COMENIUS UNIVERSITY IN BRATISLAVA
FACULTY OF MATHEMATICS, PHYSICS AND INFORMATICS

Application of Physiological Measures for User Experience Testing: A Skin Conductance Experimental Study

2018 **Bc. Andrej Brinkač**

COMENIUS UNIVERSITY IN BRATISLAVA
FACULTY OF MATHEMATICS, PHYSICS AND INFORMATICS

Application of Physiological Measures for User Experience Testing: A Skin Conductance Experimental Study

Master's Thesis

Study Programme: Cognitive Science
Field of Study: 2503 Cognitive Science
Department: Department of Applied Informatics
Supervisor: RNDr. Barbora Cimrová, PhD.

2018 **Bc. Andrej Brinkač**

Comenius University in Bratislava
Faculty of Mathematics, Physics and Informatics

81582856

THESIS ASSIGNMENT

Name and Surname: Bc. Andrej Brinkač
Study programme: Cognitive Science (Single degree study, master II. deg., full time form)
Field of Study: Cognitive Science
Type of Thesis: Diploma Thesis
Language of Thesis: English
Secondary language: Slovak

Title: Application of Physiological Measures for User Experience Testing: A Skin Conductance Experimental Study

Annotation: User experience (UX) testing in traditional human-computer interaction studies relies mostly on task performance and self-report data. Despite recent findings indicating that physiological measures, such as skin conductance, are a good index of cognitive involvement and emotional arousal, they are not a common part of UX measurements. This thesis should help to evaluate suitability of these measures for UX testing.

Aim: To explore the potential association between physiological measures (in particular, the skin conductance) and traditional usability measurements. The goal is to design and perform an experiment enabling to assess whether the level of skin conductance correlate with the subjective evaluation of task difficulty and objective task performance data.

Literature: Yao, L., et al. (2014). Using physiological measures to evaluate user experience of mobile applications. Int. Conf. on Engineering Psychology and Cognitive Ergonomics (pp. 301-310). Springer, Cham.
Foglia, P., Prete, C.A., Zanda, M. (2008), Relating GSR signals to traditional usability metrics: Case study with an anthropomorphic web assistant. Instrumentation and Measurement Techn. Conf. Proc. (pp. 1814-1818). IEEE.
Liu, S., Zhang B.Y., Liu C. (2016). Research on the Application of GSR and ECG in the Usability Testing of an Aggregation Reading App. Int. Journal Bioautomation 20.3.

Supervisor: RNDr. Barbora Cimrová, PhD.
Department: FMFI.KAI - Department of Applied Informatics
Head of department: prof. Ing. Igor Farkaš, Dr.

Assigned: 27.08.2017

Approved: 17.09.2017 prof. Ing. Igor Farkaš, Dr.
Guarantor of Study Programme

.....
Student

.....
Supervisor

Univerzita Komenského v Bratislave
Fakulta matematiky, fyziky a informatiky

ZADANIE ZÁVEREČNEJ PRÁCE

Meno a priezvisko študenta: Bc. Andrej Brinkač
Študijný program: kognitívna veda (Jednoodborové štúdium, magisterský II. st., denná forma)
Študijný odbor: kognitívna veda
Typ záverečnej práce: diplomová
Jazyk záverečnej práce: anglický
Sekundárny jazyk: slovenský

Názov: Application of Physiological Measures for User Experience Testing: A Skin Conductance Experimental Study
Použitie fyziologických meraní pre vyhodnotenie používateľskej skúsenosti: experimentálna štúdia snímania kožnej vodivosti

Anotácia: Merania používateľskej skúsenosti (UX) sa v štúdiách tradične spoliehajú na výkon v zadaných úlohách a na subjektívne hodnotenie vlastných dojmov. Napriek výsledkom viacerých štúdií, ktoré ukazujú, že fyziologické miery, ako kožná vodivosť, sú dobrými ukazovateľmi kognitívnej angažovanosti a emočného nabudenia, v súčasnosti netvoria bežnú súčasť meraní UX.

Cieľ: Experimentálne preveriť predpokladaný vzťah medzi fyziologickými mierami (konkrétne hodnotami kožnej vodivosti) a tradičnými meraniami, zaužívanými v oblasti používateľského testovania. Navrhnúť a uskutočniť experiment umožňujúci posúdiť súvis (koreláciu) hodnôt kožnej vodivosti a subjektívneho hodnotenia obtiažnosti úloh, ako aj objektívne nameraného výkonu v úlohe.

Literatúra:
Yao, L., et al. (2014). Using physiological measures to evaluate user experience of mobile applications. Int. Conf. on Engineering Psychology and Cognitive Ergonomics (pp. 301-310). Springer, Cham.
Foglia, P., Prete, C.A., Zanda, M. (2008), Relating GSR signals to traditional usability metrics: Case study with an anthropomorphic web assistant. Instrumentation and Measurement Techn. Conf. Proc. (pp. 1814-1818). IEEE.
Liu, S., Zhang B.Y., Liu C. (2016). Research on the Application of GSR and ECG in the Usability Testing of an Aggregation Reading App. Int. Journal Bioautomation 20.3.

Vedúci: RNDr. Barbora Cimrová, PhD.
Katedra: FMFI.KAI - Katedra aplikovanej informatiky
Vedúci katedry: prof. Ing. Igor Farkaš, Dr.

Dátum zadania: 27.08.2017

Dátum schválenia: 17.09.2017

prof. Ing. Igor Farkaš, Dr.
garant študijného programu

.....
študent

.....
vedúci práce

DECLARATION

I hereby confirm that this master's thesis is entirely the result of my own work and I have faithfully and properly cited all sources used in this thesis.

Date: ..... Signature: .....

ACKNOWLEDGEMENTS

I would like to express my gratitude to my supervisor Barbora Cimrová for the useful comments, remarks and engagement throughout the learning process of this master's thesis. Also, I like to thank the participants in my experiment, who have willingly shared their precious time during the process of user testing.

Abstrakt

Vzhľadom na to, že používateľsky orientovaný vývoj webových aplikácií sa stáva čoraz väčším zdrojom konkurenčnej výhody, koncept používateľskej skúsenosti získal v poslednom desaťročí veľkú pozornosť nielen medzi odborníkmi, ale aj medzi akademickými výskumníkmi. Potreba, aby boli namerané údaje čo najpresnejšie, sa zvyšuje, aby sa pomohlo pri vytváraní užívateľsky priateľskejších rozhraní. Zatiaľ čo väčšina štúdií sa zameriava na "typické" metódy hodnotenia použiteľnosti rozhrania, akými sú čas dokončenia, úspešnosť alebo dotazníky, subjektívna povaha týchto metód spôsobuje obavy, pokiaľ ide o presnosť údajov. Touto diplomovou prácou sa preto snažíme prispieť k limitovanej literatúre týkajúcej sa nových, objektívnejších a efektívnejších metrík na testovanie použiteľnosti a navrhujeme, že merania fyziologickej odozvy možno aplikovať na oblasť používateľskej skúsenosti. Kombinujeme typické úlohovo-objektívne merania (úspešnosť dokončenia úlohy, čas dokončenia úlohy) a subjektívne hodnotenie účastníka (hodnotenie obtiažnosti úloh, dotazník používateľskej skúsenosti) s fyziologickým prístupom kožnej vodivosti. Naším hlavným cieľom v rámci tejto štúdie je teda analyzovať vzťah medzi fyziologickými meraniami a tradičnými meraniami použiteľnosti. Aby bolo možné vyhodnotiť, či medzi týmito spôsobmi meraní existujú štatisticky významné vzťahy, navrhujeme to experimentálne overiť korelačnou analýzou. Zisťujeme, či reakcia na vodivosť kože účinne identifikuje špecifické emócie v interakciách používateľov s webovou aplikáciou a či metóda merania kožnej vodivosti môže obohatiť alebo podporiť tradičné metriky použiteľnosti.

Kľúčové slová: UX, Používateľská skúsenosť, Používateľské testovanie, Kožná vodivosť, Interakcia človek - počítač, fyziológia

Abstract

With user-oriented web and application development becoming an ever-greater source of competitive advantage, the concept of user experience has gained wide attention over the past decade not only among practitioners, but also among academic researchers. The need to have measured data that is as accurate as possible arises, in order to aid in the process of creation of more user-friendly interfaces. While most studies resorts to 'typical' methods of evaluating an interface's usability, via measures such as completion time, success rate, or questionnaires, the subjective nature of the methods generates concerns as to the reliability of the data. Consequently, this thesis aims to contribute to the limited literature concerning new, more objective and efficient metrics for usability testing, by proposing that physiological response measures can be applied to the field of user experience. We combine typical task-objective measures (task success rate, task completion time) and subjective participant evaluation (task difficulty evaluation, user experience questionnaire) with the physiological approach of skin conductivity. Our main objective within this study is hence to analyze the relationship between physiological measures and traditional usability measures. In order to evaluate the relationships between these measures, we design an experiment with a subsequent correlation analysis. We investigate whether skin conductance response is effective in identifying specific emotions in users' interactions with a web application, and examine if skin conductivity measures can enrich or support traditional usability metrics.

Keywords: UX, User Experience, Usability testing, Skin conductance, Human computer interaction, physiology

Table of Contents

List of figures and tables

Introduction

User Experience (UX) is increasingly becoming a vital part of technology development processes, especially in the context of web and mobile application development. In addition, UX is a core element of the increasingly popular concept of user-centered design, which primarily focuses on the usability of a product, rationalized by the fact that users are the ones who will interact with the technology and therefore the interface should be designed to suit their needs. Under such a user-centric view, capturing, measuring and analyzing data on the interactions between users and applications becomes very important for a wide array of actors in the web or application development process. This data is often helpful for web designers, UX designers, as well as web developers, as it aids to identify the most crucial usability issues and helps enhance the information architecture or navigation with the aim to eliminate user frustration from the considered user interface. Furthermore, online marketers are also benefit from UX data, as such information helps them to grasp a better understanding of customer behavior within online environment.

Despite this observed growing importance of the field, UX studies still primarily rely on traditional measures such as time needed for task completion, number of clicks, task success rate, surveys, questionnaires, or observations. However, if we want to measure and predict the levels of stress users encounter during the interaction with particular tasks and applications, the reliance on the use of such traditional measures may become problematic. In general, all of the above-listed traditional measures are defined as 'subjective measures', as to a considerable extent, they depend on the user's memory and biases created by distributional assumptions. This may lead to wrong data recordings and subsequently flawed interpretations of the collected data. Therefore, if we want to obtain a better understanding of the nature of Human Computer Interaction (HCI) and the experiences faced by the user, it becomes desirable and necessary to extend the research into the area of neurology and physiology. In line with this proposition, several bodies of research have been recently published, which use physiological data for identifying the stress levels of users while interacting with an online interface (Foglia, 2008; Yao et al., 2014). These studies indicated that human physiology responds to emotional arousal, and found a correlation between users' stress levels and their physiological responses. In this thesis, a foundational overview on the area

of contemporary UX theories and physiological measurements will firstly be provided. Secondly, the methodology adopted in this research will be laid out, before presenting the findings of the experimental study. These findings will subsequently be discussed and key implications will be considered. Doing so, the aim of the study is to find whether a correlation exists between traditional UX methods and skin conductance measures (SCR) during user testing.

Motivation

To measure and interpret the physiological responses throughout the process of HCI is considered important, as it may help better the understanding of how people truly feel while interacting with various interfaces such as web pages, mobile applications, entertainment machines among others. The main motivation behind this study is to gain better insight into UX and physiological responses, and to identify the relationship between the level of skin conductivity and traditional UX measures while conducting a user testing of a mobile application.

Psychophysiological responses are considered a sub-area of psychology, researching emotional responses and their relation to the human behavior. Emotional responses of humans are classified into six essential categories, which are: happiness, surprise, sadness, anxiety, disgust and anger (Kim, 2004). These emotions can be measured by using psycho-physiological sensors, based on behavior acts—within usability testing, these include attention, perception, cognitive load, as well as stress reaction to a given stimulus. For the purposes of this research, the levels of skin conductivity of participants will be measured per each task performed, and will be then compared with users' respective self-evaluations of the task difficulty, and with other task performance-related data.

Interdisciplinary character of the study

This master's thesis lies in the category of UX research, which is part of a bigger research area of HCI. This area is concerned with studying how people act while interacting with computer-based systems through different types of interfaces. This area contains two main research streams, namely computer science and psychology. Computer science, because it is necessary to develop systems with which people can interact, and psychology, as the end users of these systems are a diverse set of people requiring the design process to consider the needs and characteristics of

every potential user. In addition, the topic of this presented thesis further finds overlap with the field of psychology, since psycho-physiological measures of electro-dermal activity reflect the current emotional states of users, as well as their levels of stress and cognitive load.

Research background

Over the past decade, rapid advances have been witnessed in the area of user interface and among HCI methods and technologies. Current development is more and more focused on the users themselves, and is aiming to clear the gap between humans and computers in the context of a user-machine interface. In certain research communities, this is sometimes taken as far as no longer distinguishing between the user and the computer as different or separate entities, but rather viewing it as a collaborative human-machine system, or a joint cognitive system (Hollnagel, 2003). Such propositions suggest the need for further progress in the research of HCI and underlies the importance of understanding of human emotional states while interacting with technology (Hudlicka, 2003).

It is well known in the HCI literature that the physiological responses of users typically correlate with their task performance (Zhai, 2006). UX, a unique part of computing research, relates to the processes that are influenced by or related to emotions. In other words, UX is the art and science of generating emotions among people who interact with products or services.

As mentioned earlier, emotions are a phenomenon present in every human being, and the differences between emotions experienced by individuals are observable on the basis of factors such as triggering event, intensity and duration. Every emotion, however, has its specific characteristics relating to physiological responses of the peripheral nervous system. Emotions are viewed as having evolved through their adaptive value in dealing with fundamental life-tasks. Each emotion has core distinct features: signal, physiology, and antecedent events. In addition, each emotion also has certain characteristics that are common with other emotions: rapid onset, short duration, unbidden occurrence, automatic appraisal, and coherence among responses. These shared and unique characteristics are seen as a product of the evolution of the human race, and distinguish emotions from other affective phenomena (Ekman, 1992).

One biosensor that is able to measure the emotional state of arousal is that of the skin conductance sensor, which measures the electrical changes in skin what is caused by sweat. Skin conductance (SC) sensor, therefore, measures the electro-dermal activity of the skin and any increases of SC can be observed during more stressful events or with higher attention, and lower conductivity while relaxing, sleep and less stressful events. Hence, skin conductance provides a functional signal of emotional response by measuring electro-dermal changes of the skin. This reaction can be measured with electrodes non-invasively placed on the index and middle fingers. While designing computer interfaces, one of our main goal is to make it as easy to use for users as it is possible. If a user is interacting with an interface for the first time, is dealing with a rather robust and complex website or a mobile application, or the usability of the interface is not satisfactory, users generally tend to express the emotion known as stress. Jacob Nielsen (1994), one of the most respected and most cited author within the UX stream of literature, suggests the evaluation of stress levels of users in order to understand how they perceive the interface's usability via direct methods such usability testing and talking to participants, conducting focused groups or by using questionnaires.

Research questions and objectives

For purpose of this thesis, two research questions are formulated:

- Do skin conductance measures vary according to task performance in a usability testing setting?
- Do skin conductance measures correlate with self-reported data from usability testing?

To be able to answer this question, the following objectives arise:

(i) To survey the existing stream of literature on user experience research and human physiological responses;

(ii) To design and execute a usability testing, necessary for data acquisition;

(iii) Conduct an analysis of results, comparing task performance data with the self-evaluation data and skin conductance data.

Methodology

To obtain answers to the prior-specified research questions, we execute a UX usability testing, based on a merger of two existing experimental studies. We take the task design and all of the experimental set up from the Yao et al. (2014) study, and enrich the methodology with a more robust SCR data analysis procedure undertaken in the study conducted by Foglia (2008). For the usability testing component, we use subjective evaluation methods like self-reporting of task difficulty and user emotional questionnaire, alongside with objective measures of the task completion rate and time needed for task completion. Results from these two methods are then evaluated through a comparative analysis of data from the skin conductance metrics.

A theoretical background

With competition getting fiercer in the mobile applications market, usability has become a very important factor for securing a competitive advantage over rivals. Usability testing is therefore a crucial tool for developing and designing more usable applications, and is increasingly a becoming a standard part of the application creation and optimization process. Yet, it is necessary not only to conduct usability testing, but also to opt for the right selection of testing methods and measurements, in order to correctly identify the biggest usability barriers. The use of time efficient, accurate and at the same time effective methods is hence needed. Numerous quantitative and qualitative UX usability testing methods exist— for instance user testing, focus group, heuristic evaluation, interviews, surveys, or questionnaires. Of these methods, which are generally considered 'traditional', user testing claims the most common status. One commonality between all of these measures can be found in their downside—data collection is difficult to be done in the real time and this data highly depends on the users' working memory capacities and understanding capabilities, making this data very difficult to verify. Data distortion may be also caused by the wrong interpretation of user feelings on purpose or unintentionally, due to different interpretations of the test conductor's guidelines (Hornbæk, 2006; Czerwinski et al., 2001). Notably, problems with purposefully not admitting to true expressions can be seen often among participants of older age groups, as in this type of participants, it was observed that they often attribute the struggles which occur during usability testing as their own failure rather than that of the poor usability of the tested interface.

Usability testing

The vast majority of the above-mentioned traditional, physiological and other measures used in the field of UX can be classified according to four different dimensions, which every researcher should consider beforehand, in order to be able to apply the most suitable selection for their particular research. These dimensions, presented by Bergstrom et al. (2014) are as follows:

I. Subjective vs Objective
II. Real time vs Delayed
III. Natural Context vs Artificial lab

IV. Invasive vs Non-invasive

Subjective vs Objective

One of the most commonly used UX methods is the self-evaluation of task difficulty on a Likert scale, where the participant evaluates for instance on scale of 1 to 5 the tested interface, with 1 being the least difficult and 5 the most difficult. As asserted earlier, this type of measuring method is rather subjective, as it highly depends on personal perceptions and on the capacity of the participant's working memory, which may be lead to distorted results if the users do not correctly remember their interaction with the interface. On the other hand, objective measures do not rely on the user's direct evaluation of the conducted tasks and their perceptions of said tasks. Rather, their actions are recorded and the necessary data may then be later extracted for the analysis. Such objective metrics include for instance time on task, number of clicks, completion rate or number of errors. Using this categorization, skin conductance measures (which will be used in this study) fall under objective methods, as they do not rely on the users' evaluation of their emotional arousal, but instead measure the electro-dermal activity in order to determine which HCI event triggered greater arousal.

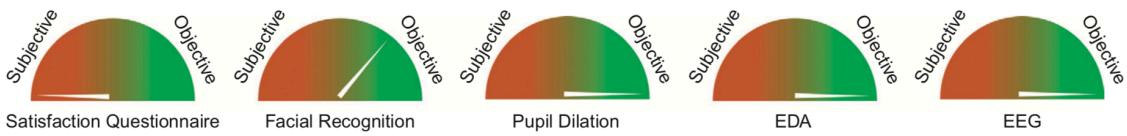


Figure 1 The objectivness scale of physiological response measurements (The more objective a measurement is, the better)

Real time vs Delayed

The next considered attribute of UX methods is concerned with the dimension of time. More specifically, if the data is readable and possible to be interpreted while the user is participating in usability research, it is classified as a real-time method. On the other hand, if the results can only be used after a certain period of time passes, the method falls under the category of delayed. Here, it is worthy to note that physiological measures are better in recording real time data, since this recording can be done without disturbing the participants. On the other hand, the frequently used user emotional questionnaires are given to the participant only after they finished a series of tasks, and therefore the response given in this questionnaire may not reflect

precisely and accurately the user's opinion on the task that was performed several minutes ago.

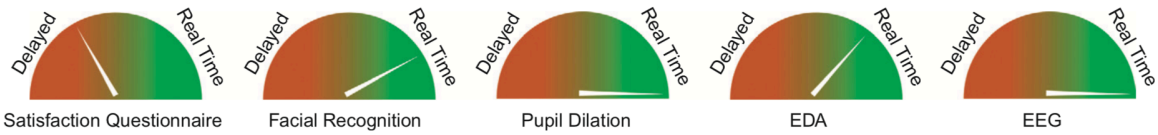


Figure 2 Real-time scale of physiological response measurements (The more real-time a measurement is, the better)

Natural context vs Artificial lab

Not a long time has passed since the first personal computers were made. In the era of desktop computers, it was acceptable to conduct usability measures in a lab space, which was very similar to the natural environment where individuals typically used computers. Yet, with the dawn of portable electronics such as laptop computers, smartphones and other devices, it became insufficient to conduct UX testing only in sterile lab environments. For instance, if usability tests are being conducted on GPS navigation systems, which are mainly made for drivers, it is evidently more sensible to conduct these tests in a natural context, in this case inside a vehicle and exposed to real road traffic. The ability to measure users in their natural enviroments brings about data with greater value, as it is more reflective of reality. Developments marked in measrument devices have also enabled researchers to obtain more precise data of higher quality— for example, using a portable eye tracker helps with the measuring of a vast majority of mobile interfaces outside of a lab setting.

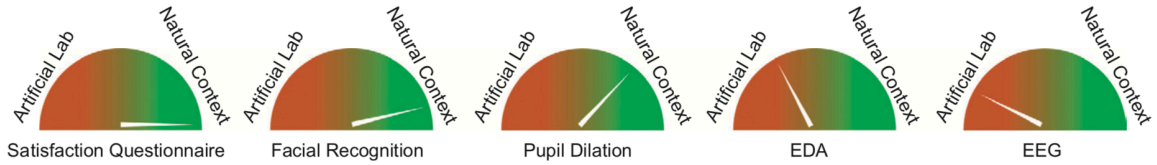


Figure 3 Natural context scale of physiological response measurements (The more natural context a measurement is, the better)

Invasive vs Non-invasive

When only traditional UX methods are considered, all of them can be considered to fall under the category of non-invasive methods. However, when we think of adding neurophysiological, psycho-physiological or other pharmaco-biological-based methods, we have to be aware about their levels of invasivness and try to choose the one which offers a good compromise of being less invasive on the

one hand, yet at the same time will deliver good data on the other. Most of UX studies conducted in the past do not use any particularly invasive methods, but in certain instances when researches wish to find neural correlates for a specific HCI event using neuroimaging techniques (like magnetoencephalography (MEG), functional magentic resonance imaging (fMRI) or possytron emmision tomography (PET)), this dimension may become a relevant consideration. Fortunately, the resolution currently obtainable from electro encephalography (EEG), which is a non-invasive technique, is sufficient also for the most complex of usability testings. Hence, this dimension presents less critical in comparison with the others in the context of UX research.

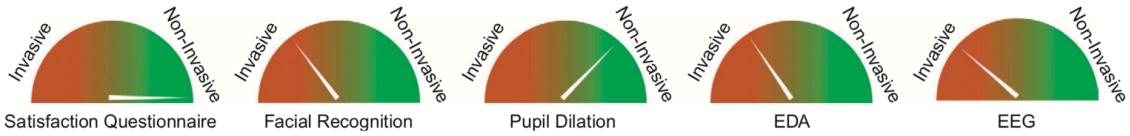


Figure 4 Invasiveness scale of physiological response measurements (The less invasive a measurement is, the better)

Physiological measures within UX

UX is founded upon human perceptions and interactions with interfaces. User emotional experience (UEX), is then focused on the physiological responses occurring during HCI events. Discovering various relationships between emotions and physiological responses over the course of HCI events allows researchers to go deeper in the area of UEX. A few years ago, it was only possible to identify major emotional responses to stimuli, such as experiencing software failure, listening to a favorite song or watching a scary video clip (Scheirer, 2002). All of these stimuli induced strong physiological responses or facial expressions, and hence were able to be evaluated using appropriate methods. However, these methods are not suitable for measuring rather subtle HCI events, which are more typical in a day-to-day life context in comparison with strong emotional stimuli. Therefore, finding appropriate physiological research methods continues to pose a great challenge in the stream of UX research (Ward et al., 2004).

Electrophysiology, the measuring of the physiological responses of our body to perceived reality, has developed significantly over the last decades. Hence, such new effective and efficient electrophysiological techniques allow us to tackle the challenge

faced by UX researchers in a novel manner—it enables the measures of UX in real time and presents an affordable way of conducting relatively robust usability testing.

In user research, several electrophysiological measures may be used, including electro-dermal activity (EDA), commonly referred to as skin conductance, electrocardiogram (ECG), electroencephalogram (EEG), electromyogram (EMG) and blood pressure (BP). Of these, EDA and EEG are identified as techniques which are most easily applied for the purpose of data acquisition, and offer an effective evaluation of physiological arousal, making them the most suitable for usability testing research (Ge, 2014). Consequently, the electrophysiological measuring of skin conductance is discussed next in this chapter, as it is this measure that has been selected to be used in this particular user research experimental setup.

Skin conductance

Skin conductance measurement is a non-invasive psycho-physiological method, where the signal of the skin conductance response is related to the changes in sympathetic nervous system (Picard, 2003). Electro-dermal activity is the most frequently used physiological measure in the area of HCI, as skin conductance is found to have a linear relationship with the level of arousal (Ganglbauer et al., 2009). In this study, we select this method to complement traditional usability metrics, in order to deeply examine the stress that users experience during performing tasks in a user testing. However, one has to be very careful with skin conductance measures while conducting UX research, as there are some discrepancies among findings in the existing body of research, making it is relatively easy to misinterpret recorded data. Some research has found the level of skin conductance to be positively correlated with instances when users were performing user testing on low usability interfaces (Ward, 2003), when several system errors appeared during testing (Pfister, 2011) or when other UX issues were artificially added by the test conductors (Liapis et al., 2015). Other studies, however, report a positive correlation between the skin conductance levels and the user's interaction with more usable interfaces in comparison to ill-designed ones. Lean (2012), for instance, finds such positive relationship in the case of mail server websites.

Another study, which finds evidence for the existence of a relationship between physiological responses and UX, is that of Wilson (2001). The study

measured the skin conductance of users with the aim to understand the subjective responses to a video conference software, while different video qualities were presented— specifically, the comparison entailed 25 frames per second versus 5 frames per second. Furthermore, Wilson (ibid) investigated for significance between skin conductance and other physiological measures, including the heart rate (HR), the blood volume pulse (BVP). Wilson's (ibid) research had an unanticipated outcome, as the correlation between physiological measures and video quality surprisingly did not exhibit the same relationship with traditional usability measures—the majority of users did not realize the worsening of the video quality. This suggests that by rather using psycho-physiological measuring methods, as in the case of Wilson's (ibid) study, we are able to get more precise and objective data about the users' true experiences when interacting with computers, in a manner that we are not able to do by using solely traditional UX methods.

All of the above-discussed research confirms the significance of a relationship between the user's arousal level and electro-dermal activity response. To be able to apply these measures to UX, however, there is an inherent need to well understand the context of the tested interface, in order to be able to correctly interpret the data. In addition, it is important to keep in mind that a lack of consensus prevails among the interpretations of the relationship between traditional usability testing measures and skin conductance in the existing body of knowledge (Foglia, 2008).

Neurophysiological basis of skin conductance

Electro-dermal activity consists of two components, namely tonic and phasic activity. Tonic changes in skin conductance occur slowly and are measured on a relatively long timescale. For instance, tonic levels of emotional arousal can be measured at the beginning of a task and at the end of a task, after which the difference is calculated. This difference has to be normalized to users' baseline tonic levels. Phasic activity, on the other hand, is a quick response in skin conductance to sudden stimuli. The difference in the levels of tonic skin conductance is therefore measured over time, while phasic activity is measured as a reaction to subsequent stimuli.

In the area of consumer neuroscience, researchers are interested in measuring both the tonic and phasic activity, yet the operationalization of this might be rather challenging due to the difficulty of correct measurement. A main problem lies in

clearly specified threshold values often lacking in published reports, bringing about a level of ambiguity (Bergstrom, 2014). Researchers also often forget to bear in mind the fact that the rise time of skin conductance response is shorter than the recovery time, and this information may be important when analyzing results, especially when an activity overlaps (Bergstrom, ibid).

Limitations of the skin conductance measuring method

While skin conductance is a valid and a useful tool for measuring human emotional arousal, it is not without its limitations, as the previous section began to illustrate. One of the most critical shortcomings is that skin conductance can only measure the power of the emotion, meaning the data cannot be used to distinguish the types of emotion (or combinations of emotions) that were recorded. Therefore, it is difficult to discriminate if the heightened emotion that was observed was stress, arousal or anxiety. In other words, similar to other physiological measurement techniques, skin conductance response is able to show only the amplitude of arousal, but fails to identify the valance. This is the main reason why using the skin conductance method alone, without complementing it with other methods, will in a majority of cases not be justifiable, as the method will be not able to determine if emotional arousal is positive or negative. Hence, it is important to note that while using the skin conductance method for UX research is rising in popularity, the accuracy of the results remains to be a challenge. Another limitation of the skin conductance response method lies in the huge individual differences between skin conductance responses to the same stimuli (Foglia, 2008). Due to this presence of subjectivity in signal data, one has to be very cautious when designing experimental studies, as well as during data interpretation. For instance, it makes little sense to directly compare two data sets from two different users, as they may have distinctly different basis thresholds and peaks (Foglia, ibid). Thus, the data collected from two different users are highly difficult to compare, not only due to individual inconsistencies in the skin conductivity response, but also because some users may have a very low and flat response in EDA. In order to systematize this discrepancy, Picard (2003) observed that a flat response of EDA is present in highly extroverted users, while introverted user produce EDA traces that peak more sharply. In addition,

Picard (ibid) also identified the following six fundamental patterns, which are found to hold for skin conductivity response traces:

I. Learning effect: an experienced event has a lesser effect than a novel one;

II. Summation effect: one single big event can be less powerful than many smaller events;

III. Time variant effect: same event can cause different effects on the same user at different times;

IV. Recurrent pattern of emotions: an emotion typically causes a steep increase in the physiological signal (peak) followed by a smooth decrease;

V. Subjective effect: same event can cause different effect on different users;

VI. Relaxation pattern: a relaxed user shows a trace that decreases gracefully and continuously.

The final major problem faced in using the skin conductance method, as well as physiological measurements in general, is the latency of users' physiological responses. The ability to eliminate the latency problem from a recorded data set is still very limited, mainly due to unexpected latency intervals of responses (Andreassi, 2000). Nonetheless, measuring of subjects' skin conductance levels remains a useful tool with a wide array of applications, and if applied correctly and with a degree of caution, can provide researchers with valuable insights.

Summary

With the importance of UX in web and application design gaining momentum over the past years, it was asserted that using physiological response measurements as part of UX research can enable researchers to obtain more objective data from participants about their emotional involvement during task performance. By considering four key dimensions classifying UX measurement methods, it was shown that non-traditional methods, such as skin conductance response, could provide more precise data. Instead of users subjectively evaluating via a questionnaire whether the interaction with a computer interface does or does not evoke stress, interest or excitement, physiological measures can reveal whether emotional arousal indicating stress, interest and excitement were actually present during the usability testing stage, by capturing the change in electro-dermal activity. While such method is not without its

limitations, it can be considered a useful complementary technique to gain deeper understanding of users' emotional processes during HCI events.

Methodology

As outlined earlier, in this thesis, the physiological response measure of skin conductance will be applied to a HCI setting through an experiment. The aim is to find whether these measures give consistent results to traditional UX research methods and hence can serve to enrich the UX toolkit. Using an interdisciplinary approach melding knowledge from physiological and HCI research, we conducted usability tests, in order to quantify how a user's skin conductivity behaves consistently with task performance reports and self-reported data. This study bases itself on the limited number of existing studies that have examined the problem in the past (Foglia, 2008; Yao et al., 2014), modifying their adopted methodologies as will be specified in this chapter. Emotional stimulation in artificial settings such as the laboratory (Liapis et al., 2015) proved a great challenge in the previous studies, and hence the modifications applied in this experiment aim to reduce these shortcomings.

Hypothesis formulation

The primary objective of this study is to evaluate using statistical analyses the relationship between physiological measures and traditional usability measurements, and whether SCR measures may help to produce more robust usability evaluations. Consequently, the hypotheses to be tested in this thesis are laid out as follows:

H_1: There is a positive relationship between SCR levels and objective UX measures.

H_2: There is a positive relationship between SCR levels and subjective, self-reported UX measures.

Experimental design

In order to contribute to the existing body of knowledge in a consistent manner, allowing for superior discussion, the study adopts a comparable methodology to that of Foglia (2008) and Yao et al. (2014). Hence, the correlations between SCR results and traditional UX measures are evaluated by designing a within-group experimental setting, whereby all participants complete 5 tasks in the same order.

Foglia (2008) asserts that SCR measures are important to calibrate in the beginning of the experimental setup due to very high individual differences in physiological responses. Hence, consistent to his recommendations, in order to

improve between-user comparisons, a stable baseline is defined for each participant. Participants' individual stress and emotional levels may vary due to several reasons, such as stress from the usability lab premises, unpleasant feelings from being connected to SCR measuring equipment among many others. Due to these underlying potential stress factors, we start the experimental session with welcoming the participants and will try to induce a friendly atmosphere in the lab. Afterwards, we will inform them about the course of the experiment, all of its parts and will emphasize to them that we are not going to test them, that this is not a test and that there is no correct way how to interact with the web interface. Rather, we clearly explain that it is a usability testing and therefore we are focusing on evaluating the user friendliness or unfriendliness of a tested web application. This information should help them to not be overly stressed during the testing period and not to blame themselves if they fail at a task. Next, to better calibrate the SCR and to make the participant feel more relaxed, a short 3 minute emotionally neutral video clip is presented to them. After participants finish this video clip, there is a 1 minute waiting period, right after which the first of the five tasks is presented to the participants.

Starting from the moment the short video is played, all activities and corresponding SCR levels are recorded to be analyzed. After finishing each of the required tasks, a participant is asked to evaluate the difficulty and amount of mental effort put into the given task on a Likert scale. In addition, 15-second resting phases are given to each participant between tasks, in order to stabilize their SCR levels. At the end of the testing, a participant is asked to fill out the widely-used user experience questionnaire (UEQ-Online, 2018) for a standardized usability evaluation. The overall duration of the usability testing was on average around 30 minutes.

Participants

For the purposes of this experiment, 7 healthy participants were recruited. The group was balanced by gender (3 males and 4 females) and the age range of participants was between 24 and 37 (with the mean age of 28). All participant can be described as tech-savvy, but none of them have previous working experience in the area of HCI nor UX. Furthermore, these participants have not had the experience of participating in a usability testing in the past. While the participants received no compensation for their time devoted to the usability testing process, lower quality of

data does not pose a major concern, as these participants voluntarily signed up for the experiment, knowing there will be no tangible reward.

Tasks

As for the tasks given to the participants in the testing phase of the experiment, we used the same tasks as was used in the Yao et al. (2014) study. Participants were asked to perform five tasks. All tasks were characteristically different and were ordered in the same manner to minimize the learning effect. The only difference from the study of Yao et al. (ibid) is in the use of a different mobile application, since the original interface is in the Chinese language, with no English nor Slovak translation, making it unsuitable for the selected group of participants. Hence, we have instead selected a more appropriate application for our needs—in this case the travelling review and guide application of Tripadvisor was used. The five required tasks to be performed by the participants were to book a hotel (Task1), find a restaurant according to specific criteria (Task 2), find directions from a hotel to a given restaurant (Task 3), give a review to this restaurant (Task 4) and to search for a travel guide with the best reviews (Task 5). More specifically the tasks were worded as follows:

1. Find the cheapest a hotel in the Piccadilly part of London for 2 nights and 3 persons that lies in the price range of 100-150 euro per night.
2. Find a restaurant that serves fondue in Geneva, Switzerland that has the most reviews.
3. Comment under this restaurant to write a one-word satisfied commentary and star the restaurant.
4. Find a best route to this restaurant from the President Wilson hotel, Geneva, using public transport.
5. Find a travel guide for the city of Paris, France with the best reviews.

Data collection

The above presented tasks were performed on the Tripadvisor mobile application using the widely-used smartphone iPhone 8 Plus, with a screen size of 5,5 inches and a screen resolution of 1920x1080 pixels. Three types of data were

collected for the purposes of this thesis, as will be described in greater detail in this section.

(i) Task performance measures

Two task performance measures were collected, namely the success rate (on a binary pass or fail basis) and the task completion time. To get accurate data for both of these measures, a screen recording application was running throughout the course of the usability testing, starting from the first task. This method of data collection was used, since the use of a screen recording device makes the collection less disturbing for participants, in comparison to collecting the success rate by questioning the participants for answers or manually counting the time which was needed for task completion.

(ii) Self-reported data

To gather the subjective data needed for the experiment, a task difficulty and the UX questionnaires were distributed (See Appendix A for a preview of these questionnaires). Both questionnaires were filled out online by using the Google Forms web application. The task difficulty questionnaire was filled out by participants after each task and took the format of a 5-point Likert scale, where 1 stand for the least difficult and 5 for the most difficult. The User Experience Questionnaire was derived from an online source (UEQ-Online, 2018) and was consistent with the widely used format consisting of 26 usability attribute pairs. Participants were asked to choose on a Likert scale of 1 to 7, each end of the scale represented an extreme of a specific attribute (for instance pair: easy to learn; difficult to learn).

(iii) SCR data

Skin conductance data was recorded using the g.GSRsensor 2 sensor, which contains two electrodes fixed without gel on the index and middle finger of the left hand. The sensor has a range from 0 - 30 μS. For receiving the SCR data, a CE-certified g.USBamp bio-signal amplifier was used, for transferring the collected data and g.Recorder software was used. Calibration of the sensor was done by using the 1μS calibration button. Furthermore, the sampling rate was set up to 512Hz. Preprocessing of data consisted of the application of the Notch filter at a range of 48-

52 Hz, the low-pass filter at a value of 100Hz and the high-pass filter at 0.1Hz. To time-lock the data from SCR to task performance data, a set of 6 markers were used within the g.Recorder software, five for the beginning of each task and one for the finished task. With the aid of these markers, we were able to identify the exact times of the beginning and end of each task afterwards.

Data analysis

The data collected from the user testing was then subject to analysis using three different approaches—task performance approach, self-evaluation approach and physiological approach. As has been asserted earlier, in this study, we investigate the relationship between traditional usability testing data (task performance data, self-evaluated data) and skin conductance measures. In order to do this, four measures were compared with SCR levels, namely the task success rate, task completion rate, task difficulty evaluation and the user experience questionnaire.

Since the physiological data have a degree of subjectivity and are influenced by an array of factors, a set of metrics were considered for every participant, to obtain their tonic skin conductance level from the SCR recording. This was done by using the Matlab software, where a script can be used to transform raw data in HDF5 format to a Matlab-readable form. Subsequently, vectors of SCR for each of the five tasks were identified by using markers, and were also used for identifying the baseline. The baseline was calculated as the average SCR value from 30 seconds before the first task. Furthermore, SCRmin, SCRmax, SCRmean were calculated for the following three the physiological metrics applied in this study, consistent with the approaches taken on by numerous similar studies (e.g. Foglia, 2008; Lin, 2005; Moore, 2004; Shi, 2007):

I. average galvanic skin conductivity on single task;

II. average galvanic skin conductivity on single task normalized with the minimum;

III. SCR ratio on single task $SCRratio = \frac{SCR_{max} - SCR_{min}}{SCR_{min}}$.

For all of the analyses, we primarily relied on the use of the SCR ratio, as it is suggested to be the most reliable, diminishing the differences between individuals (Mirza-Babaei, 2017). However, all three SCR measures were used for the analysis relating to the data on task difficulty for purposes of superior comparison.

Firstly, we performed a task performance analysis, where the task success rate and task completion time values were compared with the participants' SCR ratio values. For the analysis of the task success rate, we compared the average value of the SCR ratio data obtained during tasks with the successfulness of participants. For more precise data, the analysis only considered data from participants that have at least one successfully finished tasks and one failed task (Foglia, 2008). We assessed the successfulness of each of the tasks and per each participant by watching screen recorded data. Then we filtered all participants that had at least one successful and one failed task during user testing. Subsequently, the SCR ratio values were obtained by using the analytical software Matlab. Next, to identify the relationship between these two measures, we used a repeated ANOVA analysis—as done in a comparable study by Yao et al. (2014), to identify the relationship between the task success rate and the level of skin conductance.

Secondly, Spearman's correlation analysis was conducted for the time needed for completing a task and the corresponding SCR ratio for each of the tasks. The task completion time was measured from a participant's first interaction with the mobile device for each task, until they finished it.

For the analysis of the self-reported data, we considered the task difficulty and user experience questionnaire. The task difficulty analysis was again done using Spearman's correlation analysis where the first variable was the user evaluation of task difficulty and the second variable was the SCR value. We conducted this analysis three times in order to gain more robust findings, each time with a different SCR measure (i.e. SCR mean, SCR mean standardized with the minimum, and SCR ration). Furthermore, the task difficulty measured by the questionnaire was put to Spearman's correlation analysis to investigate the relationship between task difficulty measures and SCR data as well as for identifying the difference between three different SCR type of data.

To evaluate the overall UX of the mobile application, a user experience questionnaire was used (Laugwitz et al., 2008). This questionnaire is a widely used tool for the evaluation of the UX of all kinds of computer interfaces. The questionnaire consists of 26 bipolar items, of which each tuple is evaluated on a seven-point Likert scale. These items are then processed for evaluation and transferred via factor analysis into six key dimensions of attractiveness, perspicuity, efficiency,

dependability, stimulation and novelty. An overall score for each dimension is then calculated on a scale of -3 to +3. Next, for the comparison of the overall UX of the tested mobile application and SCR data, a Spearman's correlation analysis was conducted to assess the nature of the relationship between the SCR ratio and UX questionnaire results.

Limitations

While the methodology adopted in this thesis has been carefully designed and is based primarily on the conducted in the past, it is not without its limitations. Firstly, the statistical analyses required for the evaluation of the physiological measures may be problematic with our recruited sample of 7 participants. The small number of participants may result in the significance power being adversely affected. On the other hand, the selection of this number of participants is justified by the range that is typically recommended and widely used in usability testing— the standard number in this context is 6-8 participants (William, 2013). Hence, in a usability testing setting, larger numbers of participants do not tend to lead to superior data, as the findings produced are relatively homogeneous.

Secondly, as the chapter providing a theoretical background has outlined, the SCR measure is only capable of uncovering the intensity of an emotion, not the attribute of the emotion. Hence, while SCR measures changes in the amplitude, it cannot distinguish whether the user is experiencing is that of stress or that of arousal. By obtaining SCR data and comparing it with traditional measures of UX, we can thus only infer that the emotion was likely stress. Furthermore, the relatively high differences in SCR levels between individuals can often pose a challenge. While effort has been put in to mitigate the effects of this issue by taking the SCR ratio instead of other SCR measures, the difference cannot be diminished completely and one must be cautious when handling this kind of data.

Thirdly, as has been discussed earlier when looking at the dimensions of UX methods by Bergstrom et al. (2014), it was established that natural settings are superior to that of the laboratory, as it produces more realistic data. However, the tools required for measuring of SCR levels do not allow for testing in a natural environment, bringing in a degree of artificialness. Furthermore, the carrying out of tasks by participants was not a result of their inherent need to obtain certain information. Rather, the 'need' to find information was artificially created by the

experiment conductors. This again may lead to slight distortions in the SCR levels, as the true need may have produced different data.

Findings

The aim of this chapter is to present the findings of our study, in order to provide a pilot exploration of the concept, rather than to provide robust claims. As discussed over the previous sections, when examining the suitability of a tested interface, task performance metrics and subjective user evaluations form an important part of the data acquisition process in a traditional user testing setting. In order to contribute to the knowledge in the field of UX and HCI in a consistent and incremental manner, this study compares, rather than substitutes, these traditional metrics with a physiological response measure of skin conductance. Consequently, there are two sections in this chapter, of which the purpose is to present the results from the prior-specified experiment. In the first section, the results from objective measures of usability testing, specifically task success rate and task completion time, are provided, and supplemented by a comparative analysis of these task performance results with skin conductance measures. The second section consists of an analysis of the subjective self-reported data, i.e. task difficulty evaluation and user emotional questionnaire, also paired with a comparison to the physiological response results. Matlab was used for data processing purposes, and subsequent analysis was carried out with the aid of Microsoft Excel.

Task performance analysis

The objective measures used in this experiment are task success rate and task completion time, and are defined for the context of the study as follows:

Task success rate: Count of tasks that a participant was able complete, in relation to the ones which he/she could not

Task completion time: Time that each participant needed to finish a task, measured from the first interaction with the interface after they read the task instructions.

(i) Task success rate

To investigate the relationship between task performance measures and SCR data, the SCR ratio is calculated and compared for each participant and for both successful tasks and failed tasks alike, as shown on Figure 5. The SCR data pairs used compose of that from 6 participants, as one participant was successful in all tasks and

therefore a measure from this participant was not able to be used, leaving no data for comparison. A repeated ANOVA measure shows a significant relationship on the SCR ratio measures and between successfully finished tasks and failed tasks, as evidenced by $F(1,6)=8.76$ and $p=0.05$. Furthermore, the correlations presented in Figure 5 shows that participants have higher levels of SCR if they failed at a task. This result is consistent with other studies (e.g. Yan, 2014; Lin et al., 2005), which arrived at the same results and therefore works to support the proposition that stress levels rise when users are unable to complete an assigned task, and confirms that this phenomenon can be measured via an increased level in SCR. Hence, this finding serves to support the propositions of the hypothesis H_1, where a positive relationship between objective measures and SCR levels was anticipated.

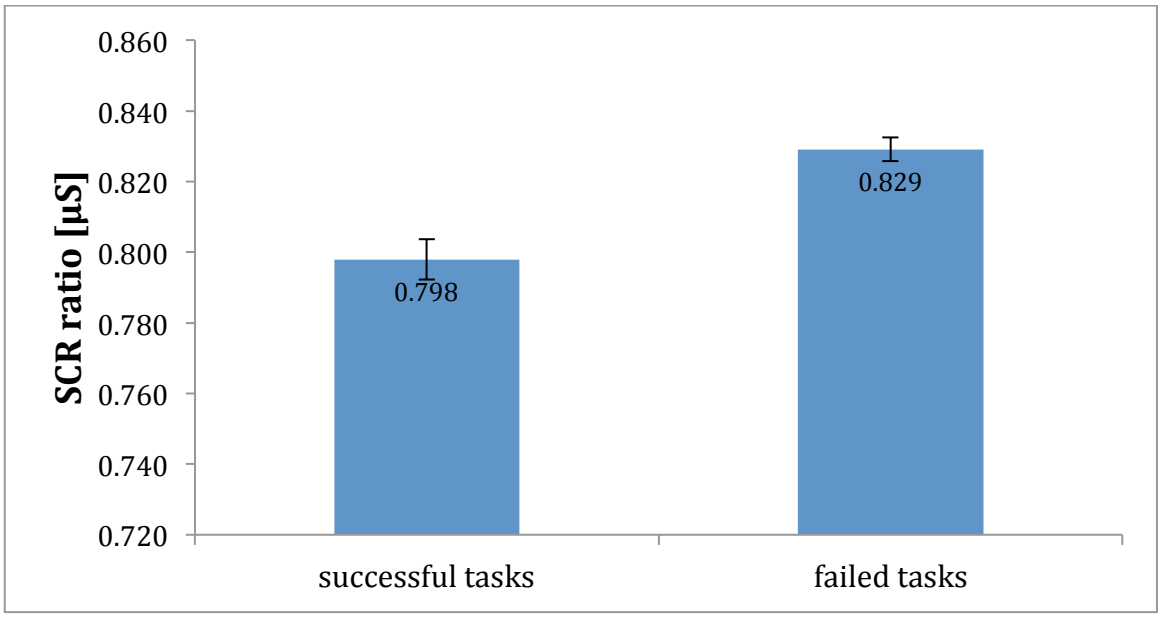


Figure 5 SCR ration in successful and failed tasks. Error bar stands for standard error

(ii) Task completion time

The physiological data were also analyzed in relation to the time that a participant spent on each task. Overall, the presented results show that tasks with lower completion times tended to result in lower measured SCR, as can be seen on Figure 6 below. However, this relationship failed to exhibit statistical significance at a 5% level when using Spearman's correlation analysis ($r = 0.31$, $p > 0.5$), making confident assertions about the examined relationship difficult. Therefore, the positive relationship hypothesized by H_1 does not seem to hold in a significant manner when

task completion time, rather than the task success rate, is considered. There are two potential reasons that may explain why the relationship lacks significance. Firstly, it may be caused by a relatively small sample (in this case the number of participants), which resulted in subtle SCR differences not being distinguishable. The Second potential reason might be the huge disparities between participants' total completion times, where the shortest time needed to complete all five tasks was 6 minutes 37 seconds, in comparison to the longest time of 12 minutes 07 seconds. These individual differences increase the variance of the considered dataset, and may adversely act to hinder the significance of the relationship between task completion time and SCR.

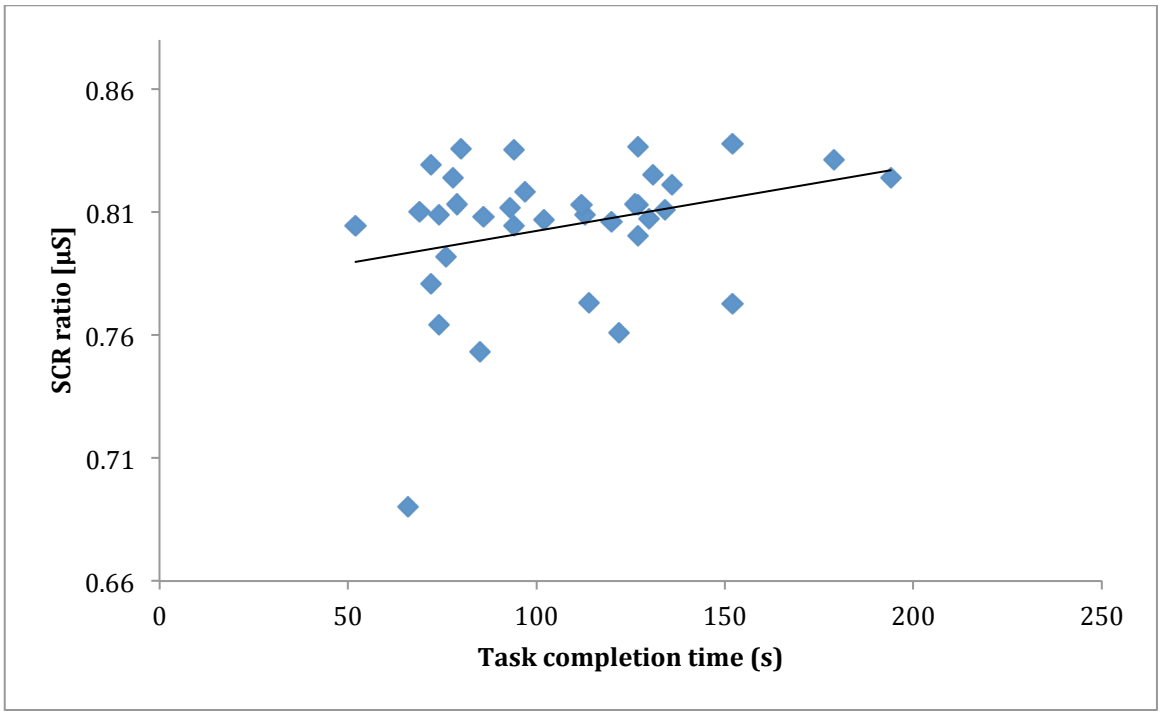


Figure 6 Spearman's correlation analysis of SCR ratio and task completion time

Self-reporting data analysis

After processing the collected SCR data, the SCR ratio values were also analyzed in relation to the subjective task difficulty measures, in order to assess the validity of the hypothesis H_2. The subjective UX measuring methods used for the purposes of this experiment are individual task difficulty evaluation and the user experience questionnaire. These are to be understood in the following manner:

Task difficulty evaluation: Users' individual evaluations of each task's difficulty, whereby participants are asked to assess the task on a Likert scale of 1 to 5 after every completed task. 1 stands for not difficult and 5 for most difficult.

User experience questionnaire: A 26 item questionnaire filled out by each participant after the completion of all 5 tasks. The aim is to measure the six dimensions of UX, namely attractiveness, perspicuity, efficiency, dependability, stimulation and novelty.

(i) Task difficulty evaluation

In order to assess whether a relationship exists between subjective evaluations of task difficulty and measured physiological responses of participants, correlations were analyzed across each task. Specifically, Spearman's correlation analysis was conducted, evaluating the relationship between SCR ratio of skin conductance levels and the participants' Likert scale evaluations of task difficulty. The data used for this analysis is graphically summarized in Figure 7 and 8, which compare the two sets of results- subjective and physiological. As one can intuitively infer from the lack of observable link between the two figures, the correlations did not exhibit statistical significance—evidenced by the statistical metrics of r=0.2, p=0.74.

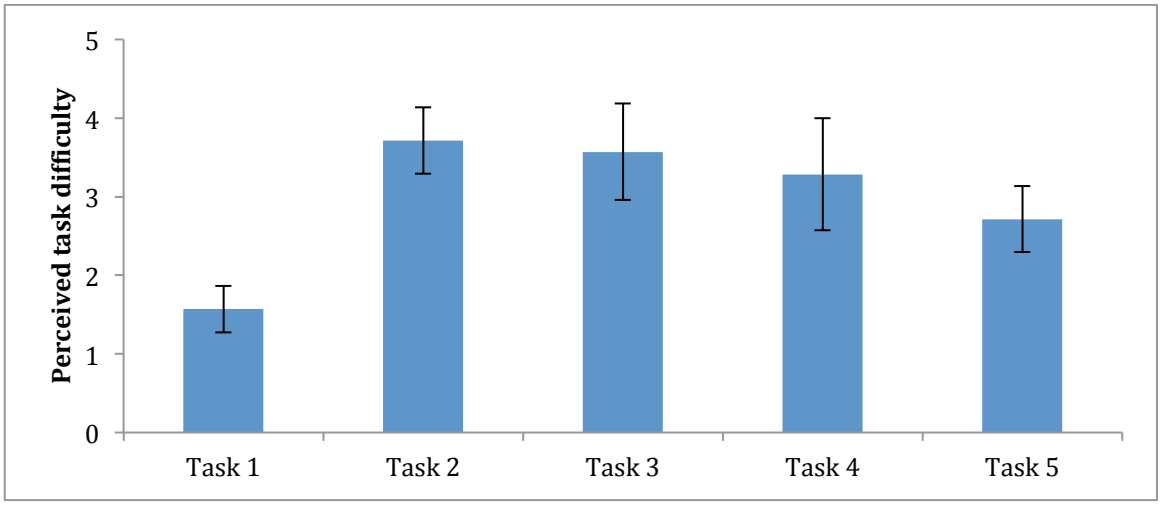


Figure 7 Percieved task difficulty (average score on a five-point Likert scale - 1 = not difficult, 5 = most difficult). Error bar stand for standart error.

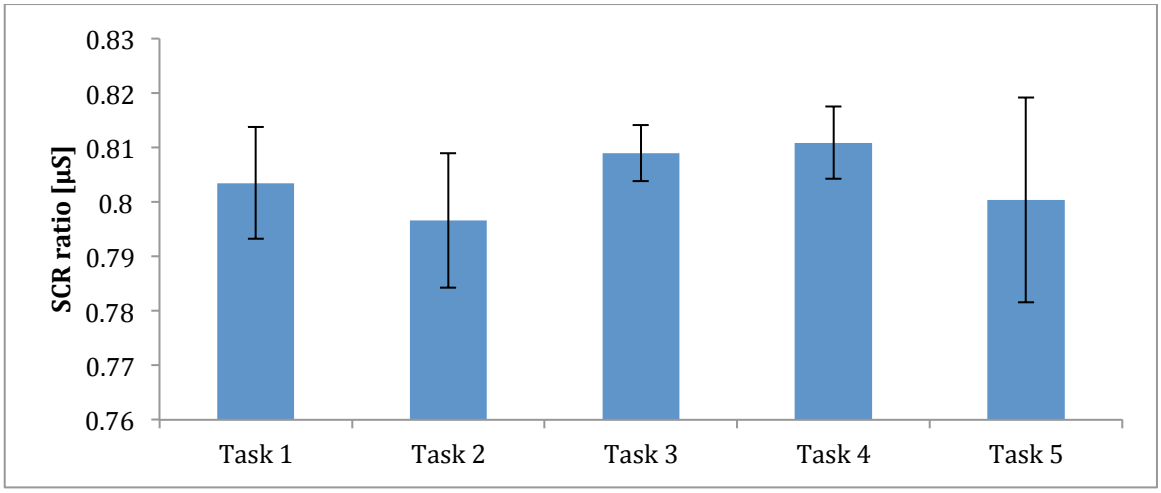


Figure 8 SCR ratio of the five tasks. Error bar stands for standard error.

In order to mitigate the impact of potential biases caused by differences in SCR levels among tested individuals, the relationship with the perceived task difficulty has been examined using three different SCR measures—using the mean SCR, mean SCR normalized with the minimum, as well as the SCR ratio defined in the Methodology chapter. While all three measures exhibited a positive relationship, all of them nonetheless lacked in statistical significance (SCR mean r= 0.13, p=.45; SCR mean normalized with the minimum r= 0.25, p=.34; SCR ratio r= 0.42, p= .13). Hence, by looking at this subjective measure, we cannot accept the proposition of the hypothesis H2 predicting a positive relationship between SCR levels and subjective UX research methods.

(ii) User experience questionnaire

As established earlier, the user experience questionnaire used for the purpose of this study has been evaluated by transforming the data via a factor analysis into six key dimensions of attractiveness, perspicuity, efficiency, dependability, stimulation and novelty. Hence, it is important to keep in mind that the questionnaire does not intend to produce an overall score of the usability and UX of the tested interface. Because of the way the questionnaire is constructed, it would not make much sense to attempt to combine the data into an overall score (for instance by calculating the mean across all six dimensions).

The UX of the mobile application tested in this study was seen positively by users in four out of six dimensions, more specifically in the dimensions of attractiveness, perspicuity, efficiency and dependability. This finding is inferred from

the mean being greater than zero for these four dimensions, as Table 1 shows. Stimulation was the only dimension that resulted in negative user ratings and the dimensions of novelty was overall perceived to be neutral. In order to solidify these findings, the consistence of the scale is measured and the Cronbach Alpha-Coefficients are calculated. While there is no generally accepted value for the coefficient, many authors refer to alpha >0.7 as the threshold value for significant results. As the alpha values in our results were found to be greater than 0.7 for all of the six considered dimensions, the presented assertions can be deemed to hold with a relative degree of confidence.

UEQ score	Mean	SD
Attractiveness	0.28	1.18
Perspicuity	0.15	1.06
Efficiency	0.15	1.17
Dependability	0.33	1.28
Stimulation	-0.15	0.94
Novelty	0	0.94

Table 2 UX questionnaire results transformed by a factor analysis into 6 UX dimensions (data is presented as scaled means of a 7-point Likert scale into a -3 to +3 index)

In addition to the traditional analyses of the UX questionnaire presented above, Spearman's correlation analysis was conducted, in order to if there is a significant relationship between each one of the six UX dimensions and SCR ratio values, as depicted on Figure 9. Of the six dimensions, positive relationships with a statistical significance at a 5% level was found in the stimulation dimension ($r=0.35$, $p< .05$) and the novelty dimension ($r = 0.39$, $p<.05$). This implied that greater the perceived stimulation and novelty of an interface by the user, the greater the skin conductivity of the user. The remaining four dimensions lacked significance in their relationship with the measured SCR levels. Therefore, in this case, we can only assert that the hypothesis H_2 has partial validity.

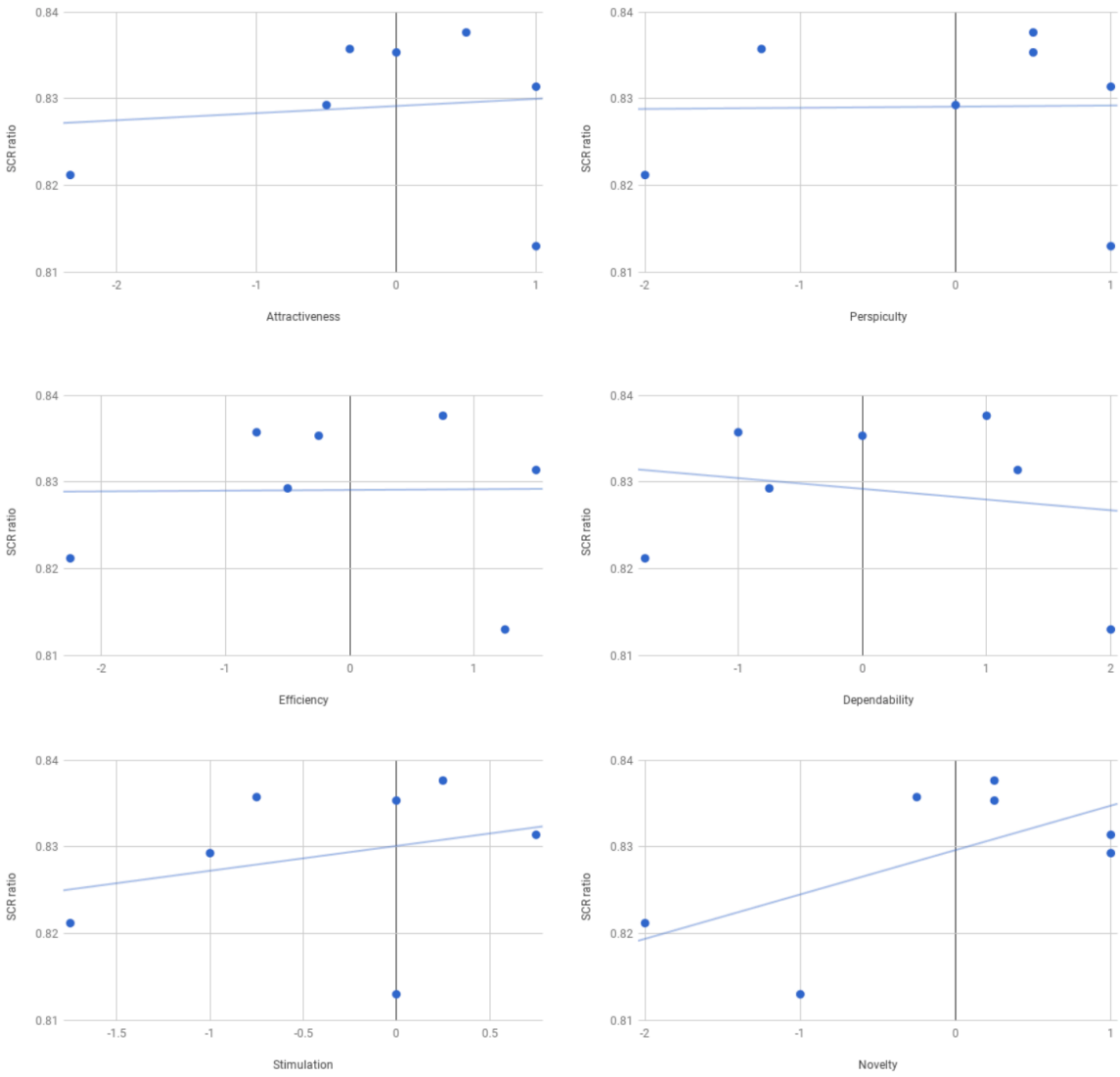


Figure 9 Spearman's correlation analysis of UX questionnaire results against SCR ratio values

Discussion, conclusions and possibilities for future research

The presented study has assessed whether a link or a disconnect exists between traditional UX research methods and physiological response measures, in order to explore whether these measures can be used to enrich usability testing. The analyses of the previous chapter have shown that indeed, participants exhibit a higher level of SCR when unsuccessful at a task, implying the stress associated with a user-unfriendly interface can be monitored with this metric. This finding supports and fortifies the conclusions reached by Yao et al. (2014), expanding the knowledge to a sample characterized by a different demographic. Furthermore, the statistically significantly positive correlations found with the stimulation and novelty attributes within the user experience questionnaire may be due to physiological response measures also capturing emotions other than stress. These two particular attributes being significant and the remainder not may have perhaps been caused by the cognitive load of the task impacting the correlation, rather than frustration and stress. Additionally, while there was a positive trend visible between task completion time and the SCR levels of participants, the lack of statistical significance has hindered us from making confident conclusions regarding this relationship. Such insignificance is predicted to stem from the limitation of the study of the small sample considered, and hence future research is called for in this aspect.

The knowledge gained from the conducted experiment can aid us to answer the two overarching research questions set at the outset of this thesis. First, skin conductance measures can be concluded with some confidence to vary according to task performance in a usability testing setting. This proposition is based on the positive relationship identified with the task success rate. Yet, as task completion time seems to have stimulated lesser frustration among participants, it may be worthwhile in future stream of research to examine which metrics within the traditional, objective methods correlate with SCR levels. Second, skin conductance measures are found, to most extent, to not be correlated with self-reported, subjective measures of usability testing. The biases and a lack of precise recollection among participants may be the force causing this disconnect, and it may rather be the case that the objective nature of SCR measurement serves to produce superior data as to the usability of an interface. However, as the SCR data is incapable of distinguishing the types of emotions experienced by users, it must be complemented by different

methods, such as subjective questionnaires. In this essence, the lack of a relationship suggests that further research should attempt to identify which complementary measures can be used in combination with the monitoring of physiological responses, if they are to become a novel method for assessing the UX of an online platform. For instance, future research could try to examine the method of combining the SCR data with the monitoring of facial expressions to deduce the usability of a platform. Furthermore, one of the limitations of the applied methodology was the artificial setting of the user testing, whereby the need to complete a task arises in a non-natural manner. In this essence, the high SCR ratio we observed in the first task (despite the low perceived task difficulty, as seen on Figures 7 and 8) may suggest that it is rather the stress experienced from starting an unfamiliar activity that is user testing, distorting our data. Finally, additional research may be worthwhile in exploring new types of SCR quantification metrics in order to obtain more objective data. This may entail for example measuring the peaks in SCR levels and the associated slopes, which might provide a superior image of the stress component of SCR.

Overall, while the findings of this thesis suggest that physiological response measures are a useful and enriching tool for UX research, and further investigation is necessary for it to have the potential of becoming an established technique in this field. Nonetheless, this thesis may serve as a proof of concept for future research.

Bibliography

Yao, L., Liu, Y., Li, W., Zhou, L., Ge, Y., Chai, J. and Sun, X., 2014, June. Using physiological measures to evaluate user experience of mobile applications. In International Conference on Engineering Psychology and Cognitive Ergonomics (pp. 301-310). Springer, Cham.

Foglia, P., Prete, C.A. and Zanda, M., 2008, May. Relating GSR signals to traditional usability metrics: Case study with an anthropomorphic web assistant. In Instrumentation and Measurement Technology Conference Proceedings, 2008. IMTC 2008. IEEE (pp. 1814-1818). IEEE.

Lin, T., Omata, M., Hu, W. and Imamiya, A., 2005, November. Do physiological data relate to traditional usability indexes?. In Proceedings of the 17th Australia conference on Computer-Human Interaction: Citizens Online: Considerations for Today and the Future (pp. 1-10). Computer-Human Interaction Special Interest Group (CHISIG) of Australia.

Moore, M.M. and Dua, U., 2004, October. A galvanic skin response interface for people with severe motor disabilities. In ACM SIGACCESS Accessibility and Computing (No. 77-78, pp. 48-54). ACM.

Shi, Y., Choi, E.H., Ruiz, N., Chen, F. and Taib, R., 2007. Galvanic Skin Response (GSR) as an index of cognitive workload. In ACM CHI Conference Work-in-progress.

Kim, Kyung Hwan, Seok Won Bang, and Sang Ryong Kim. "Emotion recognition system using short-term monitoring of physiological signals." *Medical and biological engineering and computing* 42.3 (2004): 419-427.

Zhai, Jing, and Armando Barreto. "Stress detection in computer users based on digital signal processing of noninvasive physiological variables." *Engineering in Medicine and Biology Society, 2006. EMBS'06. 28th Annual International Conference of the IEEE*. IEEE, 2006.

Hollnagel, Erik. "From human factors to cognitive systems engineering: Human-machine interaction in the 21st Century." (2001).

Hudlicka, Eva. "To feel or not to feel: The role of affect in human–computer interaction." *International journal of human-computer studies* 59.1-2 (2003): 1-32.

Ekman, Paul. "An argument for basic emotions." *Cognition & emotion* 6.3-4 (1992): 169-200.

Czerwinski, Mary, Eric Horvitz, and Edward Cutrell. "Subjective duration assessment: An implicit probe for software usability." *Proceedings of IHM-HCI 2001 conference*. Vol. 2. 2001.

Hornbæk, Kasper. "Current practice in measuring usability: Challenges to usability studies and research." *International journal of human-computer studies* 64.2 (2006): 79-102.

GE, Yan, et al. "Electrophysiological Measures Applied in User Experience Studies." *Advances in Psychological Science* 22.6 (2014): 959-967.

Ganglbauer, Eva, et al. "Applying psychophysiological methods for measuring user experience: possibilities, challenges and feasibility." *Workshop on user experience evaluation methods in product development.* 2009.

Lean, Ying, and Fu Shan. "Brief review on physiological and biochemical evaluations of human mental workload." *Human Factors and Ergonomics in Manufacturing & Service Industries*22.3 (2012): 177-187.

Pfister, Hans-Rüdiger, Sabine Wollstädter, and Christian Peter. "Affective responses to system messages in human–computer-interaction: Effects of modality and message type." *Interacting with Computers* 23.4 (2011): 372-383.

Ward, Robert D., and Philip H. Marsden. "Physiological responses to different WEB page designs." *International Journal of Human-Computer Studies* 59.1-2 (2003): 199-212.

Marek, Tadeusz, Waldemar Karwowski, and Valerie Rice, eds. *Advances in understanding human performance: Neuroergonomics, human factors design, and special populations*. CRC Press, 2010.

Liapis, Alexandros, et al. "Recognizing emotions in human computer interaction: studying stress using skin conductance." *Human-Computer Interaction*. Springer, Cham, 2015.

Ward, Robert D., and Philip H. Marsden. "Affective computing: problems, reactions and intentions." *Interacting with Computers* 16.4 (2004): 707-713.

Scheirer, Jocelyn, et al. "Frustrating the user on purpose: a step toward building an affective computer." *Interacting with computers* 14.2 (2002): 93-118.

Picard, Rosalind W. "Affective computing: challenges." *International Journal of Human-Computer Studies* 59.1-2 (2003): 55-64.

Wilson, Gillian M. "Psychophysiological indicators of the impact of media quality on users." *CHI'01 extended abstracts on Human factors in computing systems*. ACM, 2001.

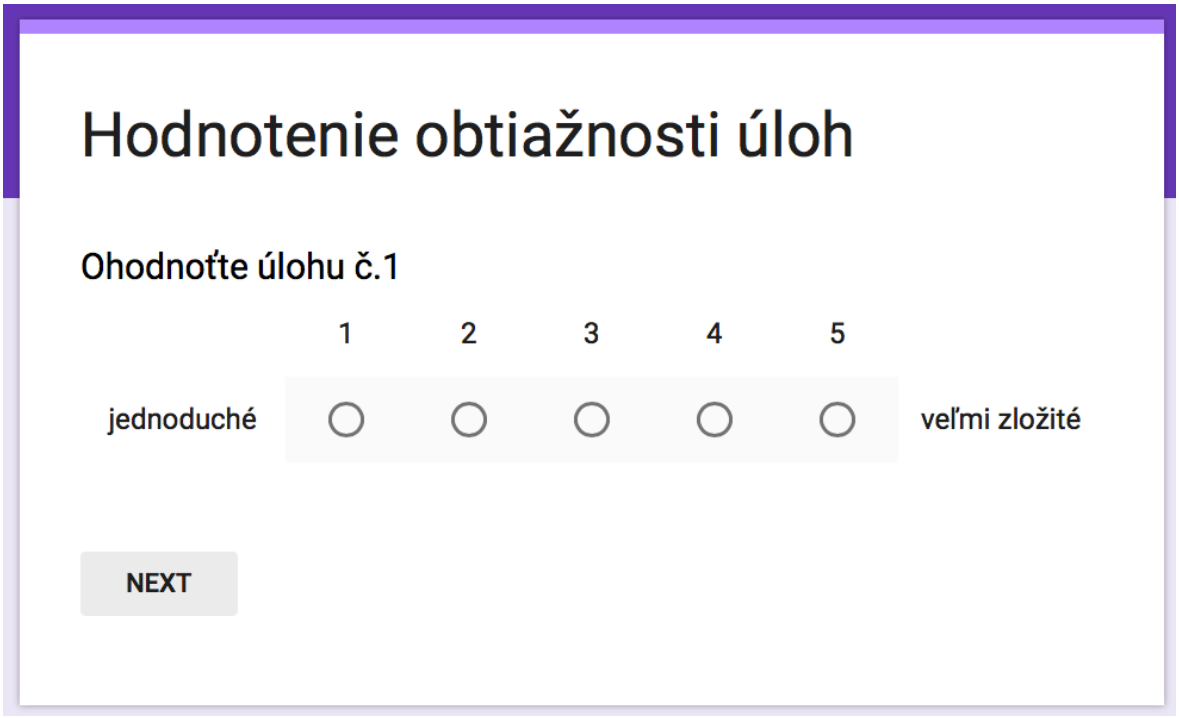
Andreassi, J. L. "Psychophysiology: human behavior & physiological response. 2000 Mahwah."

Bergstrom, Jennifer Romano, et al. "Physiological Response Measurements." *Eye Tracking in User Experience Design*. 2014. 81-108.

Laugwitz, Bettina, Theo Held, and Martin Schrepp. "Construction and evaluation of a user experience questionnaire." *Symposium of the Austrian HCI and Usability Engineering Group*. Springer, Berlin, Heidelberg, 2008.

Albert, William, and Thomas Tullis. *Measuring the user experience: collecting, analyzing, and presenting usability metrics*. Newnes, 2013.

Appendix A



Sample of the task difficulty evaluation

User Experience Dotazník

Prosím vykonajte vaše hodnotenie

Pre ohodnotenie produktu vyplňte nasledujúci dotazník.

Dotazník pozostáva z dvojíc kontrastných vlastností ktoré sa týkajú aplikácie. Krúžky medzi vlastnosťami predstavujú odstupňovanie medzi protikladmi. Môžete vyjadriť váš súhlas s vlastnosťami zakliknutím krúžku, ktorý najviac odráža váš dojem. Po vyplnení všetkých odpovedí odošlite dotazník stlačením tlačítka 'submit'.

Príklad:

	1	2	3	4	5	6	7	
attractive (atraktívne)	○	◉	○	○	○	○	○	unattractive (neatraktívne)

Táto odpoveď by znamenala, že aplikáciu hodnotíte viac atraktívnu ako neatraktívnu.

Rozhodujte sa spontánne. Nepremýšlajte príliš dlho nad vašou voľbou, aby ste označili váš pôvodný dojem

Niekedy si nemusíte byť istý či súhlasíte s danou vlastnosťou, alebo sa vám môže zdať že daná vlastnosť sa na daný produkt nevzťahuje. Napriek tomu zaškrtnite odpoveď v každom riadku.

Toto je váš osobný názor. Pamätajte, že nie je žiadna nesprávna alebo správna odpoveď!

	1	2	3	4	5	6	7	
annoying (otravné)	○	○	○	○	○	○	○	enjoyable (príjemné)
not understandable (nepochopiteľné)	○	○	○	○	○	○	○	understandable (pochoopiteľné)
creative (kreatívne)	○	○	○	○	○	○	○	dull (fádne)
easy to learn (ľahko naučiteľné)	○	○	○	○	○	○	○	difficult to learn (ťažko naučiteľné)

Sample of the user experience questionnaire