Asserting Real-Time Emotions through Cued-Recall: Is it valid?

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@leicester.ac.uk

Poul Svante Eriksen
Aalborg University
Aalborg Oest, Denmark
svante@math.aau.dk

ABSTRACT
Asserting emotions through free-recall is commonly used to evaluate user experience (UX) of interactive systems. From psychology we know that free-recall of emotions leads to a significant memory bias where participants rely on a few of the most intense episodes when asserting an overall experience. It is argued that cued-recall can reduce the memory bias in UX evaluations. Yet, this has not been studied empirically. We present a systematic empirical study based on 38 participants. We measured emotions in terms of objective galvanic skin responses (GSR) and subjective Self-Assessment Manikin (SAM) ratings. We found significant correlations between emotions experienced in real-time and those experienced during cued-recall. This validates the use of cued-recall for UX evaluations. An implication is that HCI researchers and practitioners now have cued-recall as an alternative that significantly reduces the memory bias and enables highly detailed measurements of emotions while not disturbing participants during system interaction.

Author Keywords
User Experience; Emotion; Cued-recall; Real-Time.

ACM Classification Keywords
H5.2. Information interfaces and presentation (e.g., HCI): User Interfaces: Evaluation/methodology.

INTRODUCTION
Emotion is key in assessing UX of interactive systems [14,35], yet measuring emotions is challenging. Kahneman et al. found a memory bias between emotions experienced in real-time and retrospective assertions made by participants [18]. Retrospection based on free-recall reflects the most intense and final emotions of an event; a phenomenon known as the peak-end effect [18].

This implies that we cannot rely on free-recall assertions to reflect actual experiences and recent studies confirm the existence of the peak-end effect in HCI contexts [4,8,19]. The common approaches to eliminate the memory bias in UX studies are based on measuring emotions in real-time rather than in retrospect. As an example, Mahlke et al. applied multiple physiological sensors to measure emotional responses during interaction [23]. This included e.g. sensors for measuring Galvanic Skin Response (GSR) and software for analyzing facial expressions, which measure arousal and valence, respectively. An alternative approach applied by e.g. Hassenzahl and Ullrich [16] is to let participants fill in subjective ratings of emotions after completing each task in an interaction sequence. Given their immediate temporal proximity to the actual emotional reaction, these two approaches provide valid measures of emotions experienced in real-time [30].

The research interest addressed in this paper is to examine the extent to which cued-recall (as opposed to free-call) reduces the memory bias when asserting emotions in retrospect. But why study retrospective methods? Why not rely on the validated methods where emotions are asserted in real-time? The reason is that real-time methods have the shortcoming of only providing quantitative data. This is useful for summative purposes, but in formative assessments we need qualitative insights in order to make relevant changes to a particular interaction design [3,32]. Qualitative insights are challenging to gather in real-time as this interferes with participant behavior [4]. Thus, retrospection is relevant for gathering the data needed in formative UX assessments based on emotions.

This paper presents a systematic empirical study comparing real-time assertion of emotions and emotions asserted in retrospect through cued-recall. The contribution is the validation of alternative methods for assessing UX. An implication is that HCI researchers and practitioners now have a real alternative to free-recall that significantly reduces the memory bias while enabling highly detailed measurements of emotions relevant to formative evaluations of UX.

THEORETICAL BACKGROUND

In this section we present the theoretical underpinnings of the study of measuring emotions and how memory of specific episodes can affect emotional responses. The latter is relevant since cued-recall relies on memories of past events.

What are Emotions?

The classical James- Lange theory of emotions asserts that physical changes in autonomic and motor functions result in emotions [17]. When an event happens in our environment such as being attacked, we get instant physiological reactions in the form of e.g. muscle tension, increased perspiration etc. We interpret unique combinations of physiological reactions as specific emotions [17].

Defining emotions has been a widely debated topic and our intention here is not to provide a thorough review of the literature on this. That said, basic assumptions behind the classical James-Lange theory are still supported to this day. Recently, Scherer denotes emotions as the mobilization and synchronization of physiological subsystems responding to external and internal stimuli of *"major concerns for the organism"* [29]. When an event of major concern occurs, the bodily response (the emotion) is elicited through activation of the physiological subsystems [29].

Measuring Emotions

In the related literature two major approaches to measuring emotional reactions are subjective self-reported ratings and objective psycho-physiological measurements.

Subjective Self-reported Ratings

Subjective ratings of emotions are collected through questionnaires, which consist of standardized labels or pictograms expressing different emotions [29]. The majority of UX studies on emotions are based on such subjective rating methods where the Self-Assessment Manikin (SAM) is the most widely applied [2]. In SAM, participants assess their emotional states through graphical scales and it is based on a dimensional model of emotion denoted as PAD [21]. PAD represents emotions based on the three dimensions of Pleasure, Arousal and Dominance. *Pleasure* indicates how pleasant an emotion is, i.e. its valence spanning from negative to positive. *Arousal* indicates the intensity of an emotion spanning from calm to excited. Finally, *Dominance* indicates how much in control a person feels and spans from low to high level of control. Participants are asked to rate emotions based on the PAD dimensions, typically on a 9-point Likert-type scale.

Objective Psychophysiological Measurements

More recently, physiological sensors have been applied to measure emotions [29]. Rather than rating emotions subjectively one or more sensors are attached on the body. Different sensors exist for measuring different dimensions of PAD. One example is Galvanic Skin Response (GSR) sensors, which have been shown to correlate well with emotional arousal [22]. A GSR sensor measures skin conductance through electrical resistance and reacts on varying levels of perspiration produced by the sweat glands. We know that a physiological reaction to an excited emotional state is to start perspiring, and, conversely, we do not perspire in calm states. Hence a GSR sensor enables real-time measurements of arousal [4,29]. Other sensors are electromyography (EMG), heart rate and electroencephalography (EEG). A comprehensive overview of sensors and their performances in detecting emotions can be seen in [1]. Scherer notes that it is not feasible to gather all types of physiological measurements [29]. Studies have also shown varying reliability of these sensors in measuring emotions, yet, GSR sensors are proven to consistently correlate with arousal, which also includes the few studies of UX applying physiological sensors (see e.g. [37]).

Relation between Memories and Emotions

In the following we make a theoretical outline of how free-recall and cued-recall affect assertion of emotions. This leads to our hypotheses of what to expect when measuring emotions based on memories.

Free Recall and the Peak-End Effect

Scherer argues that the purpose of emotions is to handle episodes that are of major concern to the organism. Physiological subsystems are activated such that we can deal with these events successfully, but such immediate and intense activation requires lots of resources [29]. Due to this intensity, which we cannot endure over prolonged periods, emotions are short-lived. Scherer also argues that emotions are tied to specific events [29]. In the introduction we stated our research interest in examining retrospective assessments of emotions. But when emotions are short-lived and tied to specific events, what do we measure in retrospective assessments based on recall?

Kahneman et al. discovered a memory bias during their studies of how emotions are experienced and recalled. They found that participants retrospectively rated an entire experience based on the highest emotional intensity (the peak) and the emotions experienced towards the end [28]. This free-recall memory bias has since become known as the peak-end effect. The peak-end effect has recently been shown to occur in UX evaluations [4,8].

The Effect of Cued-Recall on Emotions

Studies in psychology have dealt with the bi-directional relation between memory and emotion: 1) Emotions influence how experiences are encoded into memory and 2) remembering an experience affects our emotions. In terms of encoding into memory, studies have shown that, e.g. emotionally arousing experiences lead to increased attention. This in turn creates vivid memories of these episodes [26]. Note that the study in [4] also confirmed the tendency of emotions with negative valence being more

prevalent in memory than those with positive valence. However, the effect of emotional arousal on memory encoding seems larger than the effect of emotional valence. Thus, the level of excitement has a greater effect on encoding than whether an event is experienced as negative or positive [26].

The second direction (memory influences emotions) has been widely used as a regulatory mechanism in studies where researchers induce particular emotional states in participants. Researchers in psychology primarily apply cues based on general autobiographical events such as *"when I was in college"* to induce emotional states [26]. It is argued that specific events such as *"taking my English midterm"* (as opposed to general events) lead to higher levels of emotional intensities upon recollection [27]. In relation to this study, cued-recall would reflect specific events from an interaction sequence by, e.g. showing video clips of participants' interactions with a system.

Affect priming theory also deals with the relationship between emotion and memory. According to this, emotions are stored in a network of associations [5]. It is argued that the activation of an emotional state leads to some level of activation among other components of the network. In practice this means that memories associated with the current affective state are more likely to be recalled than memories not associated with that state [5].

Thus, it has been argued that recalling previous events can facilitate emotional reactions similar to the emotions experienced originally. For our study this would mean that emotions experienced during cued-recall are correlated to the experiences from actual interaction with a system. We have not found any studies examining how emotions asserted on the basis of retrieved memories are compared to emotions experienced in real-time. Murray et al. mention that future work is needed in order to examine the efficacy of e.g. cued-recall to induce emotional reactions [26]. This is also supported by researchers within neuropsychology such as Smith et al. [31]:

"There is considerable evidence that encoding and consolidation of memory are modulated by emotion, but the retrieval of emotional memories is not well characterized".

Our study contributes by further understanding the extent to which cued-recall assists in retrieving emotional reactions. As argued in the introduction, this is particularly relevant for formative UX evaluations.

Hypotheses

Based on the theoretical background described above we put forth the following hypotheses on retrospectively asserting emotions through cued-recall in UX evaluations:

H1. Retrieval of specific memories through cued-recall induces emotions (asserted through SAM and GSR) that are significantly *correlated* with emotions experienced during real-time system usage.

H2. Retrieval of specific memories through cued-recall induces emotions with *similar intensities* (asserted through SAM and GSR) as emotions experienced during real-time system usage.

RELATED WORK

In this section we present commonly used summative and formative methods for assessing emotions in UX studies.

Summative Methods: Real-Time and Quantitative

The Self-Assessment Manikin (SAM) [21], Emocards [12] and PrEmo [11] questionnaires are the most widely applied methods for measuring emotions in UX studies [2]. More recently the HCI research community has also gathered emotional data through physiological sensors, e.g. for measuring galvanic skin response, heart rate, etc. [4,23]. Questionnaires such as SAM and physiological measurements are applied during interaction where physiological sensors measure emotions in real-time. Questionnaires are typically filled in after completing each task, see e.g. [16,24]. While these standard methods reduce the memory bias (peak-end effect), they have the shortcoming of only providing quantitative data. To be used in formative evaluations, however, they need to be combined with other qualitative data sources such as interviews or user diaries [3,4,34]. Qualitative insights are necessary for uncovering the causes leading to specific experiences, and to direct the changes needed to improve a particular interaction design [3,32].

Formative Methods: Retrospective and Qualitative

Methods suitable for formative evaluation purposes can be divided according to the underlying form of recall, i.e. they are either based on free-recall or cued-recall.

Methods Relying on Free-Recall

The Day Reconstruction Method (DRM) has been applied in HCI contexts, e.g. by Kujala and Miron-Shatz [19]. In DRM participants are asked to qualitatively reconstruct the main episodes experienced during the day and to evaluate their related emotions. This process is based on free-recall and is conveniently done at the end of the day. From studies in psychology we know that DRM suffers from a significant memory recall bias [13]. It is therefore well suited for longitudinal studies with an interest in estimating predominant emotional reactions over time. On the other hand this free-recall approach also makes DRM ill-suited measuring emotional reactions at specific time points during interaction [7].

The same limitation is raised against the UX Curve and iScale methods [20]. These are very similar and built on the basis of participants' asserting their emotional reactions by drawing a curve of their experiences on a timeline at selected periods of evaluation. In this sense they reflect the reconstruction approach underlying DRM. Like DRM, UX Curve and iScale are also longitudinal. Qualitative

comments are given for each change in curve direction and slope steepness. While both methods can provide rich qualitative data, they lead to a significant memory bias, as they rely on free-recall.

The Experience Sampling Method (ESM) [10] reduces the memory recall bias by letting participants describe events and assert emotions multiple times during the day (rather than once at the end of the day) [30]. Thus, asserting emotions through free-recall is done in closer temporal proximity to when the actual emotions occurred. ESM is also applied in longitudinal studies typically spanning a period of 1-2 weeks [30]. The quality of ESM data is optimal when subjects assert emotional reactions immediately after an event occurred [30]. However, there is a trade-off as more responses are obtained when participants are allowed to report experiences at a more convenient, later time [30]. In cases where participants are allowed to make delayed assertions, researchers set a limit of no more than 20-30 minutes after the actual experienced emotions, see e.g. [10,33]. Even if participants report within a relatively short timeframe, a memory bias can still occur. The UX studies by Cockburn et al. [8] and Bruun and Ahm [4] showed that the peak-end effect was significant, even within 10 minutes of interaction with a system. So, the memory bias is present even after very short time-frames.

While the above methods are aimed at longitudinal studies, the method 3E (Expressing Experiences and Emotions) is for short-term assessment; participants are asked to fill in a template to assert emotions through drawings on a stick figure [36]. This could e.g. be a smiling or an angry face. Participants also describe their experience in a speech bubble located next to the stick figure. Like the above studies, this also relies on free-recall. Findings in [36] show that emotions elicited through 3E are comparable to ratings made through SAM and Emocards. However, since all ratings were based on free-recall, the memory bias would likely be significant.

Methods Relying on Cued-Recall

The Valence method [6] and Cued-Recall Debrief (CRD) [4] are two similar short-term methods based on gathering real-time quantitative data. Qualitative data are collected in retrospect via cued-recall, which contrasts the free-recall approach applied in most methods. In the Valence method, participants use a keypad and press "+" or "-" depending on their emotional reactions during interaction [6]. In [4], CRD was used with a GSR sensor attached on the hand of participants, hereby measuring arousal in real-time. In both methods, retrospective interviews are applied to collect qualitative data. Interviews are based on cued-recall as participants view video clips of their specific actions at selected points in time. These points are selected based on GSR peaks in CRD and for each +/- instance in the Valence method. For each clip the participants describe what happened and why.

Two critical questions are: Would these video clips re-immerse participants into their past experiences? Would cued-recall enable them to assert the same emotional reactions as experienced during interaction? There are two arguments as to why the answer should be yes: 1) There is relatively close proximity to the actually experienced emotions and 2) Video clips enable cued-recall, hereby providing specific memories of experienced emotions [4,6]. These claims are intuitive and in line with the theoretical background outlined above. Yet, we have found no studies that empirically examined these claims.

Finally, the Affective Diary is also based on cued-recall [34]. It combines the element of recalled reconstruction seen in DRM, UX Curve and iScale while also including real-time GSR measurements and logging of system events. These data are presented on a timeline that provide memory cues. Participants can annotate each event and describe what have happened at these specific moments. Findings in [34] show that the four participants were able to explain what happened at specific points in the log, but the quality of such recall was not examined.

METHOD

The aim of the study was to compare real-time assertion of emotions and emotions asserted through cued-recall.

Experimental Conditions

We designed two experimental conditions:

1) A **baseline** condition building on the standard method of participants filling in a questionnaire after each task, see e.g. [16,24].
2) A **CRD** condition based on the **cued-recall debrief** method where participants retrospectively fill in a questionnaire on the basis of cued-recall, see e.g. [4].

The purpose was to compare the subjective ratings obtained through these two conditions. In order to triangulate data we also included objective physiological measurements in both conditions. This was measured through a GSR sensor. Such triangulation and the fact that we selected to include the baseline condition enabled us to determine the quality of CRD.

Why CRD?

In the cued-recall condition we chose CRD over the Valence method. This is because we were interested in testing the similarity of emotional intensities between real-time and cued-recall experiences (hypothesis H2). The intensity is possible to measure through a GSR sensor [4]. In the Valence method participants indicate "+" and "-" during interaction, but the intensity of these positive and negative emotions is not measured. As mentioned in the theoretical background above, emotional arousal leads to more vivid memories than emotional valence, so we chose to leverage a method based on measuring arousal.

Procedure

Figure 1 outlines the procedure in the two conditions.

Step 1: Introduction and GSR setup

In both conditions participants were initially introduced to the study. Here we explained that we conducted a UX evaluation of a statistics website (detailed below). They were not told the purpose of the experiment. The GSR sensor was then attached in the palms of the participants' non-dominant hand. This was done so that participants would not be impeded during interaction with the system as the GSR readings were sensitive to hand movements.

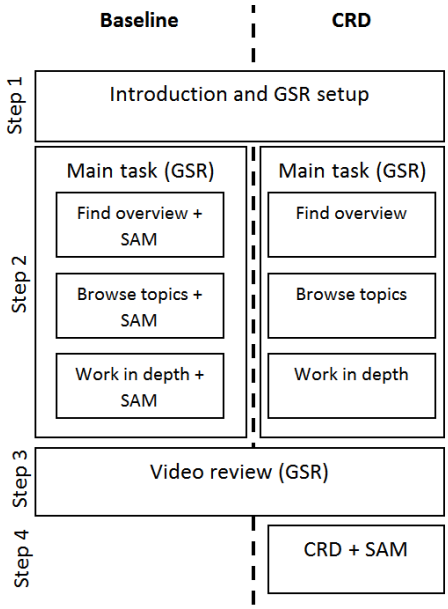

Figure 1: Outline of procedure for the experimental conditions. Baseline = standard UX method to assert emotions, CRD = Cued-Recall Debrief method to assert emotions. GSR = Galvanic Skin Response senor, SAM= Self-Assessment-Manikin questionnaire.

Step 2: System Interaction

After the initial introduction and GSR setup, participants in both conditions were given the same main task to solve. In the baseline condition participants were asked to solve it through three subtasks and to give SAM ratings after each subtask as suggested in e.g. [16,24]. Note that participants in the CRD condition worked uninterrupted, i.e. no subtasks were given. However, to enable comparison between conditions, the baseline subtasks reflected the task solving strategy of participants in the CRD condition (see details in section "Tasks" below). A time constraint of 15 minutes was imposed for a single session in both conditions.

Participants were asked not to think aloud while solving the tasks as this would interfere with the GSR data [4]. The experimenter sat in an adjacent control room.

As the GSR sensor reacts on arousal we had to get participants into a relaxed state before task solving. In both conditions, a blank screen was shown for the first four minutes while playing a relaxing piece of music. As in [4], we used the song "Weightless" by Marconi Union. After the four minutes, the website was shown automatically and task solving began. All sessions were video recorded.

Step 3: Video Review

After completing the main task (or the time was up), they relaxed four minutes listening to the same song as in the previous step. To induce cued-recall all participants (in both conditions) viewed video clips of their interaction. All wore the GSR sensor while viewing the video. This enabled us to compare correlations (hypothesis H1) and intensities (hypothesis H2) between GSR data from actual interaction and the cued-recall experience induced by the video.

Step 4: CRD & SAM

After watching the video, participants in the baseline condition had completed the experiment, i.e. we had gathered both subjective SAM ratings after each of the three subtasks and objective GSR data. For participants in the CRD condition we still needed to collect subjective SAM ratings. In CRD cues in the form of video clips were shown to participants. Participants viewed one clip at a time and gave SAM ratings for each of these to reflect their emotional experiences when the event occurred. This was done for all video clips. For each GSR peak, we showed participants a video clip of their interaction. This procedure is in line with [4].

Participants

We recruited a total of 40 participants for the study. All were university students in their 2nd or 4th semester taking the same courses in Informatics (mean age = 23, sd = 3.2). A total of 32 used the web application on a monthly or yearly basis and 8 had no prior experience with the system. Participants were distributed randomly across both experimental conditions.

Most participants (n=35) accepted to take a Big-Five personality test, cf. [15]. To rule out a potential personality bias in our study, we tested whether or not participants in the baseline and cued-recall conditions had comparable personality traits. Independent samples t-tests revealed no significant differences on any of the five personality traits between conditions (Extraversion, t=1.4, df=32, p=.2, pwr=.5; Agreeableness, t=1.5, df=32, p=.23, pwr=.8; Conscientiousness, t=1.8, df=32, p=.1, pwr=.5; Emotional stability, t=1.9, df=32, p=.1, pwr=.5; Intellect/Imagination, t=.1, df=32, p=.9, pwr=.9).

System

A web application for data dissemination was assessed in the study (www.dst.dk). The system provides public information on various kinds of national statistics from

Denmark. This includes information such as level of education, IT knowledge and skills, employment rates etc. The web application is targeted towards students that need quantitative data to support their school or university assignments. Participants thus represented actual end users.

GSR Sensor

For our study we used the Mindplace Thoughtstream GSR sensor. This measures skin resistance in kOhm between two electrodes, which are attached to the underside of the palm.

Tasks

Participants in the baseline and cued-recall conditions were asked to solve the same main task:

How many hotels and restaurants were there in Vejen [small town in Denmark] in 2012, with one person employed?

The standard procedure for the baseline condition is to let participants fill in SAM after each task. We therefore needed to introduce a set of subtasks after which SAM data could be collected. We gave participants three subtasks, which collectively led to answering the main task. In the CRD condition participants did not get interrupted after each task as in the baseline one. Therefore, to enable comparisons of SAM data between conditions, we derived the three baseline subtasks by observing participants' task solving strategies in the CRD condition. We observed that 18 of 20 participants of the CRD group used the following strategy: 1) Find an overview outlining all statistics topics, 2) Browse through potential relevant topics (few seconds per topic) and 3) Work in depth with a selected topic (several minutes) to find the answer for the main task. The strategy demonstrated by the CRD participants was reflected in the three subtasks given to their baseline counterparts (see Figure 1).

RESULTS

In the following we present our findings related to objective GSR measurements and subjective SAM ratings.

Objective Measures based on GSR data

Data from the GSR sensor was used as a real-time measure to compare emotional reactions during interaction and emotions experienced during cued-recall (video review).

GSR Data Correlations

Figure 2 shows an example of the two GSR graphs obtained from one of the participants in the CRD condition.

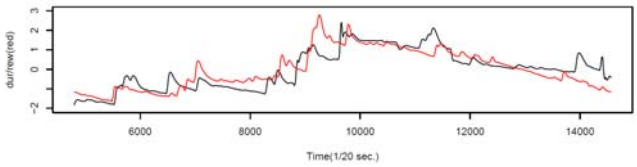

Figure 2: Example graphs (normalized) showing GSR data from one of the participants in the CRD condition. Black: GSR during interaction, red: GSR during video review.

The black graph shows GSR data during actual interaction with the system (step 2) while the red graph represents GSR data from the video review (step 3). All GSR data was normalized and all data points above ±5 standard deviations were removed. The latter was done in order to eliminate artefacts in the data that led to slopes with a steepness that is physiologically impossible to obtain. Such artefacts were typically caused by participants moving their hand with the GSR sensor attached. We removed two participants from the CRD condition as the GSR sensor failed to collect data in these instances.

By visual inspection the graphs in Figure 2 look similar as GSR peaks occur at approximately the same timestamps and they tend to increase and decrease at the same locations. We calculated the level of similarity through Pearson Product-Moment correlations (after removing outliers, based on a mean of 15977 GSR data points, SD=2700). The mean correlation coefficient was higher for participants in the CRD condition (r=.56, SD=.24, p=.01) than in the baseline condition (r=.38, SD=.28, p=.01). All identified correlations in both conditions were significant. However, the difference between conditions is not significant as revealed by an independent samples t-test (t=1.7, df=24, p=.1, pwr=.41).

The above findings suggest that there are significant within-subject correlations between GSR measurements taken while interacting with the system and afterwards when reviewing the recorded videos. This was the case in both experimental conditions. Thus, objective measures of emotions during cued-recall correlate with real-time data. This level of correlation is comparable between conditions.

GSR Data Intensity

Table 1 shows the mean GSR intensities measured in both experimental conditions.

Condition	Mean GSR intensity (SD)	
	During interaction	Video review
CRD (n=18)	36.7 (21)	37.5 (21.4)
Baseline (n=20)	46 (25.4)	48.7 (24.9)

Table 1: Mean intensities of normalized GSR data measured during interaction and during video review. "n" denotes the number of test participants.

Note that data is normalized following the formula given in [25]: Normalized GSR (i) = ((GSR(i) - GSRmin)/(GSRmax - GSRmin))*100. Using this formula, all GSR readings are between 0-100 where 0 means no skin resistance and 100 means high resistance. Low skin resistance means that test subjects' palms are very sweaty, i.e. sweat increases conductance and leads to lower resistance. This also implies a highly aroused state. Conversely, a high skin resistance means a dry surface and a more calm emotional reaction.

In the CRD condition we found a mean GSR intensity during interaction of 36.7 while the video review intensity was 37.5. In the baseline condition the GSR intensity was 46 during interaction and 48.7 during video review.

Thus, across both conditions, we generally see a lower skin resistance (higher level of arousal) during the interaction step compared to the video review step. However, the within-subjects difference is minor and non-significant as revealed by a repeated measures Wilks' Lambda test (CRD: F=.017, df=17, p=.9, pwr=.9; Baseline: F=.16, df=19, p=.7, pwr=.9).

Furthermore, we found no significant differences in intensities between conditions, neither during interaction (t=-1.22, df=35, p=.23, pwr=0.51) nor in terms of the video review (t=-1.45, df=35, p=.15, pwr=.5).

From the above we see that the GSR intensity during interaction is comparable to the intensity during the video review. This applies for within-subjects condition measures of both the CRD and baseline conditions. Additionally, the GSR intensities are comparable between conditions.

Subjective Measures based on SAM ratings

In the CRD condition, SAM ratings were taken retrospectively during step 4, which follows the CRD method (see Figure 1). In the baseline condition, SAM ratings were taken after completing each subtask (step 2). This enabled us to compare SAM data between conditions.

In the following we compare how SAM ratings developed during interaction in the baseline condition and in the CRD condition. We then compare the emotional intensities reflected in SAM ratings between conditions.

SAM Correlations

Across both conditions the level of pleasure started out high after which it gradually dropped towards the end of the interaction sequence. Thus, participants reacted more negatively towards the end of the interaction. This corresponds to the observation that only two out of 40 participants (5 %) ultimately solved the main task. As suggested in the previous literature, we captured three data points during interaction in the baseline condition (cf. [16,24]) and computed the averages over them for each of three emotional dimensions measured by SAM per participant. In the CRD condition we captured a mean of 10.1 (SD = 3.3) data points (i.e., the number of video clips to be viewed) and computed the averages of the three dimensions over the respective number of data points per participant. Pearson correlation coefficients between these data of two conditions were then calculated.

For the pleasure ratings between the CRD and baseline condition (r=.75, p=.01) was significant. In terms of arousal we found no significant correlation (r=.11, p=.76), i.e. the baseline and CRD conditions did not correlate. This is because the subjective assertions of arousal in the baseline condition fluctuated to a greater extent (higher variance, var=2.2) than in the CRD condition (var=0.9). We return to this issue in the discussion section.

In case of dominance a pattern similar to the pleasure ratings was observed, where ratings started out high followed by a steady decrease over time. As with the pleasure ratings, this development was expected since most participants did not succeed in solving the main task. This would intuitively lead to a feeling of not being in the control of the situation. In this case we did find a significant correlation between conditions (r=.72, p=.01).

SAM Intensities

Table 2 provides an overview of the mean intensities in SAM ratings between conditions. This expresses the overall intensities across the whole interaction sequence in the baseline condition (step 2) and in the whole CRD sequence of the CRD condition (step 4).

Condition	Mean SAM intensity (SD)		
	Pleasure	Arousal	Dominance
CRD (n=17)	4.5 (1.1)	4.7 (1.3)	4.5 (1.2)
Baseline (n=18)	4.1 (1.1)	5.2 (1.3)	3.8 (1.2)

Table 2: Mean intensities of SAM data between CRD and baseline conditions. "n" denotes the number of test participants. Note that three outliers have been removed.

An independent samples t-test revealed no significant difference in Pleasure between the two conditions (t=1.3, df=33, p=.27, pwr=.7). Similarly, a non-significant difference in arousal between the two conditions was found (t=-1.04, df=33, p=.31, pwr=.7). There was also no significant difference in dominance (t=1.77, df=33, p=.09, pwr=.7).

The above findings indicate that the mean intensities in SAM ratings are similar between cued-recall and the baseline conditions across all three SAM dimensions.

DISCUSSION

In the following we discuss our findings in relation to our hypotheses and point to areas of future work.

H1: Correlation of Cued-recall and Real-Time Emotions
"Retrieval of specific memories activates emotions correlated (asserted through SAM and GSR) with emotions experienced during real-time system usage".

Our findings lead us to confirm hypothesis H1. We found significant within-subjects correlations between GSR measurements taken while interacting with the system and afterwards when reviewing the recorded videos. This was the case in both experimental conditions. Thus, objective measures of emotions obtained during cued-recall correlate with data from actual interaction.

This level of correlation is comparable between the baseline and CRD conditions. However, the statistical power of comparing GSR correlations between the baseline and CRD conditions was 0.41. This indicates that there is some risk of not correctly rejecting the null hypothesis where correlations are considered different. It is known that people have individual physiological reactions with, e.g. varying levels of perspiration [29]. This explains the lowered power of these between-subjects measures. To increase the statistical power there is a need to include a higher number of participants. Yet, we stress that the within-subjects correlations were between moderate and high for this kind of study [9].

In terms of subjective measurements, we also found strong [9] significant correlations between the two conditions. This was the case for the SAM dimensions of pleasure and dominance. On the other hand, we found no significant correlation between subjective ratings of arousal between the baseline and CRD conditions. This finding can be explained by the higher level of fluctuation in the baseline condition compared to CRD. Such fluctuation can be explained by the difference in experimental settings where participants in the baseline condition were interrupted three times during interaction (once for each of the three subtasks). Note that these interruptions follow the standard method of gathering SAM data during interaction and represent an approximation of real-time measurements, see e.g. [16,24]. By breaking the interaction flow, participants can, e.g. go from an excited emotional state of trying to locate an answer in the system to a more calm state while filling in the SAM questionnaire.

The insignificant correlation in SAM arousal could also be because we compare SAM ratings between subjects rather than within subjects. Emotional experiences are individual and may differ from one participant to the next, i.e. we can expect differences between subjects. To reduce this bias we asked participants to take a personality test, which also included elements related to emotional stability. We found comparable personality traits between participants in the baseline and CRD conditions. To be more conclusive, there is a need to further increase statistical power on differences/similarities between conditions. This can be done by increasing the number of participants.

In sum, with the exception of subjective SAM ratings of arousal, all objective and subjective measurements represent significant correlations between emotions experienced in real-time and emotions experienced during cued-recall. Therefore, we verify hypothesis H1.

H2: Intensity of Cued-recall and Real-Time Assertions
"Retrieval of specific memories activates emotions with similar intensities (asserted through SAM and GSR) as emotions experienced during real-time system usage".

Results from this study confirm hypothesis H2. In terms of the objective GSR data, we found comparable within-subject intensities when comparing data obtained during interaction with data from the video review. This within-subjects comparison had a high level of statistical power (.9). We also found comparable intensities in GSR data between the baseline and CRD conditions. However, similar to the between-subjects comparison of correlation mentioned earlier, the statistical power in this respect is relatively low (.5).

Subjective data also show comparable intensities between conditions across all three SAM dimensions. In this case statistical power was high (.7). Given that all of our findings point towards comparable intensities between emotions experienced during interaction and in cued-recall, we confirm hypothesis H2.

Should we discard free-recall methods?
Our aim with this study was to examine the extent of which cued-recall (as opposed to free-call) reduces the memory bias when asserting emotions in retrospect. But why study retrospective methods rather than relying on validated methods where emotions are asserted in real-time? As mentioned in the introduction of the paper, then real-time methods have the shortcoming of only supporting summative assessments. Retrospective methods are suitable for formative purposes as they enable researchers and practitioners to gather qualitative insights without disturbing participants. Such data is relevant for directing changes in a particular design [3,32]. As an example of targeting specific design elements, CRD revealed comments like *"I didn't find a relevant menu in the statistics overview that would lead to the answer, so I had to look into several options"*.

Our findings show that emotions asserted through retrospective cued-recall provide valid measures of emotions experienced in real-time. Thus, cued-recall does not suffer from the memory bias to the same extent as methods based on free-recall.

So, do our findings mean that free-recall methods such as the Day Reconstruction Method, Experience Sampling Method and UX Curve are not useful? No, we stress that the critique raised within this paper is not an attempt to scrutinize free-recall methods. We do find these highly relevant and applicable for identifying predominant

emotions from an entire experience. They are particularly useful for uncovering emotions that stand out so strongly that participants can recall them over time. Furthermore, free-recall methods have at least two qualities: 1) Studying UX over longer periods of time, e.g. after a product has reached the market and 2) They give qualitative insights on the influential product experiences that users remember. The latter is critical in terms of consumer recommendations and product loyalty [4,20]. Yet, there is also a need for more detailed insights when conducting formative evaluations. This is critical when targeting specific elements in a design, which need to be changed [3,32].

Future Work

This study represents initial efforts in understanding the relationship between emotions experienced in real-time and emotions asserted through cued-recall. There are some limitations to this study.

The system applied as a test case is relevant for work situations, e.g. students that need statistical data for their semester projects. However, studying the validity of cued-recall on other types of systems for e.g. leisure and games would be relevant. It is also relevant to study the validity of cued-recall based on a higher number of participants with other demographic profiles.

During our work we also identified other relevant areas for future studies. We noticed that some of the GSR peaks seemed slightly displaced between real-time and cued-recall measurements. This phenomenon can be seen in Figure 2 where some of the GSR peaks during cued-recall appear before the corresponding peak obtained during actual interaction. This suggests that, when participants reviewed the video data of their own interaction, they remembered what was about to happen. Therefore they had a premature emotional reaction compared to what really happened during the actual interaction sequence. In other parts of the graph, we see the opposite, i.e. participants forgot what happened during interaction and got reminded while reviewing the video data. It would be relevant to systematically study this effect in order to understand triggers of premature and delayed displacements.

Secondly, our findings suggest that emotions asserted through cued-recall reflect the emotions experienced during interaction. But how soon after interaction should participants partake in cued-recall in order for their responses to be valid? A memory bias such as the peak-end effect could perhaps dominate cued-recall measurements after a longer period of time. So where is the potential "sweet spot"? In relation to the Experience Sampling Method, Scollon argued that researchers should consider whether differences in the data exist between timely and tardy responses [30].

CONCLUSION

In this paper we presented a systematic empirical study comparing real-time assertion of emotions and emotions asserted in retrospect through cued-recall. Findings show that emotions experienced during cued-recall are significantly correlated with emotions experienced in real-time. We also found that the intensity of emotions experienced during cued-recall is comparable to the intensity experienced in real-time. Findings are triangulated based on objective measurements of Galvanic Skin Response and subjective measurements of the Self-Assessment-Manikin questionnaire.

Thus, cued-recall does not suffer from the memory bias to the same extent as retrospective methods based on free-recall. The contribution of the study is the validation of alternative methods to assess UX that leverage the key dimension of emotions. Implications are that HCI researchers and practitioners now have a real alternative to free-recall that significantly reduces the memory bias and enables high detail measurements of emotions. This is particularly relevant in formative evaluations of UX with a need for targeting specific elements in the design space. Furthermore, methods based on cued-recall can be applied to gather rich input from participants without disturbing them during system interaction.

ACKNOWLEDGEMENTS

Removed for blind reviewing.

REFERENCES

1. Andreassi, J.L. *Psychophysiology: Human Behavior and Physiological Response*. Lawrence Erlbaum, Mahwah, 2000.
2. Bargas-Avila, J.A. and Hornbæk, K. Old wine in new bottles or novel challenges. *Proc. CHI*, ACM (2011), 2689–2698.
3. Bevan, N. Classifying and Selecting UX and Usability Measures. In E.L.-C. Law, N. Bevan, G. Christou, M. Springett and M. Larusdottir, eds., *VUUM 2008: Proceedings of the International Workshop on Meaningful Measures: Valid Useful User Experience Measurement (VUUM)*. Institute of Research in Informatics of Toulouse, Toulouse, 2008.
4. Bruun, A. and Ahm, S. Mind the Gap! Comparing Retrospective and Concurrent Ratings of Emotion in User Experience Evaluation. *15th IFIP TC13 Conference on Human-Computer Interaction (INTERACT)*, Springer-Verlag (2015).
5. Buchanan, T.W. Retrieval of Emotional Memories. *Psychological bulletin 133*, 5 (2007), 761–779.
6. Burmester, M., Mast, M., Jäger, K., and Homans, H. Valence Method for Formative Evaluation of User Experience. *Proc. DIS*, ACM (2010), 364–367.
7. Bylsma, L.M., Taylor-Clift, A., and Rottenberg, J.

Emotional reactivity to daily events in major and minor depression. *Journal of abnormal psychology 120*, 1 (2011), 155–167.

8. Cockburn, A., Quinn, P., and Gutwin, C. Examining the Peak-End Effects of Subjective Experience. *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, ACM (2015), 357–366.
9. Cohen, J. *Statistical power analysis for the behavioral sciences*. Lawrence Earlbaum Associates, Hillsdale, NJ, 1988.
10. Csikszentmihalyi, M. and Larson, R. Validity and Reliability of the Experience-Sampling Method. In *Flow and the Foundations of Positive Psychology SE - 3*. Springer Netherlands, 2014, 35–54.
11. Desmet, P. Funology. In M.A. Blythe, K. Overbeeke, A.F. Monk and P.C. Wright, eds., Kluwer Academic Publishers, Norwell, MA, USA, 2004, 111–123.
12. Desmet, P.M.A., Overbeeke, K., and Tax, S. Designing products with added emotional value; development and application of an approach for research through design. *The Design Journal 4*, 1 (2001), 32–47.
13. Diener, E. and Tay, L. Review of the Day Reconstruction Method (DRM). *Social Indicators Research 116*, 1 (2014), 255–267.
14. Forlizzi, J. and Battarbee, K. Understanding Experience in Interactive Systems. *Proc. DIS*, ACM (2004), 261–268.
15. Goldberg, L.R. The development of markers for the Big-Five factor structure. *Psychological Assessment 4*, 1 (1992), 26–42.
16. Hassenzahl, M. and 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*, 4 (2007), 429–437.
17. James, W. What Is An Emotion? *Mind os-IX*, 34 (1884), 188–205.
18. Kahneman, D., Fredrickson, B.L., Schreiber, C.A., and Redelmeier, D.A. When More Pain Is Preferred to Less: Adding a Better End. *Psychological Science 4*, 6 (1993), 401–405.
19. Kujala, S. and Miron-Shatz, T. Emotions, Experiences and Usability in Real-life Mobile Phone Use. *Proc. CHI*, ACM (2013), 1061–1070.
20. Kujala, S., Roto, V., Väänänen-Vainio-Mattila, K., Karapanos, E., and Sinnelä, A. {UX} Curve: A method for evaluating long-term user experience. *Interacting with Computers 23*, 5 (2011), 473–483.
21. Lang, P.J. Behavioral treatment and bio-behavioral assessment: computer applications. In J.B. Sidowski, J.H. Johnson and T.H. Williams, eds., *Technology in Mental Health Care Delivery Systems*. Ablex, Norwood, 1980, 119–137.
22. Lang, P.J. The emotion probe: Studies of motivation and attention. *American Psychologist 50*, 5 (1995), 372–385.
23. Mahlke, S., Minge, M., and Thüring, M. Measuring Multiple Components of Emotions in Interactive Contexts. *CHI EA*, ACM (2006), 1061–1066.
24. Mahlke, S., Minge, M., and Thüring, M. Measuring Multiple Components of Emotions in Interactive Contexts. *CHI EA*, ACM (2006), 1061–1066.
25. Mandryk, R.L. and Atkins, M.S. A Fuzzy Physiological Approach for Continuously Modeling Emotion During Interaction with Play Technologies. *Int. J. Hum.-Comput. Stud. 65*, 4 (2007), 329–347.
26. Murray, B.D., Holland, A.C., and Kensinger, E.A. Episodic Memory and Emotion. In M.D. Robinson, E.R. Watkins and E. Harmon-Jones, eds., *Handbook of Cognition and Emotion*. Guilford Press, 2013, 156–175.
27. Philippot, P., Baeyens, C., and Douilliez, C. Specifying emotional information: Regulation of emotional intensity via executive processes. *Emotion (Washington, D.C.) 6*, 4 (2006), 560–571.
28. Redelmeier, D.A. and Kahneman, D. Patients' memories of painful medical treatments: real-time and retrospective evaluations of two minimally invasive procedures. *Pain 66*, 1 (1996), 3–8.
29. Scherer, K.R. What are emotions? And how can they be measured? *Social Science Information 44*, 4 (2005), 695–729.
30. Scollon, C.N., Kim-Prieto, C., and Diener, E. Experience Sampling: Promises and Pitfalls, Strengths and Weaknesses. *Journal of Happiness Studies 4*, 1 (2003), 5–34.
31. Smith, A.P.R., Henson, R.N.A., Rugg, M.D., and Dolan, R.J. Modulation of retrieval processing reflects accuracy of emotional source memory. *Learning & Memory 12*, 5 (2005), 472–479.
32. Springett, M. Assessing user experiences within interaction: experience as a qualitative state and experience as a causal event. In E.L.-C. Law, N. Bevan, G. Christou, M. Springett and M. Larusdottir, eds., *VUUM 2008: Proceedings of the International Workshop on Meaningful Measures: Valid Useful User Experience Measurement (VUUM)*. Institute of Research in Informatics of Toulouse, Toulouse, 2008.
33. Stone, A.A., Schwartz, J.E., Neale, J.M., et al. A comparison of coping assessed by ecological momentary assessment and retrospective recall. *Journal of personality and social psychology 74*, 6 (1998), 1670–1680.
34. Ståhl, A., Höök, K., Svensson, M., Taylor, A.S., and Combetto, M. Experiencing the Affective Diary.

Personal Ubiquitous Comput. 13, 5 (2009), 365–378.

35. Thüring, M. and Mahlke, S. Usability, aesthetics and emotions in human–technology interaction. *Int. J. Psychology 42*, 4 (2007), 253–264.
36. Tähti, M. and Arhippainen, L. A Proposal for collecting Emotions and Experiences. *Interactive Experiences in HCI 2*, (2004), 195–198.
37. Ward, R.. and Marsden, P.. Physiological responses to different WEB page designs. *Int. J. Human-Computer Studies 59*, 1-2 (2003), 199–212.

The columns on the last page should be of approximately equal length.
Remove these two lines from your final version.