Analyzing the Benefits of Integrative Multi-Dimensional Assessments of Usability Features in Interaction-Centered User Studies

Verena Ermes
RWTH Aachen University
Templergraben 55
52062 Aachen, Germany
+49 241 80 23867
verena.ermes@rwth-aachen.de

Armin Janß, Klaus Radermacher
Chair of Medical Engineering
RWTH Aachen University
Pauwelsstr. 20
52074 Aachen, Germany
+49 241 80 23871
janss@hia.rwth-aachen.de

Carsten Röcker
Human-Computer Interaction Center
RWTH Aachen University
Campus-Boulevard 57
52074 Aachen, Germany
+49 241 8049222
roecker@comm.rwth-aachen.de

ABSTRACT
Today, usability measures for the evaluation of systems and interfaces are mostly assessed in an isolated way. This paper addresses the question whether an integrative multi-dimensional feature evaluation can lead to different and more holistic results. We combined traditional measures (e.g., time to task completion) with advanced measures, like eye tracking, biosignal data logging and assessment of user emotions. For the evaluation of emotions, we used verbal methods (PAD Semantic Scale and a questionnaire) and a nonverbal method with EmoCards. The overall goal was to document and analyze the interaction as completely as possible (including effectiveness, efficiency and user satisfaction), focusing especially on objective measurements. Furthermore, the recording and assessment of emotions, which are part of the user experience, should give insights into user satisfaction.

General Terms
Human Factors, Usability Engineering, Measurement, Design, Experimentation, Security, Performance, Ergonomics, Risk Management

Keywords
Usability Measures, User Experience, Eye Tracking, Biosignal Data Logging, Emotion Evaluation, Risk Analysis, Medical Devices, Integrated Usability Evaluation

1. INTRODUCTION
Efficient information management gains increasingly more importance in a lot of working areas. Also, medical engineering depends more and more on data processing: automation in process control, a continuously growing number of telemedicine applications and monitoring systems for patient surveillance solutions can provide processed and integrated data, resulting in a data presentation for optimized treatment [3].

However, consequences of the described developments should not be underestimated. First, users have to manage the complex cognitive challenges when using the new systems [4][5]. Second, it becomes apparent that a risk-oriented and performance-based control in the field of medicine is barely possible if the design process is not user-centered. This aspect is even more important regarding the broad range of users, for example, for telemedicine applications which are not only used by skilled staff but also by the patients themselves [6][7][8].

To reduce the increased risk of use errors, the European Union determined the integration of usability engineering through the international standards IEC 62366 and IEC 60601-1-6 in 2008. Beside product-safety, which is one of the most critical requirements for medical devices, there is a development beyond usability: users not only demand accessible functionality but also search for products that are enjoyable [9]. This raises the question whether this tendency equally applies to safety-critical applications such as medical devices as it does to products for the entertainment industry.

2. RELATED RESEARCH

2.1 Objective Evaluation Methods
To reach the goal of an integrative Human-Machine-System, Lin and Imamiya [11] demand a method which is capable of showing users' thought patterns. For this purpose, subjective measures cannot provide sufficient results. Lin and Imamiya searched for an extensive image of users for which they need a large and reliable data volume. Their analysis focused on human experience which is measured by objective appraisal procedures. Lin and Imamiya reach the conclusion that a single physiologic measure cannot provide sufficient information about either a complete interaction or the isolated measuring of eye tracking. Therefore, they suggest combining the measurements of eye- and hand-movements, e.g., with a quotient of mouse clicks over fixations [11].

2.1.1 Biosignal Data Logging (Heart Rate Variability)
Through physiological data, it should be possible to give objective answers regarding to the user's experience. Physiological data are inherently multi-dimensional and provide different perspectives to

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
User Centered Design 2014, May 20-22
Copyright © 2014 ICST 978-1-63190-011-2
DOI 10.4108/icst.pervasivehealth.2014.255142

the physical state of the users. Especially the Heart Rate Variability (HRV) should reflect the work load while interacting with an interface. The HRV measures the variability of the interval between two continuous heart beats. The measured results allow to draw conclusions about modulations of the vegetative nervous system [11].

2.1.2 Eye Tracking

Data from eye tracking are, just as physiological data, multi-dimensional. From the recorded data it is expected to gain information about the user's focus of attention and consequently, to find the sources of occurring usability problems. Furthermore, eye tracking data can give hints about the work load [11]. In the areas of usability investigation as well as commercial and cognitive neurosciences, so-called gaze tracking systems are used to record eye movements. These movements can be divided into three aspects: fixation, pursuit, and saccade. During a fixation, which means the focusing on a specific point, a number of cognitive processes take place Contrary to that, saccades are rapid eye movements during which cognitive processes should be suppressed [11].

2.2 Subjective Evaluation Methods

With subjective evaluation methods, information about users' attitudes, satisfaction, and preferences are gained. To collect these data, questionnaires are state of the art, but it is also possible to gather information through the measurement of emotions [13].

2.2.1 Appraisal of Emotion

Agarwal and Meyer [7] assessed to what extend emotions can affect the field of human-computer interaction (HCI). So far, only few researchers in the HCI field took the role of emotions into account. However, many psychologists argue that it is impossible to interact without emotions, whether they are conscious or unconscious. Prior research points out the fact that emotions are closely connected to user acceptance and user satisfaction. Moreover, it is conceivable that the usability of a product influences the emotional state of its users.

In this paper, we lean on the definition of emotion by Agarwal and Meyer: "Emotion is comprised of psychological, affective, behavioral, and cognitive components" [12, p. 2920]. The field of research concerning emotions is very complex, which is the reason why there are various kinds of measurement tools. Firstly, verbal tools allow insights into emotional reactions of which the subjects are aware. Secondly, nonverbal tools are used to show unconscious emotional reactions. For sufficient evaluation reliability, it is recommended to combine both tools [12].

Agarwal and Meyer chose the PAD Semantic Scale by Mehrabian and Russel as verbal tool and EmoCards by Desmet as a nonverbal tool. Additionally, they used two traditional usability measures: time to task completion and error rate. To prove the suitability of the PAD Scale for the field of HCI, Agarwal and Meyer calculated the reliability of the PAD Scale with Cronbach's Alpha. The result of .960 indicates a high reliability and allows the assumption that the PAD Scale appears as a useful verbal tool in the context of HCI [12].

As a result of their study, Agarwal and Meyer discovered that the analysis of the usability measures (effectiveness, efficiency) suggests that the user interfaces are nearly identical concerning the performance whereas the results of PAD Scale and EmoCards confirmed that there are significant differences between the emotional reactions of the subjects for each interface [12].

2.2.2 User satisfaction assessed with questionnaire

Standardized questionnaires are often used in the field of usability for the collection of subjective perceptions. The usage of questionnaires is not very cost-intensive. On the one hand, there can be questionnaires with open questions that offer the option to gather bigger amounts of information. On the other hand, there are possibilities for standardization which allow a statistical analysis of the data. Especially for usability purposes, different standardized questionnaires have been developed which include questions for specific problems (e.g., AttrakDiff2, ISONORM 9241-10 or SUMI).

3. INTEGRATIVE SOLUTION STRATEGY

3.1 Planning and Navigation System 'McMinn'

The planning and navigation 'McMinn' system has been developed at the Chair of Medical Engineering, Helmholtz Institute for Biomedical Engineering at RWTH Aachen University. It is supposed to assist orthopedic surgeons with the correct planning of a surface replacement of the femoral head and therefore accurate drilling of the implant cavity. Due to the necessity of very precise acting and reduced visibility during keyhole surgery, a careful planning of the operation is highly important [8]. Finally, the usability of the system dialogs 'Position Cylinder' and 'Position Implant' has been investigated. For this examination, an integrative, multi-dimensional analysis has been conducted. Here, especially the question whether a new type of information for usability evaluation can be generated that could not be accessed through an isolated analysis of usability-attributes.

3.2 Usability Attributes and Evaluation Methods

Initially, the usability attributes efficiency and effectiveness have been evaluated, since the surgeon should be able to plan the operation as faultless and with as little effort as possible. Beyond that, it would be worth striving for a high user-satisfaction, so that the criteria of efficiency and effectiveness are easier to achieve. In order to measure user-satisfaction, the mental workload has been considered as well. With regard to the goal of analyzing integrative and multi-dimensional, it could be an advantage to measure in an integrative way from the start. Therefore, the appraisal procedures of Lin and Imamiya as well as the chosen tools by Agarwal and Meyer are integrated in this paper.

3.2.1 Effectiveness

The effectiveness is analyzed in respect to four exercises: choice of x-ray images, positioning of the cylinder, positioning of the shaft axis, and positioning of the implant. Furthermore, it should be checked if the quotient fixations over mouse clicks, proposed by Lin and Imamiya [11], can provide information regarding effectiveness. Lin and Imamiya claimed that it may give information about the hand-eye coordination which is conceivable to influence the effectiveness.

3.2.2 Efficiency

The efficiency is analyzed on the basis of the results of the traditional attribute time to task completion. In addition, it should be checked if any results of the eye tracking or measurements of the HRV can provide indications of efficiency.

3.2.3 User-Satisfaction

In this study, the identification of user-satisfaction should exceed a usual questionnaire. The HRV, which allows conclusions to be drawn about the vegetative nervous system, should be associated with measurements of emotional reactions. In the following, this association should be compared to answers of a questionnaire. For this purpose, an ECG should be recorded and EmoCards as well as a PAD Scale, tools used by Agarwal and Meyer, should be included and compared. Additionally, a questionnaire measuring user-satisfaction with rating scale has been created. Furthermore, the mental workload has been evaluated by with NASA-TLX, which is suggested by Bevan [10]. In addition, the measurements of eye tracking and HRV have been included. According to Lin and Imamiya [11], there is a correlation between HRV and NASA-TLX.

4. RESULTS

Two female and four male university students, aged 20 to 26, participated in the user tests. The duration of each user test varied from 25 to 50 minutes. At first, the data has been analyzed regarding the usability attributes efficiency and effectiveness. Results of the measure time to task completion vary strongly from 1 to 14 minutes. Looking at the four exercises, the first exercise (choosing two 90° alternated x-ray images for optimal virtual 3D presentation) is nearly equally completed successfully by every participant; likewise is task three (positioning the shaft axis). But significant differences can be seen in the execution and completion of the tasks two and four. According to the effectiveness evaluation of task two (positioning the cylinder) there are difficulties in positioning the implant (exercise four). These difficulties are seen as well in the time to task completion. If a participant was not very effective in task two, the subject needed more time to finish task four successfully. The measure 'Average Count of Mouse Clicks' provides similar results of 20-30 clicks per task and person. The 'Average Summative Amplitude' per subject, on the other hand, shows different heights among the subjects and a standard deviation of 12254°. The results of the quotient building 'Fixations over Mouse Clicks' suggested by Lin and Imamiya are shown in table 1.

Table 1. Comparison of exemplary usability features

Subjects	Ø Quotient Fixations over Mouse Clicks	Total Time to Test Completion [Minutes]	Total Fixation Count	EmoCard Index
Subject 1	17,353	9:09	1,629	0.656
Subject 2	8,975	13:27	1,087	-2.75
Subject 3	9,525	14:51	1,150	0.719
Subject 4	2,725	32:14	1,320	-2.219
Subject 5	14,475	11:06	1,111	3.844
Subject 6	16,45	19:01	2,253	-1.625

For a more detailed examination, possible dependencies are calculated in a correlation analysis. The average count of mouse clicks strongly correlates positively with the average amplitude of saccades (.990). A medium negative correlation (-.719) exists between the average quotient of fixations over mouse clicks and the average amplitude of saccades. Another strong correlation (.949) exists between the average count of mouse clicks and the total test completion time. There is an equally strong correlation (.962) between the total time to test completion and the average amplitude of saccades. The total time to test completion correlates as well with the quotient fixations over mouse clicks.

Furthermore, the data has been analyzed with regard to the usability attribute user-satisfaction. The measurement of this attribute consists of the features and methods EmoCards, PAD Scale, questionnaire, and HRV. The subjects' choices and evaluation of single EmoCards as well as the individual classification of PAD Scale items show already many different emotional appraisals. These differences are easily accessible in the graphic analysis of EmoCards and PAD Scale. The analysis shows that subjects who chose predominantly negative EmoCards also show this prevailing mood in the ratings of the PAD Scale.

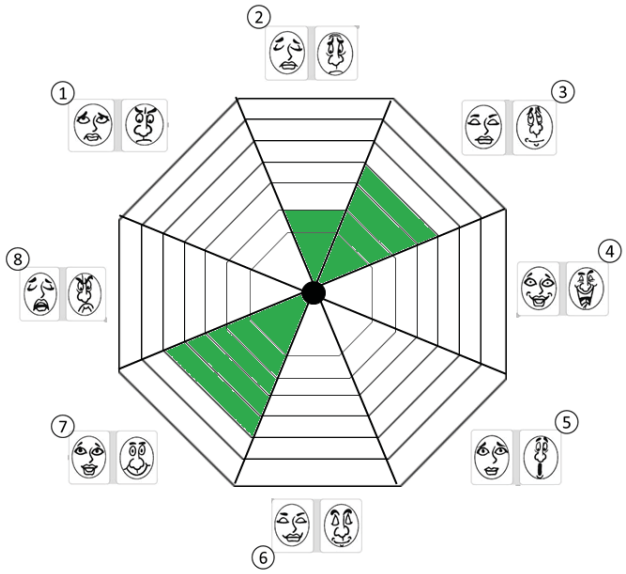

Figure 1. Exemplary EmoCard ranking after exercise 1.

For the analysis of user-satisfaction, the four adjective pairs 'unsatisfied-satisfied,' 'tense-relaxed,' 'annoyed-pleased,' and 'despairing-hopeful' are of particular importance. For examining the association between these four adjective pairs and the EmoCard Index, which is calculated for each subject, a second correlation analysis is built. It shows that a strong positive correlation between those adjective pairs exists.

Beside the presented subjective evaluation methods, data from the HRV measurements should be included and may serve as validation of the subjective results. Especially the frequency bands, calculated with the Fast Fourier Transform algorithm, VLF (very low frequency), LF (low frequency), and HF (high frequency) should be taken into account. The basis of the interpretation of the frequency bands should be as sound as possible. Therefore, a correlation analysis between the frequency bands and selected statements of the questionnaires is computed.

As a result, the following meanings of the frequency bands can be interpreted: the VLF gives evidence on user-satisfaction and a positive mood in general. The more satisfied and cheerful the user, the higher the energy in the VLF. The LF describes the user's effort completing tasks. If the LF power is low, the user

does not need much effort to finish a task. High energy in the HF frequency band indicates annoyance, despair, and tension.

Additionally, the mental workload has been examined. The values of NASA-TLX item 'Frustration' are the highest for the subjects who show an EmoCard index which has been rated mostly negative. The subject with the most positive EmoCard index rated the NASA-TLX item 'Mental Demand' the lowest. But the correlation analysis between NASA-TLX and the other measures does not show many associations. Most notably, there is no correlation between the HRV frequency bands and NASA-TLX. However, the quotient LF/HF shows a positive correlation with the items 'Total Workload,' 'Mental Demand,' 'Temporal Demand,' 'Performance,' and 'Effort.'

5. DISCUSSION

The integrated analysis of different attributes and methods concerning the four selected usability measures efficiency, effectiveness, user-satisfaction, and mental workload has shown that an extension of the attributes does provide additional information. Thus, (in our tests) the efficiency could be determined not only by the time to task completion, but also on the average amplitude of saccades, which is measured by eye tracking. Additional information on the effectiveness provides the quotient fixations over mouse clicks. The quotient gives information on the eye-hand coordination and allows insights into the information processing of individuals.

Measuring user satisfaction can be very diverse. With the presented approach of measuring emotions to conclude user-satisfaction or dissatisfaction, there is a possibility to collect this data in a more objective way. The study shows that the nonverbal appraisal procedure EmoCards is equally well suited as the verbal PAD Semantic Differential Scale. To validate the subjective results, the heart rate variability can be used to a certain degree. The tendencies of the emotional state of a subject can be fairly well read on the basis of the frequency bands.

As a result, an extensive picture of the usability of an interface can be drawn which provides new potential. The more accurate the information about usability, the better errors can be eliminated at an early stage of development. Referring to the research question, it can be stated that broader conclusions can be drawn about an examined usability criterion by a multidimensional feature analysis.

Furthermore, an answer could be given to the question of the relevance of emotions in the context of risk-sensitive applications such as medical devices. It can be of benefit to examine the emotions of users. Knowing them allows an implicit assessment of user-satisfaction. In addition, it has proven to be worthwhile to carry out the single surveys of emotion after each task. Thus, it can be accurately tracked at which points difficulties have arisen and triggered dissatisfaction or annoyance.

For future studies, a larger sample should be sought to ensure that the derived results can be validated. In addition, an accurate evaluation of Areas of Interest (AOI) could be meaningful. The number of mouse clicks within an AOI could be of importance for measuring efficiency.

6. REFERENCES

[1] Holzinger, A., Ziefle, M., and Röcker, C. 2010. Human-Computer Interaction and Usability Engineering for Elderly (HCI4AGING). In: *Computers Helping People with Special Needs*. Springer, Heidelberg, Germany, 556-559.

[2] Ziefle, M., Röcker, C., Wilkowska, W., Kasugai, K., Klack, L., Möllering, C., and Beul, S. 2011. A Multi-Disciplinary Approach to Ambient Assisted Living. In: *E-Health, Assistive Technologies and Applications for Assisted Living: Challenges and Solutions*. IGI, Niagara Falls, NY, 76-93.

[3] Röcker, C., Ziefle, M., and Holzinger, A. 2014. From Computer Innovation to Human Integration: Current Trends and Challenges for Pervasive Health Technologies. In: A. Holzinger, M. Ziefle, C. Röcker (Eds.): *Pervasive Health - State-of-the-Art and Beyond*. Springer, London, UK.

[4] Ziefle, M., Röcker, C., Kasugai, K., Klack, L., Jakobs, E.-M., Schmitz-Rode, T., Russell, P., and Borchers, J. 2009. eHealth – Enhancing Mobility with Aging. In: *Roots for the Future of Ambient Intelligence*, 25-28.

[5] Röcker, C. 2012. Universal Access to Awareness Information: Using Smart Artefacts to Mediate Awareness in Distributed Teams. In: *Universal Access in the Information Society*, 11(3), Springer, Heidelberg, Germany, 259-271.

[6] Ziefle, M., Röcker, C., and Holzinger A. 2011. Medical Technology in Smart Homes: Exploring the User's Perspective on Privacy, Intimacy and Trust. In: *Proceedings of COMPSACW'11*, IEEE Press, 410-415.

[7] Röcker, C. 2013. User-Centered Design of Intelligent Environments: Requirements for Designing Successful Ambient Assisted Living Systems. In: *Proceedings of CECIIS'13*, 4-11.

[8] Röcker, C. 2013. Intelligent Environments as a Promising Solution for Addressing Current Demographic Changes. In: *International Journal of Innovation, Management and Technology* (IJIMT), 4(1), 76-79.

[9] Kasugai, K., Heidrich, F., Röcker, C., Russell, P., and Ziefle, M. 2012. Perspective Views in Video Communication Systems: An Analysis of Fundamental User Requirements. In: *Proceedings of PerDis'12*, ACM Press.

[10] Bevan, N. 1995. Measuring usability as a quality of use. *Software Quality Journal*, 4, 115-130.

[11] Lin, T. and Imamyia, A. 2006. Evaluating Usability Based on Multimodal Information. An empirical study. In: *Proceedings of ICMI'06*, ACM, 364–371.

[12] Agarwal, A. and Meyer, A. 2009. Beyond Usability: Evaluating Emotional Response as an Integral Part of the User Experience. In: *Extended Abstracts of CHI'09*. ACM, 2919-2930.

[13] Young, S. R. 2005. Development of Usability Questionnaires for Electronic Mobile Products and Decision Making Methods. http://scholar.lib.vt.edu/theses/available/etd-08212005-234205/unrestricted/ETD_Ryu_Final.pdf. Accessed 21 February 2014.