Finnish language region?	Negative audio
What is the typical height of an African buffalo?	Positive video
Look up a video on the first air raid of the Winter War in Helsinki. How many Finns were killed or lost during the Winter War?	Negative video

Audio was played back to the participants at a constant comfortable volume level. The video elements were embedded in the encyclopedia and the resolution of the video display areas was 320 x 240 pixels.

Procedure

In this research, an experimental usability testing method developed by the authors was tested. The working title of the method was *combined walkthrough*, because it is partly based on the cognitive walkthrough method and because of the aim of combining expert evaluation and laboratory testing as well as measurements of behavioral, affective, and cognitive information.

One participant at a time participated in the test. The sessions were carried out in a silent laboratory, and they lasted for about an hour for each participant. The participant was first seated in front of a computer desk and asked to fill in a demographic data form. After that the researcher read aloud the instructions for the test. The participants were made aware that task times were measured, and they were instructed to complete the tasks without any unnecessary breaks. The participants were told that answers to post-hoc questions (based on the cognitive walkthrough) were recommended to be short yes or no answers, but could be extended with more detailed comments, if necessary.

In the evaluation of affective experiences, methods and instructions typically used in basic research (e.g., Partala & Surakka, 2003) were used. Nine-point rating scales for valence and arousal were shown and explained to the participant with examples of rating affective experiences. The participants were told to try to get through the tasks independently without thinking aloud. They were instructed to tell the researcher the answer to the task question when they thought they had found the answer. If the answer was incorrect, the researcher quickly indicated that to the participant, who continued solving the task.

The participants were then presented with seven information retrieval tasks, one at a time, in a randomized order (different for each participant). During the task performance the researcher observed the completion of the task and made notes about phases in which the participant chose an incorrect action or had difficulties in finding the right action. More specifically, the researcher paid special attention to two typical actions: the participant takes a wrong action and starts following a wrong path or stays on the path without returning immediately, or the participant uses a significantly long time in trying to find a correct function from a screen.

If the participant had not found the correct answer in four minutes (twice the expected average task time), the tasks were classified as incomplete tasks. The tasks were designed so that if the participants had not completed the tasks in four minutes, they were typically stuck in a problematic situation that they could not solve themselves. In these cases, time measurements were no longer valid, but the participants completed the task to be able to fill in the questionnaire on affective experiences. To avoid long experimental sessions, the researcher gave hints to the participant about the next correct move if the participant was lost in the interface and time measurements were not valid any more.

After finishing the task (telling aloud the correct answer for the task) the participant rated his or her overall user experience related to the performed task by filling in an affective experience rating form. The ratings of the participants' affective experiences were carried out using a form containing a 9-point rating scale for both valence and arousal. On the valence scale, number 1 indicated a very negative affective experience, number 5 indicated a neutral experience, and number 9 indicated a very positive affective experience. On the arousal scale, number 1 indicated a very calm experience, number 5 indicated a neutral experience, and number 9 indicated a very highly aroused experience. If the recently completed task contained an audio or video media element, the participants also rated their experienced valence and arousal in response to this media element (the media element was played once again before this evaluation). Special attention was paid to ensure that the participants understood that the target of the first evaluation was the holistic affective user experience related to the interaction with the system when completing the task and that the target of the second evaluation was the experience evoked by the particular media element alone.

Before moving on to the following task, the problematic situations detected by the researcher during the completion of the task were examined interactively by the researcher and the participant. For this, a method was developed based on the cognitive walkthrough method (Wharton et al., 1994). The researcher and the participant together revisited each detected problematic point using the software. The researcher asked the participant three selected questions based on the original cognitive walkthrough. Out of the four original questions of the cognitive walkthrough, we selected three questions that were the most appropriate to be used in the context of usability testing. We changed the wording of the questions so that the researcher could directly ask the participant. The following were the three questions asked:

1. Did you notice that this action was available? (Researcher points to the correct action to the participant.)
2. Did you associate this (the correct) action with your goal at this point?
3. (After choosing the correct action) Did you notice that you progressed towards the completion of your task?

The original question "Will the user try to achieve the right effect?" was left out, because the tasks in the current experiment were straightforward and the possibilities for misunderstanding the tasks were minimized. The researcher wrote down the participant's answers and asked further questions if the participant's initial answers were ambiguous. The trials were video recorded and analyzed afterwards in order to understand the nature of usability problems found and to evaluate the performance of the researcher.

Data Analysis

The differences between the categories of subjective ratings of emotional experiences were analyzed using Friedman's tests and Wilcoxon's matched pairs signed ranks tests. These tests were selected due to the nonparametric nature of the evaluation data. Spearman correlations were used to analyze the relationships between the different usability indicators. Data from one task out of total 112 (0.9%) had to be dropped from analysis due to missing data.

Results

The following sections discuss the identified usability problems, the emotional experiences, task times and completion rates, correlations of measures, and experiences of using the method.

Identified Usability Problems

By observing the participants during the experiment, the researcher identified a total 146 usability problems. Of these, 136 (93.2%, on average 8.5 per participant and 1.2 per task) were also judged as usability problems based on the participants' answers to the cognitive questions (if the participant answered "No" to any of the cognitive walkthrough questions for the identified problem, or he or she could not give a simple yes or no answer, then the researcher judged that the participant's comments indicated the presence of a usability problem). For the three questions, the participants' answers indicated that there was a problem related to the first question (awareness) 60 times, to the second question (association) 78 times, and to the third question (feedback) 15 times. Further analyses revealed that 116 out of the 136 problems were identified by only question 1 and 2. All the usability problems indicated

by question 3 were also indicated by either or both of the other questions. The distribution of found usability problems by unique question is presented in Table 2.

Table 2. The Number of Usability Problems Detected by Each Question or Combination of Questions.

Question	1 only	2 only	3 only	Two questions	All three questions	Total
No. of problems	50	66	0	18	2	136

The figures presented in Table 2 contain multiple instances of the same usability problems. When analyzing the found usability problems, it was in some cases difficult to determine which of the found problems were unique because of the differences in the participants' behavior and answers to the cognitive questions concerning the usability problems. Thus, a formal analysis of the number of unique usability problems was not fully feasible, but an estimate about one fourth of the problems presented in Table 2 (35 cases) were unique.

Among the incomplete tasks, there were only five cases in which the researcher's questions related to the participant's cognition did not give any indication of the nature of the usability problem resulting in a failure to complete the task. In the other 31 cases the cognitive questions found at least one potential explanation for the critical problem.

The found usability problems were related, for example, to the following aspects of the media software:

- communication using metaphors (e.g., the participants could not infer the meaning of an icon based on its appearance)
- choice of concepts (e.g., the participant did not associate the information he or she was looking for with the label of the category that contained the information)
- interaction styles (e.g., the participants did not notice that a field was scrollable)
- media interface (e.g., when a list of available multimedia files appeared, the first file on the list started playing automatically)
- navigational structure of the application (e.g., the path to the desired information was too complex)
- navigation and information presentation in the map-based interface to information (e.g., the participant did not notice that place names in a map were hyperlinks)

Emotional Experiences

The average overall task-related emotional experiences of the participants for all tasks were 4.6 on the valence scale and also 4.6 on the arousal scale. However, there was a significant amount of variation in the ratings of emotional experiences. The average overall task-related emotional experiences for the first three tasks (which did not include dominant media elements) were 4.9 on the valence scale and 4.5 on the arousal scale. The overall task-related user experiences for the four tasks (which contained a dominant media element) are presented in Figure 1.

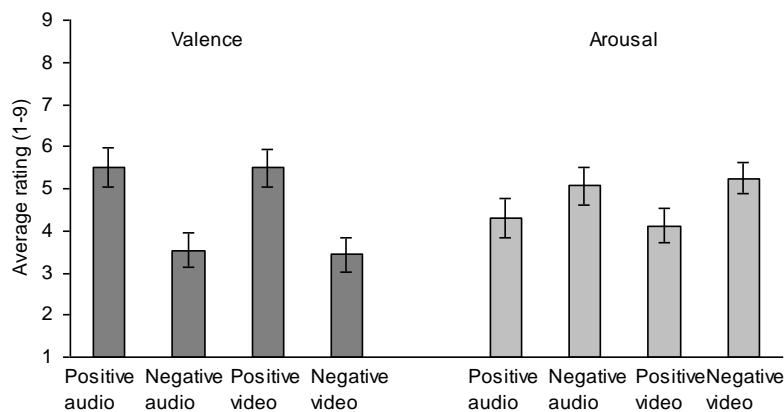


Figure 1. The average overall task-level affective experiences for the tasks containing media elements

The statistical analysis showed that the participants evaluated their task-level user experience as more positive in valence in response to the tasks containing a positive audio element than to tasks containing a negative audio element $Z = 3.2, p < 0.01$. The other pairwise differences for the evaluations of task-level affective experiences were not statistically significant. The average emotional experiences of the participants in relation to the dominant audio and video elements alone are presented in Figure 2.

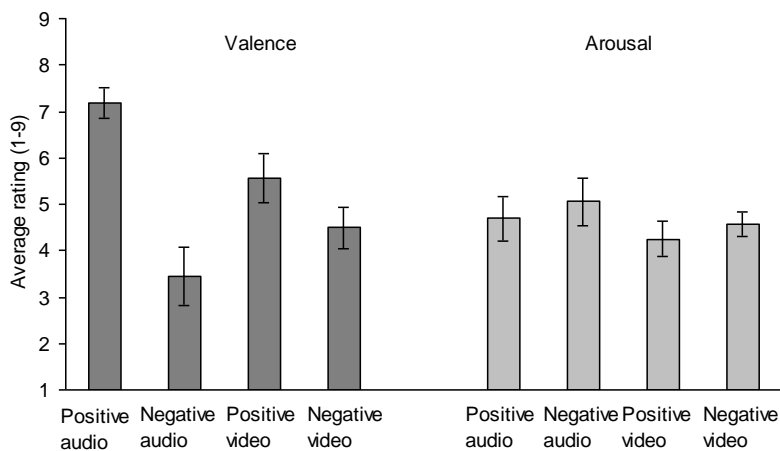


Figure 2. Average emotional experiences for the video and audio clips alone

As expected, the participants rated their experienced affective responses to the positive audio as significantly more positive in valence than to the negative audio $Z = 3.0, p < 0.01$. Similarly, the participants rated their experienced affective responses to the positive video as significantly more positive in valence than to the negative video $Z = 2.9, p < 0.01$. The differences in experienced arousal were not significant, but they were approaching statistical significance ($p < 0.07$ for both audio and video).

Task Times and Completion Rates

Of the 111 analyzed tasks, in 36 cases (32.4%) the participants did not reach the goal in four minutes (double the estimated average task time); these cases were considered as incomplete tasks. Average task time for the tasks that were successfully completed (N=75, 67.6%) was 125.2 seconds. The average task times for the different tasks varied from 96.5 seconds to 198.8 seconds.

Correlations of Measures

The correlations between the different measures were first calculated for all the completed tasks. For comparison, correlations were also calculated for completed tasks 1-3 (the tasks that did not contain dominant media elements, N=36). In order to also obtain a view on the incomplete tasks, correlations were calculated separately for that data (N=36) excluding the data for task times. The correlations for the different measures are presented in Table 3.

Table 3. Correlations of Different Measures Measured for the Completed Tasks (N=75)

	Time	Usability problems	Valence	Arousal
Time	1			
Usability problems	.70*	1		
Valence	-.50*	-.43*	1	
Arousal	.30*	.33*	-.47*	1

*Correlation is significant, $p < 0.01$

For the completed tasks 1-3, the correlations were as follows: task times: usability problems .64, task times: valence -.48, task times: arousal .43, and valence: arousal -.52, all significant ($p < 0.01$). The following correlations were also significant ($p < 0.05$): usability problems: valence -.40 and usability problems: arousal .35. For the tasks that the participants could not finish by themselves (N=36), the correlations were as follows: usability problems: valence .01, usability problems: arousal .19, and valence: arousal -.53 (the last correlation is significant, $p < 0.01$). Task times were not included in this analysis.

Experiences of Using the Method

A researcher in usability conducted the experiment. She had previous experience of using the cognitive walkthrough method as a usability inspection method. Before the experiment began, the creator of the combined walkthrough method thoroughly instructed the researcher on the method. The method and arrangements were tested in pilot studies before the actual experiment occurred. The researcher analyzed introspectively her experiences of using the method in practice. According to her, the method was clearly defined and efficient. However, using the method required some practice. Firstly, it was a challenge for the researcher to note all the points in which a test participant takes an incorrect action during a task. This was especially true for the interactive media software used in the experiment in which there is more than one way to finish the tasks. Secondly, the researcher has to be able to interpret the test participants' answers to the cognitive walkthrough if the test participants cannot give a simple yes or no answer, but instead explains his or her thoughts with many sentences. However, information obtained from free comments related to a particular problematic point with the user interface can be valuable in further developing the user interface.

It is essential for a novice researcher to have detailed instructions for using the combined walkthrough method. In addition, a researcher has to be familiar with the software that is the object of evaluation in order to be able to design the tasks and detect incorrect actions. In the case of this experiment, the creator of the method assisted the researcher in designing the tasks.

Discussion

The results of the experiment suggest that the approach developed for this research succeeded in producing useful information about the participants' behavior, affects, and cognition in single usability testing sessions of interactive media software. In this experiment, we used quantitative scales to study the participants' emotions. This approach was easily understood by the participants and produced significant variations in experienced affective valence and arousal for different tasks. The participants were given detailed instructions validated in psychological basic research, and it seemed that the participants had no difficulties in distinguishing between emotional valence and arousal or overall emotional experience and experience related to a particular media element alone. The valence ratings were useful in evaluating the severity of the usability problems found. The neutral middle point of the valence dimension could be used as a critical value. If the test participants' subjective ratings of valence fall on the negative side on average, an intervention in the form of redesigning (a part of) the user interface is clearly necessary. Arousal ratings gave further information about the depth of the participants' responses, especially in conjunction with negative emotional valence.

Significant correlations were found between the different usability measures. Many other studies have reported lower correlations. In a recent meta-analysis of correlations among usability measures in 73 human-computer interaction studies, Hornbæk and Law (2007) found mostly low (absolute value $<.3$) correlations between measures of effectiveness (e.g., errors), efficiency (e.g., task times), and subjective satisfaction (e.g., satisfaction rating or preference). The correlations from this experiment are comparable to those presented by Hornbæk and Law, because in both studies the correlations were calculated at the least aggregated level possible (in the case of this experiment, at the level of single tasks). The higher correlations obtained in this experiment may be due to the context of *ordinary* usability testing, while studies reviewed by Hornbæk and Law were in many cases experiments comparing innovative user interfaces or interaction techniques. Even though the between-measure correlations were higher in the present study than those reported by Hornbæk and Law (2007), the current results still suggest that it is beneficial to measure different aspects of usability in order to obtain a more coherent view of the usability of software. The coefficients of determination between the objective and subjective measures obtained were below $.5$, which suggests that each measure gave information that complements the other measures. By measuring, for example, task times or experienced affective valence alone, the other variable cannot be predicted reliably.

The use of positive and negative affective media elements significantly affected the valence ratings of overall subjective experiences related to the interaction when completing the tasks. In the context of a positive media element, the participants evaluated their overall user experience related to the information retrieval task as more positive than in the context of a negative media element. The media elements used in the experiment were deliberately chosen as clearly positive and negative. These results suggest that user experiences of interactive media can be guided in the desired direction by using dominant audio and video media elements with emotional valence. These findings have some design implications. In order to design for coherent emotional user experiences, the designers have to take into account both the affective effects of the media elements used in the product and the effects of different user interface solutions and usability problems.

When compared to other existing usability evaluation methods, the current approach has some advantages. First, it is an empirical method and the data gathered is based on observations of real system usage. Second, it is an integrated method capable of producing cognitive, behavioral, and affective information in the same test trials. Third, it combines some of the benefits of usability testing and expert evaluation. It is possible to get objective data based on the participants' performance in the tasks, but it also enables using expert judgment to detect problems when observing the participants carrying out the tasks. In the current experiment, the instructions for the researcher were detailed, but experienced usability experts could use their judgment more freely in identifying usability problems when observing the participant. In addition to the cognitive questions developed based on cognitive walkthrough, they could also ask the participant additional questions in a more open-ended fashion.

The method also has some challenges. First, learning the method demands some practice and the researcher also has to be familiar with the software to design representative tasks. On the

other hand, the number of test participants does not have to be very large. Because of our research goals, we carried out this study with 16 participants, but in practice a smaller number of participants would have been enough to find a clear majority of the problems and to get a realistic idea about task times and the users' experiences and processes on the affective and cognitive levels. The second challenge is related to a possibility for human errors, for example, in this experiment, the evaluator accidentally skipped a couple of problems that she was supposed to react to. On the other hand, there were a few occasions when participants did not agree with the evaluator about the existence of a usability problem. These disadvantages can be complemented with good instructions, practice, and a post-hoc video analysis to make it possible to detect missed usability problems afterwards.

Recommendations

Based on this experiment, we suggest that combining expert evaluation and usability testing, and also the measurements of behavioral, affective, and cognitive aspects of interaction is a noteworthy and useful approach for usability testing. In the future, the evaluation of the users' affective responses could also include gathering and analyzing qualitative interview data about the nature of their user experiences more systematically than in the current experiment in which the participants already gave some valuable comments. Further information about the users' emotions could be also gathered, for example, using a discrete emotions framework (e.g., based on the model by Ekman (1992), consisting of anger, disgust, fear, joy, sadness, and surprise or the model by Frijda (1986), consisting of desire, happiness, interest, surprise, wonder, and sorrow). In the future, the third cognitive question (asking whether the participant notices that she or he has made progress towards the task goal) could be dropped, at least when the target of evaluation is similar to the interactive media software of this experiment. Finally, methods for systematically combining data from the different sources should be developed.

Conclusions

In all, the results of this experiment suggest that different kinds of information (behavioral, affective, and cognitive) can be successfully captured in practice in one usability testing session using the proposed combined walkthrough method. Especially the affective subjective ratings could be used successfully in capturing the variations in users' affective states. The different measures were correlated, but a large coefficient of determination was found only between time and the number of usability problems. The effect sizes between the other variables ranged from small to medium. These results support the view that it is worth measuring user interaction from different aspects in order to gain a more multifaceted understanding of the interaction. This approach may act as a starting point for usability testing methods that aim at producing different types of information, but are nevertheless designed for cost-effectiveness.

Practitioner's Take Away

The following were the main findings of this experiment:

- Behavioral, affective, and cognitive aspects of computer system usage can be cost-effectively studied together in usability testing.
- The information obtained by the behavioral, affective, and cognitive measurements can contribute to a more multifaceted understanding of user interaction with the system.
- Variations in the users' emotional experiences (valence and arousal) related to completing a task using an interactive system can be efficiently measured using bipolar scales. Systematic measurement of emotional experiences broadens the scope of subjective measures beyond traditional satisfaction measures.
- The use of highly positive or negative media elements influences overall ratings of task-related affective experiences in interactive media applications.
- Ideas underlying the cognitive walkthrough can be useful in retrospective analysis of usability problems together with the user.

Acknowledgements

The authors would like to thank all the voluntary test participants for their participation and the reviewers and the editor for their comments on the manuscript. This research was partly supported by European Structural Funds and State Provincial Office of Oulu. In addition, Timo Partala was funded by Tampere University of Technology when finalizing the research and by University of Tampere when developing the first version of the presented method.

References

- Bradley, M. M., & Lang, P. J. (1994). Measuring emotions: the self-assessment manikin and the semantic differential. *Journal of Behavioral Therapy and Experimental Psychiatry*, 25(1), 49-59.
- Dillon, A. (2001). Beyond usability: process, outcome and affect in human-computer interactions. *Canadian Journal of Library and Information Science*, 26(4), 57-69.
- Edwardson, M. (1998). Measuring emotions in service encounters: An exploratory analysis. *Australasian Journal of Market Research*, 6(2), 34-48.
- Ekman, P. (1992). An argument for basic emotions. *Cognition and Emotion*, 6(3/4), 169-200.
- Frijda, N. H. (1986). *The emotions*. New York: Cambridge University Press.
- Frøkjær, E., Hertzum, M., & Hornbæk, K. (2000). Measuring usability: Are effectiveness, efficiency and satisfaction really correlated? In *Proceedings of CHI 2000: vol. 1*. (pp. 345-352). The Hague, The Netherlands. ACM Press.
- Hassenzahl, M., Platz, A., Burmester, M., & Lehner, K. (2000) Hedonic and ergonomic quality aspects determine a software's appeal. In *Proceedings of CHI 2000: vol 1*. (pp. 201-208). The Hague, The Netherlands. ACM Press.
- Hassenzahl, M., & Tractinsky, N. (2006). User Experience - a research agenda. *Behavior & Information Technology*, 25(2), 91-97.
- Hazlett, R. L., & Benedek, J. (2007). Measuring emotional valence to understand the user's experience of software. *International Journal of Human-Computer Studies*, 65(4), 306-314.
- Hornbæk, K. (2006). Current practice in measuring usability: Challenges to usability studies and research. *International Journal of Human-Computer Studies*, 64(2), 79-102.
- Hornbæk, K., & Law, E. L. (2007). Meta-analysis of correlations among usability measures. In *Proceedings of CHI 2007: vol 1*. (pp. 617-626). San Jose, CA. ACM Press.
- Isen, A. M. (2006). Positive affect and decision making. In M. Lewis & I. Haviland (Eds.) *Handbook of emotions* (pp. 417-435). New York, NY: Guilford Press.
- ISO 9241-11. (1998). Ergonomic requirements for office work with visual display terminals (VDTs) – Part 11: Guidance on usability. ISO/IEC.
- Kirakowski, J. (1996). The software usability measurement inventory, background and usage. In P.W. Jordan, B. Thomas, B.A. Weerdmeester & I.L. McClelland (Eds.), *Usability Evaluation in Industry* (pp.169-178). London, UK: Taylor & Francis.
- Mahlke, S., Minge, M., & Thüning, M. (2006). Measuring multiple components of emotions in interactive contexts. In *Proceedings of CHI 2006: vol. 2*. (pp. 1061-1066). Montreal, Canada. NY. ACM Press.
- Mahlke, S., & Thüning, M. (2007). Studying antecedents of emotional experiences in interactive contexts. In *Proceedings of CHI 2007: vol. 1*. (pp. 915-918). San Jose, CA. ACM Press.
- Matera, M., Costabile, M.F., Garzotto, F., & Paolini, P. (2002). SUE inspection: An effective method for systematic usability evaluation of hypermedia. *IEEE Transactions on Systems, Man and Cybernetics- Part A: Systems and Humans*, 32(1), 93-103.
- Nielsen, J. (1994). Heuristic evaluation. In J. Nielsen and R. L. Mack (Eds.), *Usability Inspection Methods*(pp. 25-64). New York: John Wiley and Sons, Inc.

- Nielsen, J., & Levy, J. (2003). Measuring usability: Preference vs. performance. *Communications of the ACM*, 37(4), 66-75.
- Osgood, C.E. (1952). The nature and measurement of meaning. *Psychological Bulletin*, 49, 197-237.
- Partala, T. (2002). The combined walkthrough: Combining cognitive, behavioral, and cognitive methods in human-computer interaction. In *Proceedings of WWDU 2002: Vol. 1* (pp. 458-460). Berlin: ERGONOMIC Institut für Arbeits- und Sozialforschung.
- Partala, T., & Surakka, V. (2003). Pupil size variation as an indication of affective processing. *International Journal of Human-Computer Studies*. 59(1-2), 185-198c
- Partala, T., & Surakka, V. (2004). The effects of affective interventions in human-computer interaction. *Interacting with Computers*, 16(2), 295-309.
- Partala, T., Surakka, V., & Vanhala, T. (2006). Real-time estimation of emotional experiences from facial expressions. *Interacting with Computers* 18(2), 208-226.
- Picard, R. (1997). *Affective computing*. Cambridge, MA: MIT Press.
- Sauro, J., & Dumas, J. S. (2009). Comparison of three one-question, post-task usability questionnaires. In *Proceedings of CHI 2009: vol. 1*. (pp. 1599-1608). Boston, MA. New York, NY. ACM Press.
- Ward R. D. (2005). An analysis of facial movement tracking in ordinary human-computer interaction. *Interacting with Computers*, 16(5), 879-896.
- Wharton, C., Rieman, J., Lewis, C., & Polson, P. (1994). The cognitive walkthrough method: A practitioner's guide. In J. Nielsen and R. L. Mack (Eds.) *Usability Inspection Methods* (pp. 105-140). New York: John Wiley & Sons.

About the Authors



Timo Partala

Timo Partala is an assistant professor at Tampere University of Technology in Finland. He received a Ph.D. degree in interactive technology from University of Tampere in 2005. His current research interests include user experience, location-aware systems, and navigation.



Riitta Kangaskorte

Riitta Kangaskorte is a project engineer at Oulu Southern Institute, University of Oulu, Ylivieska, Finland. She holds a M.Sc. degree in computer science since 2008. Her research interests are human-computer interaction, usability and user experience.