It's not Complicated

A Study of Non-Specialists Analyzing GSR Sensor Data to Detect UX Related Events

Bruun, Anders

Published in:
NordiCHI 2018

DOI (link to publication from Publisher):
10.1145/3240167.3240183

Publication date:
2018

Document Version
Accepted author manuscript, peer reviewed version

It's not Complicated: A Study of Non-Specialists Analyzing GSR Sensor Data to Detect UX Related Events

Anders Bruun
Aalborg University
Aalborg Øst, Denmark
bruun@cs.aau.dk

ABSTRACT
Emotion is a key factor in understanding user experiences (UX) of interactive systems. An emerging trend within HCI is to apply physiological sensors for uncovering emotions. Previous studies rely on various sophisticated analysis techniques and specialized knowledge to interpret sensor data. While commendable for increasing accuracy at fine grained latencies (to detect events within seconds), this can be challenging for UX practitioners without specialized knowledge. This study contributes in two ways. Firstly by understanding the level of sensor accuracy in detecting UX related events. Secondly by applying a basic analysis approach where sensor data is interpreted by 21 non-specialist participants (no previous experience in doing this). Their performance is compared to random guessing. Findings show that sensor data analyzed by non-specialists are significantly more accurate in capturing subjectively reported UX events than random guessing. An accuracy level of 60-80% was obtained at granularities within 3.5-11 seconds of UX related events.

Author Keywords
Emotion; Non-specialists; Physiological sensors; GSR; Sensor Data Analysis; Orienting responses; Subjective;

ACM Classification Keywords
H.5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous;

INTRODUCTION
Emotion is a fundamental factor in measuring and understanding user experience (UX) of interactive technologies [2,21,59]. Emotional states of users have typically been elicited through subjective ratings of arousal/valence dimensions in, e.g. well validated questionnaires such as the Self-Assessment-Manikin [7]. An emerging alternative is to apply real time sensor data to measure and understand the emotional dimension of UX [24,60]. Data from physiological sensors indicate emotional states of users where e.g. Galvanic Skin Response (GSR) sensors in particular are proven reliable for measuring changes in arousal [7,23]. UX researchers have argued that physiological sensors have limited applicability for practice due to their extensive costs, see e.g. [34]. However, sensors are now more commonplace in smart watches [51] and have also been introduced in smart fabrics [28] and therefore hold considerable potential as data sources to measure and understand UX during actual interactions with technologies.

Based on experiences from the software development industry, Georges et al. argue for a need to include more data driven recommendations based on emotional reactions [24]. However, there are at least two critical challenges for using physiological sensors in practice: 1) data analysis requires specialized knowledge and 2) the level of sensor accuracy in detecting events of interest is currently unknown in HCI contexts.

In terms of data analysis, the challenge is to identify UX related events in the sensor data at specific points in time. Such analysis is challenging [22,23,25,34], partly because physiological data are fluctuating within seconds [63].

There is also a need to study the accuracy of sensors in revealing UX related events during interaction. This is particularly relevant for formative purposes, i.e. in order to identify design elements leading to particular emotional experiences [9,24]. Studies within psychology have found correlations between sensor data and external stimuli, yet, those studies primarily rely on presenting distinct stimuli such as pictures. This may not translate to the interactive nature of technology use in HCI contexts [20,63].

The contribution of this study is to demonstrate practical feasibility of using physiological sensors for assessing the emotional dimension of UX in real time. To this end no participants had prior knowledge in the use of physiological sensors and the analysis of sensor data. Within this practice related constraint, the following research question is examined: *How accurate are physiological sensors in detecting emotional reactions related to specific UX events during interaction?*

Firstly, the theoretical background and how to measure emotions is outlined. This is followed by a walkthrough of related work within HCI. Next, the method, findings, discussion and finally conclusions are presented.

NordiCHI'18, September 29-October 3, 2018, Oslo, Norway

ACM ISBN 978-1-4503-6437-9/18/09…$15.00
https://doi.org/10.1145/3240167.3240183

THEORETICAL BACKGROUND

This section gives an overview of the theoretical background on defining and measuring emotions as well as challenges in analyzing sensor data.

Defining Emotions

Defining emotions is a highly debated topic within psychology and providing an exhaustive walkthrough is outside the scope of this paper. That said, the century old theory of William James and Karl Lange has influenced more recent work. James and Lange theorized that emotions occur as a result of bodily changes in autonomic and motor functions, which in turn are activated by perceived stimuli [30]. As an example of more recent accounts supporting this, Klaus Scherer's appraisal theory consider emotions as adaptive responses to appraisals of environmental features [45]. These responses are also referred to as orienting responses. Emotions are considered as the result of a process involving physiological and cognitive components of which external or internal stimuli act as catalysts. Once the process is initiated, appraisals lead to mobilization and synchronization of physiological subsystems. According to Scherer, we respond to external or internal stimuli in relation to the well-being of our organism, which is done through appraisals that then lead to physiological reactions, i.e. orienting responses [45,54].

Describing emotions typically falls within discrete and dimensional models. Ekman's set of basic emotions is an example of the former [18,19] whereas Lang and Russel consider emotions as a vector in multidimensional space of valence and arousal [5,37,53]. However, regardless of the model applied, there still exist an inherent challenge in eliciting emotions due to their multifaceted and ephemeral states [11]. This is furthermore complicated by considering related nuances such as mood and affect (see [17] for a review). The study presented in this paper is informed by the dimensional model of valence-arousal using subjective accounts and objective measures for eliciting emotional responses (see e.g. [43,50] for further discussions).

Measuring Emotions

The following provides an outline of subjective and objective approaches in eliciting emotions as well as a discussion of merits and limitations.

Subjective Measures

Gathering subjective accounts of emotional reactions are typically done through questionnaires such as the Self-Assesment-Manikin (SAM) [5] or Emocards [16]. To date this is the primary method for eliciting emotional reactions within HCI studies, and data is typically gathered after task completion, cf. [2,7]. Given the ephemeral nature of emotions, one should be careful in letting study participants provide such retrospective accounts of real time experiences as this leads to memory biases. An example of such a bias is the peak-end effect, cf. [7,9,32,44]. Therefore it is critical that emotions are measured as close as possible to the point in time they occurred.

A more recent alternative within HCI dealing with the peak-end effect is the Valence method proposed by Burmester et al. [10]. Users experience positive or negative emotions while interacting with a product and these are captured by the users marking a "+" (plus) or "-" (minus) sign on an external keypad. A timestamp for each mark is logged and used in a follow-up retrospection phase to interview users about their experiences. These +/- markers are referred to as valence markers and thus indicate points in time where users experience either positive or negative emotions in relation to the interaction.

Objective (psychophysiology)

Aligned with Scherer's appraisal theory, an emerging trend in HCI studies is to elicit emotions through objective data obtained from physiological sensors [24,60]. Several studies have e.g. applied heart rate (HR), galvanic skin response (GSR) and other sensors to gain insights on emotional states of participants in real-time. Orienting responses causing changes in electro dermal activity can be measured with GSR sensors, which have been shown to correlate well with subjective accounts of arousal across several independent studies in varying contexts, see e.g. [7,22,23,34,36]. GSR sensors are reported to be less sensitive to noise and less ambiguous than electromyography (EMG) and HR sensors [34].

Analyzing Physiological Data

It is challenging to analyze data signals from physiological sensors [23,25,34], partly because physiological data are changing rapidly (within seconds) [63]. To understand emotional experiences related to interaction designs, it is important to identify orienting responses in the sensor data reflecting when UX related events occur. This is necessary to couple specific events within an interaction sequence to an emotional experience [9]. Orienting responses could e.g. be observed as fluctuations in arousal, which would show up as changes in the skin conductance level with varying magnitudes and durations (see Figure 1 for an example), cf. [29,38].

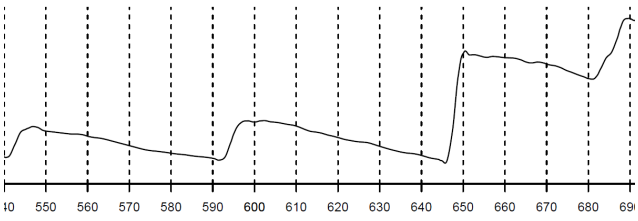

Figure 1. Part of a GSR graph from one test user within this study. Y-axis represents skin conductivity, x-axis represent the time in seconds.

Theoretically, the continuous and dimensional model of describing emotions leads to an infinite number of valence/arousal combinations. Fairclough also discuss the potential many-to-many relationship between physiological and psychological manifestations, where e.g. several physiological signals can be associated with more psychological elements [20]. This level of nuance requires

advanced data analysis in order to classify responses based on signals (and the fusion of several signals). Previous studies suggested using machine intelligence to correlate signal patterns with events representing some form of ground truth. Such approaches have enabled researchers to create algorithms that classify emotions at two or three levels (e.g. high, neutral and low valence/arousal) with varying degrees of accuracy (40-90%), cf. [12,26,31]. However, given the individual nature of emotions, such precision requires that machines are trained up front and that a training session should be conducted for each individual participant in order to reach accuracy levels of 90% as shown by Calvo et al. [12]. When the same training set was used across multiple participants', Calvo and colleagues found that accuracy dropped to 40% [12]. Thus, much effort is required to obtain high accuracies even for relatively crude two and three level classifications.

Another analytic issue is the latency of sensor signals in relation to the moment an event occur. This latency is explained through Scherer's appraisal theory, in particular that physiological manifestations occur after appraisals of an event are made. Reports of latency levels vary between previous studies (and depends on the sensor type used). Forne [22] states that it is hard to link observed responses to particular points in time and claims that GSR sensors have latency times of up to 6 seconds, which is also supported by Park et al. [49]. Kivikangas et al. report lower latencies between 1-4 seconds for GSR [34], while Ward and Marsden operated with latencies of 10 seconds [63]. Park [49] and Stern et al. [58] argue that it is challenging to pinpoint exact moments in time of an event leading to physiological reactions. They therefore suggest using signal averages spanning over 10 second periods.

Given the level of nuance in classifying the level of valence/arousal of emotions, individual differences and discrepancies in signal latency, it is challenging for UX practitioners without specialized knowledge to analyze physiological data [22].

RELATED WORK

The use of physiological measurements in HCI is an emerging trend [20,42,46,60] with a primary emphasis on understanding UX in digital game applications. Studies within the gaming domain have dealt with method validations, measuring social gaming experiences, adaptive interfaces, studying the effect of particular game features and events, see [20,34] for more details. Less attention is given to non-gaming applications in terms of measuring UX physiologically.

Generally, there are two approaches to utilize physiological data in HCI. One is to consider averaged signals across entire periods of time (long-term changes in psychophysiological states, typically within several minutes). This could cover averaged signals over an entire interactive session or across entire tasks, which is in line with Park [49] and Stern et al.'s [58] recommendations outlined above. The alternative approach (short-term changes) is to identify individual orienting responses within seconds, i.e. discrete events. This e.g. covers using physiological data signals to detect specific points in time where particular events occur, cf. [63].

The most common tendency in HCI studies is to analyze physiological signals over entire periods of time, e.g. a full interactive session or across entire tasks rather than isolating signals around specific events within a session or task. Ward and Marsden present one of the earliest HCI studies exploring how GSR and heart rate is affected when interacting with well- and ill-designed versions of a web site. That study includes findings related to both long- and short term signal changes. Findings show that changes in skin conductivity over the entire interaction sequence was higher for participants using the ill-designed version of the website (indicating a higher level of arousal) compared to the well-designed version. Ward and Marsden also observed more short-term effects on skin conductivity related to specific events (pop-up windows). Here they found significant differences in conductivity changes 10 seconds before and 10 seconds after the events.

Wilson [64] presents a study based on measuring long term changes. The purpose is to examine the effect of media quality on the user experience within a video conferencing context. This is done by measuring GSR and heart rate in relation to varying video quality grades (5 vs. 25 fps). Results are reported as overall averages across entire interactive sessions and reveal that physiological signals indicate higher stress levels in lower video quality conditions compared to the higher video quality conditions.

Another, and more recent study based on long term signal changes is presented by Yao et al. [65]. They showed correlations between task performance and physiological data obtained from GSR and heart rate data. That study is based on comparing data within the entirety of tasks.

Similarly, Novak et al. [47] use GSR, heart rate and other sensors to study the effect of mental workload when performing single vs. dual-tasks on a computer. Physiological measurements, e.g. skin conductivity change, were based on averaging signals over entire task periods. Findings from that study show that physiological signals are sensitive to mental workload caused by different task types.

While averaging long term signals is justifiable for summative assessment purposes, it is arguably insufficient for formative assessments where interaction designers need to identify user experiences in relation to specific design elements [7,9]. In sum, long term changes (over minutes) is the predominant approach to analyzing physiological data in HCI studies. There is a need to understand the extent to which UX practitioners can utilize physiological data to identify specific points of interest (short term) related to positive or negative user experiences.

METHOD

The overall procedure of this study was divided in two parts. The first part was a user test in which participants were asked to solve one task using a web application while gathering GSR sensor data. The second part of the study was the analysis of sensor data in which participants were asked to identify orienting responses from the GSR graphs gathered from the first part of the study.

Participants

The study is empirically based on 21 participants acting in different roles during the study. Initially they acted as the test users interacting with a web application (outlined below). The same participants also acted as analysts and were asked to interpret sensor data.

All participants were first year students enrolled in the Informatics education at our university. They participated on a voluntary basis. Their mean age was 21 years (sd=1.6), all male. The profile of this education relates to the design of IT systems in general with an emphasis on HCI and systems development. At the time of the study they had just finished their foundational course on the topic of designing and evaluating user interfaces. Thus, although participants were students, their profiles reflected that of novice UX designers without specialized knowledge or experience in working with physiological sensors.

Part 1: User Test

This part of the study was conducted in a dedicated usability lab at the university department. Participants were introduced to the study as having the purpose of being a UX assessment of a web application (detailed below). The purpose of the experiment was not revealed.

Participants wore a GSR sensor during the evaluation in order to capture orienting responses. The sensor was attached in the palms of the participants' non-dominant hand before starting the task. This hand was chosen since GSR sensors are sensitive to physical movements, which cause artifacts in the data [23]. They were also asked to provide valence markers while interacting with the interface (see e.g. the section "Measuring Emotions" above). The markers were used as the ground truth to which the GSR data was compared.

After the introduction and GSR setup, participants were provided with the task to be solved using the web application. At this point the experimenter left the test room and went into an adjacent control room to observe when participants finished. As the GSR sensor responds to arousal, participants had to get into a state of relaxation before task solving began. To this end a blank screen was shown for the first three minutes after which the web application automatically started. They had a maximum of 10 minutes to complete the task and were asked not to think aloud during interaction as this could also interfere with the GSR data, cf. [7].

System and Task

The web application Statistics Denmark (www.dst.dk) was used as the case system for the user test. This provides publicly available statistics on income- and educational levels, employment rates and many other statistics related to the Danish society. Test users were given the following task:

"Your sister considers opening a communications agency in Vejen [Danish town]. How many communication agencies were there in 2015 in Vejen with one employee?"

This task could only be solved using different advanced search filters in the application.

Valence Markers as the Ground Truth

It is typical practice within psychology to use subjectively reported emotional experiences as the ground truth. As an example, the widely applied IAPS set of pictures have been verified using subjective SAM ratings, see e.g. [4]. Studies in HCI have also used subjective ratings through questionnaires for this purpose, e.g. [27,39–41,59]. The valence method presented in [10] was used as this seems less obtrusive than filling in questionnaires during interaction.

A timestamp was logged every time test users pressed the +/- keys. This was done through a simple software application running in the background, i.e. not visible to the user. The timestamps were synchronized with timestamps obtained from the GSR sensor data.

Physiological Sensor

It was opted to use a GSR sensor in this study as this type of sensor is widely covered within previous research and is less sensitive to noise than e.g. EMG or HR sensors. See the previous discussion on this in the "Measuring Emotions" section above.

The Mindplace Thoughtstream GSR sensor was used to measure skin conductivity. This is registered in terms of resistance (kOhm) between two electrodes attached to the underside of the palm.

Part 2: Analysis of Orienting Responses

If physiological sensor data is to be applied in UX design and evaluation practices, the data analysis prerequisites must not exceed available analytic resources of practitioners. Therefore, the study presented in this paper is not based on sophisticated techniques as those outlined previously. For the purpose of the study, participants (now acting as analysts) were asked to identify orienting responses through visual inspection of the GSR graphs obtained from the first part of the study. For each graph participants were asked to identify points in time where the GSR graph, in their opinion, reflected sudden fluctuations in conductivity.

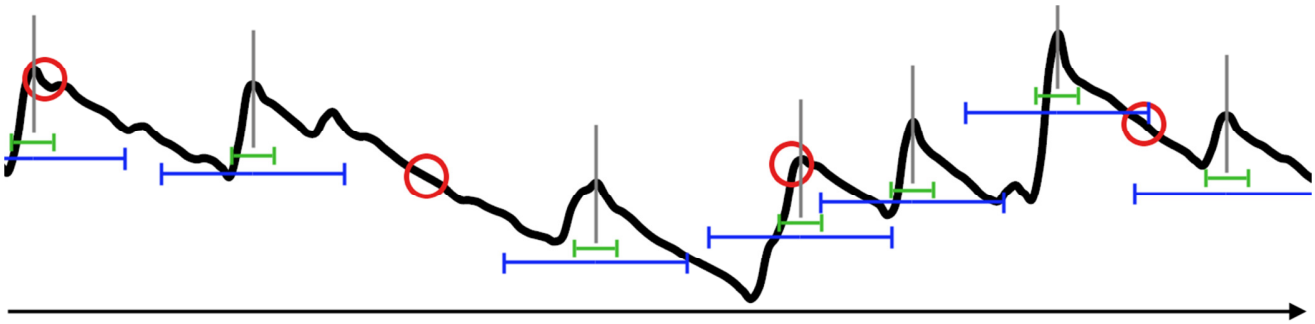

Figure 2. Example graph showing metrics for measuring accuracy and noise. Vertical lines (grey) = GSR orienting responses, Circles (red) = Valence markers, Narrow horizontal lines (green) = Latency, Broad horizontal lines (blue) = Latency

The GSR example in Figure 1 was given along with four timestamps to illustrate the idea. They were then asked to perform the analysis individually, and when done they submitted a list of timestamps for each test user. Thus, the 21 participants interpreted data for each of the 21 test.

Prior to sending the data to participants, the GSR graphs were smoothed by excluding large abrupt changes in skin conductivity levels (SCL), which can be caused by physical movement. Following [8], this was defined as 5 standard deviations larger the mean SCL occurring within 1 second.

Metrics

This section presents the accuracy, latency and noise metrics used to analyze findings. This meta-analysis was done by the authors of this paper.

Accuracy

Since participants (from now on referred to as non-specialists) had no prior experience in using and analyzing GSR data, it was relevant to study how accurate they were in identifying orienting responses. In this study accuracy is based on the level of agreement between:

1) Points in time of *valence markers* made during interaction, and
2) Points in time of *GSR orienting responses* as identified by the non-specialists

Figure 2 illustrates overlap between 2 valence markers (red circles) and 2 GSR orienting responses (grey vertical lines). Accuracy is defined as:

$$\text{Acc.} = \frac{\text{Valence markers} \cap \text{GSR orienting responses}}{\text{Total no. of valence markers}}$$

Since the total number of valence markers in Figure 2 is 4, the agreement is 2/4=.5.

Noise

Opposite to accuracy, noise is defined as the extent of non-overlapping points between valence markers and orienting responses. However, to get a sense of how many false positives there are registered by the sensor, noise is considered in relation to the number of GSR orienting responses.

Noise is therefore defined as:

$$\text{Noise} = \frac{\text{GSR orienting responses} \notin \text{Valence Markers}}{\text{Total no. of GSR orienting responses}}$$

Using the example in Figure 2, there is a total of 7 GSR orienting responses of which 5 do not overlap with valence markers. In this case the noise is 5/7=.71.

Latency

Appraisal theory states that physiological sensors inherently introduce latencies in registering orienting responses. Such latencies vary from 1-10 seconds thus making it difficult to link observed responses to the exact point in time of an event [22]. However, we do know that orienting responses occur after a triggering event. Adding to this uncertainty is the use of valence markers as we do not know whether these conscious events occur before or after orienting responses. This study therefore considers latencies surrounding both sides of GSR orienting responses as illustrated in the green horizontal lines of Figure 2.

The results section presents accuracy and noise as a function of latency. As an example, the accuracy in Figure 2 is .5, but allowing for a larger latency interval would eventually increase overlap between GSR orienting responses and valence markers. Using larger latency intervals, as illustrated by the wider horizontal (blue) lines in Figure 2, would allow for overlaps between 3 valence markers and 3 GSR orienting responses. In that case the accuracy would increase from .5 to 3/4 = .75. In turn this would also reduce noise from .71 to 4/7 = .57. In practice this larger latency interval means that specific UX related events are not pin pointed in time as precisely as when restricting to lower latencies.

Random guessing

Finally, the study examines the level of accuracy and noise of non-specialists' analyses compared to naïve (random) guessing of GSR orienting responses. This is relevant since, intuitively, interacting with a web application and doing search tasks would result in subtle emotional reactions compared to e.g. interacting with more engaging gaming environments. Additionally, as noted in [7], it is necessary to better understand the feasibility of using of physiological sensors in HCI contexts with subtle emotional stimuli.

Naïve guessing provides a worst-case scenario as this is based on the assumption of independency between experienced emotions and orienting responses obtained through the GSR sensor. For the naïve guessing it was chosen to randomly generate suggested points in time for each permutation of non-specialist and test user GSR graphs. This gave 21×21=441 sets of random data. The data distribution of non-specialist interpretations for each permutation was taken into account. This was done in order not to overly deflate naïve performance.

FINDINGS

The following sections present findings in terms of overall descriptive statistics followed by comparisons on accuracy and noise levels between non-specialists' analyses and random guessing.

Number of Orienting Responses and Valence Markers

Figure 3 (left) shows an overview of the number of GSR orienting responses that the non-specialists registered in their analyses. The mean number of orienting responses per test user is 15.4 (sd=9.41).

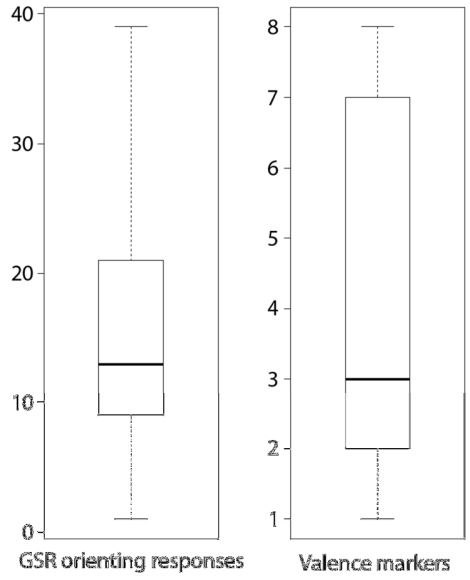

Figure 3. Boxplots showing the number of GSR orienting responses and valence markers of the test users.

Findings also show that the mean number of valence markers per test user is considerably lower than the number of GSR orienting responses (Figure 3 right side, μ=4.21, sd=2.63). An independent samples t-test (assuming unequal variances) shows this difference is significant (t = 9.97, df = 24, p-value < 0.0001).

Accuracy at Different Latency Intervals

Figure 4 shows the level of accuracy obtained assuming different latency intervals. Latencies span 0-20 seconds before and after a point has been identified in a graph (as illustrated in Figure 2). The solid line is based on participants' analysis, i.e. the accuracy of non-specialist interpretations in terms of GSR orienting responses. The dashed line is based on random guessing the points in time of orienting responses. Both graphs represent mean accuracies across participants and random guesses, both of which are functions of latency intervals.

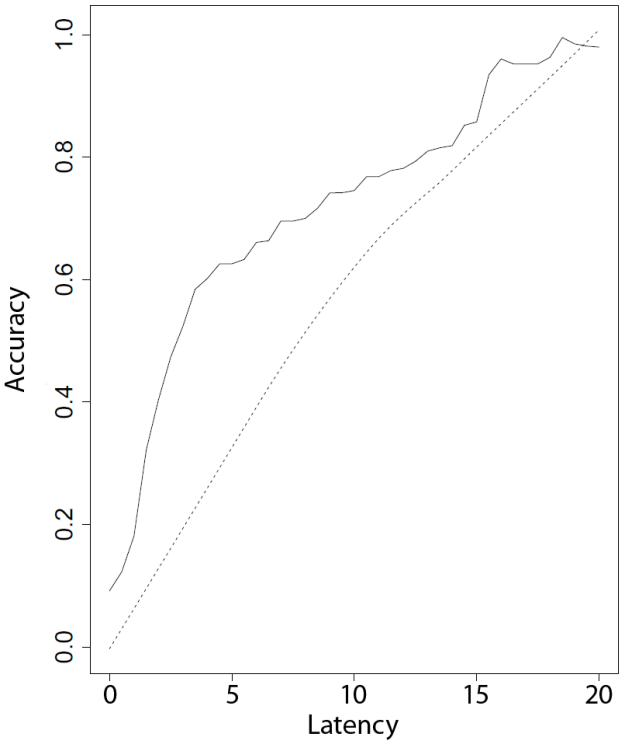

Figure 4. Accuracy as a function of latency levels. Solid line = Accuracy of GSR orienting responses, based on non-specialists' analysis, dashed line = Accuracy based on random guessing.

Assuming a latency of e.g. 2 seconds reveals an accuracy of .18 in case of the non-specialist analysis. In comparison, random guessing has an accuracy less than .1. An independent samples t-test (equal variance assumed) reveals significant difference in this respect (t = -8.3266, df = 38, p-value < 0.0001). The largest difference between accuracies of non-specialist and random guessing is at the 3.5 second latency interval, i.e. 3.5 seconds before or after a valence marker. At this point the accuracy of non-specialist interpretations is .58 and .23 in case of random guessing.

Generally, non-specialist interpretations represent significantly higher accuracies than random guessing in latency intervals from 0-11 seconds. As shown in Figure 4, the two graphs begin to converge at latency intervals above ~5 seconds, which culminate when assuming latencies of 11 seconds and beyond. At that point there is no significant difference between analyst interpretations and random guessing (t = 7e-04, df = 38, p = .99). This applies for the remaining latency interval from 12-20 seconds with p values between .05 and .99 ($pwr_{1-\beta}$ = [0.05;0.49], μ_{pwr}=.25, SD_{pwr}=.19). Assuming the coarse grained latency of 20

seconds, non-specialist interpretations and random guessing both reach an accuracy level of ~1.

Noise at Different Latency Intervals

Figure 5 shows the level of noise obtained at different latency intervals. Like the accuracy graphs, noise latencies are presented spanning 0-20 seconds.

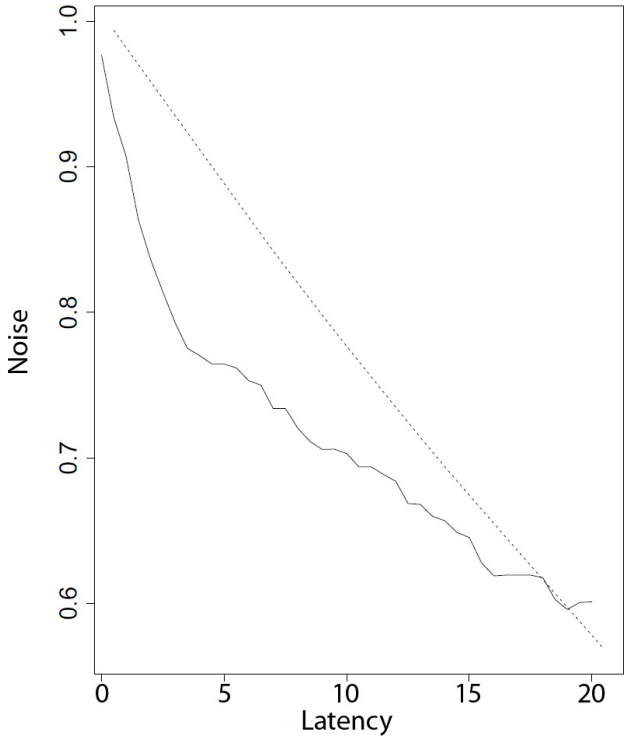

Figure 5. Noise as a function of latency levels. Solid line = Noise of GSR orienting responses, based on non-specialists' analysis, dashed line = Noise based on random guessing.

Assuming 0 seconds latency, the noise level for non-specialists' interpretations is .98 while random guessing is .99. Although similar (and likely of limited practical impact), an independent samples t-test (equal variances assumed) indicate that the noise level of analyst interpretations is significantly lower than random guessing ($t = 8.2537$, $df = 19.209$, p-value 0.0001). Similar to the accuracy graph in Figure 4, the noise graphs start to converge after assuming ~5 seconds latency. In the latency interval from 0-12 the noise level of non-specialist interpretations is significantly lower than that of random guessing.

When assuming 13 seconds latency, non-specialist interpretations and random guessing show noise levels of .68 and .72 respectively. At this latency interval, there is no longer significant differences in noise between non-specialists and random guessing ($t = 0.9378$, $df = 21.238$, p-value = 0.36). Thus, there are no significant differences when assuming latencies between 13-20 seconds with p values between .06-.92 ($pwr_{1-\beta}$ = [0.05;0.47], μ_{pwr}=.14, SD_{pwr}=.11).

DISCUSSION

This section discusses findings in terms of accuracy, noise and outline implications for research and practice. Finally study limitations and pointers for future study directions are presented.

It is not Complicated Obtaining High Level Accuracies

Findings from this study are encouraging as UX analysts without prior experience in working with physiological sensors reached a significantly higher accuracy level in their interpretations than naïve random guessing. It was found that non-specialists analyzing GSR sensor data were able to detect close to 100% of all subjective events marked by the participants (valence markers). This is the case when allowing for a latency interval of 20 seconds, i.e. GSR peaks occurring 20 seconds before and after a valence marker. However, non-specialist interpretations are not significantly higher than random guessing at that coarse grained level of latency.

Accepting only a very fine grained latency interval, e.g. 0-3 seconds before or after a valence marker, reveals a non-specialist accuracy of 10-18%. While this is a significantly higher level of accuracy compared to random guessing, it still seems rather low. Findings reveal that a 3.5 second latency interval before and after a marked event denotes the point with the largest accuracy difference between non-specialist interpretations (58%) and random guessing (23%). In general, considering latency intervals from 0-11 seconds before and after valence markers showed a significantly higher accuracy in case of non-specialist interpretations over random guessing. This resonates well with related work where GSR sensors are reported to have orienting response latencies between 1-10 seconds after an event occur, cf. [22,34,49,63].

In sum, UX analysts with no previous experiences in working with GSR sensors are able to capture between ~60-80% of subjectively reported UX related events during interaction. They even do so within a relatively fine grained latency interval spanning 3.5-11 seconds before and after a valence marker.

There is a High Level of "Noise"

Although the above findings are promising, there is also a considerable amount of noise in the data. To illustrate this point, users made an average of four valence markers on UX related events during interaction. In comparison there were 15 orienting responses (GSR peaks), on average. So, although three of four valence markers overlapped with a corresponding number of orienting responses (and indicated a high level of accuracy), this still leaves an average of 12 orienting responses unaccounted for.

Allowing for a wider latency interval results in a higher degree of overlap between valence markers and GSR orienting responses. This in turn increases accuracy. Yet, even if the 20 second latency interval is considered and all valence markers overlap with a GSR responses, there is still

a noise level of about 60% as shown in Figure 5. Regardless of latency granularity, these remaining responses can be considered false positives. Note that the valence markers used in this study denote points in time when test users become consciously aware of a UX related event occurring (discussed later). In relation to consciousness of events, Fairclough [20] notes that many events happen unconsciously, i.e. not explicitly prevalent by participants. Therefore these remaining orienting responses should not necessarily be dismissed as noise or false positives. Chances are that unconscious UX related events are indeed captured by the GSR sensor, which is also supported by Ward and Marsden in one of the early HCI studies examining the efficacy of physiological measures [63]. Yet, it is not trivial to identify such unconscious events as this requires inclusion of more data sources, e.g. observations and various types of automatically logged clickstream-like data.

Dealing with such noise by including more data sources have previously been done using relatively sophisticated data analysis techniques based on machine intelligence classifiers. Classifiers like support vector machines or multilayer perceptrons are used to e.g. fuse signals from multiple data sources to classify emotions, see e.g. [12,61]. While such techniques can reduce the level of noise, they require a set of training data in order to match individual orienting response patterns. Machine intelligence allows the fusing of signals from multiple data sources, but is susceptible to the curse of dimensionality. This denotes the situation where the ratio of dimensions (e.g. data sources) to training data is so high that it results in an overly fitted model requiring more training data. The need for training data to support a model often grows exponentially with the number of dimensions to be included.

Thus, having to include various contextual differences further complicates the use of such sophisticated analysis techniques. Ganglbauer et al. [23] for instance applied GSR to gain initial insights on using physiological sensors to assess UX in a mobile context. They note that physical movements in mobile contexts causes peaks in GSR sensor signals, which were unrelated to participants' state of arousal. False positives caused by such movements could be filtered out by combining GSR data with data from e.g. an accelerometer, but there will be a plethora of other contextual dimensions to consider, hereby adding to the curse of dimensionality in analyzing data.

In relation to the noise caused by the context, it is also important to note that previous work within psychophysiology is based on highly stringent settings. These typically allow for lengthy baseline periods, control of temperature and humidity, using special conductivity gels on sensor electrodes and skin abrasion considerations, see [1] for more examples. While increasing accuracies and keeping noise levels to a minimum is commendable, this is not feasible to control in UX design and evaluation practices. Ward and Marsden argue for a need to study the use of physiological sensors in HCI without these tightly controlled constraints [63]. The approach of including multiple data sources and reducing noise through machine intelligence is challenging in practice. Such an approach requires the machine to train on elaborate training sets based on usage patterns from a plethora of contextual permutations using various technologies.

Implications for Research and Practice

Returning to the motivation of this paper, the software industry calls for a need to include more data driven recommendations based on emotional reactions, cf. [24]. Currently, emotional data is primarily gathered through questionnaires, but capturing emotional responses in real time, e.g. through physiological sensors, allows for a finer granularity in detecting positive or negative experiences. This allows designers to identify points of interest during an interaction sequence, which potentially leads to insights valuable for making design changes. Analyzing data from physiological sensors is, however, reported to be challenging [23,25,34] and designers may not possess the competences necessary to perform such analysis [22].

The essence of the above discussions is two-fold:

1) Non-specialists are able to analyze and interpret GSR sensor data through which they are able to detect ~60-80% of all valence markers made by test users within an interval of 3.5-11 seconds
2) The level of sensor noise (whether considered as false positives or responses to unconscious events) at the same latency interval is ~70-80%

This study demonstrates that it is practically feasible to use physiological sensors and analyze the data based on a basic approach of simply considering whether or not there are GSR peaks at particular moments in time. Furthermore, events where users experience UX related problems can be pinpointed within seconds. Given the key UX dimension of emotions [2,21,59], this data is foundational for understanding and assessing user experiences during actual interaction, e.g. in order to reduce the effect of memory biases [7]. Using physiological data also offers great potential of capturing experiences while users interact with technologies without the presence of evaluators, hereby increasing ecological validity. This is fully realizable as physiological sensors are now commonplace in smart watches, see e.g. [51].

However, qualitative insights are also needed in order for practitioners as well as researchers to understand experiences and to make informed decisions on what to redesign and how [3,49,55,56,62]. These insights include knowledge as to why orienting responses occur at particular moments in time [23]. To this end, Cued-Recall Debriefing (CRD) is one way of collecting qualitative insights based on physiological sensor data [9]. In CRD test users are asked to retrospectively comment on their experiences

based on a series of video clips showing their interactions with a particular technology. These video clips are chosen based on timestamps obtained from orienting responses, i.e. when sensor data suggest that users experience emotional reactions. Alternative methods include the Affective Diary [57] and UX Curve [35], which are based on Kahneman et al's Day Reconstruction Method [33]. In using those methods, study participants are asked to reflect on their experiences at the end of the day. Methods such as the Affective Diary and UX Curve could be supplemented by physiological data, which can provide study participants with further reminders of particular moments with increased arousal that occurred during the day. Since this study shows that non-specialists are able to make sense of sensor data to an extent beyond naïve guessing, such data may also be interpreted by study participants and not necessarily UX practitioners or researchers.

Findings are framed as having potential implications for *assessing* and *understanding* UX in practice and research. Yet, the discussion on detecting conscious and unconscious events is also relevant for the area of affective computing in which an interface is adapted to the emotional states of users in real time. From the application area of recommender systems it has been shown that transparency as to why recommendations are given increases user understanding and system acceptance [13]. This leans well against real time adaptations of an interaction design, which should occur at points in time where users are conscious about their experiences, i.e. the reasons for changing the design are transparent. Findings from this study suggest that such adaptations should occur within a window of 3.5 to 11 seconds in order to capture the majority of events related to conscious user experiences. At least this could potentially be transferred to contexts similar to those of this study.

Limitations

This section discusses limitations of this study related to the experimental setting and the use of valence markers as the ground truth to assess accuracy and noise.

Setting is Relatively Controlled

This study deals with the constraint of not having highly specialized knowledge in working with physiological sensors. While this is step towards understanding the feasibility of using physiological sensors in practice, this study is still limited in relation to using sensors in controlled settings.

Ward and Marsden criticized previous work in psychophysiology for being "*observed in stringently controlled experimental situations using pure distinct stimuli, with other possible confounding sources of variability held constant*" [63]. Findings from such studies do not necessarily translate well into UX practices occurring under less tightly controlled settings. This is why this study was designed for measuring GSR data during interaction with a real system and not to use distinct stimuli such as IAPS [4] or GAPED [14]. That said, test users were interacting with the system in a lab setting with reduced environmental interferences, which could otherwise have impacted orienting responses registered by the GSR sensor. This is particularly critical to consider when studying sensor data in relation to systems in mobile contexts, see e.g. Ganglbauer et al.'s study [23].

The system in this study is designed for use in more static settings, e.g. using a desktop or laptop in an office environment. This increases validity given that findings are transferred to systems aimed for use in similar contexts.

Using Self-Reporting as the Ground Truth

The accuracy measure used in this study is based on the extent of overlap between orienting responses registered by the GSR sensor and valence markers. The valence markers were based on subjective data. These subjective markers are based on conscious acts by the users. Fairclough argues that physiological sensors respond to both conscious as well as unconscious processes [20], and discusses several inherent challenges in using self-reporting of emotions:

- Self-reports may interfere with the target behavior
- Artifacts in the physiological data may occur as participants have to do physical movements
- Physiological sensitivity is blunted by only studying correspondence with psychological states that are consciously reported

Self-reporting may interfere with participant behavior, which is a general challenge when conducting controlled studies. See the classical work by Orne [48] and a more recent study within HCI [15] for more lengthy discussions on demand characteristics. Self-reporting in this study was based on the valence method presented in [10], and is arguably less obtrusive and straining than filling in the SAM questionnaire during interaction, which has been done in related HCI studies (e.g. [27,39–41,59]). Also, we have previously used the same statistics website for another study, cf. [6]. In that study we gave participants the same task but they were not required to self-report emotions during interaction. In terms of behavior, we see comparable task completion times and rates between the previous and current studies. Thus, it cannot be dismissed that self-reporting influenced participant behavior. However, the valence method is arguably less straining than SAM.

Artifacts in the sensor data typically occur if participants move physically, which leads to unrealistic peaks in the GSR data. This was filtered out using a simple algorithm before the non-specialists analyzed the data by excluding abrupt changes in skin conductivity level.

The case of blunting physiological sensitivity by considering conscious events only will lead to increases in noise. Therefore, several GSR orienting responses may not be false positives, but rather reflect unconscious events. But the extent of this is unclear. See the above discussion on noise for the challenges for dealing with such bluntness in practice.

So, even if there are several limitations in using self-reports as the ground truth, this is still reported to be the best viable option for measuring accuracy of physiological sensors: *"Despite these disadvantages, subjective self-reports represent the best available approximation of the private experience of the individual"* [20].

Using Non-Specialist Study Participants

The study compared the performance of novice UX designers in analyzing GSR sensor data to that of naïve guessing. Another relevant direction would be to also include expert designers having more experience in analyzing physiological data. However, the merit of this study lies in illustrating a worst case scenario. A key finding is that non-specialists were able to obtain 60-80% accuracy within time intervals of relatively few seconds. Arguably, experts would be better at reducing the noise levels in terms of filtering out insignificant peaks in e.g. GSR data during analysis.

In terms of participant expertise, it would also be relevant to study the feasibility of including participants with even less expertise than those employed in the current study. For instance when conducting HCI studies to understand users' activities outside the confinements of the laboratory. As noted above, physiological data may also be interpreted by study participants, who could apply physiological data to assist in daily reflections.

Future Directions

This study extends previous work by examining the feasibility of using physiological sensors to assess UX in practice. This was done by using a more lenient analysis technique of data gathered from a less restricted experimental setting than related work. Although fewer constraints were posed on the setting, this was still relatively controlled. Findings from this study should be considered as a proof-of-concept allowing for physiological sensor data to go further into natural settings. It is recommended to follow the recent trend in HCI research of studying UX and interaction design "in-the-wild" [52] and to examine sensor feasibility in such uncontrolled settings. Using physiological data from e.g. smart wearable sensors would be highly beneficial as data sources to supplement e.g. ethnographic methods. Using physiological data in conjunction with methods such as Cued-Recall Debriefing, the Affective Diary or UX Curve could provide researchers with real time data on experiences, and it could help study participants to better reflect on and report daily experiences.

Physiological sensors seem to be more sensitive than "numb" in revealing *potential* UX related events during interaction. This study indicates that emphasizing a simple analysis approach enables non-specialists to identify most events within GSR sensor data. Thus, the prevalent challenge in using GSR data (and likely data from other sensors) is not to detect points of interest in relation to UX, but rather to filter away noise in the form of false positives. The main challenge in relation to this is deciding when orienting responses are indeed false positives and when they reflect unconscious events.

CONCLUSION

This study revolved around the emerging trend of using real time physiological sensor data to measure and understand emotions in relation to interactive user experiences. Related work suggests that UX practitioners may not possess the specialized knowledge required to analyze sensor data. Also, there is a need to understand the extent to which physiological sensors can detect UX related events within an HCI context. The following question was examined with an emphasis on studying non-specialists' abilities to analyze sensor data: *How accurate are physiological sensors in detecting emotional reactions related to specific UX events during interaction?*

A controlled study was conducted with 21 test users wearing a GSR sensor while interacting with a web application. Test users also subjectively marked when UX related events occurred during interaction (this was the ground truth). The level of accuracy with which GSR data could pinpoint these subjectively marked events was studied. To this end, the same 21 participants analyzed the 21 GSR data sets obtained through the user tests (a total of 441 analysis data sets). Analyses were done using a simple approach to support the lack of specialized knowledge in interpreting sensor data. Study participants had no previous experience in using or analyzing sensor data.

The study demonstrates that it is feasible to let non-specialists analyze physiological data as they uncovered 60-80% of all UX related events on average. Furthermore, these events could be pinpointed within a latency interval of 3.5 – 11 seconds. Since this study shows that non-specialists are able to make sense of physiological data to an extent significantly beyond naïve guessing, sensor data can be used to source interactive user experiences during real-time use. Note that this at least seems to apply for a use case in which participants were asked to search for a specific piece of information in a web application. This may not represent UX related events within other types of systems.

Regarding future work, the results on practical feasibility should be considered as a proof-of-concept allowing for physiological sensor data to be utilized for studying UX in natural settings. In this regard it would be relevant to follow the recent trend in HCI research of studying UX and interaction design "in-the-wild". Using physiological data from e.g. smart wearable sensors would be highly beneficial data sources to complement e.g. established ethnographic methods. While findings on accuracy showed promise, a considerable amount of "noise" or false positives was also identified, i.e. sensor data suggested that more UX related events occurred than those marked by the test users. An avenue for future work would be to further study the noise and the extent to which this noise actually represents false positives or unconscious events related to user experiences.

REFERENCES

1. John L. Andreassi. 2000. *Psychophysiology: Human Behavior and Physiological Response*. Lawrence Erlbaum, Mahwah.
2. Javier A. Bargas-Avila and Kasper Hornbæk. 2011. Old wine in new bottles or novel challenges. *Proc. CHI*, ACM, 2689–2698. http://doi.org/10.1145/1978942.1979336
3. Nigel Bevan. 2008. Classifying and Selecting UX and Usability Measures. In *Proc. Int. WS on Valid Useful User Experience Measurement (VUUM)*. IRIT, Toulouse. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.177.7123&rep=rep1&type=pdf#page=62
4. Margaret M. Bradley and Peter J. Lang. 2007. The International Affective Picture System (IAPS) in the study of emotion and attention. In *Handbook of Emotion Elicitation and Assessment*, James A. Coan and John J. B. Allen (eds.). Oxford University Press, 29–46.
5. Margaret M Bradley and Peter J Lang. 1994. Measuring emotion: The self-assessment manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry* 25, 1: 49–59. http://doi.org/http://dx.doi.org/10.1016/0005-7916(94)90063-9
6. A. Bruun and J. Stage. 2015. *An empirical study of the effects of three think-aloud protocols on identification of usability problems*. http://doi.org/10.1007/978-3-319-22668-2_14
7. Anders Bruun and Simon Ahm. 2015. Mind the Gap! Comparing Retrospective and Concurrent Ratings of Emotion in User Experience Evaluation. *15th IFIP TC13 Conference on Human-Computer Interaction (INTERACT)*, Springer-Verlag. http://doi.org/http://dx.doi.org/10.1007/978-3-319-22701-6_17
8. Anders Bruun, Effie Lai-Chong Law, Matthias Heintz, and Lana H A Alkly. 2016. Understanding the Relationship Between Frustration and the Severity of Usability Problems: What Can Psychophysiological Data (Not) Tell Us? *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, ACM, 3975–3987. http://doi.org/10.1145/2858036.2858511
9. Anders Bruun, Effie Lai-Chong Law, Matthias Heintz, and Poul Svante Eriksen. 2016. Asserting Real-Time Emotions through Cued-Recall: Is it Valid? *Proceedings of the 9th Nordic Conference on Computer-Human Interaction (NordiCHI)*, ACM. http://doi.org/http://dx.doi.org/10.1145/2971485.2971516
10. Michael Burmester, Marcus Mast, Kilian Jäger, and Hendrik Homans. 2010. Valence Method for Formative Evaluation of User Experience. *Proc. DIS*, ACM, 364–367. http://doi.org/10.1145/1858171.1858239
11. J T Cacioppo and W L Gardner. 1999. Emotion. *Annual review of psychology* 50: 191–214. http://doi.org/10.1146/annurev.psych.50.1.191
12. Rafael A Calvo, Iain Brown, and Steve Scheding. 2009. Effect of Experimental Factors on the Recognition of Affective Mental States through Physiological Measures. In *AI 2009: Advances in Artificial Intelligence: 22nd Australasian Joint Conference, Melbourne, Australia, December 1-4, 2009. Proceedings*, Ann Nicholson and Xiaodong Li (eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 62–70. http://doi.org/10.1007/978-3-642-10439-8_7
13. Henriette Cramer, Vanessa Evers, Satyan Ramlal, et al. 2008. The effects of transparency on trust in and acceptance of a content-based art recommender. *User Modeling and User-Adapted Interaction* 18, 5: 455. http://doi.org/10.1007/s11257-008-9051-3
14. Elise S Dan-Glauser and Klaus R Scherer. 2011. The Geneva affective picture database (GAPED): a new 730-picture database focusing on valence and normative significance. *Behavior research methods* 43, 2: 468–477. http://doi.org/10.3758/s13428-011-0064-1
15. Nicola Dell, Vidya Vaidyanathan, Indrani Medhi, Edward Cutrell, and William Thies. 2012. "Yours is Better!": Participant Response Bias in HCI. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM, 1321–1330. http://doi.org/10.1145/2207676.2208589
16. Pieter M.A. Desmet, Kees Overbeeke, and Stefan Tax. 2001. Designing products with added emotional value; development and application of an approach for research through design. *The Design Journal* 4, 1: 32–47. http://doi.org/http://dx.doi.org/10.2752/146069201789378496
17. Panteleimon Ekkekakis. 2013. *The Measurement of Affect, Mood, and Emotion*. Cambridge University Press, Cambridge.
18. Paul Ekman. 1972. Universals and Cultural Differences in Facial Expressions of Emotion. *Nebraska Symposium on Motivation 19*, University of Nebraska Press, 207–282.
19. Paul Ekman. 1978. FACS - Facial Action Coding System. Retrieved August 1, 2016 from http://www.cs.cmu.edu/~face/facs.htm
20. Stephen H. Fairclough. 2009. Fundamentals of

physiological computing. *Interacting with Computers* 21, 1–2: 133–145. http://doi.org/10.1016/j.intcom.2008.10.011

21. Jodi Forlizzi and Katja Battarbee. 2004. Understanding Experience in Interactive Systems. *Proc. DIS*, ACM, 261–268. http://doi.org/10.1145/1013115.1013152

22. Malin Forne. 2012. Physiology as a Tool for UX and Usability Testing.

23. Eva Ganglbauer, Stephanie Deutsch, and Manfred Tscheligi. 2009. Applying psychophysiological methods for measuring user experience: possibilities, challenges and feasibility. *User Experience Evaluation Methods in Product Development (UXEM)*. http://doi.org/10.1.1.189.3410

24. Vanessa Georges, François Courtemanche, Sylvain Sénécal, Pierre-Majorique Léger, Lennart Nacke, and Marc Fredette. 2017. The Evaluation of a Physiological Data Visualization Toolkit for UX Practitioners: Challenges and Opportunities. *Workshop on Strategies and Best Practices for Designing, Evaluating and Sharing Technical HCI Toolkits (HCI Tools)*, ACM Press.

25. Vanessa Georges, François Courtemanche, Sylvain Sénécal, Pierre-Majorique Léger, Lennart Nacke, and Romain Pourchon. 2017. The Adoption of Physiological Measures as an Evaluation Tool in UX. In *HCI in Business, Government and Organizations. Interacting with Information Systems: 4th International Conference, HCIBGO 2017, Held as Part of HCI International 2017, Vancouver, BC, Canada, July 9-14, 2017, Proceedings, Part I*, Fiona Fui-Hoon Nah and Chuan-Hoo Tan (eds.). Springer International Publishing, Cham, 90–98. http://doi.org/10.1007/978-3-319-58481-2_8

26. Rishabh Gupta, Khalil ur Rehman Laghari, and Tiago H. Falk. 2016. Relevance vector classifier decision fusion and EEG graph-theoretic features for automatic affective state characterization. 174: 875–884. http://doi.org/10.1016/j.neucom.2015.09.085

27. Marc Hassenzahl and Daniel Ullrich. 2007. To Do or Not to Do: Differences in User Experience and Retrospective Judgments Depending on the Presence or Absence of Instrumental Goals. *Interact. Comput.* 19, 4: 429–437. http://doi.org/10.1016/j.intcom.2007.05.001

28. J Healey. 2011. GSR Sock: A New e-Textile Sensor Prototype. *2011 15th Annual International Symposium on Wearable Computers*, 113–114. http://doi.org/10.1109/ISWC.2011.36

29. J A Healey and R W Picard. 2005. Detecting stress during real-world driving tasks using physiological sensors. *IEEE Transactions on Intelligent Transportation Systems* 6, 2: 156–166. http://doi.org/10.1109/TITS.2005.848368

30. William James. 1884. What Is An Emotion? *Mind* os-IX, 34: 188–205. http://doi.org/10.1093/mind/os-IX.34.188

31. E H Jang, B J Park, S H Kim, M A Chung, M S Park, and J H Sohn. 2014. Emotion classification based on bio-signals emotion recognition using machine learning algorithms. *2014 International Conference on Information Science, Electronics and Electrical Engineering*, 1373–1376. http://doi.org/10.1109/InfoSEEE.2014.6946144

32. Daniel Kahneman, Barbara L Fredrickson, Charles A Schreiber, and Donald A Redelmeier. 1993. When More Pain Is Preferred to Less: Adding a Better End. *Psychological Science* 4, 6: 401–405. http://doi.org/10.2307/40062570

33. Daniel Kahneman, Alan B Krueger, David A Schkade, Norbert Schwarz, and Arthur A Stone. 2004. A Survey Method for Characterizing Daily Life Experience: The Day Reconstruction Method. *Science* 306, 5702: 1776 LP-1780. Retrieved from http://science.sciencemag.org/content/306/5702/1776.abstract

34. Mathias Kivikangas, Inger Ekman, Guillaume Chanel, et al. 2010. Review on Psychophysiological Methods in Game Research. *Nordic Digital Games Research Association*.

35. Sari Kujala, Virpi Roto, Kaisa Väänänen-Vainio-Mattila, Evangelos Karapanos, and Arto Sinnelä. 2011. {UX} Curve: A method for evaluating long-term user experience. *Interacting with Computers* 23, 5: 473–483. http://doi.org/http://dx.doi.org/10.1016/j.intcom.2011.06.005

36. Peter J. Lang. 1995. The emotion probe: Studies of motivation and attention. *American Psychologist* 50, 5: 372–385. http://doi.org/10.1037/0003-066X.50.5.372

37. PETER J. LANG, MARK K. GREENWALD, MARGARET M. BRADLEY, and ALFONS O. HAMM. 1993. Looking at pictures: Affective, facial, visceral, and behavioral reactions. *Psychophysiology* 30, 3: 261–273. http://doi.org/10.1111/j.1469-8986.1993.tb03352.x

38. Alexandros Liapis, Christos Katsanos, Dimitris Sotiropoulos, Michalis Xenos, and Nikos Karousos. 2015. Recognizing emotions in human computer interaction: Studying stress using skin conductance. Springer Verlag, 255–262. http://doi.org/10.1007/978-3-319-22701-6_18

39. Sascha Mahlke and Gitte Lindgaard. 2007. Emotional Experiences and Quality Perceptions of Interactive Products. In *Human-Computer Interaction. Interaction Design and Usability SE - 19*, JulieA. Jacko (ed.). Springer Berlin Heidelberg, 164–173. http://doi.org/10.1007/978-3-540-73105-4_19

40. Sascha Mahlke, Michael Minge, and Manfred Thüring. 2006. Measuring Multiple Components of Emotions in Interactive Contexts. *CHI EA*, ACM, 1061–1066. http://doi.org/10.1145/1125451.1125653

41. Sascha Mahlke and Manfred Thüring. 2007. Studying Antecedents of Emotional Experiences in Interactive Contexts. *Proc. CHI*, ACM, 915–918. http://doi.org/10.1145/1240624.1240762

42. Regan L Mandryk, M Stella Atkins, and Kori M Inkpen. 2006. A Continuous and Objective Evaluation of Emotional Experience with Interactive Play Environments. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM, 1027–1036. http://doi.org/10.1145/1124772.1124926

43. Regan L Mandryk, Kori M Inkpen, and Thomas W Calvert. 2006. Using psychophysiological techniques to measure user experience with entertainment technologies. *Behaviour and Information Technology* 25, 2: 141–158. http://doi.org/10.1080/01449290500331156

44. Talya Miron-Shatz, Arthur Stone, and Daniel Kahneman. 2009. Memories of Yesterday's Emotions: Does the Valence of Experience Affect the Memory-Experience Gap? 9, 6: 885–891. http://doi.org/10.1037/a0017823

45. Agnes Moors, Phoebe C Ellsworth, Klaus R Scherer, and Nico H Frijda. 2013. Appraisal Theories of Emotion: State of the Art and Future Development. *Emotion Review* 5, 2: 119–124. http://doi.org/10.1177/1754073912468165

46. Lennart Erik Nacke, Michael Kalyn, Calvin Lough, and Regan Lee Mandryk. 2011. Biofeedback Game Design: Using Direct and Indirect Physiological Control to Enhance Game Interaction. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM, 103–112. http://doi.org/10.1145/1978942.1978958

47. Domen Novak, Matjaž Mihelj, and Marko Munih. 2012. Dual-task performance in multimodal human-computer interaction: a psychophysiological perspective. *Multimedia Tools and Applications* 56, 3: 553–567. http://doi.org/10.1007/s11042-010-0619-7

48. Martin T Orne. 1962. On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. *American Psychologist* 17, 11: 776–783. http://doi.org/10.1037/h0043424

49. Byungho Park. 2009. Psychophysiology as a Tool for HCI Research: Promises and Pitfalls. In *Human-Computer Interaction. New Trends: 13th International Conference, HCI International 2009, San Diego, CA, USA, July 19-24, 2009, Proceedings, Part I*, Julie A Jacko (ed.). Springer Berlin Heidelberg, Berlin, Heidelberg, 141–148. http://doi.org/10.1007/978-3-642-02574-7_16

50. Rosalind W. Picard. 1997. *Affective Computing*. MIT Press, Cambridge.

51. Reza Rawassizadeh, Blaine A Price, and Marian Petre. 2014. Wearables: Has the Age of Smartwatches Finally Arrived? *Commun. ACM* 58, 1: 45–47. http://doi.org/10.1145/2629633

52. Yvonne Rogers and Paul Marshall. 2017. Research in the Wild. *Synthesis Lectures on Human-Centered Informatics* 10, 3: i-97. http://doi.org/10.2200/S00764ED1V01Y201703HCI037

53. James A Russell. 1980. A circumplex model of affect. *Journal of Personality and Social Psychology* 39, 6: 1161–1178. http://doi.org/10.1037/h0077714

54. Klaus R Scherer. 2005. What are emotions? And how can they be measured? *Social Science Information* 44, 4: 695–729. http://doi.org/10.1177/0539018405058216

55. N Sadat Shami, Jeffrey T Hancock, Christian Peter, Michael Muller, and Regan Mandryk. 2008. Measuring Affect in Hci: Going Beyond the Individual. *CHI '08 Extended Abstracts on Human Factors in Computing Systems*, ACM, 3901–3904. http://doi.org/10.1145/1358628.1358952

56. Mark Springett. 2008. Assessing user experiences within interaction: experience as a qualitative state and experience as a causal event. In *Proc. Int. WS on Valid Useful User Experience Measurement (VUUM)*. IRIT, Toulouse. Retrieved from http://eprints.mdx.ac.uk/2140/

57. Anna Ståhl, Kristina Höök, Martin Svensson, Alex S Taylor, and Marco Combetto. 2009. Experiencing the Affective Diary. *Personal Ubiquitous Comput.* 13, 5: 365–378. http://doi.org/10.1007/s00779-008-0202-7

58. Robert M. Stern, William J. Ray, and Karen S. Quigley. 2000. *Psychophysiological Recording*. Oxford University Press, New York. http://doi.org/10.1093/acprof:oso/9780195113594.001.0001

59. Manfred Thüring and Sascha Mahlke. 2007.

Usability, aesthetics and emotions in human–technology interaction. *Int. J. Psychology* 42, 4: 253–264. http://doi.org/10.1080/00207590701396674

60. Erin Treacy Solovey, Daniel Afergan, Evan M Peck, Samuel W Hincks, and Robert J K Jacob. 2015. Designing Implicit Interfaces for Physiological Computing: Guidelines and Lessons Learned Using fNIRS. *ACM Trans. Comput.-Hum. Interact.* 21, 6: 35:1--35:27. http://doi.org/10.1145/2687926

61. Gyanendra K. Verma and Uma Shanker Tiwary. 2014. Multimodal fusion framework: A multiresolution approach for emotion classification and recognition from physiological signals. *102*, 162–172. http://doi.org/10.1016/j.neuroimage.2013.11.007

62. Arnold P O S Vermeeren, Effie Lai-Chong Law, Virpi Roto, Marianna Obrist, Jettie Hoonhout, and Kaisa Väänänen-Vainio-Mattila. 2010. User Experience Evaluation Methods: Current State and Development Needs. *Proceedings of the 6th Nordic Conference on Human-Computer Interaction: Extending Boundaries*, ACM, 521–530. http://doi.org/10.1145/1868914.1868973

63. R.D Ward and P.H Marsden. 2003. Physiological responses to different WEB page designs. *Int. J. Human-Computer Studies* 59, 1–2: 199–212. http://doi.org/10.1016/S1071-5819(03)00019-3

64. Gillian M Wilson. 2001. Psychophysiological Indicators of the Impact of Media Quality on Users. *CHI '01 Extended Abstracts on Human Factors in Computing Systems*, ACM, 95–96. http://doi.org/10.1145/634067.634125

65. Lin Yao, Yanfang Liu, Wen Li, et al. 2014. Using Physiological Measures to Evaluate User Experience of Mobile Applications. In *Engineering Psychology and Cognitive Ergonomics SE - 31*, Don Harris (ed.). Springer International Publishing, 301–310. http://doi.org/10.1007/978-3-319-07515-0_31