Mind the Gap! Comparing Retrospective and Concurrent Ratings of Emotion in User Experience Evaluation

Anders Bruun and Simon Ahm

Aalborg University, Department of Computer Science
bruun@cs.aau.dk, x.simon@gmail.com

Abstract. User experience (UX) is typically measured retrospectively through subjective questionnaire ratings, yet we know little of how well these retrospective ratings reflect concurrent experiences of an entire event. UX entails a broad range of dimensions of which human emotion is considered to be crucial. This paper presents an empirical study of the discrepancy between concurrent and retrospective ratings of emotions. We induced two experimental conditions of varying pleasantness. Findings show the existence of a significant discrepancy between retrospective and concurrent ratings of emotions. In the most unpleasant condition we found retrospective ratings to be significantly overestimated compared to concurrent ratings. In the most pleasant condition we found retrospective ratings to correlate with the highest and final peaks of emotional arousal. This indicates that we cannot always rely on typical retrospective UX assessments to reflect concurrent experiences. Consequently, we discuss alternative methods of assessing UX, which have considerable implications for practice.

Keywords: User Experience, Emotion, Memory-Experience Gap, Peak-End Rule.

1 INTRODUCTION

Emotion is considered a fundamental factor in determining user experience (UX) of interactive technologies [1, 2]. Forlizzi and Battarbee state: *"Emotion affects how we plan to interact with products, how we actually interact with products, and the perceptions and outcomes that surround those interactions"* [3].

The most frequent method to assess users' emotional states is subjective ratings where users fill in questionnaires such as the Self-Assessment Manikin (SAM) [1]. However, studies outside a UX context have shown a critical caveat in subjective ratings of emotions. Recently, Scherer argued that emotions are fleeting, i.e. they are short term, intensive peaks of experiences which may be difficult to recall at a later point in time [4]. This is supported by empirical observations of a discrepancy between overall retrospective ratings of an episode and the concurrent experiences during an episode [5]. This discrepancy is denoted the *memory-experience gap* and is verified by several studies in e.g. pain research. Notably, Redelmeier and Kahneman found retrospective ratings of pain to correlate with the highest and final intensities of

pain experienced [6]. Those observations led to what is now known as the *peak-end rule*. Thus, studies outside a UX context suggest that retrospective ratings of emotions will likely not reflect concurrent experiences of an entire event.

This caveat also seems to be recognized within UX where studies measure emotions concurrently. SAM is typically filled in after completing each task in an interaction sequence, see e.g. [7–9]. However, very few studies within UX have been engaged with more detailed measurements of emotions [1]. Kujala and Miron-Shatz present initial insights on the presence of the memory-experience gap in a Human-Computer Interaction (HCI) context. The emphasis in [10] is on longitudinal assessments of emotions based on the Day Reconstruction Method (DRM). In DRM participants evaluate the set of experienced emotions by the ending of each day [10]. Given the fleeting nature of emotions, DRM suffers from a recall bias and is therefore well suited for longitudinal studies with an interest in estimating overall averages of emotional reactions [11]. It is, however, ill-suited for measuring emotions at specific time points, e.g. at specific points in an interaction sequence [12].

Consequently, we still know very little of the existence and extent of the memory-experience gap in an HCI context. Arguably, such a context is less extreme than Redelmeier and Kahnemans studies of pain.

This study contributes to the HCI community by showing that we cannot *always* rely on retrospective UX assessments to reflect concurrent experiences, even when they are assessed immediately after interacting with a system. This is critical as UX is primarily assessed retrospectively [1]. We discuss alternative approaches to assess UX, which provide more accurate accounts of concurrent experiences than, e.g. the Day-Reconstruction Method.

In the remainder of this paper we provide a theoretical overview introducing potential caveats of retrospective ratings as well as a set of hypotheses. We then present related work, experimental method and results. Finally, we discuss and conclude on our findings.

2 THEORETICAL BACKGROUND

In this section we introduce theoretical stances dealing with the relationship between concurrent and retrospective ratings. We start by giving an account for the concept of emotions and how emotions can be measured. We then go through the theoretical background concerning the discrepancy between retrospective and concurrent ratings, i.e. the memory-experience gap and the peak-end rule. We conclude this section by presenting our hypotheses.

2.1 What are Emotions?

In the classical theory of James-Lange from 1884, emotions are defined as the result of physical changes in autonomic and motor functions. Input from our senses creates a range of responses in our body, and our awareness of these changes is what constitutes an emotion [13]. Individual emotions are distinguished based on their unique

bodily expression. The theory states that when an event happens in our environment (e.g. being attacked by a predator) we instantaneously get physiological reactions like muscle tension, increased sweat production etc. We interpret these unique combinations of physiological reactions as being a specific emotion [13].

Defining emotions is a topic of much debate and we do not attempt to provide an exhaustive walkthrough of the literature. However, basic assumptions behind the classical James-Lange theory are still supported after more than a century. Recently, Scherer described emotions as a mobilization and synchronization of five organismic subsystems as a response to a cognitive evaluation of external or internal stimulus events that are *"relevant to major concerns of the organism"* [4]. When an event happens that is of major concern, the event is evaluated through a comparison to innate prototypical responses, and a response (the emotion) is elicited through activation of the organismic subsystems [4]. It is important to note that this "evaluation" is not a time consuming and conscious process as is often associated with the term in HCI. Instead it relies on fast subconscious processes [4].

2.2 Measuring Emotions

Throughout research, two approaches have been used to elicit emotions: Subjective ratings and objective measurements.

Subjective Ratings. Subjective ratings of emotions are typically collected through questionnaires. These typically consist of standardized labels or pictograms representing emotions [4]. Most studies of emotions in UX are based on such subjective rating methods where the Self-Assessment Manikin (SAM) is the most widely applied technique [1]. SAM enables participants to assess their emotional states through graphical scales, cf. [14]. It is based on a dimensional model of emotion denoted PAD (Pleasure, Arousal, Dominance) that uses three dimensions to represent emotions:

- *Pleasure:* Indicates how pleasant an emotion is, i.e. its valence. It spans from negative to positive.
- *Arousal:* Indicates how intense an emotion. It spans from relaxed to excited.
- *Dominance:* Indicates how dominating an emotion is. It spans from low to high dominance.

Participants are asked to rate emotions based on these three dimensions. Thus, they assess the level of Pleasure, Arousal and Dominance, each on a 1-9 point Likert scale.

Objective Measurements. Psychological research has recently seen an advancement in applying physiological sensors to objectively assess emotions [4]. In doing this, participants do not rate emotions subjectively, but instead a sensor (or sometimes several) is attached on the body. A range of different sensors exist, each of which typically measures one dimension of the PAD model. As an example, Galvanic Skin Response (GSR) sensors have in particular been shown to correlate with arousal [15]. In essence, a GSR sensor reacts on changes in skin resistance (measured in mOhms)

through varying levels of perspiration in sweat glands. A GSR sensor enables real-time measurements of arousal. Other sensors are Electromyography (EMG), Heart Rate (HR) and Electroencephalography (EEG). For a more comprehensive overview of measurement sensors and their relative performances, see [16]. As noted by Scherer, it is currently not feasible to collect all types of measurements [4], and studies have also shown varying reliability of these in measuring emotions. However, GSR sensors are consistently correlated with arousal across different studies, this also includes the few studies of UX applying physiological sensors, see e.g. [17].

2.3 Memory-Experience Gap

According to Scherer, the purpose of emotions is to deal successfully with an event that is of direct concern to the organism, and the activation of physiological subsystems require lots of resources to do so [4]. Due to the level of intensity, which cannot be endured over a longer period of time, emotions are short-lived mental states. Furthermore, Scherer argues that emotions are always tied to a specific event [4]. So, if emotions are short-lived and tied to a specific event, what is then measured in retrospective ratings based on recall?

Several studies in psychology have demonstrated a discrepancy between the average of actual experienced emotions and the overall retrospective assessment of an experience [5]. Thus, there is a gap between concurrent emotions and the recollection of these after a given episode. This discrepancy is denoted the *memory-experience gap* [5]. Studies in psychology have shown that the memory-experience gap leads to overestimated retrospective ratings, i.e. that retrospective ratings reflect a higher emotional intensity compared to concurrent experiences [5].

It is argued that the memory-experience gap is present for both positive and negative stimuli [5]. However, a study conducted by Baumeister et al. indicated a larger memory-experience gap when experiencing negative stimuli compared to positive stimuli [18].

Transferring the above into an HCI context we could expect that 1) Retrospective ratings of UX are overestimated compared to concurrent ratings and 2) Retrospective ratings would especially be overestimated in episodes of negative experiences.

2.4 Peak-End Rule

The discrepancy between actual experienced emotions and retrospective ratings has also been studied in detail by Redelmeier and Kahneman. They made a striking discovery during their studies of how pain is experienced and recalled. They found that test subjects preferred a longer duration of pain over a shorter duration of pain [19]. In the condition with the longest duration, pleasantness was increased towards the end, although still being painful. In the other condition, the level of pain was kept constant, but the duration was shorter. Test subjects primarily recalled the experience of pain towards the end of the experimental conditions, i.e. they preferred the longest duration of pain with increasing pleasantness. In a later study it was found that test subjects retrospectively rated the entire experience based on the highest intensity of pain (the

peak) and the pain experienced towards the end [6]. This has since become known as the *peak-end rule*.

In relation to HCI, it is plausible that retrospective ratings of emotions correlate with the highest peak of emotional intensity and the intensity measured towards the end of an interaction sequence.

2.5 Hypotheses

We have formulated the following hypotheses based on the above considerations of the memory-experience gap and peak-end rule:

H1. *Retrospective ratings of emotions are higher than concurrent ratings of emotions.*

H2. *The memory-experience gap is larger for episodes of negative emotions than for episodes of positive emotions.*

H3. *Concurrent and retrospective ratings of emotions correlate following the peak-end rule.*

3 RELATED WORK

A recent literature survey confirmed that most UX assessments are conducted retrospectively [1]. Based on the theoretical background above, this led us to question the extent to which retrospective ratings reflect an entire experience. As mentioned in the introduction of this paper, emotion is a key dimension in determining UX of interactive technologies. Therefore we emphasize this particular dimension and now provide an overview of previous UX studies of emotions.

Most studies of emotions in UX research are based on questionnaires, typically SAM, through which users provide subjective ratings [1]. From related work we primarily know that SAM ratings are affected by instrumental factors such as the level of usability, but we know very little of the discrepancy between concurrent and retrospective ratings. Hassenzahl and Ullrich studied the effect of giving users predefined tasks versus open ended interaction [7]. SAM ratings were measured concurrently three times during interaction followed by a retrospective AttrakDiff rating. Primary findings showed that the type of tasks affected SAM and AttrakDiff ratings. As a byproduct of the entire experiment Hassenzahl and Ullrich found a correlation between concurrent SAM ratings and retrospective AttrakDiff ratings [7]. Mahlke and Thüring conducted a similar study in which they examined the effect of high/low usability and high/low aesthetics on emotional responses [9]. SAM ratings were taken three times during interaction followed by retrospective ratings of usability and visual aesthetics. Findings showed that concurrent SAM ratings differed significantly between high/low usability and to a minor extent between high/low aesthetics. In another study, Mahlke and Thüring studied the feasibility of measuring emotional states through a multiple components approach [20]. This was done by combining objective psychophysiological measurements and subjective ratings. In that study SAM was

filled in after each task while collecting real-time physiological data. Findings also showed that concurrent SAM ratings differed significantly between the high/low usability versions of the system. Other UX studies of emotions apply SAM to extract emotions during use and with similar findings, see e.g. [2, 8].

Hassenzahl and Sandweg studied how concurrent experiences of mental effort related to retrospective assessments of perceived usability [21]. Participants were given seven tasks and the SMEQ questionnaire (for measuring mental effort) was filled in after each task. Findings showed significant correlation between the last rating of mental effort and the perceived level of usability. In their longitudinal study, Kujala and Miron-Shatz [10] applied the Day-Reconstruction Method (DRM) to elicit participant emotions of mobile phone usage. Participants were asked to report specific episodes at the end of each day and their related emotions over a five day period. On the sixth day, participants were asked to provide a summary rating of emotions. Findings showed that participants significantly overestimated the experienced positive emotion in their summary assessments.

From the above we know that ratings of emotions are affected by instrumental factors such as providing tasks or not, the level of usability and partly by the non-instrumental factor of aesthetics. It also seems that the potential memory-experience gap is recognized as participants are typically asked to provide SAM ratings concurrently rather than retrospectively. Yet, we know little of the potential discrepancy between retrospective and concurrent ratings. Notably, a single study showed that concurrent SAM ratings correlated with retrospective AttrakDiff ratings, cf. [7]. Findings from that study could indicate that the memory-experience gap is insignificant in an HCI context. Yet, this finding is contradicted in [10] and [21]. Nevertheless, across all studies, we found that concurrent ratings were measured relatively few times, in some cases as few as three times per participant. Arguably, this may not reflect the whole set of short-lived emotions experienced during an entire interaction sequence. This critique has also been raised against DRM [12, 22], on which the study by Kujala and Miron-Shatz [10] is based. Thus, there is a critical need for more detailed studies of emotions in UX research, as is also stressed by Bargas-Avila and Hornbæk [1]. In the following section we describe how we studied the memory-experience gap and peak-end rule in an HCI context.

4 METHOD

The objective of our study was to examine the memory-experience gap and peak-end rule in subjective ratings of emotions. Therefore we needed participants to elicit several emotional states reflecting concurrent experiences, i.e. during interaction, as well as retrospective ratings. However, collecting concurrent ratings during interaction is not as straightforward as collecting retrospective ratings. Following related work, concurrent ratings can be collected after each task, this, on the other hand provides relatively few measurement points. We collected concurrent ratings based on the Cued-Recall Debrief (CRD) method as outlined below. This enabled a higher density in measurements while avoiding interruptions during interaction.

4.1 Cued-Recall Debrief

CRD is a method based on situated recall. The method was developed by Omodei et al. [23] to elicit emotional experiences while not interfering with participant behavior 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 [23]. To foster re-immersion, it is crucial that video recordings resemble a first-person point of view. In Omodei et al.'s study, video recordings were collected by head-mounted cameras positioned on helmets of firefighters [23]. CRD essentially builds on retrospection but several studies have validated the approach. In [23] it was found that CRD leads to considerably more detailed responses compared to retrospective ratings based on free recall [23]. Furthermore, Bentley at al. [24] found correlation between CRD ratings and real-time physiological measurements in an HCI context. Also, a more recent study published in the renowned TOCHI outlet 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. Furthermore, it does so without causing interference during interaction.

In practice CRD is conducted by selecting a set of video clips, each showing a specific episode of an entire event. Participants view one clip at a time for which they provide subjective ratings of emotions. This is then done for all video clips. In our case we selected video clips on the basis of real-time GSR data, i.e. we selected video clips surrounding peaks of arousal. Fig. 1 illustrates this principle.

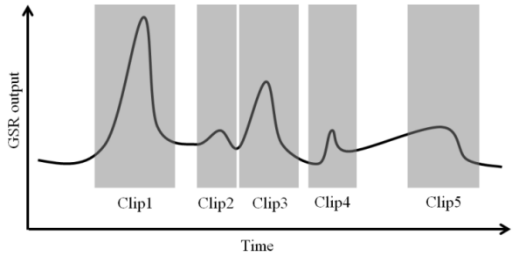

Fig. 1. Selecting video clips based on GSR peaks.

4.2 Experimental Conditions

Related work show that instrumental goals such as tasks/no-tasks and usability problems affect user emotions when interacting with technology, see e.g. [7, 9]. We were interested in studying the memory-experience gap and, based on previous studies, it was relevant to consider how this was affected by the level of usability.

We developed two versions of a software system (described in section 4.4 "System"). The two versions differed as we in one version seeded a set of five usability problems into the user interface. The other version was developed to the best of our abilities. Thus we had two experimental conditions: 1) *Seeded* and 2) *Unseeded*.

4.3 Participants

A total of 20 university students participated (between subjects) with 10 in each condition. Ages ranged from 19-28 years (mean 22.85). In each condition we had 2 females and 8 males. All subjects were kept unaware of the actual premise of the experiment and none of them had previous experience in using the system.

4.4 System

We designed a system to create and edit posters which was specifically developed for the experiment. The functionality included simple text editing and styling for creating bold, italic and underlined text as well as a color picker. It was also possible to include images etc. The unseeded and seeded versions were built on the exact same platform and source code. Versions differed in the user interface only. In the seeded version we introduced the following five usability problems: 1) Standard buttons for text styling (known from typical word processors) were replaced with checkboxes, 2) no font preview was available, 3) font required for the task was unavailable, 4) color picker was based on text rather than actual colors and 5) inserting an image led to an incomprehensible error message.

4.5 Task

The task was to use the system to create an exact copy of a poster pictured on paper. Participants were informed that there was a 10 minute time limit to solve the task. They could click a submit button at any time, if they felt the task had been solved before time the limit.

4.6 Setting and Data Collection

The experiment was conducted in a university classroom. The user was placed in front of a laptop with an external mouse, and given the paper describing their task as well as the poster to replicate. The student was alone in the room for the full duration.

We collected concurrent and retrospective ratings of emotions via the SAM questionnaire. All sessions were recorded using screen capture software. This was a picture-in-picture setup primarily showing participants' interaction with the system and a small picture of the face, which was recorded with a high-resolution webcam. The desktop recording resembled a first-person point of view, which is in line with the Cued-Recall Debrief (CRD) method outlined previously. Finally, a Galvanic Skin Response (GSR) sensor was placed in the palm of participants' non-primary hand. This was done in order to determine specific points in the interaction sequence where participants experienced peaks of arousal. That information was used to select video clips on which to base concurrent SAM ratings.

4.7 Procedure

This section describes the CRD procedure applied in the study:

1. *Introduction and setup:* Participants were directed to the room where they received a piece of paper with the task and the poster to replicate using either the unseeded or seeded version of the software. The task was also described verbally by a researcher (one of the authors). The GSR sensor was then attached to their palm. Participants were not informed of the purpose of the study and the GSR sensor until they completed the experiment. After the user had confirmed that they understood the task at hand, the researcher started the software and left the room.
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. This baseline was measured by showing a blank screen for the first 4 minutes, while playing a relaxing piece of music. The song "Weightless" by Marconi Union was chosen. Previous studies have shown that this particular piece of music has a relaxing effect on participants [26].
3. *Completing the task:* After the 4 minutes of relaxation, the user interface for the system appeared and the task could begin. Participants were allowed 10 minutes for this. All recordings were automatically stopped after 10 minutes and the researcher returned to the room.
4. *Retrospective SAM rating:* Participants were asked to subjectively rate the overall emotion immediately after the 10 minutes.
5. *Concurrent SAM ratings (via CRD):* The video of the users face was superimposed over the lower right corner of the screencast. The GSR data was visualized as a graph and superimposed over the timeline of the video player. This allowed for the researcher to visually identify peaks in the GSR data, and fast-forward to about 5 seconds before these points. The user was then shown this part of the video (screencast as well as their own facial expressions), and asked to freely describe their own thoughts as to what they may have reacted to, and how. If the user was able to deduce a reaction, the peak was noted along with their description of the event. They were also asked to rate the emotional state at the time using SAM.

5 RESULTS

In this section we provide several measures, which will form the basis of our later discussion on whether to verify or falsify our three hypotheses.

5.1 Differences between Retrospective and Concurrent Ratings

According to theory, the memory-experience gap should cause retrospective ratings to be overestimated in comparison to concurrent ratings (hypothesis H1). Below we compare the mean of concurrent and retrospective ratings in both experimental conditions, see Table 1.

Table 1. Mean SAM ratings distributed by condition and PAD dimensions. *=significant, n=No. of participants.

		Seeded (n=10)	Unseeded (n=10)
SAM - Pleasure	Retrospective	6 (1.6)*	5.6 (1.8)
	Concurrent	4.6 (1.8)*	5.7 (1.5)
	Overall mean	**4.7 (1.8)**	**5.6 (1.5)**
SAM - Arousal	Retrospective	4.6 (2.2)	3.3 (2.1)
	Concurrent	4.4 (1.8)	3.9 (2.1)
	Overall mean	**4.5 (1.8)**	**3.8 (2.1)**
SAM - Dominance	Retrospective	5.3 (2.1)*	5.5 (1.5)
	Concurrent	4.4 (1.7)*	5.4 (1.3)
	Overall mean	**4.5 (1.8)**	**5.4 (1.4)**

In the seeded condition the mean retrospective rating of Pleasure is 6 (SD=1.6) while the mean concurrent rating is lower with 4.6 (SD=1.8). A repeated measures Wilks' Lambda (.05 level) shows a significant difference as indicated by the * in Table 1 (Wilks' Lambda = .41, F=11.3, p = .01). In the unseeded condition the retrospective rating of Pleasure is 5.6 (SD=1.8), which is similar to the concurrent rating of 5.7 (1.5). This difference is not significant (Wilks' Lambda = .99, F=.008, p = .93).

In terms of Arousal, the seeded condition reveals a retrospective rating of 4.6 (SD=2.2) and a similar concurrent rating of 4.4 (SD=1.8). A Wilks' Lambda test reveals no significant difference (Wilks' Lambda = .99, F=.1, p = .77). In case of the unseeded condition we also observe comparable retrospective and concurrent ratings of Arousal with 3.3 (SD=2.1) and 3.9 (SD=2.1) respectively. No significant difference is found (Wilks' Lambda = .95, F=.44, p = .52).

Considering Dominance, we see that the retrospective rating in the seeded condition is 5.3 (SD=2.1), which is higher than the concurrent rating of 4.4 (SD=1.7). This difference is significant (Wilks' Lambda = .41, F=11.3, p = .01). In case of the unseeded condition we see no significant difference between retrospective and concurrent ratings (Wilks' Lambda = .99, F=.008, p = .93).

In sum we found that all retrospective ratings in the seeded condition were higher than concurrent ratings. This was significant in terms of Pleasure and Dominance. In the unseeded condition we did not find significant differences between any of the retrospective and concurrent ratings. This indicates a larger memory-experience gap in the seeded condition.

5.2 Differences Between Conditions

Hypothesis H2 suggests that the memory-experience gap is larger for episodes of unpleasant emotions than for episodes of pleasant emotions. Below we seek to verify that the seeded condition leads to a worse experience compared to the unseeded condition. We do this by using two metrics: 1) Differences in ratings between conditions and 2) The level of emotional fluctuation experienced.

Differences in Ratings. In Table 1, the overall mean refers to the mean of retrospective and concurrent ratings combined. In the seeded condition the overall mean rating of Pleasure is 4.7 (SD=1.8) while the overall mean is 5.6 (SD=1.5) in the unseeded condition. An independent samples t-test at the .05 level reveals a significant difference between conditions (t=-4.1, df=226, p=0.0).

Participants rated the experienced level of Arousal to be higher in the seeded condition (overall mean = 4.5, SD=1.8) than in the unseeded condition (overall mean = 3.8, SD=2.1). An independent samples t-test indicates that this difference is significant (t=-2.4, df=226, p=0.02).

Finally, participants expressed a lower level of Dominance in the seeded condition (overall mean = 4.5, SD=1.8) compared to the unseeded condition (overall mean = 5.4, SD=1.4). This difference is also significant (t=-4.1, df=226, p=0.0).

In sum, emotions experienced in the seeded condition were more negative with a higher level of arousal compared to the unseeded condition. Emotions in the seeded condition were rated as less dominant than emotions in the unseeded condition.

Differences in Emotional Fluctuation. As mentioned in section 2.2 "Measuring Emotions", GSR data shows fluctuations in arousal. Intuitively, the system version that was seeded with usability problems will lead to a higher level of fluctuation in arousal as the stress level increases when encountering a usability problem. Fig. 2 shows two typical examples of GSR data obtained, one example from the unseeded condition (top) and one from the seeded (bottom). By visual inspection, the above examples appear to fluctuate differently with the seeded example having more peaks and a higher level of variance in arousal.

The number of peaks in the GSR graphs was counted on the basis of visual inspection by one of the researchers. Participants in the unseeded condition experienced a mean of 17.22 (SD=3.8) peaks while participants in the seeded condition experienced 22.2 (SD=5.6) peaks. An independent samples t-test revealed a significant difference between conditions (t=-2.21, df=15, p=0.042).

We also calculated the mean variance for both conditions. The mean variance for Pleasure is 1.58 (SD=1.48) in the unseeded condition and 2.93 (SD=0.77) in the seeded. An independent samples t-test reveals a significant difference in this respect (t=2.31, df=15, p=0.034). The mean variance for Arousal is 1.63 (SD=2.17) in the unseeded condition and 3.63 (SD=1.1) in the seeded, which is also significant (t=2.44, df=15, p=0.03). A similar pattern is observed in case of Dominance where the mean variance is 1.04 (SD=1.24) and 3.02 (SD=1.55) in the unseeded and seeded conditions respectively (t=2.92, df=15, p=0.01).

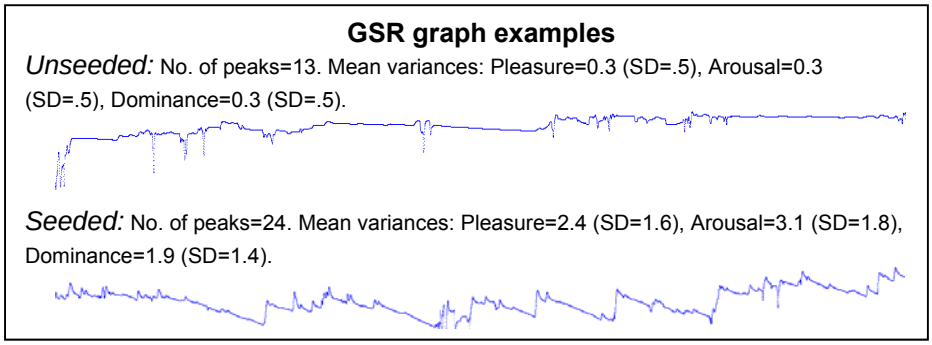

Fig. 2. Two examples of GSR graphs. Top: Unseeded condition, Bottom: Seeded condition

Thus, we found significant differences in emotional fluctuations between the two conditions. In the seeded condition participants experienced significantly more peaks and the variance in ratings within this group was 2-3 times larger compared to the unseeded condition. Based on the overall mean ratings in Table 1 and differences in fluctuation of arousal, we find that the seeded condition leads to a worse experience compared to the unseeded condition.

5.3 Peak-End Rule Correlations

In the following we consider the extent to which retrospective and concurrent ratings correlate with the peak-end rule (hypothesis H3). Table 2 shows the correlation table between retrospective, concurrent, highest peak and last peak ratings for the unseeded condition. This is based on computation of Pearson r correlation coefficients.

In the unseeded condition there is significant correlation between retrospective ratings of Pleasure and ratings of Pleasure made towards the end of the interaction sequence, i.e. at the last peak ($r=0.642$, $n=10$, $p=0.045$). Additionally, we find significant correlation between retrospective ratings of Arousal and ratings of Dominance made at the highest peak ($r=0.776$, $n=10$, $p=0.008$) and the last peak ($r=0.9$, $n=10$, $p=0.000$). Finally, in the unseeded condition we see a significant correlation between retrospective and concurrent ratings of Arousal ($r=0.774$, $n=10$, $p=0.009$).

In the seeded condition (not shown in Table 2) we observe significant correlation between retrospective and concurrent ratings of Pleasure ($r=0.677$, $n=10$, $p=0.031$). Retrospective ratings of Pleasure also correlate with concurrent ratings of Dominance ($r=.688$, $n=10$, $p=0.028$).

Thus, in the unseeded condition we primarily see significant correlations between retrospective ratings and ratings given at the highest and last GSR peaks. This differs from the pattern in the seeded condition where we see correlations between retrospective and concurrent ratings.

Table 2. Correlation table of retrospective and concurrent SAM ratings. Unseeded condition. *= significant, n=No. of participants.

	SAM correlations (Unseeded condition, n=10)		**Pleasure (retrospective)**	**Arousal (retrospective)**	**Dominance (retrospective)**
Concurrent	**Pleasure**	Pearson	-.072	-.062	-.366
		Sig. (2-tailed)	.844	.866	.299
	Arousal	Pearson	-.249	.774*	.273
		Sig. (2-tailed)	.489	.009	.445
	Dominance	Pearson	-.359	.542	-.054
		Sig. (2-tailed)	.309	.106	.882
Highest Peak	**Pleasure**	Pearson	-.184	.593	-.187
		Sig. (2-tailed)	.611	.071	.604
	Arousal	Pearson	-.035	.268	.447
		Sig. (2-tailed)	.924	.454	.196
	Dominance	Pearson	-.103	.776*	.367
		Sig. (2-tailed)	.777	.008	.296
Last Peak	**Pleasure**	Pearson	.642*	.109	.555
		Sig. (2-tailed)	.045	.765	.096
	Arousal	Pearson	-.597	.545	-.161
		Sig. (2-tailed)	.068	.103	.657
	Dominance	Pearson	-.395	.900*	.058
		Sig. (2-tailed)	.259	.000	.873

6 DISCUSSION

In the following we discuss our findings in relation to the three hypotheses and related work. Based on our findings we discuss how to apply retrospective ratings of user experience, which has high relevance for practice.

6.1 Significant Memory-Experience Gap

Based on the theoretical background of emotions we formulated hypothesis H1: "*Retrospective ratings of emotions are higher than concurrent ratings of emotions"*. Looking across the two experimental conditions shows that we are able to verify hypothesis H1 in the case where the system had a relatively higher number of usability problems. Findings from our study revealed that retrospective ratings were higher than the mean of concurrent ratings. This finding applied to the system version that was seeded with five usability problems. This overestimation was especially apparent in terms of Pleasure and Dominance. In relation to this we also formulated hypothesis H2: "*The memory-experience gap is larger for episodes of unpleasant emotions than for episodes of pleasant emotions"*. We were also able to verify hypothesis H2. The seeded system version led to a relatively higher number of usability problems, which in turn caused users to experience more negative emotions compared to the unseeded

version. We also found that the seeded condition led to a significantly higher level of fluctuation in arousal. The more negative ratings and higher level of arousal indicates that the seeded condition was experienced as the most unpleasant episode. As mentioned previously, we found that retrospective ratings of Pleasure and Dominance were significantly higher than concurrent ratings in this condition. Thus, there was a larger memory-experience gap in the seeded condition compared to the unseeded condition.

6.2 Retrospective Ratings Follow the Peak-End Rule

In terms of the peak-end rule we formulated hypothesis H3: "*Concurrent and retrospective ratings of emotion correlate following the peak-end rule"*. We validate H3 in the case where the system had a relatively lower number of usability problems. In the unseeded condition we identified significant correlations between retrospective ratings of Pleasure and ratings of Pleasure located at the last GSR peak measured. Furthermore, we found correlations between retrospective ratings of Arousal and ratings of Dominance at the highest and last peaks. This suggests that H3 could be verified. However, findings from the seeded condition show the opposite where e.g. retrospective ratings of Pleasure correlated with concurrent ratings. Thus, in the unseeded condition we primarily identified significant correlations following the peak-end rule while this was not the case in the seeded condition.

The difference between conditions can be explained by the level of emotional fluctuation experienced. In the seeded condition we observed significantly more peaks of arousal via the GSR sensor compared to the unseeded condition. The variance in SAM ratings in the seeded condition was also significantly higher on all Pleasure-Arousal-Dominance (PAD) dimensions. Consequently the highest and final peaks are interwoven with several other peaks of similar intensities. Arguably, such fluctuation obscures the peak and end experiences in the seeded condition. As a result, the most intensive and end experiences are evened out, hereby leading to correlations between retrospective and concurrent ratings. The opposite is the case for the unseeded condition where the lower level of fluctuation does not obscure peak and end experiences. Although the peak-end rule was not verified in the seeded condition we still stress that retrospective ratings were significantly overestimated in this case.

This is in line with the findings of Hassenzahl and Sandweg who found that the last concurrent rating of mental workload correlated with the retrospective rating of perceived usability [21]. Similarly, Kujala and Miron-Shatz found weekly ratings of emotions to be overestimated compared to the day-to-day ratings [10]. Thus, it seems that the peak-end rule is present for multiple UX dimensions. In contrast, Hassenzahl and Ullrich found correlations between concurrent SAM ratings and retrospective measurements of AttrakDiff [7]. This is in line with findings from our seeded condition, but contradictory of findings from the unseeded condition. As discussed above, peak-end correlation seems dependent on the level of fluctuation. Maybe the system applied in [7] led to a comparable variance in fluctuations as experienced by participants in our seeded condition. This brings us to discuss the main caveat of retrospective UX assessments.

6.3 Bottom Line: Mind the Gap in Current Practices

We found that participants significantly overestimated the level of Pleasure and Dominance in retrospective ratings, at least in the system with a relatively higher number of usability problems. Conversely, retrospective ratings were not overestimated in the system version with fewer problems. These findings show that retrospective assessments of emotions do not always reflect the concurrent experience and that this memory-experience gap depends on the level of usability.

Although we in this study emphasized the UX dimension specifically related to emotions, we believe that our findings also apply to more general assessments of UX. This is also supported by the findings in [21], which are based on measurements of mental workload. From related work we know that overall measurements of UX, e.g. AttrakDiff ratings, depend on emotions [7], and Forlizzi and Battarbee state: *"Emotion affects how we plan to interact with products, how we actually interact with products, and the perceptions and outcomes that surround those interactions"* [3].

Given that such overall ratings are primarily collected in retrospect [1], we now pose the following question: *How do we know when we can rely on retrospective UX assessments to provide accurate accounts of concurrent experiences?* In the unseeded condition we found no significant difference between retrospective and concurrent ratings. So an answer could be to apply current retrospective methods on a system with relatively few usability problems. This could e.g. be towards the end of a development process going into a phase of summative assessments. However, it is not trivial to decide when there are "few enough problems" such that retrospective ratings accurately reflect concurrent experiences.

Arguably, the current practice of applying questionnaires in retrospective UX assessment is defensible due to its simplicity. Others have justified this approach through the relative ease and low cost of use [20]. From previous studies we also know that people do not remember all their experiences, which lead to the memory-experience gap. Yet, retrospective evaluations are very important for people's later decisions, e.g. they buy products based on their memory of them [27]. Therefore, UX is not *only* about what happens concurrently but also how people remember their experience of products. However, the above discussion leads us to say that retrospective ratings (of emotions and overall UX alike) should be handled with care. Findings from this study have considerable implications for HCI practice and research, which we discuss in the following.

6.4 Moving Forward

As an alternative to retrospective questionnaire assessments we encourage practitioners to apply multiple methods in order to reflect concurrent experiences. In line with Scherer [4], we suggest to apply an approach relying on a combination of subjective ratings and objective psychophysiological measures. This is furthermore supported by Avila-Hornbæk who state that: *"In emotional psychology there are many established and validated ways of measuring emotions that provide more detailed and richer*

data... Future research might benefit from these to do in depth studies of affective sates in UX" [1].

Psychophysiological measurements are widely used within emotional psychology and we believe that these can contribute in providing the detailed measurements suggested in [1]. As an example of an alternative method, our application of Cued-Recall Debrief (CRD) comprises a combination of subjective ratings (SAM) and objective psychophysiological measurements (GSR). We applied CRD by measuring real-time GSR data during interaction. We re-immersed users by showing video clips of their interaction at selected points, clips that were selected based on GSR peaks of arousal. Although ratings in CRD are retrospective in nature, CRD has been validated for measuring concurrent experiences [23–25]. Validity is also indicated through the verification of our initial hypotheses, i.e. we are able to explain our findings.

Our study also points to a more general implication for HCI research, which is in line with Karapanos et al. [28]. In [28] a highly relevant discussion of the reliability and veridicality of UX studies is brought up. Reliability in a temporal sense denotes the consistency with which people recollect emotional experiences while veridicality deals with the consistency between actual and recalled experiences. Within our study (and in using CRD in general) the aim is to increase veridicality. This is much more difficult to control in naturalistic and/or longitudinal studies such as the study by Kujala and Miron-Shatz [10]. In those types of studies, the emphasis should be on reliability. We believe the HCI community needs further discussions and studies to understand veridicality and reliability in UX assessments and the challenges herein. Being aware of the different challenges inherent in different methods will assist HCI researchers in selecting appropriate methods in relation to study aims.

6.5 Limitations

A clear limitation in our study is the sample size where we had 10 participants in each condition. Also, our study is based on a system for creating posters. A larger sample and more diverse systems would increase generalizability. However, as stated in the introduction, the contribution of this study is to show that we cannot *always* rely on retrospective UX ratings to reflect concurrent experiences. We found e.g. that retrospective ratings of pleasure were significantly higher than concurrent ratings, even on this relatively small dataset. This is sufficient to prove the point of the paper, but further studies are needed to make claims about the extent of the memory-experience gap across a wider population and other systems.

Another limitation is the artificial laboratory setting. Several UX studies are appearing based on "in the wild" methods as this reflects naturalistic system usage, e.g. the study by Kujala and Miron-Shatz [10]. However, the strength of the lab is its ability to control for outside interferences such as being startled by a ringing phone, being interrupted by a colleague, emails etc. Factors like these can impact physiological measurements. It was crucial for us that all measurements reflected the usage of the system and not the presence of confounding factors from the context.

A third limitation is the retrospective nature of the Cued-Recall Debrief method. We applied this to enable approximations of concurrent experiences as participants

were asked to rate their emotions immediately after the interaction took place. Although not truly concurrent, we argue that the recall bias is considerably reduced compared to the Day-Reconstruction Method applied in other studies, e.g. [10].

7 CONCLUSIONS

The empirical study presented in this paper contributes to the HCI community by showing the existence of a memory-experience gap between concurrent and retrospective ratings of emotions. Findings showed that in the most unpleasant event, participants significantly overestimated retrospective ratings. In a more pleasant episode we found that retrospective ratings correlated with the highest and final emotional peaks of an entire experience. Thus, retrospective ratings do not always reflect concurrent experiences. This shows a potential caveat in current practices, which primarily are based on retrospection.

Alternatively we encourage practitioners to apply multiple methods to get insights on concurrent experiences. In the future it is critical to conduct further studies with larger sample sizes and different systems. This should be done in order to determine the extent of the memory-experience gap across other populations and systems.

8 References

1. Bargas-Avila, J.A., Hornbæk, K.: Old wine in new bottles or novel challenges. Proc. CHI. p. 2689. ACM, New York (2011).
2. Thüring, M., Mahlke, S.: Usability, aesthetics and emotions in human–technology interaction. Int. J. Psychol. 42, 253–264 (2007).
3. Forlizzi, J., Battarbee, K.: Understanding Experience in Interactive Systems. Proc. DIS. pp. 261–268. ACM, New York (2004).
4. Scherer, K.R.: What are emotions? And how can they be measured? Soc. Sci. Inf. 44, 695–729 (2005).
5. Miron-Shatz, T., Stone, A., Kahneman, D.: Memories of yesterday's emotions: does the valence of experience affect the memory-experience gap? Emotion. 9, 885–891 (2009).
6. Redelmeier, D.A., Kahneman, D.: Patients' memories of painful medical treatments: real-time and retrospective evaluations of two minimally invasive procedures. Pain. 66, 3–8 (1996).
7. Hassenzahl, M., Ullrich, D.: 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, 429–437 (2007).
8. Mahlke, S., Lindgaard, G.: Emotional Experiences and Quality Perceptions of Interactive Products. In: Jacko, J. (ed.) Human-Computer Interaction. Interaction Design and Usability SE - 19. pp. 164–173. Springer Berlin Heidelberg (2007).
9. Mahlke, S., Thüring, M.: Studying Antecedents of Emotional Experiences in Interactive Contexts. Proc. CHI. pp. 915–918. ACM, New York (2007).
10. Kujala, S., Miron-Shatz, T.: Emotions, Experiences and Usability in Real-life Mobile Phone Use. Proc. CHI. pp. 1061–1070. ACM, New York (2013).

11. Diener, E., Tay, L.: Review of the Day Reconstruction Method (DRM). Soc. Indic. Res. 116, 255–267 (2014).
12. Bylsma, L.M., Taylor-Clift, A., Rottenberg, J.: Emotional reactivity to daily events in major and minor depression. J. Abnorm. Psychol. 120, 155–167 (2011).
13. James, W.: What Is An Emotion? Mind. os-IX, 188–205 (1884).
14. Lang, P.J.: Behavioral treatment and bio-behavioral assessment: computer applications. In: Sidowski, J.B., Johnson, J.H., and Williams, T.H. (eds.) Technology in Mental Health Care Delivery Systems. pp. 119–137. Ablex, Norwood (1980).
15. Lang, P.J.: The emotion probe: Studies of motivation and attention. Am. Psychol. 50, 372–385 (1995).
16. Andreassi, J.L.: Psychophysiology: Human Behavior and Physiological Response. Lawrence Erlbaum, Mahwah (2000).
17. Ward, R.., Marsden, P..: Physiological responses to different WEB page designs. Int. J. Human-Computer Stud. 59, 199–212 (2003).
18. Baumeister, R.F., Bratslavsky, E., Finkenauer, C., Vohs, K.D.: Bad is stronger than good. Rev. Gen. Psychol. 5, 323–370 (2001).
19. Kahneman, D., Fredrickson, B.L., Schreiber, C.A., Redelmeier, D.A.: When More Pain Is Preferred to Less: Adding a Better End. Psychol. Sci. 4, 401–405 (1993).
20. Mahlke, S., Minge, M., Thüring, M.: Measuring Multiple Components of Emotions in Interactive Contexts. CHI EA. pp. 1061–1066. ACM, New York (2006).
21. Hassenzahl, M., Sandweg, N.: From Mental Effort to Perceived Usability: Transforming Experiences into Summary Assessments. CHI EA. pp. 1283–1286. ACM, New York (2004).
22. Dockray, S., Grant, N., Stone, A.A., Kahneman, D., Wardle, J., Steptoe, A.: A Comparison of Affect Ratings Obtained with Ecological Momentary Assessment and the Day Reconstruction Method. Soc. Indic. Res. 99, 269–283 (2010).
23. Omodei, M.M., McLennan, J.: Studying complex decision making in natural settings: Using a head-mounted video camera to study competitive orienteering. Percept. Mot. Skills. 79, 1411–1425 (1994).
24. Bentley, T., Johnston, L., von Baggo, K.: Evaluation Using Cued-recall Debrief to Elicit Information About a User's Affective Experiences. Proc. OzCHI. pp. 1–10. CHISIG Australia, Narrabundah (2005).
25. Gao, Y., Bianchi-Berthouze, N., Meng, H.: What Does Touch Tell Us About Emotions in Touchscreen-Based Gameplay? ACM Trans. Comput. Interact. 19, 31:1–31:30 (2012).
26. Belford, Z., Neher, C., Pernsteiner, T., Stoffregen, J., Tariq, Z.: Music & Physical Performance: The effects of different music genres on physical performance as measured by the heart rate, electrodermal arousal, and maximum grip strength. JASS. 3, (2013).
27. Karapanos, E., Zimmerman, J., Forlizzi, J., Martens, J.-B.: Measuring the dynamics of remembered experience over time. Interact. Comput. 22, 328–335 (2010).
28. Karapanos, E., Martens, J.-B., Hassenzahl, M.: On the Retrospective Assessment of Users' Experiences over Time: Memory or Actuality? CHI EA. pp. 4075–4080. ACM, New York (2010).