Understanding the Relationship between Frustration and the Severity of Usability Problems: What can Psychophysiological Data (Not) Tell Us?

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

Effie Lai-Chong Law
University of Leicester
Leicester, UK
lcl9@le.ac.uk

Matthias Heintz
University of Leicester
Leicester, UK
mmh21@le.ac.uk

Lana H.A. Alkly
University of Leicester
Leicester, UK
lana.alkly@yahoo.com

ABSTRACT
Frustration is used as a criterion for identifying usability problems (UPs) and for rating their severity in a few of the existing severity scales, but it is not operationalized. No research has systematically examined how frustration varies with the severity of UPs. We aimed to address these issues with a hybrid approach, using Self-Assessment Manikin, comments elicited with Cued-Recall Debrief, galvanic skin responses (GSR) and gaze data. Two empirical studies involving a search task with a website known to have UPs were conducted to substantiate findings and improve on the methodological framework, which could facilitate usability evaluation practice. Results showed no correlation between GSR peaks and severity ratings, but GSR peaks were correlated with frustration scores – a metric we developed. The Peak-End rule was partially verified. The problematic evaluator effect was the limitation as it confounded the severity ratings of UPs. Future work is aimed to control this effect and to develop a multifaceted severity scale.

Author Keywords
Psychophysiological; GSR; Emotion; Frustration; Severity; Usability problem; Evaluator effect; Cued-Recall Debrief;

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

INTRODUCTION
In the literature on usability research and practice published in the 1990s, the severity of a usability problem (UP) is primarily defined in terms of cognitive and performance-based impacts that the UP exerts on users or on developers (e.g., [17, 34, 60, 77]). Levels of UP severity are typically defined as the extent to which a user or a developer needs to spend additional time and effort to achieve a task given or to fix the UP identified. In most of the severity scales created more than a decade ago, an experiential criterion is rarely addressed. Even when included, it is just briefly mentioned, for instance, in Jeffries's [34] three-level severity scheme, an affective term is only stated in one of the three levels: Level 2 "Create significant delay and *frustration*" (our emphasis). Similarly, in [73], frustration is mentioned in the serious level, but not in the cosmetic or critical level.

Since the shift of focus from usability to user experience (UX) in the field of HCI at the turn of millennium, very few attempts have been made to revise UP severity scales in the wake of this new emphasis. Hassenzahl [27] proposed a relevant model of judgment-driven and data-driven UP severity estimates (cf. [50]) in which the notion 'psychological cost' instantiated as human stress is included. However, the model is not formalized as a severity scale. Meanwhile, while some severity scales have been shared informally in social media (e.g., [4]), all remain largely non-experiential except Sauro's [70]. Accordingly, the severity (minor, moderate, critical) of a UP increases with the degree of irritation (slight, moderate, extreme) it causes in users. But this emotion-oriented criterion seems secondary to the performance-based one in the two-part definition of each of the three levels (e.g., "3 = critical, leads to task failure or causes user extreme irritation") [70]. Like the previous severity scales, no operationalization of irritation is given. Consequently, the degree of irritation is primarily based on evaluators' judgment of users' verbal as well as non-verbal behaviours.

According to [9, 31], there are at least three different systems for measuring emotional responses – affective (self) reports, physiological reactivity, and observable behaviours. UPs occur when users are interacting with a system. In evaluations users are normally not interrupted to complete an affective report when a UP happens, and measuring physiological data is not yet a commonplace practice. Hence, the main source of data for estimating the emotional response (or 'psychologist cost' [27]) of a UP is users' behaviours from real-time observations and post-test video analyses, if available. As such a cost plays a critical role in influencing users' acceptance of the system and their sustained motivation in using it [27], we argue that a more reliable, objective assessment is necessary.

CHI'16, May 07-12, 2016, San Jose, CA, USA
 ISBN 978-1-4503-3362-7/16/05…$15.00
DOI: http://dx.doi.org/10.1145/2858036.2858511

Feeling frustrated is a negative affective response that users tend to have when experiencing UPs. Among the list of seven criteria for UP identification proposed by [35], two are related to users' emotions, expressing *surprise* or some *negative affect*; frustration is not explicitly mentioned. According to Russell's [68] circumplex model of affect, which has been applied to study a variety of phenomena in the field of HCI, including usability evaluation (e.g., [51, 72,78]), frustration is positioned in the quadrant categorized as high arousal (activated) and negative valence (unpleasant). Usability evaluation feedback in terms of subjective self-reports and objective measures indicate that users tend to expend more cognitive as well as emotional resources in dealing with UPs than in interacting with a UP-free system. From both the theoretical and empirical perspectives, we hypothesize that users are more aroused with negative valence (frustration) when the severity of the UP they experience is higher.

The significance of understanding the relationship between frustration and the severity of UPs lies in the implication for prioritizing UPs to be fixed [27, 29]. Such prioritization is influenced by UP frequencies and severity ratings (the review in [69]). However, as discussed earlier, UP severity is mainly determined by non-emotional criteria; we query whether such prioritization is inherently fallible as the experiential aspect of UP is not addressed systematically. Furthermore, if users' overall evaluation of an interactive system is dictated by the most severe UP evoking the strongest negative affective response, then the highest fixing priority should be assigned to such a UP. Hence, we have been motivated to explore two related questions: How does the strength of a user's *momentary* affective response vary with the severity of UPs experienced in the course of completing a task with a system? To what extent do such momentary affective responses influence the user's overall UX evaluation of the system? To answer them, we resort to the use of psychophysiological measures.

With the improved accuracy, ease of use, non-obtrusiveness and affordability, certain psychophysiological measuring devices have increasingly been deployed in HCI research. Particularly widespread is the use of galvanic skin response (GSR), which is well-recognized as a reliable tool for objectively measuring emotional arousal (e.g., [1,12]). While other measures such as heart beat rate, blood pressure, EEG, and EMG, could help understand affective responses to UPs, we have opted to focus on GSR. Our rationale is to develop a parsimonious framework for integrating users' self-reports with psychophysiological data by improving on the methodological approach used in [10] where the relationships between GSR, Cued-Recall Debrief (CRD) [61] and Self-Assessment Manikin (SAM) ([9,56]) data were examined.

Specifically, we have moved beyond [10] in three major ways: Quantifying frustration with SAM ratings; Understanding the relation between UP-related frustration and UP severity with GSR measures and SAM ratings; Incorporating eye-tracking data to enrich the cue for CRD and to facilitate the interpretation of GSR data. Our hybrid approach of integrating self-reported and psycho-physiological data augments the established usability evaluation techniques to systematically address frustration induced by UPs. Our approach can also enhance the efficiency of video data analysis by marking segments of interest as indicated by psychophysiological data.

Overall, the research goals of our work are: (i) to substantiate the emotion-oriented criterion of UP severity scales by grounding it in an empirical understanding of affective responses to UPs in terms of GSR measures; and (ii) to consolidate the methodological approach for integrating GSR measures with users' self-reports. As a corollary, we examine the evaluator effect ([30,58,59]), which is inevitably a challenge for research studies like ours involving UP identification and classification.

RELATED WORK

Measuring Frustration

The emphasis on understanding users' emotional responses before, during and after interacting with a computer system demarcates UX from usability [66]. There have been debates about the measurability of emotions at the theoretical, methodological, and practical levels (e.g., [8, 45]). The question whether it is more viable to evaluate UX as a discrete affective state (cf. six basic emotions [19]) or as a vector in a two-dimensional space defined by arousal and valence [44, 68] remains contentious. Both perspectives face the inherent challenge: emotions are *multifaceted* and *ephemeral* [11]. The dispute is further complicated by the nuanced distinctions among the related notions: affect, emotion and mood ([18] for review). While not elaborating the related arguments (e.g., [53,62]) here as it entails a separate exposition, they have informed our adoption of the arousal-valence grid to analyse frustration as a salient emotional response to UPs with a hybrid use of objective and subjective evaluation approaches. We leniently use the terms affect and emotion interchangeably in this paper.

Frustration has been studied in the field of psychology since the 1930s, evolving from being a behavioural phenomenon to a cognition-emotion issue (cf. reviews in [6,46,71]). Frustration occurs when a need is not satisfied or a reward is not delivered within an expected period of time or not at all [3,48]. While frustration is a commonly used term to describe negative affect resulting from unpleasant interactions with computing technologies, it is not included in most of the severity scales except [70] where the term 'irritation' rather than frustration is used. Nonetheless, a handful of research studies have investigated 'computer user frustration' in the context of affective computing (e.g., [21,41,71]) and gaming [54] with different goals and approaches. For instance, [6] developed a frustration model with personal and situational factors, relying on users' self-reported affective responses and performance (e.g., time

lost) in diaries and questionnaires. [21] and [71] demonstrated the possibility of automatic detection of frustration and non-frustration episodes as a binary variable, using different psychophysiological measures and sophisticated computational methods (e.g., Hidden Markov Models and fuzzy logic), but they did not analyse the relationship between the extent of frustration and the severity level of the related causes. [71] characterised frustration as a multidimensional emotional state lasting multiple seconds and argued for the robustness of GSR as a measure of frustration, but they caveated the limitation that it does not measure the *entirety* of frustration.

Galvanic skin response (GSR) & Gaze

The ability of GSR to measure arousal and thus emotion is accounted by the fact that human skin can become momentarily a better or worse conductor of electricity, contingent on the perception of external (e.g., seeing a disgusting image) or internal stimuli (e.g., thinking of an anxiety-inducing situation). Utilizing the sweat glands in palms is a common and convenient means to measure GSR. The palmar sweat increases in response to higher arousal, resulting in stronger skin conductance, which is captured by a GSR measuring device and displayed as a continuous curve. Dynamic fluctuations of GSR data can be observed in real time. Wearing a GSR device can impede the hand movement and oral communication, given its sensitivity [23,54]; the use of think-aloud is thus not possible. Several studies [51,76,78] measured GSR data in usability tests and confirmed that the skin conductance changed when the users experienced stress or some undifferentiated negative affect during interaction. But none of them examined if the arousal level is associated with the severity of a problem.

The validity of GSR in measuring emotions can be threated by certain factors. More than a decade ago [76] identified three major issues, which remain unresolved. First, a compelling concern is individual differences in skin conductivity and personality traits that influence emotional reactions; expressive people generally have a peaked curve whereas reserved people have a flatter trace [22]. Hence, it is necessary to control for personality factors (NB: we used the standardized test Big Five [26] to control it) and to normalize GSR data for inter-individual comparisons. Second, substantial changes in GSR signals (i.e. peaks) suggest emotional responses of a person under scrutiny [1,14,36]. However, there is still a lack of standardized thresholds for the latency, duration and magnitude of responses to recognize certain peaks as having significant implications. [21] proposed a technique of detrending of the GSR signal by subtracting a 10-second time-varying sample mean, but the 10-second boundary seems arbitrary. Third, the issue of many-to-one: different emotional states can result in the same physical response. The arousal-valence grid may be a viable alternative, but the two dimensions are known to be not entirely orthogonal [11,44,53].

The above review implies the need to use multiple methods to maximize their respective strengths – the physiological approach allows moment-by-moment and non-disruptive measures; the self-reporting approach allows personal interpretations of cognitive and emotional reactions.

While GSR data are often taken in tandem with heart beat rate or blood pressure, their similar characteristics cannot provide researchers or participants with extra support to infer the implications of graphical or numeric data. We argue that eye-tracking data can be a better alternative. People fix their gaze on a stimulus, which may be perceived to be interesting or challenging [16] [63]. In viewing fixation-based visualizations such as gaze plots, heat map, and scan paths, the participant is enabled to recall their activities and associated feelings when the data have been captured. By the same token, the researcher is enabled to identify with higher confidence which stimuli might have triggered certain responses in the participant.

Peak-End-Rule

Rubin [67] argued that the human minds tend to remodel past episodes because of their inability to recall them reliably. [24] showed that a few significant events of an occasion dominated what people recounted it. Subsequent work [37,38,39] confirmed that the overall experience of an episode is highly correlated with the last event and the most intense (peak) event. This phenomenon is now widely known as the peak-end rule. It defies the rule of *temporal monotonicity* whereby the magnitudes of positive and negative moments of an episode are aggregated to give its overall sum of pleasure or pain. [15] confirmed the peak-end-rule concept based on their empirical study in evaluating pleasurable experiences. However, [40] found that the end event was more effective than the peak event in influencing the overall evaluation of an incident. In the field of HCI research, several studies [10,13,28,40]) have explored the applicability of the peak-end-rule in understanding interaction experiences. Results about the relative strength of the peak and the end event are mixed. These studies relied on subjective data (questionnaires, diaries) except [10] where physiological data were used.

Evaluator effect

The phenomenon of evaluator effect is well recognized since the related work first published in the late 1990s [33]. Accordingly, usability evaluators analysing the same usability test sessions identify markedly different sets of UPs. Hence, it is improbable to attain perfect reliability of discovering UPs [49]. This issue has been investigated by a series of studies known as Comparative Usability Evaluation (CUE) coordinated by [58] and revisited more recently [30] (see also General Discussion).

METHODS

We conducted two empirical studies, designated as Study1 and Study2, in sequence about six months apart in two HCI research labs: one in Aalborg and the other one in Leicester. Both studies were aimed to address the same set of research

goals and hypotheses by improving on the research protocol of [10] to investigate the issues about the severity of UPs. The major methodological difference between Study1 and Study2 is that eye-tracking data were collected in latter but not in the former. In reviewing the procedure of Study1, we reckoned that gaze data could not only serve as an extra pointer with which a researcher might confirm GSR peaks with higher confidence and accuracy but also as a stronger cue to enable participants to recall details of the events contributing to changes in arousal as indicated by GSR peaks. We do not aim to merge the data of the two studies.

Instruments

Website

The website *Statistic Denmark* (www.dst.dk) was selected as our evaluation target. An independent empirical study shows that the website has a number of UPs with different severity, enabling us to observe how participants would emotionally respond to different UPs. The original Danish version was used in Study1 whereas its English equivalent was used in Study2. Participants were asked to complete a main search task (similar for both studies), comprising three subtasks (the same for both studies) (Figure 1).

Study1: "*Your sister considers opening a restaurant in Vejen. How many hotels and restaurants were there in 2012, with one person employed?"*
Study2: "*Your brother considers opening a construction firm in Greve (a region in Denmark). How many construction firms were there in 2013 in Greve?*"

1) Find a page that gives an overview of topics on the website.
   Note the name of the page here (the heading):[ ]
2) Choose the topic that you believe leads to the correct answer. It is ok to browse through the topics briefly before you decide. Don't go in-depth with all topics.
   Note the name of the page here (the heading):[ ]
3) Go in-depth with the chosen topic. Can you find the answer?
   If YES: What's the answer? [ ]
   If NO: Try looking under a different topic until you find the answer. What's the answer? [ ]

Figure 1. Descriptions of the search tasks

GSR sensor

The wireless *Shimmer3* GSR sensor with the following specifications was used: Exosomatic skin conductance level/response (SCL/SCR) is measured with two electrodes producing a continuous current of 60μA attached to two fingers (10kΩ-4.7 MΩ; DC-15.9Hz). The Ag/AgCl electrodes used are non-polarizing and a low DC potential can reduce the counter-electromotive force or the risk of sensor drift. We used the default setup: 0-5 Hz for tonic measurements, 0.03-5Hz for phasic ones. In Study2, a desktop Tobii T-120 eye-tracker was used to capture participants' gaze data and facial expressions (Figure 2). In both studies, the experimenter observed the data stream in real time in a location not within the participant's sight.

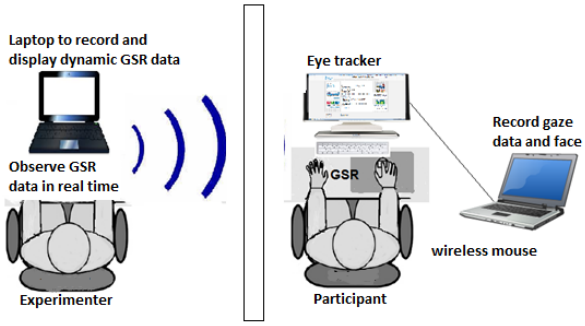

Figure 2. Experimental setup (the eye-tracker for Study2 only)

Self-Assessment Manikin (SAM)

Lang [43] (cf. [9, 56]) developed this pictorial, nonverbal instrument to assess affective responses to an object/event along three dimensions: *Pleasure (P)* is depicted by a scale ranging from a smiling, happy figure to a frowning unhappy one. *Activation (A)* (NB: to avoid confusion with the arousal measured as GSR, we rename this construct) is depicted as a scale ranging from a relaxed, sleepy figure to an excited, wide-eye one. *Dominance (D)* is depicted as a scale with an increasing image size; the larger the size, the stronger a respondent feels in control of the situation. A 9-point scale for each dimension is used (Figure 3).

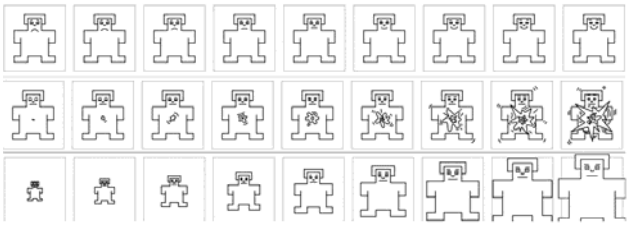

Figure 3. Three 9-point SAM scales – Pleasure (P) on the top, Activation (A) in the middle, Dominance (D) at the bottom

As mentioned in the above review that arousal and valence are not entirely independent and may influence each other [11,45,53], to address this issue of non-orthogonality and to provide a summative measure of frustration instead of inspecting three separate ratings, we propose a metric known as **frustration score** (*F*) by combining the ratings of *P*, *A* and *D* as follows:

$$F = A * [(10-P) + (10-D)]$$

The formula is based on the following assumptions: Frustration is primarily caused by UPs; Activation indicates the magnitude of Frustration; Frustration is inversely proportional to Pleasure and to Dominance. With a 9-point scale, we reverse the *P* and *D* ratings by subtracting them from 10. For a 5-point scale and different orientations of the *P*, *A* and *D* anchors, the formula should be adapted.

Eye-tracker

Earlier physiological studies demonstrated that an eye tracker could be a powerful tool to gain insights into people's cognitive processes in problem-solving such as information search in a website [16,52,63]. Fixation

duration, a common eye-tracking metric, can indicate the difficulty level of the information at which a user fixates. Based on the GSR peaks identified, a corresponding area of interest (AOI) can be derived by forwarding the peak time five seconds and back-warding five seconds to create some scenes or video segments.

Big-Five Personality Questionnaire

A 50-item questionnaire derived from [26] was used. It comprises five factors: Extraversion, Agreeableness, Conscientiousness, Emotional stability and Intellect/ Imagination. Each item was rated with a 5-point Likert scale from 'very inaccurate' to 'very accurate'.

Procedure

Cued-Recall Debrief (CRD) is a method based on situated recall. The method was developed by [61] to elicit emotional experiences while not interfering with participant behaviour in naturalistic settings. The overall approach is to provide cues that enhance participants' ability to recall specific emotions after an event has occurred. This is done by re-immersing participants through replay of several snippets of video recordings, each showing a specific episode of an entire event [61]. To foster re-immersion, it is crucial that video recordings resemble a first-person point of view. CRD essentially builds on retrospection, and several studies have validated the approach. In [61] it was found that CRD leads to considerably more detailed responses compared to retrospective ratings based on free recall. Furthermore, [5] found correlation between CRD ratings and real-time physiological measurements in an HCI context. Also, a more recent study applied CRD to elicit participants' emotions (cf. [25]). Thus, although CRD builds on retrospection, it has been shown to provide valid approximations of concurrent emotions. Additionally, it does so without causing interference during interaction.

In this study, we selected video clips on the basis of real-time GSR data where peaks were observed (Figure 4). Participants were presented one clip at a time and asked to provide a running commentary when viewing it and to complete the SAM scales (SAM_{CRD}) when the clip was finished. This process was repeated for all video clips identified for individual participants.

1. *Introduction and setup:* The experiments were conducted on an individual basis, involving one participant at one time. All Study1 test sessions were run by one experimenter and all Study2 ones were run by another one. A participant was directed to the room where she received a piece of paper with the description of the search task. The GSR sensor was then attached to her hand. After the participant had confirmed that she understood the task, the researcher started the sensor.

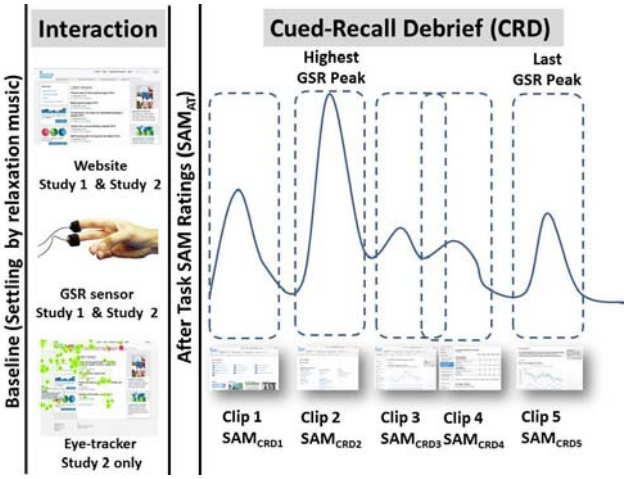

Figure 4. Selecting video clips based on GSR peaks.

2. *Creating a baseline:* Since the GSR sensor reacts on arousal we needed to identify the relaxed state of each participant, i.e. peaks of arousal are observed relative to a baseline [54,76]. This baseline was measured by showing a blank screen for the first 4 minutes, while playing a relaxing piece of music.
3. *Completing the task:* After the 4 minutes of relaxation, the user interface of the system appeared and the task could begin. A time limit of 15 minutes was imposed. The participant performed the task in silence without being asked to think aloud. Upon completing the task, she was asked to fill in the SAM scale (SAM_{AT}).
4. *Making verbal comment with CRD:* The experimenter inspected the GSR data to visually identify peaks. Three parameters are factored in our visual peak detection [7]:
   - Latency: From experiencing a UP to SCR: 1s-4s;
   - Rise time: From initial deflection to peak: 0.5s-5s. If rise time is less than 1s, the abrupt increase is unrealistic and the peak is ignored as artifactual;
   - Recovery: A UP-induced SCR lasts ~10s [21, 76], but more than 1 SCR can happen within a 10s window, caused by the same or different UPs. If the recovery limb of a wave declines to 63% of the initial SCL before the SCR and the wave rise again, then we count them as 2 peaks, else they are merged as one.

   We did not use phasic SCR amplitude, which ranges from a threshold of 0.05μS (micro-Siemens) to 8μS and varies much with context. Rather than using an arbitrary value, we aim to derive it from a larger dataset in future. Upon detecting a peak, the experimenter would then move backward 5s before and forward 5s after each peak point. This 10s window[1] allowed the latency between the perception of triggering stimuli and the manifestation of

[1] In the related literature, the time lapse between an onset of a stimulus and the expression of emotional response is inconsistent and seems arbitrary, ranging from 3s [76], 5s [54] to 10s [21].

emotional responses and also provided more contextual information for the participant to recall the situation. The participant was then shown this part of the video (the screencast of the website with the user's face superimposed in the lower corner; in case of Study2 the gaze plots were additionally displayed), and asked to freely describe her own thoughts as to what she might have reacted to, and how. If the participant was able to deduce a reaction, the peak was noted along with her description of the event. In other words, some high GSR peaks would not be further analysed if the participant could not relate them to any internal or external stimuli.

5. *Completing the Big-Five Personality Questionnaire:* The participant was asked to complete the paper-based questionnaire after the task had been done.

Participants

Study 1

A sample of 36 MSc students in Computer Science was recruited. All had experience in usability evaluation. They were randomly assigned to one of the two roles: a) *Participants* (N=21, Age: mean=23.4, SD=2.67) carrying out individually a search task with the website and subsequent evaluative activities. Five of them have never used the website under evaluation, four have used it on a monthly basis, and 12 on a yearly basis. None of them have ever searched for the specific information required in the task. There were no significant differences in any of the five factors of the Big-Five questionnaire, suggesting that personality traits are comparable between participants. In theory they could all have been (equally) emotionally sensitive. b) *Evaluators* (N=15, Age: mean = 22.7, SD=2.82) analysing video-clips to identify UPs and rate their severity, first individually and then collaboratively.

Study 2

A sample of 20 postgraduate students in different disciplines was recruited. They were approached randomly on campus without any specific selection criterion. Their participation was voluntary without any reward. Their average age was 26.5 years (SD = 5.41). None of them have used the website before or received any training in usability evaluation. No significant differences in any of the five factors of the Big-Five questionnaire were found.

Hypotheses (H)

The severity of a UP, the level of arousal, and the level of frustration are constructs that are measured as severity ratings, GSR peaks, and frustration scores, respectively.

H1: The higher the severity of a UP, the higher the level of arousal experienced by a user; the correlation between the severity ratings of UPs and GSR peaks is significant.

H2: The level of arousal increases with the level of frustration a user experiences when dealing with UPs; the correlation between GSR peaks and frustration scores is significant

H3: The level of frustration increases with the severity of a UP; the correlation between the severity ratings of UPs and frustration scores is significant.

H4: The level of arousal increases with the level of *positive* affect a user experiences when interacting with the website; the correlation between GSR peaks and SAM_{CRD} ratings is significant;

H5: The highest and last peaks of arousal influence the overall user experience evaluation of the website after the task; the correlations between the highest peak and SAM_{AT} ratings and between the last peak and SAM_{AT} are significant.

Note that H1, H2, H3 and H5 are evaluated in both Study1 and Study2. H4 is evaluated in Study2 only.

DATA ANALYSIS

Normalization of GSR data

The GSR data were smoothed by excluding large abrupt SCL changes (i.e., 5 standard deviations outside the group SCL mean occurred in < 1s) apparently caused by artifacts (e.g., moving hand). As individual differences in physiological data are known to be immense [12], it is necessary to normalize the GSR data to perform a group analysis. In reviewing the literature, different computational methods for normalization have been applied in HCI research studies (e.g., [14, 51, 54, 76, 78]). We adopted the following method as it is logically sound:

Normalized $GSR_{(i)}$:
$= [(GSR_{(i)} - GSR_{min}) / (GSR_{max} - GSR_{min})] * 100$

where GSR_{min} and GSR_{max} are global minimum and maximum GSR of all the data collected for individual participants, from the start of the baseline period till the end of the test session [54]. Arguably GSR_{min} is likely to be close to the average of the signals during the baseline period when a participant is assumed to be in a relaxed and emotionally neutral state.

UP Identification and Severity Ratings

Study1 aimed to demonstrate the practicality of training novice evaluators to identify and rate the severity of UPs with our hybrid approach based on GSR, SAM and CRD. In Study2, only expert evaluators were involved.

Study1

All video data together with CRD transcripts were analysed individually. Each evaluator made a list of the UPs detected along with severity ratings. To enhance the reliability of UP identification, all the evaluators were asked to apply predefined problem criteria based on [35] and were asked to describe each UP using a structured template [32] to facilitate UP matching. The evaluators were then grouped into a three-member team (Team 1-5) to merge their lists of UPs by applying the User Action Framework [2].

UP severity was rated using Sauro's scale [70]:

1: "Minor" - Causes some hesitation or slight irritation

2: "Moderate" - Causes occasional task failure for some users; causes delays and moderate irritation
3: "Critical" - Leads to task failure; Causes user extreme irritation

A non-UP-based category '4' is labelled as "Insight/ Suggestion/Positive" where users mention an idea or observation that may enhance the overall experience or express positive affect. When no UP was detected, '0' was given. As the evaluators of Study1 focused on identifying and rating UPs, they did not report any category '4'. Note that the evaluators had no access to GSR peaks or SAM ratings when assessing severity. The emotional criterion "irritation" was evaluated based on participants' comments and facial expressions captured in CRD clips. Those clips were made by the experimenter after the test sessions.

Each team was randomly assigned to evaluate the clips of at least 10 participants. Each clip had the severity ratings given by 2 to 4 teams of evaluators except the first participant (P1) whose clips had the ratings given by all 5 teams. Pairwise weighted Cohen's kappa (Kw) was computed for the UP severity ratings to estimate inter-rater reliability. In cases with more than two teams, Kw of each pairwise combination was calculated. The pair with the highest Kw was selected and the average of the severity ratings of those two teams was computed. The values of Kw so obtained ranged from 0.07 (poor) to 0.78 (excellent). These results can further confirm the notoriety of the evaluator effect [30]. As all teams were involved in rating the clips of P1, the rating ability of each team could be compared by adding up pairwise Kw_s for each team (e.g., accumulated pairwise Kw_s of Team 1 is the sum of Kw of Team 1-2, 1-3, 1-4, 1-5). Team 2 had the highest accumulated Kw_s, followed by Team 4. For those cases with Kw below 0.4, the severity ratings made by Team 2 or Team 4 were used, as at least one of these two teams was involved in rating each of the participants' clips.

To compare with the teams' findings, two HCI experts with more than 10 years of experience in usability research and practice were involved. The experts independently conducted usability tests of the website with the same search task *prior to* this study with the traditional think-aloud approach and produced a list of UPs with severity ratings. Based on the empirical data of that earlier and this study, the experts matched the two lists of UPs. Pairwise Kw_s were computed between the experts' and novices' severity ratings; the Kw_s ranged from 0.1 to 0.6. We report the findings of both types of evaluators.

Study2

Two HCI experts with 15 and 8 years of experience in usability research were involved to evaluate the severity of UPs independently. As mentioned earlier, the peaks for which the participants could not recall any specific emotional response were eliminated. Each expert inspected individual clips. The superimposed gaze plots were used to analyse the search behaviours. The transcripts of the verbal comments made during CRD were referenced to consolidate the severity ratings. A step beyond Study1 is the use of the category '4' – positive affect, for instance, when the search was observed to proceed smoothly along with positive comments such as 'I was confident I found the answer", "I felt in control". The Kw for 5 categories was 0.6. The two experts discussed discrepant ratings to reach consensus, consolidating the ratings with Kw of 0.79.

RESULTS

Descriptive Statistics

Study1

As the GSR data of two of 21 participants were corrupted and thus discarded, the results were based on the remaining 19 datasets. Altogether 192 GSR peaks were identified (Mean=10.11; SD=3.35). Table 1 shows how expert and novice evaluators were discrepant in categorising the peaks.

	'1,2,3'= UP	**'0' =Non-UP***
Expert	149 (Mean = 7.84, SD =3.04)	43 (Mean=2.26, SD=1.69)
Novice	154 (Mean = 8.11, SD=3.13)	38 (Mean=2.0, SD=1.41)

*27 of the peaks were commonly identified by experts and novice evaluators as non-UPs.

Table 1. Distribution of the GSR peaks over the four categories of the severity scale (Study1).

Four participants were able to accomplish all three subtasks with correct answers; the task completion rate was 21%;

Study2

The GSR data of one participant were corrupted and thus discarded. As shown in Table 2, 184 (74%) of the GSR peaks identified are related to UPs of different severity levels (category '1, 2, 3') whereas 45 (17.5%) are related to positive causes (category '4'); 22 (8.5%) are in category '0', without any identifiable trigger.

Category	All 5	'1,2,3'= UP	'0' =Non-UP	'4'= Positive
Total	251	184	22	45
Mean	12.9	9.68	1.1	2.25
SD	4.72	4.47	1.59	1.89

Table 2. Distribution of the GSR peaks over the five categories of the severity scale (Study2)

Despite the fact that some participants answered all three subtasks, none of their responses were correct. The task completion rate was essentially zero.

GSR, Severity & Frustration Correlations

To verify H1, H2 and H3, correlations between GSR peaks, severity ratings, SAM ratings taken after each CRD episode (P_{CRD}, A_{CRD}, D_{CRD}), and the frustration score were computed.

Study1

Table 3 shows the results concerning the UP-related GSR peaks. Contrary to our hypothesis, GSR peaks are not significantly correlated with UP severity ratings (H1 rejected). On the other hand, GSR peaks are significantly correlated with frustration scores (H2 confirmed), which, however, are not significantly correlated with UP severity ratings (H3 rejected). These results can partially be

accounted for by the phenomenon of the evaluator effect [30] (see General Discussion).

	Severity		P_{CRD}	A_{CRD}	D_{CRD}	FS
UP-related GSR peaks	Expert (n=149)	.03	-.12	.07	**-.23****	**.18***
	Novice (n=154)	-.01	**-.17***	.05	**-.31****	**.22****
Severity Ratings	Expert (n =149)		.10	-.12	.08	.11
	Novice (n=154)		-.15	.02	-.15	.14

Table 3. Study1: GSR peaks correlate with UP severity ratings, SAM ratings of CRD and Frustration Scores (FS).

Negative correlations were found between GSR peaks and Pleasure (moderately significant) and between GSR peaks and Dominance (highly significant). The findings suggest that the less pleasant (more frustrated) and less in control of the situation (stronger fear) a participant felt, the higher the GSR peaks. This lends further support to our method of reversing Pleasure and Dominance ratings to derive the frustration score.

The interpretations are consistent with the results that Pleasure are negatively correlated with Activation (N =149, r = -.29; N = 154, r =-.31; p<.01) and that Pleasure are positively correlated with Dominance (N =149, r = .41; N = 154, r =.37; p<.01) for both sets of GSR peaks categorised by expert and novice evaluators. In contrast, there is no significant correlation between GSR peaks and Activation. There are no significant correlations between Activation and Dominance (N=149, r = -.07; N = 154, r = -.11; p>.05) with a negative correlation tendency.

Study2

Results similar to Study1 were obtained (Table 4): No significant correlation between GSR peaks and severity ratings is found (H1 rejected); GSR peaks are significantly correlated with frustration scores (H2 confirmed). Unlike Study1, severity ratings are significantly correlated with frustration scores (H3 confirmed). This confirmed relationship implies that UPs of higher severity as judged by the evaluators are subjectively perceived to be more frustrating by the participants. Such a concurrence is seldom assessed in usability tests as moment-by-moment evaluation is rarely performed in real-life practice.

Other than the seemingly unavoidable evaluator effect, we cannot identify any plausible reason for the non-significant relation between GSR peaks and severity ratings.

N = 184	Severity	P_{CRD}	A_{CRD}	D_{CRD}	FS
UP-related GSR peaks	.04	**-.21****	**-.16***	-.11	**.20****
Severity ratings	--------	**-.21****	-.29	-.13	**.22****

Table 4. Study2: Correlations between GSR peaks, severity ratings, SAM ratings after CRD, and Frustration Scores (FS)

No significant correlation between the GSR peaks for positive affect (category '4') and SAM_{CRD} is found (N =45; $P_{CRD,}$ r=-.20; A_{CRD} = -.09; D_{CRD} = -.12). H4 was rejected. The observation that positive affect is not related to arousal can be explained by the assumption that good usability is a hygiene factor [79].

Peak-End Rule

Study1

SAM_{AT} were taken right after the participant had completed the task or when the time limit of 15 minutes was reached. In addition, SAM_{CRDs} were taken right after each CRD episode, the number of SAM_{CRDs} varied with the number of GSR peaks identified for individual participants. The two SAM_{CRDs} that are of particular interest are those taken at the highest peak and the last (or end) peak (NB: the last peak is a valid proxy for the end point; no significant difference between the two variables). The subscales are designated as $P_{HIGH,}$ $A_{HIGH,}$ D_{HIGH} and $P_{END,}$ $A_{END,}$ $D_{END,}$ respectively. Correlations between the SAM ratings taken after the task, at the highest Peak and end Peak were computed. While most of the correlations are insignificant, the significant ones are the corresponding counterparts (Table 5): A_{AT} to A_{HIGH}; D_{AT} to D_{HIGH}; P_{AT} to P_{END}; D_{AT} to $D_{END.}$

		SAM_{AT} After the Task		
		P_{AT}	A_{AT}	D_{AT}
SAM_{CRD} at Highest Peak	P_{HIGH}	.39	-.21	.16
	A_{HIGH}	-.11	**.65****	.10
	D_{HIGH}	.20	-.24	**.48***
SAM_{CRD} at End Peak	P_{END}	**.58****	-.02	.08
	A_{END}	-.03	.42	.13
	D_{END}	.08	-.38	**.48***

Table 5. Study1 - Correlations between SAM ratings taken at different times (N = 19).

With regard to the Peak-End rule, while we caution that this may not be a causal relationship, the correlational results suggest that the post-task evaluation of Dominance (D_{AT}) could be shaped by the most recent emotional response elicited by the last UP as well as the most intense emotional response elicited by a UP. In some cases, the last and most intense peaks were related to the same UP. However, such a UP was not necessarily judged to be the most severe (i.e. level '3'). On the other hand, the post-task evaluation of Pleasure (P_{AT}) and Activation (A_{AT}) tend to be shaped by the last UP. To further investigate the Peak-End Rule, one plausible approach is to perform a multivariate analysis with the GSR measure at the highest peak and that at the end point (or the last peak) of the task being two predictors and the three SAM_{AT} subscales being criteria. However, the sample size (N = 19) is too small to allow such a statistical analysis. Summing up, the mixed patterns of the relationships between SAM_{AT}, SAM_{HIGH} and SAM_{END} suggest that H5 is partially confirmed.

Study 2

Like Study1, correlations among SAM ratings taken at different times were computed (Table 6). While eliminating the clips categorised as '0' (no identifiable trigger), those categorised as '4' (positive affect) are included for

computing the correlations. Results show that the emotional responses reported at the end peak are strongly related to the overall user experience evaluation of Pleasure and Dominance but not to that of Activation. A similar pattern, albeit to a lesser extent, is observed for the highest peaks.

		SAM_{AT} After the Task		
		P_{AT}	A_{AT}	D_{AT}
SAM_{CRD} at Highest Peak	P_{HIGH}	**.52***	-.01	.43
	A_{HIGH}	.41	.20	**.56***
	D_{HIGH}	**.50***	.36	.42
SAM_{CRD} at End Peak	P_{END}	**.56***	.06	.25
	A_{END}	**.59****	.22	**.61****
	D_{END}	**.47***	.17	**.59****

Table 6. Study2: Correlations between SAM ratings taken at different times (N = 19).

The findings suggest that the participants tended to remember the valence rather than arousal of their emotions and that the end peaks as compared with the highest peaks had a stronger impact on the overall user experience evaluation. In fact, six of the 19 end peaks belong to the category '4'- positive affect. Although these participants might have struggled during the search, they rated Pleasure high when they eventually found an answer, irrespective of its correctness, close to the end of the episode.

This observation seems consistent with that of [40], who argued that the highest peak might not always effectively influence the post-task subjective rating as compared with the last event, which could have a significant effect on the overall rating. Similarly, [28] also stated that the overall evaluation of the system's usability tended to be more strongly influenced by the mental effort enacted in the later stages of the episode (i.e., the recency effect [28]). Nonetheless, like Study1, we can only partially confirm H5 because of the insignificant correlations for Activation in both the highest and end peak conditions.

Previous research shows that retrospective affective ratings such as SAM tend to be different from their concurrent counterparts due to the memory gap or certain contextual factors [42,57]. People tend to report on an episode by referring to the most painful or most enjoyable moment, which influences their overall evaluation of the episode [38, 39,64]. In Study1 and Study2, basically SAM_{CRD} are not concurrent ratings because the participants were asked to re-immerse in the recorded interactions and evaluated their re-felt emotions.

GENERAL DISCUSSION

To recap, Study2 was conducted with the dual purpose of substantiating the empirical findings of Study1 and improving on the methodological approach of Study1.

	H1	H2	H3	H4	H5
Study1	×	✓	×	n/a	✓
Study2	×	✓	✓	×	✓

Note: ×: rejected; ✓: confirmed; ✓ : partially confirmed

Table 7. Summary of the outcomes of the five hypotheses.

Table 7 summarises the outcomes of the hypothesis testing. Our mixed results support the argument that mapping the inclusive psychophysiological response to a multifaceted emotional experience like frustration is a very challenging process, as it is complicated by a person's cognitive and motor activities and other contextual stimuli [47]. Despite strong evidence for the reliability of GSR to measure arousal, there are still inconsistent results (e.g., [20,55]).

Ordering Effect and Evaluator Effect

We found the non-significant correlation between GSR peaks and UP severity ratings baffling, and proposed as an explanation the *ordering effect* of UPs on emotion. After experiencing a UP, a user's frustration may either be intensified or dampened when the UP recurs, depending on the coping strategy the user has developed [74]. For instance, the user may choose to ignore the UP to remain calm; in this case, her GSR decreases rather than increases. The first-occurring UPs may evoke direct emotional responses with minimal cognitive interpretation. When the same problem happens again, the user tends to reflect more on it and emotionally responds differently. To evaluate this assumption, we extracted two sets of UPs from individual participants: 'first retained' (keeping UPs a user experienced for the first time and removing their recurrences); 'recurring retained' (removing the first experienced UPs and keeping the recurrences). In both cases, UPs experienced only once are retained. We hypothesized that there would be a significant correlation between GSR peaks and severity ratings in the case of 'first retained'. But there remains no significant correlation, disconfirming our hypothesis of the ordering effect.

The evaluator effect [33] is then the most plausible reason. This recalcitrant problem has recently been revisited by [30], who provided further evidence for it. They put forward several reasons of which two are more relevant to our work. First, a situation characterised by judgment and uncertainty is prone to discrepant interpretations [65]. In our study, two judgments needed to be made – whether GSR fluctuations signify relevant changes in a user's emotional states; how much cognitive and emotional cost UPs incur in a user. The lack of standardized thresholds for judging the relevance of GSR peaks [76] and the lack of operationalised criteria to judge the impact of UPs on a user's performance (how long a delay is to be considered severe) and emotion (how frustrated a user feels is deemed critical). Second, the role of domain knowledge in assessing UPs is crucial. The evaluators of Study1 should have higher familiarity with the website as compared with those of Study2. However, the e-government website we used as the evaluation target in our studies is comparable to the e-commerce website used in [30] in terms of general understanding of web-based search expected from their users, one could argue that knowledge should not play a decisive role here. Overall, we tend to agree on [30] that the evaluator effect cannot be eliminated but can only be managed to mitigate its undesirable impact.

Implications for Usability Evaluation Practice

The basic motivation driving this work was to explore whether psychophysiological measurement as accessible as GSR can facilitate usability evaluation, especially with regard to the extent of frustration induced by UPs of different severity. Using GSR peaks as markers to identify episodes of interest can support both data capture and data analysis. It can engage participants in retrospective think-aloud and relieve evaluators from sifting through an entire length of video to focus on specific segments, thereby improving the validity of verbal comments and the efficiency of analysing them. The informativeness of video data can be enhanced by superimposing them with gaze plots; the goal we attempted to achieve in Study2. While using eye-tracking for usability evaluation is not new [16], combining it with GSR can exploit the potential of both.

As discussed earlier, frustration is a multifaceted construct [71]. It can be treated as an amalgamation of anger, fear (low dominance) and unhappiness (unpleasant). Hence, we addressed this issue by deriving a formula for computing a summative frustration score based on the three SAM subscales. The confirmed relation between GSR peaks and frustration scores lends further empirical support to the utility of GSR data for identifying relevant interaction episodes in usability evaluation. When the situation does not allow post-test CRD exercises (e.g., time constraint), GSR data alone are still useful for extracting relevant video segments for further analysis.

The observation that no significant differences in the SAM ratings taken at different points of time could be found is attributable to the short duration of our test sessions. What is the longest time gap between the actual interaction and CRD that allows effective re-immersion to produce valid comments? This will be an empirical question for the future work. Findings of the SAM subscale Activation (Arousal) are puzzling. It may be attributed to the rather ambiguous figures with the 'explosive chest' (cf. the facial expressions of Pleasure and varying sizes of Dominance, which are arguably less ambiguous). Although the definition of Arousal from [9] was instructed to the participants, some might have difficulty in relating the extent of their emotional responses to those figures or even in estimating such an extent themselves. As compared with the previous studies examining the relationship between GSR and emotions (mostly stress), verbal questionnaires such as NASA-TLX and DSSQ [55] were employed. It is intriguing to find out if different results will be obtained if a verbal questionnaire is used.

Limitations

As a strength as well as a weakness is our reliance on a single type of psychophysiological measure. With our goal of developing a parsimonious framework for practice, we opted for the relatively more reliable, affordable and least intrusive device - GSR. The inclusion of other measures such as heart beat may shed some light on understanding frustration induced by usability problems. Eye-tracking data are more for understanding cognitive rather than emotive behaviour, although, in combination with other data, they can contribute to the analysis of user experience.

The involvement of student evaluators in Study1 had the purpose of assessing the practicality of the proposed hybrid approach. When it was demonstrated to be workable with novices, a stronger claim of its applicability could be stated. This inevitably complicated the procedure, and might have made the evaluator effect more acute than otherwise.

Psychophysiological data typically require a large number of datasets to derive consistent patterns, given the huge individual differences in emotional responses. The modest sample sizes of both Study1 and Study2 restrict the use of multivariate analysis and other sophisticated computational models.

We might have omitted some data by not conducting CRD with non-peak video segments. Our study was based on the relation between computer user frustration and GSR. While we cannot eliminate the odds that some users may not have GSR to UPs, it is reasonable to assume such instances are relatively low. We are aware of this limitation but struggle to address it. A peak segment is typically 10s long. If we applied this length to segment non-peak periods, the number of clip to be viewed and SAM to be filled by users would be overwhelming, evoking negative feelings that could confound the results. As our future work, we aim to develop an approach that allows us to collect such data effectively and reliably.

CONCLUSION

Can psychophysiological data tell us about the relationship between frustration and the severity of usability problems? Our answer is affirmative albeit with some caveats. While we detected the significant correlation between GSR peaks and frustration scores, the assumed relationship between UP severity ratings and GSR peaks was not verified. We argued that the inevitable evaluator effect could be the major cause. As recommended in [30] and also practised in our study, multiple evaluators and a group process to negotiate discrepancies could help mitigate the evaluator effect. But what is still missing and much needed are more structured, explicitly operationalised and multifaceted (performance- and emotion-based) criteria for a severity scale. Our research addressed this gap by proposing an approach to quantify frustration based on subjective data and correlate it with objective data. The rationale is to provide empirical evidence (GSR, self-reported data) for evaluators to cross-check severity ratings, which currently rely on subjective judgment. More research needs to be done to streamline the approach. But overall, we have made some important steps towards the development of a well-defined UX-relevant severity scale that can minimise the evaluator effect.

REFERENCES

1. John L. Andreassi. 2013. *Psychophysiology: Human behavior & physiological response*. Psychology Press.
2. Terence S. Andre, Rex Hartson, Steven M. Belz, and Faith A. McCreary. 2001. The user action framework: a reliable foundation for usability engineering support tools. *International Journal of Human-Computer Studies 54*, 107-136
3. Abram Amsel. 1992. *Frustration theory: An analysis of dispositional learning and memory, No. 11*. Cambridge University Press.
4. Thomas Baekdal. *Usability Severity Rating – Improved*. Retrieved July 12, 2015 from http://www.baekdal.com/articles/Usability/usability-severity-rating/
5. Tood Bentley Lorraine Johnston, and Karola von Baggo. 2005. Evaluation using cued-recall debrief to elicit information about a user's affective experiences." In *Proceedings of the Australia Conference on Computer-Human Interaction* (OzCHI'05), pp. 1-10.
6. Katie Bessiere, John E. Newhagen, John P. Robinson, and Ben Shneiderman. 2006. A model for computer frustration: The role of instrumental and dispositional factors on incident, session, and post-session frustration and mood. *Computers in Human Behavior,* 22, 941-961.
7. Wolfram Boucsein. 2012. *Electrodermal activity*. Springer.
8. Kirsten Boehner, Rogério DePaula, Paul Dourish, and Phoebe Sengers. 2007. How emotion is made and measured." *International Journal of Human-Computer Studies, 65*, 275-291.
9. 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.
10. Anders Bruun and Simon Ahm. 2015. Mind the Gap! Comparing Retrospective and Concurrent Ratings of Emotion in User Experience Evaluation. In *Proceedings of INTERACT 2015*.
11. John T Cacioppo and Wendi L. Gardner. Emotion. 1999. *Annual Review of Psychology, 50* (1), 191-214.
12. John T Cacioppo, Louis G. Tassinary, and Gary Berntson (Eds.). 2007. *Handbook of psychophysiology (3rd Edition)*. Cambridge University Press.
13. Andy Cockburn, Philip Quinn, and Carl Gutwin. 2015. Examining the Peak-End Effects of Subjective Experience. In *Proceedings of Conference on Human Factors in Computing Systems* (CHI'15), 357-366
14. Michael E. Dawson, Anne M. Schell, and Diane L. Filion. Chapter 7: The Electrodermal System, In *Handbook of psychophysiology*, pp. 159-181. Cambridge University Press.
15. Amy M. Do, Alexander V. Rupert, and George Wolford. 2008. Evaluations of pleasurable experiences: The peak-end rule. *Psychonomic Bulletin & Review*,15, 96-98.
16. Andrew Duchowski. 2007. *Eye tracking methodology: Theory and practice*. Springer.
17. Joseph S. Dumas and Janice Redish. 1999. *A Practical Guide to Usability Testing*. Intellect Books.
18. Panteleimon. Ekkekakis. 2013. *The measurement of affect, mood, and emotion: A guide for health-behavioral research*. Cambridge University Press.
19. Paul Ekman, Robert W. Levenson, and Wallace V. Friesen. 1983. Autonomic nervous system activity distinguishes among emotions. *Science 221,* 1208-1210.
20. Stephen H. Fairclough and Louise Venables. 2006. Prediction of subjective states from psychophysiology: A multivariate approach. *Biological Psychology, 71* (1), 100-110.
21. Raul Fernandez. 1998. *Stochastic modeling of physiological signals with hidden markov models: A step toward frustration detection in human-computer interfaces*. Master's thesis, Massachusetts Institute of Technology, Dept. of Electrical Engineering and Computer Science.
22. Pierfrancesco Foglia,, Cosimo Antonio Prete, and Michele Zanda. 2008. Relating GSR Signals to traditional Usability Metrics: Case Study with an anthropomorphic Web Assistant. In *Proceedings of IEEE Instrumentation and Measurement Technology Conference,* 1814-1818.
23. Malin Forne. 2012. *Physiology as a Tool for UX and Usability Testing*. School of Computer Science and Communication, Master. Royal Institute of Technology, Stockholm.
24. Barbara L Fredrickson and Daniel Kahneman. 1993. Duration neglect in retrospective evaluations of affective episodes. *Journal of Personality and Social Psychology,* 65.
25. Yuan Gao, Nadia Bianchi-Berthouze, and Hongying Meng. 2012. What does touch tell us about emotions in touchscreen-based gameplay? *ACM Transactions on Computer-Human Interaction* (TOCHI) 19, 31:1-31:30
26. Lewis R. Goldberg. 1992. The development of markers for the Big-Five factor structure. *Psychological assessment* 4.
27. Marc Hassenzahl. 2000. Prioritizing usability problems: Data-driven and judgement-driven severity

estimates. *Behaviour & Information Technology 19*, 29-42.

28. Marc Hassenzahl and Nina Sandweg. 2004. From mental effort to perceived usability: transforming experiences into summary assessments. In *CHI'04 Extended Abstracts on Human Factors in Computing Systems*, 1283-1286.
29. Morten Hertzum. 2006. Problem prioritization in usability evaluation: From severity assessments toward impact on design. *International Journal of Human-Computer Interaction 21*, 125-146.
30. Morten Hertzum, Rolf Molich and Jacobsen Niels Ebbe 2014. What you get is what *you* see: Revisiting the evaluator effect in usability tests. *Behaviour & Information Technology*, 33, 2, 143-161.
31. Don H. Hockenbury and Sandra E. Hockenbury. 2010. *Discovering psychology*. Macmillan.
32. Kasper Hornbæk and Erik Frøkjær. 2008. A study of the evaluator effect in usability testing. *Human–Computer Interaction, 23* (3), 251-277.
33. Niels Ebbe Jacobsen, Morten Hertzum, and Bonnie E. John. 1998. The evaluator effect in usability tests. In *Proceedings* of *CHI 98 Conference Summary on Human Factors in Computing Systems*, pp. 255-256. ACM.
34. Robin Jeffries. 1994. Usability problem reports: helping evaluators communicate effectively with developers. In *Usability Inspection Methods*, Jakob Nielsen and Robert L. Mack (eds.), New York, Wiley 25-62.
35. Bonnie E. John and Steven J. Marks. 1997. Tracking the effectiveness of usability evaluation methods. *Behaviour & Information Technology*, 16 (4-5), 188-202.
36. Christian Martyn Jones and Tommy Troen. 2007. Biometric valence and arousal recognition. In *Proceedings of Australasian Conference on Computer-Human Interaction* (OzCHI'07), 191-194.
37. Daniel Kahneman 2003. A perspective on judgment and choice: mapping bounded rationality. *American Psychologist, 58*.
38. Daniel Kahneman. 2011. *Thinking, fast and slow*. Macmillan.
39. 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.
40. Simon Kemp, Christopher DB Burt, and Laura Furneaux. A test of the peak-end rule with extended autobiographical events. *Memory & Cognition, 36*, 132-138.
41. Jonathan Klein, Youngme Moon and Rosalind W. Picard.2002. This computer responds to user frustration: Theory, design, and results. *Interacting with computers, 14* (2), 119-140.
42. Sari Kujala and Talya Miron-Shatz. 2013. Emotions, experiences and usability in real-life mobile phone use. In *Proceedings of Conference on Human Factors in Computing Systems,* 1061-1070.
43. Peter J. Lang. 1980. Behavioral treatment and bio-behavioral assessment: computer applications. In *Technology in mental health care delivery systems,* Joseph B. Sidowski, James Harding Johnson, Thomas Arthur Williams (eds.), 119-137
44. 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,* 261-261.
45. Effie L-C Law, Paul van Schaik, and Virpi Roto. 2014. Attitudes towards user experience (UX) measurement. *International Journal of Human-Computer Studies, 72*, 526-541.
46. Jonathan Lazar, Adam Jones, Mary Hackley, and Ben Shneiderman. 2006. Severity and impact of computer user frustration: A comparison of student and workplace users. *Interacting with Computers 18*, 187-207.
47. Richard S. Lazarus. 1991. Progress on a cognitive-motivational-relational theory of emotion. *American Psychologist, 46*.
48. Reed Lawson. 1965. *Frustration; the development of a scientific concept*. Macmillan.
49. James R. Lewis. 2001. Introduction: Current issues in usability evaluation. *International Journal of Human-Computer Interaction, 14* (3&4), 489-502.
50. James. R. Lewis. 2012. Usability testing. In *Handbook of Human Factors and Ergonomics* (4th ed.), Gavriel Salvendy (Ed.), New York, Wiley, 1267-1312.
51. Tao Lin, Masaki Omata, Wanhua Hu and Atsumi Imamiya. 2005. Do physiological data relate to traditional usability indexes? In *Proceedings of the Australia Conference on Computer-Human Interaction* (OzCHI'05).
52. Martina Manhartsberger and Norbert Zellhofer. 2005. Eye tracking in usability research: What users really see? In *Usability Symposium*, 198, 141-152..
53. Regan L. Mandryk. 2005. *Modeling user emotion in interactive play environments: A fuzzy physiological approach*. Ph.D. Thesis. Simon Fraser University, Burnaby.
54. Regan L. Mandryk and Stella Atkins. 2007. A fuzzy physiological approach for continuously modeling

emotion during interaction with play technologies. *Int. J. Human-Computer Studies,* 65, 329-347.

55. Gerald Matthews, Sian E. Campbell, Shona Falconer, Lucy A. Joyner, Jane Huggins, Kirby Gilliland, Rebecca Grier, and Joel S. Warm. Fundamental dimensions of subjective state in performance settings: task engagement, distress, and worry. *Emotion, 2* (4), 315 - 340.
56. Albert Mehrabian 1996. Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament. *Current Psychology, 14* (4): 261-292.
57. Talya Miron-Shatz, Arthur Stone, and Daniel Kahneman. 2009. Memories of yesterday's emotions: Does the valence of experience affect the memory-experience gap? *Emotion, 9*
58. Rolf Molich and Joseph S. Dumas. 2008. Comparative usability evaluation (CUE-4). *Behaviour & Information Technology, 23*(1), 65-74.
59. Rolf Molich, Meghan Ede, Klaus Kaasgaard, and Barbara Karyukin. 2004. Comparative usability evaluation. *Behaviour & Information Technology 23* (1), 65-74.
60. Jakob Nielsen. 1994. Heuristic evaluation. In *Usability Inspection Methods*, Jakob Nielsen and Robert L. Mack (eds.), New York, Wiley.
61. Mary M. Omodei and J. I. M. McLennan. 1994. Studying complex decision making in natural settings: using a head-mounted video camera to study competitive orienteering. *Perceptual and Motor Skills, 79*, 1411-1425.
62. Rosalind W. Picard. 1997. *Affective Computing*. Cambridge: MIT press.
63. Alex Poole and Linden J. Ball. 2006. Eye tracking in HCI and usability research. *Encyclopedia of Human Computer Interaction,* 211-219.
64. Donald A. Redelmeier and Daniel Kahneman. 1996. Patients' memories of painful medical treatments: real-time and retrospective evaluations of two minimally invasive procedures. *Pain, 66*, 3-8.
65. Wendy D. Roth and Jal D. Mehta. 2002. The Rashomon effect combining positivist and interpretivist approaches in the analysis of contested events. *Sociological Methods & Research 31*, 131-173.
66. Virpi Roto, Effie L-C. Law, A. P. O. S. Vermeeren, and Jettie Hoonhout. 2010. *User experience white paper: Bringing clarity to the concept of user experience*. Retrieved July 12, 2015 from http://www.allaboutux.org/
67. David Rubin. 1996. *Autobiographical memory*. Wiley.
68. James A. Russell. 1980. A circumplex model of affect. *Journal of Personality and Social Psychology, 39*.
69. Jeff Sauro. 2014. The relationship between problem frequency and problem severity in usability evaluations. *Journal of Usability Studies, 10* (1), 17-25.
70. Jeff Sauro. 2013. *Rating the severity of usability problems. Measuring Usability*. Retrieved July 10, 2015 from http://www.measuringu.com/blog/rating-severity.php.
71. Jocelyn Scheirer, Raul Fernandez, Jonathan Klein, and Rosalind W. Picard. 2002. Frustrating the user on purpose: a step toward building an affective computer. *Interacting with Computers, 14*, 93-118.
72. Yu Shi, Natalie Ruiz, Ronnie Taib, Eric Choi, and Fang Chen. 2007. Galvanic skin response (GSR) as an index of cognitive load. In *CHI'07 Extended Abstracts on Human factors in Computing Systems*, 2651-2656.
73. Mikael B. Skov and Jan Stage. 2007. Supporting problem identification in usability evaluations. In *Proceedings of the Australia conference on Computer-Human Interaction* (OzCHI'07).
74. Craig A. Smith and Richard S. Lazarus. 1990. *Emotion and Adaptation*. In *Handbook of Personality: Theory and Research*, L.A. Pervin (Ed.), 609-637.
75. Kim J. Vicente, D. Craig Thornton, and Neville Moray. 1987. Spectral analysis of sinus arrhythmia: A measure of mental effort. *Human Factors, 29*, 171-182.
76. Robert D. Ward and Philip H. Marsden. 2003. Physiological responses to different web page designs. *International Journal of Human-Computer Studies*, 59 (1), 199-212.
77. Chauncey Wilson. 1999. Reader's questions: Severity scales. *Usability Interface*, *5*. Retrieved July 12, 2015 from http://ww2.stcsig.org/usability/newsletter/9904-severity-scale.html
78. Lin Yao, Yanfang Liu, Wen Li, Lei Zhou, Yan Ge, Jing Chai, and Xianghong Sun. 2014. Using Physiological Measures to Evaluate User Experience of Mobile Applications. In *Engineering Psychology and Cognitive Ergonomics*, 301-310. Springer.
79. Marc Hassenzahl, Effie L-C. Law, and Ebba Thora Hvannberg. 2006. User Experience: Towards a unified view. In *Proceedings of NordiCHI'06 Workshop on Towards a Unified View of User Experience*, 1-3.