Abstract

This is an illustration of the potential value of in-depth examinations of the various dimension of User Experience. Two pieces of research are described, both carried out by the author. The first piece acts as an example of how ambiguity in User Experience may be resolved, by building an in-depth understanding of each dimension individually. The dimension of "Emotion" has been found to be the one most HCI projects focus on, and this is therefore chosen as the target of examination. An understanding of emotion, relevant to HCI, is found through a systematic process moving from the abstract world of philosophy all the way down to the physical elements. The end-result is an understanding of what emotions are, and which variables influence them.

The second piece demonstrates the practical value of gaining a deeper understanding of experiential dimensions. The piece describes the development and test of a new Usability Evaluation Method, attempting to avoid the heavy reliance on expert judgment. Through the development process, the in-depth understanding of emotion is shown to be of practical value. A theoretical method of recognizing the onset of emotion is developed, and the understanding of emotion further contributes with insights into complex nature of emotion. The end result is a method that utilizes psychophysiological measurement to determine the onset of emotional reactions, and then involves the users directly in the process of identifying usability problems. The method is tested through an experiment, and found to produce promising results.

The second piece thus demonstrates a practical application for the theoretical understanding gained through an in-depth review of a single User Experience dimension. This should act as an incentive for the HCI community as a whole to further deepen the understanding of experiential dimensions.

Practical Applications of Emotion in the Identification of Usability Problems

Simon Ahm
School of ICT,
Aalborg University,
Aalborg,
Denmark

July 2014

Preface

This master thesis is written by Simon Ahm during spring 2014.

I would like to thank all participants who took the time to participate in the experimental use of a newly developed Usability Evaluation Method. A special thank you also goes to my supervisor Anders Bruun, whose constructive feedback and assistance has been greatly appreciated.

This master thesis describes the motivation for my research, a summary of the process as well as the combined result. The two articles covering the research in detail have been appended (appendix A and B).

Contents

1 Introduction **7**
1.1 Research Question 1 7
1.2 Research Question 2 7

2 Research Papers **8**
2.1 Research Paper 1 8
2.2 Research Paper 2 9

3 Answers to Research Questions **10**
3.1 Research Question 1 10
3.2 Research Question 2 10

A Article 1 **11**

B Article 2 **22**

1 Introduction

The field of HCI is increasingly moving towards a focus on the over-all experience provided by interactive systems. There has been a tendency in the field to only concern itself with pragmatic measures of efficiency, but this is now changing. This brings a lot of new and exciting opportunities and challenges to the field. One of these challenges in this regard is to understand what "User Experience" (UX) actually entails. The term is broadly used, yet poorly defined. A dimensional approach to User Experience has been proposed by [Bargas-Avila and Hornbæk, 2011], dividing the broad field of User Experience into more manageable chunks. The division in itself is not the full solution though: each of the dimensions are in themselves complex, and ill-defined despite their wide application.

Another interesting problem is whether, and how, an added of understanding of these dimensions can contribute to the field in a practical way. Knowledge that can not be applied has limited interest. Strong examples need to be set for the practical value of investigating UX dimensions in detail.

1.1 Research Question 1

The UX dimension of "Emotion" is the most oft-studied [Bargas-Avila and Hornbæk, 2011]. This makes it a perfect candidate for detailed examination:

"What understanding of 'Emotion' is best suited for HCI-research, and what does it consist of?"

1.2 Research Question 2

The answer to the first question provides a detailed understanding of emotion relevant to HCI. As mentioned in the introduction, the application of this new-gained perspective in a way that benefits HCI practitioners in general could lead to a broader interest in further inquiry into the various dimensions of UX. The next research question is:

"How can the detailed understanding of Emotion be put to practical use?"

2 Research Papers

This chapter presents the two research papers in the thesis. The first paper is a literature review focusing on providing an understanding of Emotion that is relevant to the field of HCI. The first paper can be seen in Appendix A. The second describes the development and testing of a method, that utilizes the understanding of emotion in a practical way: identification of usability problems.

2.1 Research Paper 1

Paper Title: *Understanding User Emotion*

The growing focus on "User Experience" is providing a wealth of new opportunities as well as challenges for the field of HCI. The paper describes the problematic nature of inconsistent understandings of what "emotions" are, and sets out to alleviate the problem by making an in-depth analysis of what emotion entails. Emotion is found to be very complex in nature, and the paper therefore narrows it's field of interest to specifically seek out an understanding of emotion that is of *practical value* to the field of HCI. The paper starts out by identifying the best suited philosophical basis as *Materialism*. This view states that body and mind are the same, leading to the conclusion that emotions must be physically present and therefore measurable. Next, the body's involvement in emotion is examined, and the The Schachter-Singer theory deemed most relevant to HCI. This leads to the assumption that emotions are the result of the physical reaction to a stimuli combined with a cognitive evaluation of the context in which it is experienced. Physical reactions are therefore not seen as reliable indicators of specific emotions, and context is discovered as an important variable in emotion. The Componential Theory of Emotion is found to be strongly compatible with the choices and discoveries made in the report. It strongly contributes to the understanding of emotion through it's description of emotions as being innate responses to stimuli that is considered *"of major concern of the organism"*. Maslow's Hierarchy of Needs is set in relation to what entails "major concern", indicating a strong subjective factor into whether or not an emotional response is elicited as a response to an event. A range of Affective States all related to, but distinct from, emotion are also presented. The complex relationship between emotion and the other Affective States further emphasizes the subjective nature of emotion. A range of all independent variables of emotion identified through the review is

presented in the end. This concludes that while the combined understanding should *theoretically* allow for objective measurement, this is not feasible in practice.

2.2 Research Paper 2

Paper Title: *User Identified Usability Problems*

Emperical and Analytical Usability Evaluation Methods (UEMs) both rely heavily on expert judgment in their identification of usability problems. This paper attempts to resolve this, by looking for alternative sources of information. The user is investigated as a possible vector, and through the inclusion of "Frustration" as a characteristic of usability problems [Skov and Stage, 2005] it becomes plausible. The paper then utilizes the knowledge of emotion to theorize that a measurement of the Autonomic Nervours System (ANS) should reveal the onset of experienced emotion. The understanding of emotion is also used to shape the practical aspects of testing: the physical environment should be as close to a realistic usage situation as possible, and the task given should have a relevant relation to the subjects met and desired needs. Furthermore, the insight into the complexity of the emotional process allows for the quick conclusion that involving the user directly in the interpretation of the emotion is the only viable option. The method relies on the psychophysiological measurement to identify the timing of emotion, and then uses this to extract clips of video leading up to it. The user is then asked to describe what they experienced and why, along with reporting their emotional state at the time through self-report tools. An experiment was designed and set up to test the feasibility of the study. The developed UEM was considered a success based on the experiment, but further testing is warranted.

3 Answers to Research Questions

This chapter answers the research questions based on the research described in the articles appended as Appendix A and B.

3.1 Research Question 1

"What understanding of 'Emotion' is best suited for HCI-research, and what does it consist of?"
The best suited understanding of emotion for HCI use assumes a materialistic connection between mind and body. Emotions are understood as innate responses to events that are considered "of major concern of the organism". When an event is percieved, the body and the mind start responding to it simultaneously. The resulting emotion is based on the following range of independent variables, some of which may differ from person to person and from time to time:

- Events
- Context
- Other Affective States
- Prototypical Responses
- Organismic Subsystems
- Maslow's Hierarchy of Needs

This understanding provides valuable insight such as (1) emotional reaction can be measured, and (2) the emotional response to an event is heavily dependent on a list complex variables that differ from person to person.

3.2 Research Question 2

"How can the detailed understanding of Emotion be put to practical use?"
The study in Article 2 clearly demonstrated how the detailed understanding of emotion made it possible to identify and use an alternative data-point in an existing practice. The understanding of emotion created the foundation for an alternative approach to identify usability problems, ultimately resulting in a promising UEM.

A Article 1

Understanding User Emotion

Simon Ahm
Aalborg University, Department of Computer Science
DK-9220 Aalborg East, Denmark
sahm09@student.aau.dk

ABSTRACT

Author Keywords
User Experince; emotion; measuring emotion;

ACM Classification Keywords
H.5.2 **[Information Interfaces and Presentation]**: User Interfaces – *Evaluation/methodology, Theory and methods.*

INTRODUCTION
The field of Human-Computer Interaction (HCI) is increasingly opening up to a multifaceted understanding of a user's experience with interactive products, as opposed to traditional usability evaluations focusing only on the aspects relevant to efficiency [30]. This broadened focus is not without it's challenges though, as is illustrated through this quote from acclaimed academic Donald Arthur Norman's:

> "Yes, user experience, human centered design, usability; all those things, even affordances. They just sort of entered the vocabulary and no longer have any special meaning. People use them often without having any idea why, what the word means, its origin, history, or what it's about."
>
> *[27]*

This is problematic: the field is moving towards a broader understanding of the experience of a user, yet it lacks a clear understanding of what this entails. One approach to better understand what UX consists of is a dimensional view presented in [6]. This dimensional approach is the basis for this article, in which the most oft-studies dimension, "Emotion/Affect", is explored in greater detail. An understanding of "emotion" that can be considered relevant the field of HCI will be sought out, and comparisons made to how researchers are currently understanding and measuring it. The paper starts out by exploring emotion from its philosophical roots, moving further into the field of psychology and ending in the tangible world of neuroscience. These findings will then provide a foundation on to which currently applied UX methods of assessing emotion can be compared, and possible elaborations or alternative methods be proposed.

UNDERSTANDING EMOTION
The understanding of human emotion, and even the interpretation of the word itself, is diverse and varying across scientific fields and researchers. The brain is a very complex structure, and until we understand it in full detail, we can not expect to work with a final understanding of emotion:

> "(...) a full and accurate definition [of emotion, *sic.*] depends on the fullness and accuracy of our knowledge, and in the midst of our current ignorance we must start with approximations."
>
> *[26]*

HCI researchers should therefore focus on identifying an approximate understanding of emotion that is relevant to the field. This paper is not an attempt to create such an elaborate definition of emotion for HCI, but instead relies on initial explorations of the various understandings and underpinnings of emotion in order to examine and possibly elaborate on the existing practices of assessing emotion in HCI. We start out with an exploration of the philosophical basis of emotion, and then move on to the more tangible fields of psychology and neuroscience.

Philosophical Basis
Due to its complex nature, various ideas of the fundamental basis of human emotion has been proposed over time. The relationship between the mind and the body, known as the Mind-Body problem in Philosophy, is an important question to address in regards to HCI, as the chosen approach strongly dictates the ways in which emotion can be assessed, and additionally delimits the field of further research. This section explores three of the most recognized approaches to the Mind-Body problem in philosophy: Dualism, Materialism and Idealism:

Dualism considers mind and body as distinct - the "I" that thinks is different from the physical "I", the body [3]. This view is strongly associated with Descartes who sees the two as fully distinct, and individually whole, "substances" [12]. Some Cartesians argue that the appearance of the mind affecting the body is due to special intervention by a metaphysical force, whereas Descartes himself instead believed in a "special relationship" between the two.

Materialism is the idea, that the mind is simply an abstraction of the body [2]. Different states of mind are simply different states of the physical brain, and there is no real distinction. This view is often associated with Aristotle who described the "soul" as a type of life-force that disappears when the body dissolves. Modern knowledge on how neuron death due to neurodegenerative affects the mind supports this idea.

Idealism dismisses the existence of an external physical world [1]. The most extreme form sees all of reality as an entirely mental construction created by a non-physical "I", whereas other forms assume the existence of a physical world but holds that we are not able to observe this objective reality. Common for the idealistic approach is the idea that the world is not best understood through hard sciences such as mathematics and physics, but instead through a self-conscious mind.

The above are descriptions are merely quick summaries of the extreme forms of each of these approaches, as there are countless variations of each of them. Due to the metaphysical nature of this question, it is not possible to conclude which, if any, of these approaches is *correct*. The approach best suited for the field of HCI will therefore be chosen as the immediate foundation of the UX-dimension in this paper. Dualism depends on either an ill-defined "special relationship" between mind and body, or a metaphysical force to synchronize the two, leading to complications in trusting that the body communicates what the mind is actually thinking without interference or modification from these unknown factors. Idealism dismisses the existence of a physical world, essentially invalidating all established fields of natural science, including existing practices in HCI. Materialism on the other hand is backed up by current research into neurological disorders, assumes no outside or inexplicable forces, and suggests that emotions are physical states of the brain and body, leading to them being, at least theoretically, measurable by external observers. In conclusion, materialism seems like the best basis for a theoretical understanding of emotion in the context of HCI research.

Psychological Experimentation

With the connection between mind and body established, we move on to explore how the body as a whole is involved in emotion. According to materialism, an emotion is a certain physical state or balance of the brain - but how is this related to the various physical changes in the rest of the body? A range of theories and experiments have been conducted in the field of psychology to examine this, and they further serve as an initial approach to understand what distinguishes various emotions.

The James-Lange theory of emotion defines emotions as being the result of physical changes in the autonomic and motor functions: input from our senses creates a range of responses in our body, and our awareness of these changes is what constitutes an emotion. Individual emotions are distinguished based on their unique bodily expression [20]. The theory states that when something happens in our environment (e.g. we get attacked by a predator) we instantaneously get a physical reaction to this (e.g. muscle tension, widening of eyes, increased sweat production etc.), which we then interpret as a specific emotion based on the characteristics of this reaction alone.

The Cannon-Bard theory, on the other hand, sees physiological changes and emotions as two separate processes that begin at the same time. Input from the senses lead to *both* physiological changes *and* the experience of an emotion - one does not cause the other [11]. This theory also challenges the notion that emotions have unique bodily expressions, claiming that the physiological reaction to for example fear and anger is identical, even though the subjective experience of those two emotions are very different. In the example of the predator attacking, this theory states that there would be two simultaneous but individual processes happening: a physiological change (e.g. changed posture) and a mental change (e.g. feeling angry).

The Schachter-Singer theory, commonly known as the Two-Factor Theory of Emotion, can be seen as a combination of the two aforementioned theories. Emotions are seen as being initialized by the bodily changes occurring based on sensory input, but then argues that a cognitive evaluation of the context is performed and used to internally label the reaction as a specific emotion [28]. When a predator attacks, the person experiences a physical reaction (e.g. change of posture) and uses this physiological change, and the context in which it is experienced, as input in a cognitive process of assigning and experiencing a specific emotion (e.g. anger/fear).

The Schachter-Singer theory is based on, and backed up by, an experiment [28] in which a physiological response was induced through an injection of epinephrine, and the context varied. When the user was in a neutral context, the induced physiological change resulted in no subjective experience of emotion. In two other experimental contexts an actor was present and acted either angry or happy, leading to the subject experiencing the very same emotion - indicating that a cognitive evaluation of the environment was used to label the experienced physiological change as a specific emotion. This study, combined with experiments produced by Cannon-Bard [11], strongly suggest that the James-Lange theory is not correct, although relatively new research has shown that a specific number of emotions do have unique bodily profiles, and can therefore be classified according to these parameters [21] [26]. The experiment conducted in [28] creates strong support for the Schachter-Singer theory, meaning that emotion is a product of both physiological changes and a cognitive evaluation of the context in which they are experienced. The Schachter-Singer theory is therefore deemed the best basis for understanding the relation between changes in the brain and in the rest of the body in the context of HCI research. It is further noted that the context in which events are experienced is an important factor in the expression of emotion.

Differentiation of Emotions

It has now been established that we treat emotion as a specific physical state of the brain, and that this state occur as a product of a cognitive processing of the bodily reaction to an event, and the context in which this is experienced. The next

question is how we differentiate the various emotions a user can experience. There are two main ways of differentiating between emotions:

Discrete Emotions proposes that there is a certain set of innate basic emotions which can be distinguished based on neural, behavioral, physiological and expressive features. A set of discrete emotions which can be distinguished based on facial expressions have been proposed by Ekman and Friesen [16]: anger, disgust, fear, sadness, surprise and happiness. Ekman suggests that these emotions can be expressed in varying degrees. A recent study [5] exploring the language used to express emotions on Twitter Ekman s set of discrete emotions to be the most semantically distinct, further supporting this set of emotions as "basic". Discrete emotion theory generally sees emotions as being hard-wired "programs", shaped through the evolutionary process, for dealing with problems that pose significant consequences for the survival and well-being of the person [22]. Problems are matched to a set of innate prototypical configurations, leading to activation of an appropriate emotional response. This response is what is typically (due to evolution) best suited to solve problems of the type being experienced. It is also suggested that the emotional state is communicated out through for example facial expressions in order to help other alter their behavior accordingly.

Dimensional Models see the world of emotion as being too complex to fit into the pre-defined "boxes" or labels provided by discrete emotion theory, and instead place the emotional state of a person in a multidimensional plane. One such set of dimensions, commonly used in UX research [6], is utilized in the Self Assessment Scale (SAM) [10] which is based on the dimensions of *Pleasure, Arousal and Dominance*, known as the PAD model [25, p. 39–53]. Dimensional models are not as straightforward as discrete emotions, but has the advantage of allowing for more fine-tuned input. Dimensional models also provide beneficial insights into the degree to which emotions are felt and how various emotions relate to each other.

These two different ways of representing the landscape of emotion seem to compliment each other, rather than one necessarily being right and the other wrong. Akin to the inability of concepts such as "particle" and "wave" to fully describe light in the field of physics, it seems relevant to include both methods of representation in HCI research. Existing methods that rely on one of these representations may thus benefit from considering the other perspective.

Classification of Emotions

Besides differentiation between individual emotions, some researchers also classify emotions into broader categories. Scherer distinguishes between whether or not an emotion serve a survival or need-based purpose [29] for example:

Utilitarian emotions are based on our bodily needs, current goals, coping potential and social values. They are emotions that help us adapt to events that appear to have important consequences for our well-being. Based on the evaluation of the event, a certain behavior is then initiated to deal with the event and influence the outcome to either limit the damage or maximize the positive consequence. These emotions are the ones important for survival and well-being, and are thus often very resource-intensive as the subsystems synchronize and mobilize in order to deal with the situation.

Aesthetic emotions are emotions we experience without a utilitarian motive. Experiences such as art are not evaluated according to the aforementioned needs, goals and values. They are instead produced: *"by the appreciation of the beauty of nature, or the qualities of a work of art or an artistic performance"* [29]. A bodily reaction can still result from experiencing an aesthetic emotion, but this reaction is not in the form of preparation to adapt to, or influence, consequences. Goose pimples, shivers and moist eyes are the most common bodily symptoms of aesthetic experiences, and serve no action-oriented purpose.

This distinction is an important part of Scherer's Componential Theory of Emotion which will be described next.

Componential Theory of Emotion

The Componential Theory of Emotion [29] is an approach to understanding emotion that is highly compatible with the aforementioned choices. It assumes a materialistic connection between the mind and body, includes context as an important factor as proposed by the Schachter-Singer theory, and uses a combination of both discrete and dimensional models to distinguish between the various emotions. Emotion is described in this theory as a mobilization and synchronization of five organismic subsystems as a response to a cognitive evaluation of external or internal stimulus events that are *"relevant to major concerns of the organism"* [29]. When an event happens that is of major concern, the event is evaluated through a comparison to innate prototypical responses, and a response (the emotion) is elicited through activation of the five subsystems. It is important to note that this "evaluation" is not a time consuming conscious process as is often associated with the term in HCI, but instead relies on faster subconscious processes. The purpose of emotions is to deal successfully with an event that is of direct concern to the organism, and the activation of the subsystems require a lot of resources to do so. This means that emotions are short-acting mental states that can not be sustained for longer periods of time, and that they are always tied to a specific event.

The five organismic subsystems proposed by Scherer are: Information processing, Support, Executive, Action and Monitor. These systems are abstractions of activity happening in the Central Nervous System (CNS), Neuro-Endocrine System (NES), Autonomic Nervous System (ANS) and the Somatic Nervous System (SoNS). Based on this theory of emotions being an activation of these systems, it may be possible to capture and identify specific emotions through physiological measurements of one or more systems.

Related Affective States

Another important contribution by the Componential Theory is how emotions are distinct from, and related to, various

other affective states [29]. In order to understand what emotions are, it is relevant to understand what emotions are not. The affective states can be described as such:

Feelings are a subpart of emotions themselves. They are the subjective experience of the emotional response (the mobilization and synchronization).

Preferences, like emotions, are evaluations. They are more stable than emotions however, and are limited to simple judgments of like/dislike and preferring one thing over another. Feelings are unspecific and do not produce large behavioral changes.

Attitudes are long lasting beliefs and predispositions toward specific things, persons, events or groups. In contrast to emotions, attitudes do not need evaluation of an event to be triggered. Attitudes may make the occurrence of related emotions more likely.

Mood is an affective state that makes a range of subjective feelings related to the specific mood more dominant. Moods influence experiences and behavior. A mood does not need a clear connection to a specific event. They are of low intensity and can last for hours or even days. A bad mood makes the experience of bad emotions more likely, whereas a good mood does the opposite.

Affect Dispositions are personality traits and tendencies in a persons behavior leading to the person in question being more likely to experience certain moods. Sickness, such as depression, is an Affect Disposition that can lead to a mood being more or less permanent.

Interpersonal Stances develop either spontaneously or strategically in the interaction with people. Communication is formed to be polite or supportive for example. Often based on affect dispositions, attitudes and strategic intentions. Interpersonal Stances can also be trigged by the evaluation of an event.

By adopting this differentiation of affective states in to HCI it becomes easier to develop methods that focus on emotion specifically, while also drawing attention to how other affective states influence the emotions a user experiences. It may for example be necessary to collect data about Mood, Affect Dispositions and Interpersonal Stances in order to correctly identify the emotional state of a user, as these are all able to influence the emotional response to an event.

Maslows Hierarchy of Needs

Since the understanding of emotion we have adopted, from the Componential Theory of Emotion, states that emotions are responses to events that are of "major concern", it is relevant to understand when and why such threshold is reached. According to Maslows Hierarchy of Needs, the things that are considered of major concern may not be static, but instead relies on where in this hierarchy the individual is currently situated: According to Maslows Hierarchy of Needs [24], the needs of an individual depends on their currently met and wanted needs. A person that has already met his or her needs on the Physiological, Safety and Love/belonging levels in the hierarchy will experience negative emotions when these already-met needs are in danger, and positive emotions when currently unmet, but hierarchically close, needs show a possibility of being met. Needs that are positioned significantly higher in the hierarchy relative to a persons currently met needs therefore seem unlikely to yield a strong emotional reaction, as they can not be considered of "major concern to the organism".

This is interesting, as it introduces an important variable in the evaluation of a products emotional impact: currently met and desired needs influence whether or not specific events will cause an emotional reaction. Different people will experience different emotions based on how the events being experienced are related to their currently met and desired needs.

EVALUATION OF EMOTION IN HCI

This section describes some of the existing methods for measuring Emotion in HCI and presents a running discussion of how these methods relate to the theories of emotion presented in the previous section.

AttrakDiff

AttrakDiff is a model for evaluating the attractiveness of interactive products [18]. Emotion is mentioned in this model as a consequence of a users assessment of attractiveness, which in turn is based on the perceived pragmatic and hedonic qualities of the product. These terms can be described as follows:

Pragmatic Qualities are qualities of the product that bring the user closer to his or her goals. A program with a high pragmatic quality assists the user in task-oriented work in an efficient and easy way. These qualities seem related to Utilitarian emotions.

Hedonic Qualities are qualities of a product that are not pragmatic in nature. The hedonic qualities are not about efficient task-solving, but instead focus on how pleasant a product is to use. These qualities seem related to Aesthetic emotions.

When a user interacts with a product, a range of pragmatic and hedonic qualities are perceived by the user, and an overall assessment of the products attractiveness is made - e.g. the product is "likable". This assessment then leads to behavioral (e.g. increased use) and emotional consequences (e.g. joy). AttrakDiff evaluates the users perceived Pragmatic Quality (PQ), Hedonic Quality (HQ) and Attractiveness (ATT) of a given product by asking users to evaluate their experience through 23 7-step scales consisting of bipolar adjectives such as Good-Bad and Unusual-Ordinary.

This model describes emotion as being a consequence of a cognitive evaluation, showing a similarity to the understanding presented in the Componential Theory of Emotion. The evaluation mentioned in AttrakDiff is not of whether or not the event is of major concern to the organism though; it is clearly focusing on simpler judgments of liking or disliking certain aspects of a product. This means, that what AttrakDiff refers to as "emotions" is likely what the Componential Theory refers to as *Preferences*. This is an indication of how

broadly the concept of emotion can be understood, and is already used, in HCI.

Self Assessment Scale (SAM)

The *Self Assessment Scale* (SAM) [10] is a tool in which a user can assess his or her emotional state through graphical scales. The tool is based on a Dimensional Model of emotion called the PAD Model (*Pleasure, Arousal, Dominance*) [25, p. 39–53] that uses three dimensions to represent all possible emotions:

The Pleasure-Displeasure Scale indicates how pleasurable an emotion is. Joy would be high on the pleasure scale, while fear or anger would be in the opposite end.

The Arousal-Nonarousal Scale indicates how intense an emotion is. Rage is an example of a highly intense emotion, meaning it would be be placed towards Arousal on the scale. Boredom is an example of the opposite.

The Dominance-Submissiveness Scale indicates how controlling an emotion is. Anger is an example of a dominant emotion, while fear is is a submissive emotion.

SAM uses cartoon characters (manikins) expressing pleasure, arousal and dominance to represent these scales. Before and/or after interaction with a product, the user is asked to assess their emotional state through the choice of manikins. The tool relies heavily on the users ability to assess the severity of each dimension individually, and be able to empathize with the cartoon characters in order to choose the one best resembling their current state.

SAM is thus a dimensional model that relies on the users ability to understand their own emotional state, interpret the manikins correctly, and communicate their state through successful association between the two.

Emocards

EmoCards is a set of cartoon faces ("pictograms"), eight male and eight female, that represent eight distinct emotion categories. This method is thus an example of measurement method based on Discrete Emotion theory. The eight emotions represented in EmoCards are:

1. Excited, neutral
2. Excited, pleasant
3. Average, pleasant
4. Calm, pleasant
5. Calm, neutral
6. Calm, unpleasant
7. Average, unpleasant
8. Excited, unpleasant

The user is never shown these textual descriptions; they only ever see the cartoon faces meant to universally represent each of these emotions. Before and/or after a user is introduced to, or interacts with, a product, they are asked to choose the pictogram best representing their current emotional state. It is worth noting that this method has been elaborated upon in [14], using animated puppets for expressions instead, revealing that emotion shown through animations are easier for subjects to recognize. A practical implementation of this elaborated method is called PrEmo and is described in the next section.

As with SAM, this method relies on the users ability to understand the pictograms and use these to communicate their emotional state, but uses a specific set of discrete emotions instead of a dimensional model.

Product Emotion Measurement Instrument (PrEmo)

This method, or instrument, is presented in [13]. It is a practical implementation of an elaborated version of the aforementioned EmoCards, using animated pictures instead of static ones. The method is specifically developed to measure combinations of emotions simultaneously, as it is theorized that some of the phenomena we refer to as single emotions in everyday language may actually be a combination of multiple emotions:

> “Even more, rather than being an emotion as such, ‘having fun’ is probably the outcome of a wide range of possible emotional responses. Imagine, for example, the fun one has when watching a movie. This person will experience all kinds of emotions, such as fear, amusement, anger, relief, disappointment, and hope. Instead of one isolated emotion, it is the combination of these emotions that contributes to the experience of fun.”
>
> *[13]*

This example is particularly interesting, as "Enjoyment, fun" is the third most common UX dimension identified in [6].

The animations used in PrEmo represent 14 different emotions; 7 pleasant and 7 unpleasant ones. These emotions were chosen based on research into which emotions are most often experienced in regards to product design [13]. The emotions are:

Pleasant Emotions:

1. Desire
2. Pleasant surprise
3. Inspiration
4. Amusement
5. Admiration
6. Satisfaction
7. Fascination

Unpleasant Emotions:

1. Indignation
2. Contempt
3. Disgust
4. Unpleasant surprise
5. Dissatisfaction
6. Disappointment
7. Boredom

The above emotions are shown in a self-contained computer program as 14 different cartoon characters. When the user moves the mouse pointer on top of one of the cartoon characters, the animation is started. When an emotion is clicked,

the user is asked to choose between the following three levels of emotional involvement:

- I do feel the emotion
- To some extent I feel the emotion
- I do not feel the emotion expressed by this animation

The user is asked to click and rate every single emotion, but is free to choose the order in which this is done. This allows for the identification of states (such as the aforementioned "having fun") that are the result of multiple emotions being present at or around the same time.

This method has an interesting use of discrete emotions, as they can be represented in a multidimensional plane when the existence of multiple experienced emotions are taken into account.

Experience Clip

Experience Clip was originally developed for evaluation of general UX of mobile applications, but has been specifically applied for emotion evaluation in [19]. Two users (friends) are involved, where one is given a mobile phone with the application to be tested (Person A) and the other person a phone with the capability of recording video (Person B). Person B is then encouraged to record as many video clips of Person A using the application as possible. They have the full control over which events are recorded and which are not. Experience Clip is thus a tool to allow users themselves to reflect over the feelings and emotions evoked through the interaction with the device and application. The motivation behind using friends, instead of for example the researcher recording the video, is to let the social situation be natural for expressing and describing emotions - something that would likely be inhibited by the presence of a stranger. Another motivation is the idea of friends being experienced in interpreting and understanding emotional cues from each other, thereby creating natural, flowing conversations about Person A's inner feelings.

The method shows some interesting results: users seemed more free to use the product in innovative ways, compared to when a researcher was observing them. The presence of a researcher generally led to users being "in a hurry" and not exploring the possibilities of the application, possibly due to not wanting to waste the researchers time. Evaluations with a researcher present also led to the user following expected or typical behavior patterns, whereas they were more explorative and directly looking for unusual usage situations when it was a friend who acted as the data collector. This emphasizes the importance of the context in which the product is being experienced.

It is proposed [19], that including the users in the interpretation of the data would likely be useful. Combining Experience Clips with other non-intrusive collection methods is also suggested by the researchers, as the users could possibly leave out relevant data, as they freely choose whether or not to record specific situations.

This method does not explicitly state a specific classification method, and allows for the rich data to be interpreted in various ways based on the data itself or the preferences of the researchers.

Expressing Experiences and Emotions (3E)

This method is presented in [19] as a solution to problems with categorization of emotions by providing a structured and instructed language for expressing emotions through drawing and writing. The 3E method is usually combined with an experience diary: every time the user interacts with the product, they are asked to describe their experience by answering questions about the usage situation and the dominance (or other emotional dimensions) of the system, and fill out a drawing-template containing a stick-figure with a speech- and thought-bobble. In contrast to methods such as SAM and EmoCards, the user is not given the interpretive task of understanding and choosing the image best describing their emotional state. Instead they are allowed to express themselves freely, leaving the interpretation to the researchers. This interpretation-process is difficult, and no final method of gaining quantitative results have been proposed. The method is therefore primarily considered a qualitative method, requiring special knowledge and a lot of time for analysis.

Mobile Feedback Application

This method is specifically developed to solve problems with collecting data about emotions in mobile situations with dynamic interaction. It is implemented as a application on a mobile phones that gathers feedback from the user about his or her emotional state. In contrast to some of the other questionnaire-type methods discussed in this section, this method is a direct attempt to capture data about emotions *as they are experienced*, instead of relying on the memory of the subject afterwards. This is done as, according to [7], emotions are fleeting moments, and can be hard for the subject to recall later on. Based on this understanding of emotions, it therefore seems best to collect data about emotions as they are experienced during interaction with the product:

> "(...) collecting information about emotions during use becomes crucial not only for minimizing the time lapse between experienced emotion and data collection, but also for capturing fleeting emotions which are followed by new emotional experiences during interaction."
>
> *[19]*

This will not only give more reliable data about the emotional state of the subject, but will also directly link the experienced emotion to a specific behavior or event happening in regards to the product being tested:

> "Also, as our focus is on dynamic interaction rather than static appearance, we need to be able to link the fleeting emotions with the knowledge about the status of the application and interaction sequences at the moment of the evaluation as well as the information about the other context variables."
>
> *[19]*

The Mobile Feedback Application is installed on the device along with the mobile application being tested. Questions intended to give insights about the users emotional state is then presented based on time or application events, or can even be initiated by the user. Questions are answered through emoticons, akin to the use of pictograms in SAM and Emocards. Letting the user type in a longer explanation was proposed, but not implemented as the small keyboard on smartphones (at least at the time of this study) would not be suited for free-text entry. Voice recording was dropped due to concerns over privacy. The authors suggest, that by asking the user for input close to when emotionally relevant events happen, they will find it easier to choose a single emoticon to represent their emotional state.

The user is limited to a specific set of possible responses akin to discrete emotions, but depending on the questions asked - especially the possibility of asking follow-up questions based on earlier responses - could allow for the use of a dimensional model.

Psychophysiological Measurements

As is evident from [6], most HCI researchers have chosen to use various self-report methods including the aforementioned ones. A few studies however, have utilized psychophysiological measurements, where the emotional state of users is assessed through measurements on the body. This approach is highly dependent on the materialistic approach to the mind and body. The mobilization and synchronization of five organismic subsystems are mentioned in the Componential Theory of Emotion as being the emotion itself, and measuring changes in one or more of these systems is therefore likely to reveal the emotional state of the user. The Autonomous Nervous System (ANS) is the system most of these methods attempt to capture information from, as this system affects multiple parts of the body - from heart and respiration rate, pupil dilation and constriction of blood vessels to the amount of sweat excreted on the skin [8]. One HCI study [23] attempted to use psychophysiological measures for assessing entertainment, engagement and fun in entertainment technologies, with the argument that

> “Current subjective methods of evaluating entertainment technology aren’t sufficiently robust.” *[23]*

The study captures Autonomous Nervous System (ANS) activity, with the intention of evaluating whether these physiological measures correspond with subjective reports from the users. Based on this information, the study aims to provide a method for objective evaluation of entertainment technologies. The following 6 measures of the ANS were collected:

- Galvanic Skin Response (GSR)
- Heart Rate (HR)
- Electrocardiography (EKG)
- Respiration Amplitude
- Electromyography (EMG)
- Respiration Rate

The physiological data found both GSR and EMG scores to be significantly higher when playing against a friend compared to playing against a computer, and scoring a goal against a friend had a much larger impact on the GSR measurement. This difference once again underlines the importance of the context in which stimuli is being experienced. A significant correlation was found between GSR and subjective ratings of fun. No significant differences was found in HR, Respiration Amplitude or Respiration Rate, and no significant correlations were found between HR and subjective measures. The paper concludes that physiological measures correspond to subjective reported experiences, and may therefore act as objective indicators in evaluations.

VARIABLES OF EMOTION

Through the literature reviewed in this article, a range of independent variables all playing a role in the development of emotional responses have been identified. This section describes each of these variables, what role they play, and how existing methods attempt to account for them. Alternative methods and ideas are proposed when applicable.

The Event

Discrete Emotion Theory and Componential Theory of Emotion both agree on emotions being tied to events. When an event is registered, the user evaluates whether it is of major concern and, if so, matches the experienced event to an innate prototypical response. This ultimately leads to an emotional reaction carried out through an activation and synchronization of the five organismic subsystems.

SAM, *EmoCards*, *AttrakDiff* and *PrEmo* are usually deployed before and/or after interaction with a system. This means that information regarding events happening *during* the interaction are not registered. Using these tests before the interaction likely captures the emotional response to the event of visually perceiving the system for the first time. A test deployed after the interaction likely reflects either an average of the experienced emotions, the last experienced emotion, or a combination of the last and strongest emotion ()as proposed by the Peak/End rule [15]). *Expressing Experiences and Emotions (3E)* is intended to be used over multiple interactions, but still only captures data *after* the interaction is over. *Mobile Feedback Application* is directly targeted at capturing data about emotions as they are experienced. When set up to ask the user for input after specific events happen in the application, it is assumed that the users current emotional state is a result of this event. This is problematic, as not all events lead to an emotional reaction - only those deemed to be of major concern by the cognitive process of the individual. If one of the events assumed to create an emotional reaction does in fact not do so in an individual, the emotional data recorded will not be related to that event, but something happening prior to it. Another concern is that asking the user for input is in itself a disruptive event that may cause an emotional reaction in the user. *Experience Clip* relies on the users to identify relevant events and capture these, a long with commentary from the user experiencing it, on video. This method thus successfully captures rich data about the event leading to an emotional response.

Collecting data about the Event is relevant, not only because it an important variable when trying to determine the emotional state of a user, but because HCI researchers are generally interested in identifying the parts of a system that are triggering emotions. Parts that trigger unintended negative emotions can then be removed, and positive triggers can be maintained or further developed. *Experience Clip* manages to capture Event-data, but does not directly suggest how this data may be used to determine the emotional state of the user. It seems likely that a subjective evaluation by one or more researchers, based on their natural ability to empathize with the subject, can lead to a useful assessment.

The Context

Context is described as a major component of how events are evaluated in the Componential Theory of Emotion. The Schachter-Singer, and in particular the experiment on which it is based [28], further supports this by showing how a single contextual alteration makes the emotional response to an event vary between neutral, strongly positive and strongly negative.

SAM, *EmoCards*, *AttrakDiff*, *PrEmo* and *Mobile Feedback Application* do not in themselves capture any data about the context. The researcher may unconsciously factor in the known contextual factors while handling the data.

Expressing Experiences and Emotions (3E) may contain some contextual data, based on how the user chooses to express their emotional state in their writings and drawing. *Experience Clip* captures some contextual data due to the rich nature of video clips. There is no suggestion as to how the contextual information should be used when assessing emotion in either method though.

Context has a strong influence on how events are perceived and evaluated. This means that not only should contextual information be recorded when possible, researchers also need to beware of how their test setup affects the context. Emotions identified in a lab-environment may not translate to how users experience the same stimuli in a natural context for example. It is important to understand how broad "Context" is - it covers not only the physical environment, but also events happening prior to the currently experienced one as well as Affective States. Contextual information will not reveal the emotional state of the user by itself, but it is an important factor that needs to be included in the evaluation.

Other Affective States

The Componential Theory of Emotion proposed a range of Affective States other than Emotion, such as Attitudes, Moods and Affect Dispositions, strongly influence the probabilities of various emotions being experienced. It would make sense to collect data about a users Affective States when trying to determine emotion, not only because it is an important part of the context in which an event is being responded to, but also because it may help determine researchers determine which emotion is most probable when a research method proposes more than one.

None of the existing methods described in this paper factor other Affective States in to the evaluation of emotion. This may not be necessary in self-report methods such as SAM and Emocards, as the affective state is naturally accounted for in the users own evaluation of their emotional state. It would still be a useful measure though, as it may explain fluctuations in how different people react to the same stimuli. Methods that rely on the researchers to interpret the data - which is the case with *Experience Clip* and *Expressing Experiences and Emotions (3E)* - may however lead to wrong conclusion if data regarding the various Affective States of the user is accounted for.

It may be possible to identify a users' various Affective States through methods currently used in Psychiatry. Depression (an Affect Disposition) can for example be identified by a qualified researcher through methods such as the commonly used Hamilton Depression Rating Scale [17]. Another option would be to give the user a questionnaire such as the widely used Beck Depression Inventory (BDI) [9]. Due to the vast amount of possibilities within each of the various Affective States, it is not realistic to account for all of them. Depending on the requirements of the experiment, it may be useful to apply some of them. They would primarily be useful as a screening criteria for participants though, since even if the information was recorded it is not clear exactly how to factor it in to the analysis of emotion. One theoretical route would be for a researcher to temporarily adopt a set of Affective States matching the ones of the user - possibly through strong empathy or techniques akin to Method Acting [4], and then assess the other data (Event, Context) from this altered perspective. A more sensible approach would be to include the user in the interpretation as proposed in [19].

The Prototypical Responses

Discrete Emotion Theory and The Componential Theory of Emotion both treat emotions as innate responses, shaped through evolution, that have been shown to best handle situations that are of major concern to the organism. When an event happens, the user subconsciously compares the event to the innate prototypes, and then activates the closest matching Prototypical Response. An understanding of these prototypes would make it possible to evaluate which prototype best matches a given event, and then assess what the likely response would be. Unfortunately, we do not have such detailed understanding of the prototypes, the Prototypical Responses nor the cognitive process performing the comparison. This ultimately leaves us with two options of assessing the emotional outome: (1) our innate ability to feel empathy, and (2) the users report it themselves.

All the self-report methods, by their very definition, rely on (2). *Experience Clip* and *Expressing Experiences and Emotions (3E)* generally rely on (1). If users are included in the interpretation of the data it would arguably be a combination of (1) and (2), as the user would have to think back and remember how they felt at the time.

The best way to account for this variable seems to be inclusion of the user in the interpretation of the collected data.

The Five Organismic Subsystems

When the aforementioned process of matching an experienced event with a Prototypical Response is done, the response itself is carried out through an activation and synchronization of the Five Organismic Subsystems described in The Componential Theory of Emotion. If this process was fully reversible we would be able to determine the experienced emotion based on the activation of these systems alone. Unfortunately this is not the case, as proved in the Schachter-Singer experiment [28]: two very distinct emotions can carry the same bodily expression. Some emotions do seem to have unique bodily profiles though, and can therefore be classified through psychophysiological measurement alone [21] [26]. Another possible use of psychophysiological measures is to detect how strong an emotional response is, and *when* an emotional response happens. By knowing when the emotional response manifested itself, it is easier to identify the event that triggered it.

Maslow's Hierarchy of Needs

Maslow's Hierarchy of Needs provides an interesting answer to the question of what *"major concern of the organism"* actually entails. It indicates that the current state of met and unmet needs in the user, has a strong influence on which events will yield an emotional response. This means that not only will different people respond differently to the same stimuli, the same person may in fact do so over time as the state of current needs changes. Understanding the test-subjects current position on the Hierarchy of Needs is therefore another variable that should preferably be collected and used in the process of determining user emotion.

None of the methods reviewed in this paper attempt to directly identify the placement in Maslow's Hierarchy of Needs of the subjects. They may instead rely on carefully selecting a representative set of test-subjects from the set of end-users, as this would likely diminish the effects of variances in met and unmet needs.

CONCLUSION

Emotion is a complex concept, with many different meanings and interpretations. This article argues that a materialistic approach to the connection between mind and body is best suited for HCI, and that this leads to the possibility of - at least theoretically - objectively measuring emotion. With basis in the materialistic understanding, a review of literature has revealed a range of independent factors in emotion relevant to HCI research:

- Events
- Context
- Other Affective States
- Prototypical Responses
- Organismic Subsystems
- Maslow's Hierarchy of Needs

Due to the complex nature of these factors, a true objective measurement of emotion is seen as unrealistic at the current moment. A method has been proposed that uses the psychophysiological measurement of GSR to identify emotional responses, and then involves the user himself in the process of identifying the triggering Event and describing the experienced emotion.

REFERENCES

1. Absolute idealism. In *Encyclopædia Britannica Academic Edition (Online)*, winter 2013 ed. 2013.
2. Materialism. In *Encyclopædia Britannica Academic Edition (Online)*, winter 2013 ed. 2013.
3. Mind-body dualism. In *Encyclopædia Britannica Academic Edition (Online)*, winter 2013 ed. 2013.
4. Stanislavsky method. In *Encyclopædia Britannica Academic Edition (Online)*, summer 2014 ed. 2014.
5. Bann, E. Y., and Bryson, J. J. The conceptualisation of emotion qualia: Semantic clustering of emotional tweets. In *Proceedings of the 13th Neural Computation and Psychology Workshop (NCPW13)* (2012).
6. Bargas-Avila, J. A., and Hornbæk, K. Old wine in new bottles or novel challenges: a critical analysis of empirical studies of user experience. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM (2011), 2689–2698.
7. Battarbee, K., et al. *Co-experience: understanding user experiences in interaction*. Aalto University, 2004.
8. Bear, M. *Neuroscience : exploring the brain*. Lippincott Williams & Wilkins, Philadelphia, PA, 2007.
9. Beck, A. T., Steer, R. A., and Carbin, M. G. Psychometric properties of the beck depression inventory: Twenty-five years of evaluation. *Clinical psychology review 8*, 1 (1988), 77–100.
10. Bradley, M. M., and Lang, P. J. Measuring emotion: the self-assessment manikin and the semantic differential. *Journal of behavior therapy and experimental psychiatry 25*, 1 (1994), 49–59.
11. Cannon, W. B. Bodily changes in pain, hunger, fear and rage (ed. 2) d. *Appleton and Company, New York* (1929).
12. Descartes, R. *Descartes: Meditations on first philosophy: With selections from the objections and replies*. Cambridge University Press, 1996.
13. Desmet, P. Measuring emotion: Development and application of an instrument to measure emotional responses to products. In *Funology*, M. Blythe, K. Overbeeke, A. Monk, and P. Wright, Eds., vol. 3 of *Human-Computer Interaction Series*. Springer Netherlands, 2005, 111–123.
14. Desmet, P. M. *Designing Emotions*. PhD thesis, Delft University of Technology, 2002.
15. Do, A. M., Rupert, A. V., and Wolford, G. Evaluations of pleasurable experiences: The peak-end rule. *Psychonomic bulletin & review 15*, 1 (2008), 96–98.
16. Ekman, P., and Friesen, W. V. Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology* (1971).

17. Hamilton, M. A rating scale for depression. *Journal of neurology, neurosurgery, and psychiatry 23*, 1 (1960), 56.

18. Hassenzahl, M., Burmester, M., and Koller, F. Attrakdiff: A questionnaire to measure perceived hedonic and pragmatic quality. In *Mensch & Computer* (2003), 187–196.

19. Isomursu, M., Tähti, M., Väinämö, S., and Kuutti, K. Experimental evaluation of five methods for collecting emotions in field settings with mobile applications. *International Journal of Human-Computer Studies 65*, 4 (2007), 404 – 418. Evaluating affective interactions.

20. James, W. Ii.what is an emotion? *Mind os-IX*, 34 (1884), 188–205.

21. LeDoux, J. The emotional brain: The mysterious underpinning of emotional life. *New York* (1996).

22. Levenson, R. W. Autonomic specificity and emotion. *Handbook of affective sciences 2* (2003), 212–224.

23. Mandryk, R. L., Inkpen, K. M., and Calvert, T. W. Using psychophysiological techniques to measure user experience with entertainment technologies. *Behaviour & Information Technology 25*, 2 (2006), 141–158.

24. Maslow, A. H. A theory of human motivation. *Psychological review 50*, 4 (1943), 370.

25. Mehrabian, A. *Basic Dimensions for a General Psychological Theory: Implications for Personality, Social and Environmental Psychology*, 1st edition ed. Oelgeschlager, Gunn & Hain Inc.,U.S., 8 1980.

26. Panksepp, J. *Affective neuroscience: The foundations of human and animal emotions*. Oxford University Press, 1998.

27. Peter Merholz. Peter in Conversation with Don Norman About UX & Innovation. **http://www.adaptivepath.com/ideas/e000862/**, 2014. [Online; accessed 05-June-2014].

28. Schachter, S., and Singer, J. Cognitive, social, and physiological determinants of emotional state. *Psychological review 69*, 5 (1962), 379.

29. Scherer, K. R. What are emotions? and how can they be measured? *Social science information 44*, 4 (2005), 695–729.

30. TwinTide. COST Action TwinTide: IC0904. **http://www.irit.fr/recherches/ICS/projects/twintide/index.php?template=about.tpl**, 2014. [Online; accessed 05-June-2014].

B Article 2

User Identified Usability Problems

Simon Ahm
Aalborg University, Department of Computer Science
DK-9220 Aalborg East, Denmark
sahm09@student.aau.dk

ABSTRACT

Author Keywords
User Experience; Frustration; Usability; GSR;

ACM Classification Keywords
H.5.2 **[Information Interfaces and Presentation]**: User Interfaces – *Evaluation/methodology, Theory and methods.*

INTRODUCTION
Various methods have been proposed to identify usability problems in interactive systems. These Usability Evaluation Methods (UEMs) generally fall into two categories: Empirical UEMs, that identify rate usability problems through the inclusion of users (user testing), and Analytical UEMs that rely on guidelines and expert knowledge (e.g. Heuristic Inspections) [10]. A core difference in the results provided by each type of method is described in [10]:

> “Analytic UEMs examine intrinsic features and attempt to make predictions concerning payoff performance. Empirical UEMs typically attempt to measure payoff performance directly (e.g., speed, number of errors, learning time, etc.).” *[10]*

Analytical UEMs thus generally provide a very clear identification of where the various usability problems are located in a system. This makes it relatively easy to implement changes. The identification of these problems rely purely on expert knowledge and/or pre-defined heuristics though, and no quantifiable data on how the identified problems influence payoff performance is provided. Usability problems identified through an Analytical UEM may therefore not manifest in actual use, or only have negligible negative effects on performance, making a possibly costly fix unwarranted. Empirical UEMs, on the other hand, directly provide the end-effect on payoff performance. They too seem to depend on expert judgment to identify the problems affecting this performance:

> “None of the studies we reviewed report systematic ways of relating payoff problems to intrinsic features; all apparently rely on some form of expert judgment.” *[10]*

Identification of usability problems is thus an educated "guess" from one or more experts in both cases. Research on the performance of expert usability evaluators has shown that there is a remarkable difference in which, and how many, problems are found across evaluators. This is known as the Evaluator Effect [15] [14]. Adding more evaluators to a test tends to lead to more identified problems, but there is no guarantee these are *actual* usability problems. Even in empirical tests, the features of the program identified as problematic by the expert may not be what actually affected the user's performance in the test.

Hybrids of Analytical and Emperical methods already exist:

> “Other usability inspection methods are hybrids between intrinsic and payoff in that the analysis done during usability inspection is task driven; the expert's analysis is based on exploring task performance and encountering usability problems in much the same way users would, adding a payoff dimension to the intrinsic analysis. In this situation, a usability inspector asks questions about designs in the context of tasks to predict problems users would have.” *[11]*

This approach shares the same point of critique though: it is still heavily dependent on the experts ability to *estimate* the problematic intrinsic features.

This paper sets out to develop a method that is able to directly measure payoff performance, while at the same time providing a well-founded indication of the underlying intrinsic features that cause negative effects to it. Well-founded meaning, that it should not depend on the expert's judgment and estimates only.

METHOD DEVELOPMENT
This section describes the development of the UEM. The first part describes a set of criteria the UEM should comply to. Next, an understanding of usability problems is adopted, and a way of identifying them determined. The necessary tools for the test will then be described, and a full method is proposed.

UEM Criteria

Before developing a UEM, it is important to understand what such method entails. This method will be based on the following general understanding of UEMs:

> UEMs are used to evaluate the interaction of the human with the computer for the purpose of identifying aspects of this interaction that can be improved to increase usability.
>
> *[10]*

The method must therefore be able identify problematic *aspects* of an *interaction* between a human and a computer. The definition of "problematic aspects" and how they should be expressed also needs to be understood:

> The essential common characteristic of UEMs (...) is that every UEM, when applied to an interaction design, produces a list of potential usability problems as its output. Some UEMs have additional functionality, such as the ability to help write usability problem reports, to classify usability problems by type, to map problems to causative features in the design, or to offer redesign suggestions. We believe these are all important and deserve attention in addition to the basic performance-based studies.
>
> *[11]*

The UEM should thus *at very least* be able to output a list of usability problems. As described in the introduction, the vision of this new method is to combine the general strengths of Emperical and Analytical methods. Using the terminology of [10], this means that the method should be able to output a measure of "payoff performance", as well as the "intrinsic features" that have negatively influence this measure. This seems compatible with "classify usability problems by type" and "map problems to causative features in the design" in the above quote. In conclusion, this UEM should provide the following output:

- **Output:**
  - List of Usability Problems
  - Classification of Usability Problems
  - Causative Features in the Design

The output itself should also comply to a range of criteria. Three measures of examining an evaluation method have been identified in [4]. Two more measures, focusing on more practical aspects of usability, have since been proposed by [11]. In combination this provides the following 5 criteria for the output of the UEM:

- **Thoroughness:** The resulting output should be complete. As many of the usability problems present in the system should be identified.
- **Validity:** The list of usability problems should contain only *actual* usability problems. Problems that would be experienced by actual users in real life usage.
- **Reliability:** Results should be consistent. Two different people using the UEM should identify the same problems.
- **Downstream Utility:** The identification of usability problems is not practical by itself. The output should be of assistance in the subsequent process of finding solutions.
- **Cost Effectiveness:** Usability practitioners are usually constrained both economically and by schedule. The UEM should therefore deliver useful results without incurring too high a cost.

The Reliability criteria fits greatly into the intended goal of minimizing reliance on expert estimations. These criteria form the basis for creating and later evaluating the UEM.

Identification of Usability Problems

In order to identify usability problems without relying on an experts ability to do so, another source has to be found. As Analytical UEMs rely solely on expert judgment, an empirical approach is needed. The two ends of the interaction both act as possibles sources of information about it:

- **The User:** The person interacting with the system, perceiving the outcome and experiencing a reaction to it.
- **The System:** The system receiving input from the human, and acting accordingly based on a set of predefined rules.

Information could be gathered from The System through logging. This allows for (possibly automatic) measures on performance measures such as speed and learning time. The reasons for any identified changes in these measures would only be possible to *estimate* from within the system though, making this source no more reliable than expert judgments. The user on the other hand *experiences and reacts* to the whole interaction with the system, and would therefore register when problems arise. The identification of problematic aspects may not be a conscious process though, therefore requiring special inquiry.

This leads to the next problem of identifying usability problems: understanding their characteristics. The following identifiers of usability problems are presented in [25]:

- **Slowed Down:** The user slowed down relative to normal work speed. This may range from a few seconds delay to being fully hindered in solving the task.
- **Understanding:** The user does an action without being able to explain why, has trouble grasping the way the system functions or does not understand how the system can be used for solving the task at hand.
- **Frustration:** The user believes he has damaged something, or is clearly annoyed by some aspect of the system.
- **Test Monitor:** The person ("Test Monitor") observing the usability test while present in the same room as the user has to interfere. This ranges from the test monitor asking leading questions or giving hints, to fully assisting the user in solving a task.

"Slowed Down" and "Test Monitor" are relatively simple to objectively measure and quantify. Problems related to "Understanding" are usually exposed through the talk-aloud

technique employed in most Empirical UMEs. No objective method of determining "Frustration" is presented in [25]: it seems to a subjective evaluation by the the expert, likely based on intuitive interpretations of body language and verbal expressions from the think-aloud technique. This is unfortunate, as "Frustration" seems likely to be the most prevalent of these identifiers across all usability problems. Assuming the user is interested in optimal pay-off performance, any problems that interfere with this interest would arguably cause some form of frustration. Another interesting property of this identifier is that it is related to emotions: According to the Componential Theory of Emotion [23] emotional responses are *always* tied to a stimulus event. By identifying which event caused the emotional response of frustration, we may thus gain insight into the underlying intrinsic feature in the system, that caused the problem - the very aspect Empirical UEMs seem to be missing.

This method will attempt to identify the timing of usability problems by measuring the onset of emotional responses. An understanding of what emotions were experienced at these points in time, and their effect on overall User Experience, would further act as an interesting measure of payoff performance.

Measuring Emotion

The definition of emotion presented in the Componential Theory of Emotion states that [23]:

> “ [Emotion is, *sic*] an episode of interrelated, synchronized changes in the states of all or most of the five organismic subsystems in response to the evaluation of an external or internal stimulus event as relevant to major concerns of the organism. ”
>
> *[23]*

With "Frustration" being an indicator of at least a subset of usability problems [25], it seems plausible that identification of emotional responses would assist in the identification of usability problems. In a practice, this may be archived through psychophysiological measurement of the activity in the Autonomic Nervous System (ANS). The ANS is a constituent of the organismic subsystem referred to as "Support" in the Componential Theory, and is described as being involved in the bodily expression of emotion. The ANS mobilizes the system towards either excitation ("fight-or-flight") or relaxation ("rest-and-digest"), both affecting large parts of the body [1]. This makes it a possible objective indicator of the onset of emotions, and thus the timing of experienced usability problems. Galvanic Skin Response (GSR) is one such measure of ANS activity, that is widely used and has already been demonstrated to relate to measures of usability [24] [16]. Furthermore, GSR sensors have been successfully integrated into other things, making them ideal for continuous unobtrusive measurement. Examples of unobtrusive GSR designs include a wristband [20], a computer mouse [28] and a sock [13].

If the psychophysiological measure is able to identify the onset of an emotional response, this information could be used to extract data about the stimulus event *causing* the emotional reaction. In the case of usability problems, this event would arguably be the problematic intrinsic feature of the system. The emotional response is a complex process though, influenced by a wide range of independent variables, some of which vary between individuals [23]. The only reasonable way of analyzing the event data therefore seems to be, to let the *user* deduce what caused the response. As emotions are fleeting moments, it can be a very difficult task for a subject to recall them later on [5]. This method theorizes that this problem can be countered by providing the user with as much information regarding the event as possible. A recording of the event (Screencast) as well as the users facial expressions (Webcam) leading up to the emotional response may be enough to let the user positively identify what they felt and why. Changes in skin resistance take between 0.8-4 seconds to manifest [21], meaning the stimulus event causing the spike in GSR can be expected to be contained within the ~5 seconds of video leading up to the identified peak.

As emotions can be difficult to express verbally, the user is expected to benefit from one or more standardized self-report tools. These would also allow for easier quantification of the experienced emotions, thereby opening up for comparison of data across multiple users. Such tools are already used within the field of HCI, with the most commonly used being SAM and Emocards [3].

- **Emocards:** Emocards [8] has eight male and eight female cartoon faces ("pictograms"), representing eight distinct emotion categories. The user chooses the Emocard they find to best represent the emotion in question. The 16 cards and the dimensions they represent can be seen in Figure 1.
- **SAM:** SAM [7] is based on the PAD Model (*Pleasure, Arousal, Dominance*) [18] of emotion. These dimensions are presented as three graphical scales made up of cartoon characters ("manikins"). The user selects the manikin in each dimension that best represents the emotion in question. See Figure 2.

In summary, this method will attempt to measure the onset of emotion through the psychophysiological measure of GSR. This information is then used to extract the relevant parts of the screencast and webcam recordings. The user is then presented to the video material, and asked to describe what event caused the emotional response. Lastly, the user utilizes Emocards and SAM to report the emotion experienced at the time. Description and Emotional data is recorded by the researcher for later analysis.

Implications of Emotions

By using measures of emotion as an indicator of usability problems, a couple of new factors have to be accounted for in the method. First of all, the Componential Theory of Emotion states that emotions are only elicited when an event is *"of major concern of the organism"* [23]. Maslow's Hierarchy of Needs [17] suggest that what is regarded as "major concern" depends on a users currently met and unmet needs. A person having not yet met his or her needs on the Physiological level (e.g. food) will not elicit an emotional response to

Figure 1. The eight emotional categories of Emocards along with their respective pairs of pictograms.

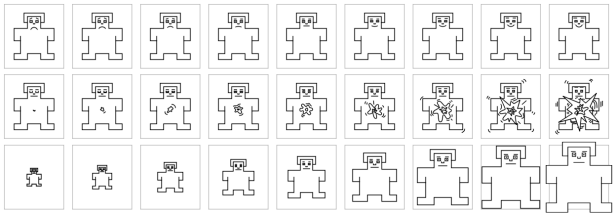

Figure 2. Example of a 9-point SAM scale. First row represents Pleasure, the second Arousal and the third Dominance.

events influencing only needs on a significantly higher place in the hierarchy for example. Only users to which the system somehow ties in to their needs, can be expected to show frustration when usability problems are encountered. Fortunately, this would be the case in most situations: people don't tend to interact with a system without some intent or purpose. It does create an important requirement for testing though: only by shaping the experimental task in a way, that is relevant to the currently met and desired needs of the test subjects, can emotional responses be expected to manifest themselves. Also, steps should be taken to ensure that test-participants are at roughly the same level in Maslow's Hierarchy of Needs. If a system has many different users, with large variations in hierarchal placement, these should be divided into hierarchically similar groups and used for separate evaluations. Each group of users may react differently to events happening in the system, thereby experiencing a distinct set of usability problems. Understanding the currently met and unmet needs of users, and the relationship these needs have with the use of the system, might assist the researcher in interpreting the descriptions of problematic events provided by the users.

The Componential Theory of Emotion also puts a lot of emphasis on context. The theory describes context as a very important factor in how events are interpreted, and which - if any - emotion is elicited as a response. An oft-cited experiment [22] gives a practical example of the influence of context: Subjects were injected with epinephrine (adrenaline) and later asked to describe their emotional state. The experiment showed that the same physical sensation (from the epinephrine) could be experienced as neutral, anger or euphoria entirely depending on the context. The context in which an interactive system is tested, therefore seems likely to influence how users react to it emotionally. As the method aims to identify *actual* usability problems, it seems relevant to make the context of the test as close to the real context of use as possible. This is difficult as it not only entails hiding cameras and other foreign objects from plain sight, but because it would likely be ethically indefensible to make such tests without first informing the subjects. As it is currently unknown how severely the context influences the type and amount of usability problems found, it might not be necessary to go to such extremes. Attempting to keep the context as close to real-life use as possible also means, that this method can not benefit from the think-aloud technique or having a usability "facilitator" in the same room as the user.

Interpretation of Qualitative Data

The first criteria set up for the UEM is an ability to output a list of usability problems. This is not a trivial task. The list of user-described problems suffers from the same problem experienced when comparing notes between two experts: the same problem may be described in substantially different ways [11]. It is possible that this problem could be mitigated by adopting a framework for describing and comparing usability problems (such as the User Action Framework [12]), and adapting it to function as a way of asking the user clarifying questions. As for now, letting the same researcher who de-briefed the user interpret the problem descriptions is hoped to suffice.

The researcher examines each noted emotional response, and uses the user's description of the event and it's impact to determine whether it qualifies as a usability problem. The list of notes describing usability problems is then examined by itself, grouping together descriptions that are assessed to refer to the same intrinsic feature of the system. Each group is then collapsed into a descriptive usability problem. This completes the criteria of outputting a "List of Usability Problems".

It seems plausible that the usability problems could be classified according to the average emotional impact they have on users. The classification system described in [25] only provides qualitative descriptions of the frustration associated with a "Serious" problem though:

> "Does not understand how a specific functionality operates or is activated.
>
> Cannot explain the functioning of the system." *[25]*

No descriptions are provided for Cosmetic or Critical problems. This complicates the translation of collected emotional measures into severity categories. Even sorting the list from "most frustrating" to "least frustrating" would require an as-of-yet unknown way of determining the relationship between reported scores in SAM and/or Emocards and the ill-defined term "frustration".

This leaves expert judgment the only viable option. By having the expert include the qualitative descriptions of each problems in their considerations, the estimated classification will hopefully reflect the actual severity of the problems quite precisely. It is possible that the self-rated emotions would also be of assistance in the process.

Summarized Method

This section condenses the considerations from the previous sections into a short and action-oriented method.

The method relies on user testing, and is thus an Emperical UEM. The participants ("users") should preferably share roughly the same place in Maslow's Hierarchy of Needs. Guidelines regarding number of participants and selection criteria may readily be adopted from existing UEMs relying on user testing, at least until studies reveal any major differences in efficiency.

The test environment and other contextual factors should be as close to real-use circumstances as possible. The user is equipped with an unobtrusive GSR sensor, and placed in front of a computer. A discrete webcam (possibly integrated in the computer screen) is recording the users face, while a screencast software is quietly recording the contents of the screen. An initial GSR recording is made while the user is relaxed, creating a baseline-reading. The user is then asked to interact with the system in a natural way, for example by solving a realistic open-ended task using the system. The full interaction happens when the user is alone - no facilitators are present in the room. When the interaction is over, the user retrieves the facilitator. The GSR data is analyzed, and the timing of all peaks extracted. The screen- and face-recordings leading up to each peak is then shown to the user one peak at a time. For each peak, the user is asked to describe what happens using their own words, and identify what they perceive as the event causing the registered emotional response. They are also asked to assess the emotional state they experienced at the time, and report it using SAM and Emocards. The facilitator records descriptions and emotion-ratings on paper.

After all users have participated in the test, the researcher examines the complete list of user-described emotional reactions. All descriptions deemed to describe problematic aspects of the interaction are extracted into a separate list. The researcher then examines this new list, and groups together descriptions that seem to describe the same problem. Each group is then collapsed into a single descriptive usability problem.

EXPERIMENTAL DESIGN

The developed UEM is based on a wide range of theoretical assumptions. It's validity and usefulness should therefore be tested through an experiment. This section describes the design of an experiment intended to assess the effectiveness of the UEM.

Selecting Test Participants

The method requires the test participants to be placed at roughly the same place in Maslow's Hierarchy of Needs, and further suggests that an understanding of the met and desired needs of the participants will aid in the design of the experiment. As this experiment is intended to test the method, rather than any specific system, a tactical choice of participants can be made. The researcher is currently a Masters student, meaning that he arguably holds a personal insight into the met and desired needs of a university student. The social circle of the researcher also consists primarily of university students, paving the way for a more generalized understanding of this group. It is assumed that "University Students at Aalborg University" can be considered within a fairly narrow space of Maslow's Hierarchy of Needs, and that this will be sufficient to determine the effectiveness of the method. As explained in the next session, the task itself will be designed to relate to needs the researcher finds plausibly universal for this otherwise broad group.

Drafting a Realistic Task

The method states that the task needs to be realistic and relevant to the users' needs. It also emphasizes the importance of keeping the *context* in which the task is solved as realistic as possible. This experiment attempts to comply with these requirements by mimicking a type of task familiar to all students in general: an exam. Students can generally be considered to have reached a place in Maslow's Hierarchy of Needs where "achievement" [17] is a currently desired need. It seems safe to assume - based on the motivations and educational background needed to become a university student - that they have a learned association between exams and the need of achievement. This experiment attempts to exploit this by mimicking the design and experience of an exam. This gives the added benefit of solving the contextual problems noted in the method: An exam puts the student under pressure, they are expected to perform well and they know they will be evaluated. These contextual factors are arguably compatible with the contextual factors introduced by the use of the UEM. In a normal case of User Testing, the facilitator would let the user know that it is the system, and not their personal ability to use it, that is being evaluated. In this experiment, the opposite is true: the user is given a task and told that *they* will be graded on how good *they* are at solving it. They are thus under pressure to perform, just like they would be in a natural exam-situation. The camera and GSR sensor only further establishes the awareness of them being evaluated. The contextual factors of the test are thus adopted to strengthen the realism of the context, as opposed to weakening it as would normally be the case. The physical environment in which the test is performed should also follow the narrative of the user being at an "exam": a university classroom seems ideal.

The task should thus be designed to look like an examination assignment. To allow the user to interact in a natural and fluid way with the system, the task is held open-ended: instead of a long list of "do this, then that", the task is presented as an open problem without guidelines on how the software can be used to solve it. The following scenario was developed for the experiment:

"The user is at an examination. Their assignment is to use a specific piece of software on the computer ("PosterMaker Pro") to create an exact copy of a poster depicted on their assignment paper. They have 10 minutes to complete the assignment, after which the program will automatically grade it, resulting in the user either passing or failing. The user can hand in their poster before the time limit by pressing a "Hand in" button placed directly in the software. This too will lead to the software automatically evaluating their work."

Informing the user that their performance will be evaluated, and in particular that they risk "failing", is intended to make them more keen to react emotionally towards problems they experience during the interaction. The inclusion of a "Hand in" button not only allows the user to finish before any artificial time limit, but is also intended as a constant reminder of the examination-context.

Seededing Usability Problems

The developed UEM is intended to identify usability problems present in an interactive system. In order to test whether it is successful in doing so, the output of the method can be compared to a Standard Usability Problem Set (SUPS) [11]. The amount of usability problems from the standard set that can be identified by the UEM is indicative of its effectiveness.

Multiple ways of producing an SUPS are presented in [11]. There are three general approaches: (1) to introduce a set of known usability problems ("seeding"), (2) to use an established UEM to determine the usability problems present in an existing system, or (3) to use a combined from multiple UEMs applied to the same existing system. As the method being tested is in a very early stage of it's development, it seems ideal to use the least costly of these methods: Seeding.

"PosterMaker Pro" was developed specifically for the experiment. The software is made as simple and purpose-specific as possible, to minimize the amount of naturally occurring usability problems. Organic usability problems may interfere with the seeded ones, making it hard to determine whether the method failed by not identifying them. Two versions of the software were made: the orginal version, intended to have High Usability (HU), and the seeded version, intended to have relatively Lower Usability (LU).

The following 5 usability problems were conceived by the researcher and implemented into the LU version:

- **Non-Standard Font Style:** The software allows for simple text styling (Bold, Italic and Underlined text). The standard buttons well-known from products like Microsoft Word were replaced with Checkboxes. Predicted to be Cosmetic.

- **No Preview Font-Selector:** The software comes with a long range of different fonts. The LU version simply states the name of the font without a preview of how it looks. Predicted to be a Serious problem.

- **Missing Font:** One of the fonts ("MS Comic Sans") that was needed to solve the given task, was removed from the font list. This makes it impossible to properly solve the task. Predicted to be Critical.

- **Colors in Combobox:** Instead of a visual color-picker, a simple combobox containing nothing but the name of each color was used. Predicted to be Serious.

- **Error Message on Image Insert:** An incomprehensible error message was set to be shown after the user inserts an image. The error is technical sounding gibberish, and asks the user to press "Yes", "No" or "Cancel". All buttons do the same thing: close the error, and let the user carry on. Predicted to be Serious.

The predictions are estimations made by the researcher. They are based on the severity scale of "Cosmetic, Serious and Critical" from [25]. Introducing problems of varied severity may indicate how sensitive the method is. The *actual* impact of these problems may be more or less severe in real use than predicted. Problems with a high intended impact are simply assumed more likely to show up in actual use, than problems with a predicted low severity.

EXPERIMENT

This section describes the actualization of the experiment.

Test Subjects: 20 university students ranging from 19-28 in age (mean 22.85) participated. The students were divided into two groups: LU and HU ("Low Usability" and "High Usability" respectively), with 8 males and 2 females in each group. All subjects were kept unaware of the actual premise of the test until after the interaction.

Software: "PosterMaker Pro", specifically developed for the experiment. The HU group was given the original version of the software, while the LU group received a version seeded with 5 usability problems.

Task: The subject was given a piece of paper entitled and described as being an examination assignment. The task was to use the "PosterMaker Pro" software to create an exact copy of a poster pictured on the paper. The user was informed that there was a 10 minute time limit to solve the task, and that the system would then automatically rate them. They would be rated as either Passed or Failed. The user could click a "Hand in" button at any time, if they felt the task had been solved before time was up. The result (Passed/Failed) was pre-determined by the researcher: half the people in each group (HU/LU) were set to fail, while the other half were set to pass.

Measurements: The screen was recorded using Camtasia Studio [27]. The users face was recorded with a webcam, which was also handled by Camtasia. The GSR of the user was recorded using a commercially available product: "Mindplace Thoughtstream USB Personal Biofeedback" [19]. The GSR sensor was placed in the palm of the users non-primary hand. A baseline for the GSR was measured by showing a blank screen for the first 4 minutes, while playing a relaxing piece of music. The song "Weightless" by Marconi

Union was chosen, due to existing studies indicating it's usefulness in relaxing human subjects [2] [6].

Environment: The experiment was conducted in a university classroom. The user was placed in front of a laptop with an external mouse, and given the paper describing their task. The task was also described verbally by the researcher. After the user had confirmed that they understood the task at hand, the researcher started the software and left the room. The first 4 minutes the software showed a blank screen while playing relaxing music, and then automatically opened the PosterMaker Pro window. This window was shown on top of a black background, and nothing but the software was visible or accessible during the interaction. The student was alone in the room for the full duration of the interaction. The users were aware that they were being recorded in various ways, but were not told the purpose of this until after the interaction had ended. The purpose of the GSR sensor placed in their hand was not explained until after the interaction either.

End of Interaction: All recordings were automatically stopped 30 seconds after the user had been shown their final result (Passed/Failed). The user was then instructed to retrieve the researcher.

Post-Interaction: The user was asked to choose the Emocard best describing their emotional state immediately after the interaction. They were then asked to report the same emotion using the three dimensions on the SAM scale. Both choices were noted by the researcher.

Identification of Usability Problems: The video of the users face was superimposed over the lower right corner of the screencast. The GSR data was visualized as a graph and superimposed over the timeline of the video player. This allowed for the researcher to visually identify peaks in the GSR data, and fast-forward to about 5 seconds before these points. The user was then shown this part of the video (screencast as well as their own facial expressions), and asked to freely describe their own thoughts as to what they may have reacted to, and how. If the user was able to deduce a reaction, the peak was noted along with their description of the event. They were also asked to state the emotional state at the time using both Emocards and SAM.

Compilation of Usability Problem Set: The complete list of qualitative descriptions of experienced emotions were reviewed by the researcher, and all related problems grouped together.

RESULTS

Both the experiment and the developed UEM is based on a range of theoretical assumptions that need to be verified before any conclusions can be drawn. The data recorded from the test is used to indicate whether the assumptions can be considered correct or not.

GSR as Indicator of Emotion

The UEM depends on the GSR measurement being able to indicate at what time the user experiences an emotional response. If true, peaks in the GSR should be indicative of emotional responses in the user. It was found that users could recognize and describe an emotional reaction at 57.75% of the registered peaks. Furthermore, there was a statistically significant correlation between GSR (relative to baseline) and self-reported Arousal in both Emocards ($p < 0.005$) and SAM ($p < 0.005$).

In conclusion, GSR seems to be a useful indicator of emotion. Users also show a high ability to successfully recognize and describe their experiences.

Consistency Across Measures

Some of the measurement-techniques applied in the experiment claim to provide data about the same dimension of emotion. It is interesting to see whether there is any correlation between the values provided by each tool, as disagreements could indicate: (1) Dimensions being different despite similar labels, (2) Failure of one or more tools in measuring the intended dimension and/or (3) Experimental errors.

SAM, Emocards and GSR are all claimed to measure Arousal. A positive correlation was found between GSR and self-reported Arousal on SAM ($p < 0.005$). A similar correlation was found between GSR and Emocard Arousal scores ($p < 0.005$). This of course implies a correlation between the two self-report methods too, which was confirmed with a significance of ($p < 0.006$).

Both SAM and Emocards claim to measure Pleasure, which the experiment seems to have confirmed through a strong positive correlation between the two measures ($p < 0.001$).

In conclusion, the measurement techniques all seem to successfully measure aspects of the actual emotional state of the user.

Successful Seeding

The 5 seeded usability problems in the LU-version of the software were conceived by the researcher, and had no empirical proof of being *actual* usability problems. As no other UME was applied to the software beforehand it is also likely to contain organic usability problems that may interfere with the seeded ones. Before any conclusions can be made on the effectiveness of the developed UME, it must first be investigated whether the seeding actually caused any more usability problems to be experienced.

If the seeding was successful, users given the LU-version should on average experience more usability problems than the HU-group. This was confirmed: a t-test indicates a significant ($p < 0.001$) difference in the amount of usability problems experienced between the two groups, with the LU-group experiencing the most problems (See Figure 3). Also, the ratio of identified problems to total registered emotional responses were found to be higher in the LU group compared to the HU group ($p < 0.001$) as depicted in Figure 4.

Identification of Seeded Problems

The UME successfully identified 4 out of the 5 seeded usability problems:

- **Identified Usability Problems (Seeded):**

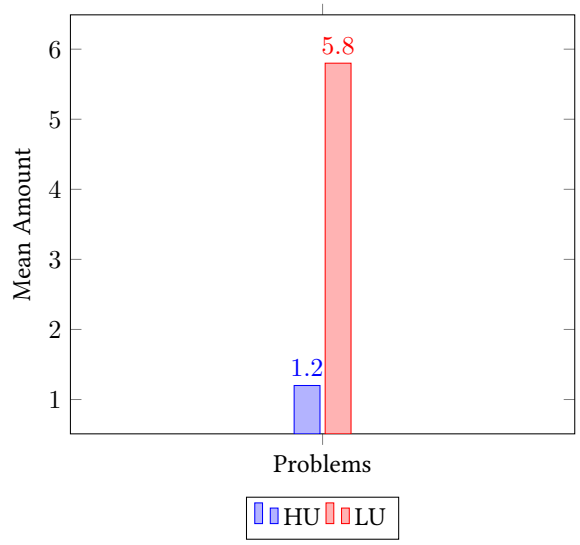

Figure 3. The compared mean amount of identified problems in the HU- and LU-version of the software.

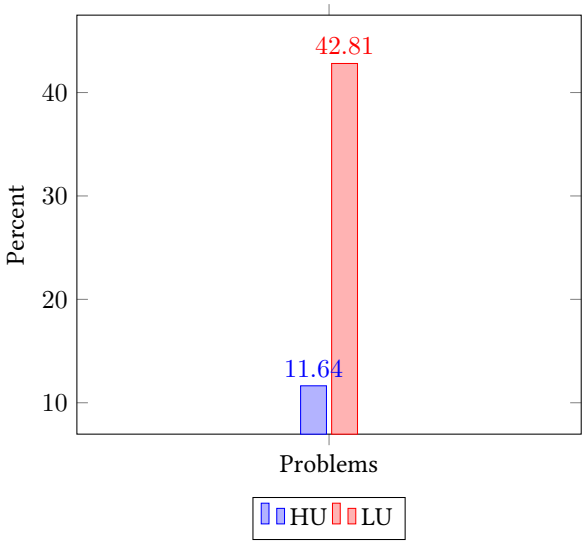

Figure 4. Comparison of the ratio between usability problems and total registrations in the HU- and LU-version of the software.

  - No Preview Font-Selector
  - Missing Font
  - Colors in Combobox
  - Error Message on Image Insert
- **Missed Usability Problems (Seeded):**
  - Non-Standard Font Style

The missed problem was predicted to be Cosmetic during the seeding. All users were able to style the text using the non-standard controls and seemed to do so without hesitation. It seems likely that this was the *seeding* failing as opposed to the UME: the predicted severeness was simply wrong. It was not an *actual* usability problem.

Another possible point of critique in the seeded problems actually led to an interesting discovery. The two first-mentioned problems (*No Preview Font-Selector* and *Missing Font*) are very similar. The users frustration is in both cases connected to the issue of not being able to identify the correct font. Whether this is due to lack of preview, or because the font has been removed, is irrelevant. In theory, this would make the two problems indistinguishable. This was also the case for most of the user-provided descriptions - but a few of them actually addressed the perceived problem as directly one or the other. This goes to show how valuable qualitative descriptions provided by the users can be.

Identification of Organic Problems

A range of organic, un-intended usability problems were also identified by the UEM:

- **Identified Usability Problems (Organic):**
  - Selector is misunderstood as text-input
  - No preview of color selection
  - Can't manually enter Color (LU), Font or Size
  - Text can't be moved while in edit-mode
  - Desc. text dissapears
  - Exceptions if Desc. is empty
  - Sudden jumping/moving of Title
  - No rulers for placement
  - Overlapping invisible border on text

The problems are sorted by the number of users experiencing them, the first in the list being the most common. No users reported any problems outside of the 4 seeded and 9 identified, but were not systematically asked to do so either. The high success-rate in identifying seeded problems, along with the fact that 9 new and unintended problems were found using the method, seems to suggest that it is a functioning UEM. An emperical study comparing its effectiveness to other UEMs seems warranted.

Usability Problems and Frustration

The relationship between usability problems and emotional responses is based on "Frustration" being described as an indicator of usability problems in [25]. The UEM has already demonstrated an ability to identify usability problems by utilizing emotional measurement, but it would be interesting to investigate whether the reported emotions are actually related to frustration. The term "Frustration" is only loosely described in [25], but the Oxford Dictionary of English provides the following two definitions:

> 1. the feeling of being upset or annoyed as a result of being unable to change or achieve something.
> 2. the prevention of the progress, success, or fulfilment of something.
>
> *[26]*

These definitions seem compatible with the general understanding of usability problems. They also allow us to assume that frustration is likely characterized by a low Pleasure-rating. Arousal and Dominance seem to have a less direct relationship to frustration: "upset" and "annoyed" would seem to describe opposing ends on both scales. This leaves the *deviation from a neutral rating* in these dimensions, as a possible indicator of frustration. The mean deviation of Arousal-ratings for identified usability problems was found to be 1.829 (SD = 1.0594), and the mean deviation of Control 1.400 (SD = 1.0594). Whether these deviations can be considered significant enough to characterize frustration is hard to tell, as no hard data has been found that distinguishes "frustration" from other emotions based on these dimensions.

Should the stated assumptions on the characteristics of frustration be true, it may show a difference between registered emotions that have been classified as usability problems (regProb) and emotions that were not recognized as such (regOther). The emotional responses in regProb were found to have a statistically significant ($p < 0.001$) lower Pleasure-rating (Mean = 3.38) compared to regOther (Mean = 5.35). Mean ratings of Arousal and Control were both higher in the regProb list, and were also found to be significantly different ($p < 0.001$). These differences have been illustrated in Figure 5.

Emocards already place high arousal on both ends of the scale, needing no modification to suit the inferred understanding of frustration. A statistical significant difference was found between the Emocard-ratings in the two lists ($p < 0.001$). The regProb list was generally rated lower (Mean = -1.38) than regOther (Mean = 0.11). This difference is illustrated in Figure 6.

These findings suggest that there is indeed a relationship between frustration and usability problems. A better model of how frustration is expressed through SAM and/or Emocards may even allow for an automatic filtration or sorting of data collected through this method. This would be beneficial, as emotions unrelated to usability problems may not be of interest in pure usability studies.

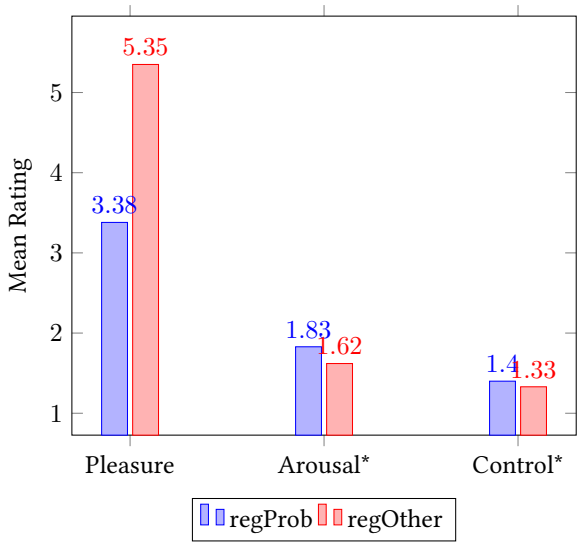

Figure 5. Mean ratings of Pleasure, Arousal and Control in regProb (usability problems) and regOther (other registered emotions).
* Arousal and Control are represented as their deviation from the neutral value of 4.5.

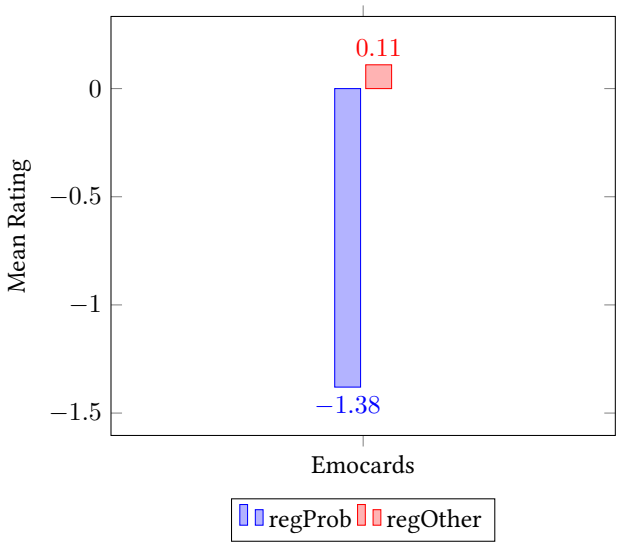

Figure 6. Mean ratings of Emocards in regProb (usability problems) and regOther (other registered emotions).

Usefulness of Post-Interaction Testing

The self-report tools used in this experiment, SAM and Emocards, are usually only deployed before and/or after the interaction with a system [3]. Deployment before the test indicates how first sight affects the user emotionally, but what does a post-interaction report indicate? Researchers may intuitively assume that it represents the overall experience of the system, and that a positive emotional score post-interaction suggests that the system overall rates high on User Experience. One argument against this is the Peak/End rule:

According to the Peak/End rule [9] an experience is judged by the strongest (Peak) and last (End) part of it. Should this hold true to this case, one would expect to find a correlation between the emotional state reported at Peak/End points and the reported state after the interaction had ended. Self-report ratings from the Peak (determined by GSR) and End (last reported emotion) experiences were compared to the ratings provided by the user post-interaction ("After") with the following results:

- **Peak:** No significant correlations were found between Peak end After in any of the measures.

- **End:** A significant correlation between End and After was found with Emocards ($p < 0.009$). Scores on the Pleasure-dimension of SAM was also found to correlate positively between the two ($p < 0.024$). The remaining dimensions did not show any significant correlations.

To test whether the post-interaction test is a good indication of the overall User Experience of a system, another correlation-test was performed. This time the mean value ("Average") of the dimensions were compared to After:

- **Average:** A significant correlation was found between the Average and After report of Arousal through SAM ($p < 0.015$). Emocards also showed a significant correlation ($p < 0.026$).

The data gathered through this experiment fails to support the validity of the Peak/End rule in interactive systems. The data instead points towards post-interaction deployment of SAM revealing something akin to the average arousal of the user. This would need to be verified through a new experiment. As Emocards showed a significant correlation in both End- and Average-scores, it is not clear which of these, if any, a post-interaction deployment of the tool is an indication of.

CONCLUSION

A UEM has been successfully developed and tested. The UEM uses a psychophysiological measure to identify the *onset* of emotion, and uses this information to extract relevant parts of a screen-recording and a video of the users face. These clips are then shown to the user post-interaction, as they are asked to describe the cause of experience in their own words and rate their emotional state at the time using self-report tools. The list of qualitative data has been shown to successfully produce a list of usability problems through expert review.

The UEM showed promising results. It successfully identified 4 out of 5 seeded usability problems, with the missed one suspected of not constituting an *actual* problem. The UME also identified 9 unintended problems in the system, further supporting its ability to find and describe problematic features of the system. The UEM complies as follows to the stated criteria of a UEM:

- **Thoroughness:** Based on the seeded problems alone, the UEM seems to comply with the criteria of Thoroughness. A study comparing the UEM to existing methods would be needed to verify this.

- **Validity:** The use of user-provided qualitative data seems to ensure a high Validity of the detected problems. The method still relies on expert judgment in some areas, but possible alternatives have been suggested.

- **Reliability:** The Reliability of the UEM has not been tested. It would be interesting to see whether the inclusion of user-provided qualitative data can mitigate the Evaluator-Effect.

- **Downstream Utility:** This UEM seems to offer a high Downstream Utility. The user-provided qualitative data provides descriptions of the problematic feature from multiple angles. Users were found to generally express their expectations when they were not met, as well as offer possible solutions to the experienced problems.

- **Cost Effectiveness:** This UEM uses specialized equipment (GSR), which incurs extra cost and acts as a bottleneck for the number of tests that can be run simultaneously. The method is faster than traditional video analysis though, and the test itself could even be performed by an employee without expert knowledge in usability due to its reliance on technology and users to identify problems.

The UEM itself seems feasible, and the usefulness of user identified usability problems warrants further exploration.

REFERENCES

1. Autonomic nervous system. In *Encyclopædia Britannica Academic Edition (Online)*, winter 2013 ed. 2013.

2. Agrawal, A., Makhijani, N., and Valentini, P. The effect of music on heart rate.

3. Bargas-Avila, J. A., and Hornbæk, K. Old wine in new bottles or novel challenges: a critical analysis of empirical studies of user experience. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM (2011), 2689–2698.

4. Bastien, J. C., and Scapin, D. L. Evaluating a user interface with ergonomic criteria. *International Journal of Human-Computer Interaction 7*, 2 (1995), 105–121.

5. Battarbee, K., et al. *Co-experience: understanding user experiences in interaction.* Aalto University, 2004.

6. Belford, Z., Neher, C., Pernsteiner, T., Stoffregen, J., Tariq, Z., and Lokuta, A. Music & physical performance: The effects of different music genres on physical

performance as measured by the heart rate, electrodermal arousal.

7. Bradley, M. M., and Lang, P. J. Measuring emotion: the self-assessment manikin and the semantic differential. *Journal of behavior therapy and experimental psychiatry 25*, 1 (1994), 49–59.
8. Desmet, P., Overbeeke, K., and Tax, S. Designing products with added emotional value: Development and application of an approach for research through design. *The design journal 4*, 1 (2001), 32–47.
9. Do, A. M., Rupert, A. V., and Wolford, G. Evaluations of pleasurable experiences: The peak-end rule. *Psychonomic bulletin & review 15*, 1 (2008), 96–98.
10. Gray, W. D., and Salzman, M. C. Damaged merchandise? a review of experiments that compare usability evaluation methods. *Human–Computer Interaction 13*, 3 (1998), 203–261.
11. Hartson, H. R., Andre, T. S., and Williges, R. C. Criteria for evaluating usability evaluation methods. *International journal of human-computer interaction 13*, 4 (2001), 373–410.
12. Hartson, H. R., Andre, T. S., Williges, R. C., and Rens, L. v. The user action framework: A theory-based foundation for inspection and classification of usability problems. In *Proceedings of HCI International (the 8th International Conference on Human-Computer Interaction) on Human-Computer Interaction: Ergonomics and User Interfaces-Volume I-Volume I*, L. Erlbaum Associates Inc. (1999), 1058–1062.
13. Healey, J. Gsr sock: A new e-textile sensor prototype. In *Wearable Computers (ISWC), 2011 15th Annual International Symposium on*, IEEE (2011), 113–114.
14. Hertzum, M., and Jacobsen, N. E. The evaluator effect: A chilling fact about usability evaluation methods. *International Journal of Human-Computer Interaction 13*, 4 (2001), 421–443.
15. Jacobsen, N. E., Hertzum, M., and John, B. E. The evaluator effect in usability tests. In *CHI 98 Cconference Summary on Human Factors in Computing Systems*, ACM (1998), 255–256.
16. Lin, T., Omata, M., Hu, W., and Imamiya, A. 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*, Computer-Human Interaction Special Interest Group (CHISIG) of Australia (2005), 1–10.
17. Maslow, A. H. A theory of human motivation. *Psychological review 50*, 4 (1943), 370.
18. Mehrabian, A. *Basic Dimensions for a General Psychological Theory: Implications for Personality, Social and Environmental Psychology*, 1st edition ed. Oelgeschlager, Gunn & Hain Inc.,U.S., 8 1980.
19. MindPlace. Mindplace Thoughtstream USB Personal Biofeedback. **http://www.mindplace.com/Mindplace-Thoughtstream-USB-Personal-Biofeedback/dp/B005NDGPLC**, 2014. [Online; accessed 07-June-2014].
20. Poh, M.-Z., Swenson, N. C., and Picard, R. W. A wearable sensor for unobtrusive, long-term assessment of electrodermal activity. *Biomedical Engineering, IEEE Transactions on 57*, 5 (2010), 1243–1252.
21. Psychlab. SKIN CONDUCTANCE EXPLAINED. **http://www.psychlab.com/SC_explained.html**, 2014. [Online; accessed 07-June-2014].
22. Schachter, S., and Singer, J. Cognitive, social, and physiological determinants of emotional state. *Psychological review 69*, 5 (1962), 379.
23. Scherer, K. R. What are emotions? and how can they be measured? *Social science information 44*, 4 (2005), 695–729.
24. Shi, Y., Ruiz, N., Taib, R., Choi, E., and Chen, F. Galvanic skin response (gsr) as an index of cognitive load. In *CHI'07 extended abstracts on Human factors in computing systems*, ACM (2007), 2651–2656.
25. Skov, M. B., and Stage, J. Supporting problem identification in usability evaluations. In *Proceedings of the 17th Australia conference on Computer-Human Interaction: Citizens Online: Considerations for Today and the Future*, Computer-Human Interaction Special Interest Group (CHISIG) of Australia (2005), 1–9.
26. Soanes, C., and Stevenson, A., Eds. *Oxford Dictionary of English*, 2 ed. Oxford University Press, 5 2004.
27. TechSmith. Camtasia Studio. **http://www.techsmith.com/camtasia.html**, 2014. [Online; accessed 07-June-2014].
28. van Nimwegen, C., and Uyttendaele, A. Unobtrusive physiological measures to adapt system behavior: The gsr mouse. *status: published* (2009).

Bibliography

[Bargas-Avila and Hornbæk, 2011] Bargas-Avila, J. A. and Hornbæk, K. (2011). Old wine in new bottles or novel challenges: a critical analysis of empirical studies of user experience. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2689–2698. ACM.

[Skov and Stage, 2005] Skov, M. B. and Stage, J. (2005). Supporting problem identification in usability evaluations. In *Proceedings of the 17th Australia conference on Computer-Human Interaction: Citizens Online: Considerations for Today and the Future*, pages 1–9. Computer-Human Interaction Special Interest Group (CHISIG) of Australia.