Do You Feel the Same? On the Robustness of Cued-Recall Debriefing for User Experience Evaluation

ANDERS BRUUN, Department of Computer Science, Aalborg University
EFFIE LAI-CHONG LAW, School of Informatics, University of Leicester
THOMAS DYHRE NIELSEN, Department of Computer Science, Aalborg University
MATTHIAS HEINTZ, School of Informatics, University of Leicester

CCS Concepts: HCI design and evaluation methods, Laboratory experiments, Empirical studies in HCI.

KEYWORDS

User Experience; Emotion; Physiological sensors; Self-Assessment Manikin; Heart Rate; Galvanic Skin Response; Peak-End Effect

1 INTRODUCTION

Retrospective think-aloud is a practical approach for evaluating user experience while avoiding interruptions in ongoing human-technology interaction. This can be done by postponing inquiries until actual interaction has ended. While retrospective inquiries are feasible to conduct, it has been argued that they may not provide valid accounts of emotional experiences that occurred during actual interaction (see e.g. [22,24] for elaborated discussions). This particularly pertains to assessing emotions, which is a central dimension in understanding user experience and interaction design [6,42]. Our research study is distinct from the work on retrospective think aloud (RTA) [47,52] by investigating the phenomenon at a deeper level. While RTA, as compared with its concurrent think aloud (CTA), was shown to produce comparable or even better results, there is a lack of explanatory insight. RTA focuses *only* on external verbalizations while ignoring underlying psycho-physiological processes. Grounded in the traditional usability paradigm, RTA typically addresses the cognitive rather than the emotional aspect of human-technology interaction. Our work attempts to fill these gaps by analyzing the relation between actual interaction and recall through the lens of emotions with physiological and self-reported data. This is not trivial due to the fleeting nature of emotions [100], which makes data collection challenging and essentially requires such data to be gathered in real time during actual interaction [24]. This is due to the gap between what we actually experience and what we can freely recall, an effect often referred to as the "memory experience gap" or "peak-end effect" [57,77,90].

Yet, emotional data on user experiences have typically been gathered on the basis of study participants having to freely recall their experiences. Within HCI studies, emotional data have largely been gathered by administering questionnaires such as the Self-Assessment-Manikin (SAM) [6]. Such recall-based subjective measures have increasingly been complemented by their real-time objective counterparts, namely psychophysiological measures gathered through, e.g. Galvanic Skin Response (GSR), Heart Rate (HR) or other sensors [22,25,44,72,73,74,103]. In experimental design, SAM is typically filled in by participants during natural breaks in the interaction flow, e.g. after completing a task [49,71] (Section 2.4). This provides more accurate reflections of emotions as compared to gathering SAM ratings at the very end of a study, given the memory-experience gap. This contrast at least holds true given that participants have to rely on free recall when expressing their emotional accounts for an interaction design.

As an alternative to free recall Omodei and McLennan suggested the Cued-Recall Debriefing (CRD) method in 1994. CRD was developed as a way of re-immersing participants into previous events through video cues with the aim of enabling them to better recall their emotional experiences [83]. More recently, CRD has been applied in HCI studies where participants have interacted with a system while wearing psychophysiological sensors that capture objective emotional data in real time. To represent cues, video data of their system interaction were also recorded [9,22,24]. Other types of cues such as screenshots and contextual cues (photos, location, nearby Bluetooth-enabled devices, usage logs etc.) have also been used for similar purposes (cf. [26,59,93,105]). Overall, it is argued that successful re-immersion through, say, video cues, whether it is the entire recording in our empirical research (Section 3) or selected clips like in [13], can

enable study participants to re-experience emotions similar to when the original event took place [83]. Some work in the domain of HCI has recently been conducted to verify this claim [9,22,25]. In this study we extend previous work by further scrutinizing the robustness of the CRD method for use in HCI. This deeper understanding of the validity of CRD is relevant as the method has potential in complementing established ethnographic approaches used to study phenomena in the wild. Rogers and Marshall [81] suggest the use of psychophysiological data when conducting such studies as it enables continuous data gathering on emotional experiences.

Overall, the main research objective driving this study is to examine the robustness of Cued-Recall Debriefing in terms of re-immersing participants to a level where emotional responses are comparable to those experienced during actual interaction. We operationalize the robustness in terms of the similarity of objective and subjective emotional responses measured during and after actual interaction with a system, i.e. if the emotional experiences between actual interaction and cued-recall are comparable, we consider CRD to be robust. To this end we applied physiological sensors measuring Galvanic Skin Response (GSR) and heart rate (HR) as well as subjective SAM ratings. Computing correlations between these data would enable us to gain further insights into the relationship between objective and subjective user experience measures.

As a retrospective approach, CRD is susceptible to memory decay that can lead to inaccurate recall of associated emotions. This inaccuracy can be more severe when the time gap between the actual interaction and CRD session is 30 minutes or longer [31,106] since other activities interfere with the memory of interaction experiences. However, the time gap is not the only factor influencing our ability to recall and re-immerse into past events. Previous studies have shown that the emotions we are currently experiencing influence how environments are perceived. The review paper by Zadra and Clore [108] highlights that fear for instance can influence low-level visual processes and sadness may change our susceptibility to visual illusions. The perceived incline of a hill as well as distance from a balcony to the ground are also influenced by emotional states. In relation to HCI, there are also indicators of particular emotional states influencing physical movement of, say, a mouse pointer, although different experiments provide varying findings on this [110]. Thus, emotions that arise between actual interaction and when entering a cued-recall debriefing session may also influence our ability to re-immerse into past events with comparable experiences.

Consequently, we investigate how the robustness of CRD can be influenced by two factors: *intervening time* and *intervening affect*. In terms of duration of intervening time, studies of the peak-end effect within psychology have shown the varied effects of different temporal gaps between emotions experienced and emotions recalled after the fact [57,91]; this finding has also been observed in HCI studies of interactive user experiences (cf. [24]). In relation to CRD, we examine the level of correlation between emotions experienced during actual interaction and those during cued recall with different lengths of intervening time between these two phases.

To further evaluate the robustness of CRD, we utilized affect priming to induce particular emotional states in participants towards the end of the intervening time period, i.e. after the actual interaction and prior to the cued recall session. The above arguments lead us to formulate the following two research questions (RQs):

RQ1: To what extent does **intervening time** impact correlations between emotional responses measured during actual interaction and those during cued recall?

RQ2: To what extent does **intervening affect** impact correlations between emotional responses measured during actual interaction and those during cued recall?

To address these RQs, we conducted two empirical studies with different instantiations of the two factors, with the intervening time ranging from 0 hour up to 24 hours and the intervening affect from negative-valence-low-arousal to positive-valence-high-arousal (Section 3). Altogether 100 participants from two countries (39 from Denmark and 61 from the UK) were involved, and valid sensor-based data from 81 of them were analysed (Section 3.4). Note that the involvement of the two experimental sites was to demonstrate the replicability of the methodological framework (with slight modifications) in different contexts, but without any intention of drawing any cross-country comparisons.

Overall, the main contribution of this paper is threefold:

- Demonstrating the robustness of the Cued-Recall Debriefing method against the intervening time and intervening affect
- Proposing the methodological framework for understanding the role of memory-experience gap in UX evaluation;
- Providing further empirical data for evaluating the relationship between subjective and objective emotional measures of user experience.

Subsequently, we outline related work in Section 2, followed by a detailed description of our experimental study design in Section 3. Next, we report on empirical results in Section 4, which are further discussed in Section 5. Finally, we reflect on the limitations of our work and draw conclusions, including our future work, in Section 6.

2 RELATED WORK

In this section, we present our reviews on four strands of existing work relevant to our study. Although our study specifically emphasizes the key UX dimension of emotions, it is related to a strand of well-established work dealing with usability evaluation methods, including Retrospective Think-Aloud (RTA). The three other strands are related to research focusing on CRD, emotion and memory as well as self-reporting of emotions. Note that CRD (Section 2.2) is essentially a type of TA method, and, more specifically, a variant of RTA (Section 2.1). CRD, as the name implies, is characterized by utilizing cues to support participants in recalling past events. Variations of CRD are defined by the type of cue (e.g. selected video clips vs. whole video) and the length of time gap (cf. no gap), which is related to memory biases (Section 2.3). Typically, for user-based studies, apart from TA, other sources of data relevant for our study are collected to triangulate findings, e.g. psychophysiological data and self-reported emotional responses (Section 2.4).

2.1 Think-Aloud (TA) Methods; Concurrent vs. Retrospective

Since think-aloud (TA) methods were introduced to the field of HCI in the early 1980s [67], they have become de facto standards for usability testing [14,51,80]. In the span of 40 years, variants of the original TA [39] have been developed, which fall into two major types: concurrent (CTA) and retrospective (RTA). In case of CTA participants are asked to verbalize their thoughts when interacting with a system to complete given tasks whereas in case of RTA participants are asked to first perform the tasks in silence and then make a verbal report on the interaction, typically immediately but reporting can also be delayed by hours or even days.

Both CTA and RTA have their respective strengths and limitations. The main argument for CTA is that real-time verbalizations reflect truly cognitive processes underlying the actual interaction with the system under evaluation. However, arguments against CTA are that it can be unnatural or even uncomfortable for participants to verbalize, especially when the task concerned is cognitively demanding, and that it can alter participants' thought processes, interfere with the primary task and thus undermine the validity of the verbal data (cf. the notion of reactivity, [96]). In contrast, RTA is recognized for providing deeper insights into the reasons behind behavioral and emotional responses to the interaction. Drawbacks, however, are that RTA is prone to memory biases, omission (forgetting) and commission (fabrication) (cf. the notion of non-veridicality, [96]) and that RTA prolongs a study session.

To facilitate recall in RTA, cues are normally used, but it can also be unaided (i.e. uncued free recall). RTA is also known as "retrospective debriefing" [107], "cued retrospective reporting" [45], "Think Afters" [20], "cued recall debrief" (CRD) [22] (Section 2.2), and some other terms. Cues can range from a plain screen video of actual interaction to an augmented screen video overlaid with gaze scanpaths. Indeed, built upon the pioneering work of Russo [97], eye-fixations have increasingly been used as cues to stimulate recall [45,54,99]. In Study 2 of our research (Section 3.3) we have adopted this promising approach.

In the context of usability evaluation quite a few studies have been conducted to compare CTA and RTA in terms of quantitative and qualitative outcomes such as number and severity of usability problems (UPs), task performance, and subjective perception of test session (e.g., mental load, comfort). A survey of the

related work (Appendix A), albeit not exhaustive, suggests that the relative cost-effectiveness of the two TA methods remained inconclusive. Despite Ericsson and Simon's [40] theoretical defense against reactivity engendered by CTA, their stance has been challenged by empirical evidence cumulated over years [46,54,85], though counter results did exist [2]. On the contrary, the pessimistic view of RTA expressed by Ericsson and Simon [40] has also been queried. While some degree of non-veridicality has been identified, it could be mitigated with the use of powerful cues such as increasingly accessible gaze scanpaths.

Overall, the benefits of mitigating reactivity and gaining deeper insights into usability issues can well justify the broader use of RTA, especially for cognitively and emotionally intensive tasks such as games.

The above review indicates that think-aloud methods have evolved into different types in the last four decades. Fig. 1 illustrates the major types. The red texted types are variants of CRD, which we elaborate in the subsequent section. Selected Cues are typically video clips with points of interest extracted from a full recording, which is used for the case of Whole Cues.

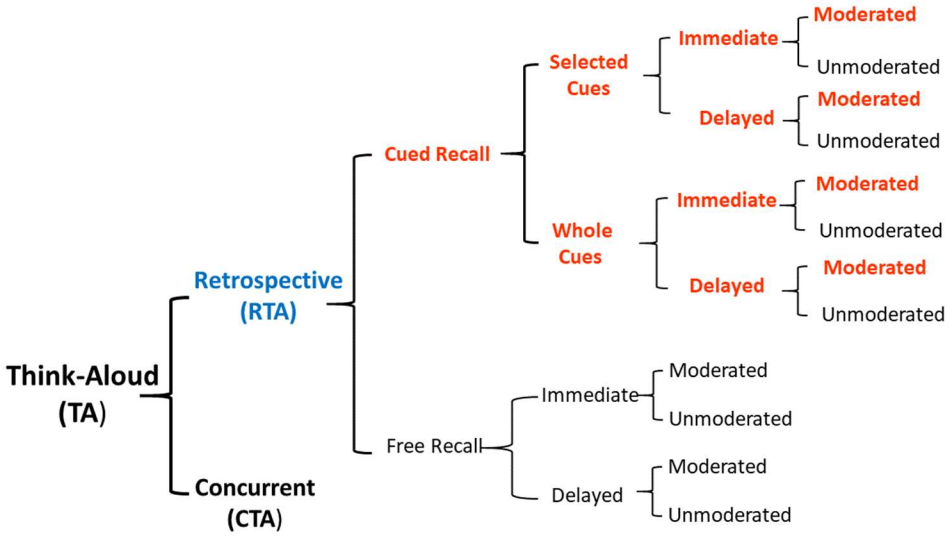

Fig. 1. The classification of different think-aloud methods. (NB: the term "moderated" is built upon [51], but the role of a moderator is not only to issue prompts to remind a participant to think aloud but also to inquire a participant to comment on points of interest).

2.2 Cued-Recall Debriefing (CRD)

Cued-Recall Debriefing is an evaluation method that attempts to re-immerse participants into past experiences. CRD is a situated recall method based on the work of Omodei and McLennan [83], who challenged the intrusive and disruptive nature of known techniques, such as think-aloud and task interruption, for studying individuals' decision making. Omodei and McLennan made use of head mounted cameras to achieve this in a field study of firefighters [83]. They used the recordings as cues during a debriefing session, which was performed within 60 minutes after completing the firefighting drills. In Omodei and McLennan's work it is relevant to notice the importance of showing data with a first-person perspective. This explains why the participants wore head-mounted cameras while also having audio recordings of footfall, breathing, and spontaneous vocalization. This is needed in order to increase the level of re-immersion [83]. In practice CRD is conducted by showing the recordings to participants while they communicate with a facilitator or moderator, such as by thinking aloud and answering or rating questions.

CRD has also been used with screen-recording and eye-tracking as cues (e.g. [45]), where it was compared to think-aloud and free recall. In that study a set of circuit board problems were presented in a

software application, which participants had to solve. Samples were taken in the form of the participants' speaking, using a code system that captured different aspects of the problem-solving process. It was found that free recall captured less actions and considerations than CRD, which in turn captured less than think-aloud [45]. Information about intervening events during time delays between the actual interaction sequence and CRD was not specified.

An example of CRD used with a long delay is Russell and Oren [93]. They logged 8 participants' browser search sessions (screenshots) for more than 6 days. After the logging, they would select three search sessions from day 2, 4, and 6. The participants would then be shown a cue in the form of a screenshot and asked about what happened next and were asked to answer if they were "reasonably confident". If they were unable to recall, the next screenshot from the same search would be shown until the test participants recalled correctly. The participants were able to accurately recall searches from two days ago. The amount of cues needed to recall confidently increased as the number of days between the searches and the recall session increased. This indicates that, while the CRD was successful, it may still be subjected to the undesirable effects of memory biases.

The validity of CRD has been tested by Bentley et al. [9], by examining whether or not CRD could successfully elicit 'true' affective information. They had ten participants play through two game sessions, both immediately followed by CRD. Heart Rate (HR), skin perfusion, and breathing rate were recorded during game play and used to confirm the validity of the comments elicited through CRD, giving more representative results. During the debrief session, they identified positive and negative affective experiences, which they found to be visible in the physiological patterns as significant increases in HR and skin perfusion variability. Neutral affect experiences did not show any changes. Bentley et al. [9] also mentioned that physiological measurements in combination with CRD enabled identification of uncommented affective responses. Such measurements can be used as cues to improve the debrief data collection by prompting participants to retrieve information about what happened at particular points of interest (e.g. spikes or troughs of a waveform).

Psychophysiological measurements were also used in a more recent study by Bruun and Ahm [24], who raised concern about the reliability of assessing the emotional dimension of UX retrospectively. For their experiment two versions of a system were created, one with seeded usability problems and one without. While interacting with the system, GSR measurements were collected to find points of interest in the video recordings, which were used as cues for CRD. Immediately after interaction the test participants were asked to rate their overall emotional state using SAM [24]. The output of the GSR sensor was presented as curves with peaks indicating aroused emotional states during actual interaction, which were related to the interactive experiences of the participants. In CRD the researchers identified peaks in the GSR data and played back the associated video clips (cues) to participants, who were asked to comment on the clips and rate their recollected emotions using SAM. Comparing the ratings from the overall emotional state with the averaged CRD ratings, they found significant differences in ratings in the seeded version, but not so in the unseeded version. This also confirms the peak-end rule in that the negative experience created a larger difference in ratings than the positive. This is in line with [7], which also found indications of a larger memory-experience gap when experiencing negative stimuli compared to positive stimuli.

2.3 Bi-Directional Relation of Emotions and Memory

Studies in psychology have dealt with the relationship between emotions and memory encoding and decoding. Studies have shown that emotionally exciting experiences with high levels of arousal enhance vividness of memories [79]. The study by Ochsner [81] showed that arousing emotional experiences, particularly of negative valence, were more prevalent in memory than experiences of positive valence and lower arousal levels. Thus, emotions impact memory *encoding*.

The memory *decoding* process is also critical for our study, since cued-recall debriefing is a retrospective method for eliciting emotions, which are related to prior interaction experiences. These cues are given through video clips with the guiding principle that participants can be re-immersed to an extent where they re-experience similar emotions as they did during actual interaction. While emotions impact encoding, decoding from memory can conversely lead to emotional reactions. Buchanan [27] suggested that exposure

to reminders of emotional events elicits brain activity similar to that taking place during the original event. Retrieving an emotional event from memory may occur through exposure to specific and complete reminders, but also through partial reminders initiating memory decoding processes. Studies from neurology have shown that cued-recall activates the amygdala and the medial pre-frontal cortex in ways very similar to the activation seen during the original exposure of an emotional event [27]. The notion that affective experiences can serve as cues for memory decoding also forms the basis of affect priming theory [15,27]. Researchers in psychology have applied affect priming to successfully induce emotional states in participants. Affect priming has been widely used as a regulatory mechanism where researchers have utilized cues on for instance autobiographical events such as "when I got married" [79]. Specific event cues such as "my math exam", in contrast to more generic events, lead to relatively higher emotional intensities once recollected [86].

An aspect to consider is that partial cues are more common than specific and complete reminders [27]. This is critical for our present study on measuring correlations between emotions experienced during original exposure to events and emotions experienced during cued-recall debriefing. In using CRD we should expect participants to elicit emotional reactions while viewing video cues before specific reminders occur, i.e. even before seeing the entirety of specific events that led to emotional reactions. As an example, participants viewing their own interactions in retrospect would retrieve memories of particular experiences moments before the specific events actually show up in the video cues. This also explains the displacement between the GSR graphs in [22] where the responses in original exposure were compared to those during recall. The study in [22] adopted a similar approach as this current study in validating the CRD method for use in HCI. In [22], these GSR graphs had similar patterns, yet displaced in time. Thus, while receiving cues from the original experience, participants may either: (a) Through partial cues anticipate what is about to happen in the clip (e.g. when pressing a particular button and the system crashes) or (b) Not anticipating what happened, but remembering the episode once the cue is specific and complete. Case (a) should be more common to observe than (b) [27], and will lead to premature emotional responses since participants remember what is about to happen.

A key point here is that we should **not** expect graphs of physiological signals obtained through actual interaction (original exposure) and retrospective viewing (cued-recall) to be directly comparable on the exact same temporal axis. Rather, in considering correlations, there is a need for **calibrating the temporal axes** by shifting the axes in order to obtain a firmer basis for comparison. In this respect it is of course also relevant to examine the level of axis shifting necessary to obtain optimal correlations between graphs.

Furthermore, Murray and colleagues mention that more research is needed in order to examine the effectiveness of cued-recall in inducing emotional reactions [79]. That position is supported by Smith et al. [104] stating the retrieval of emotional memories is studied to a limited extent.

2.4 Self-reported Emotional Measurement

In the field of HCI, measuring emotions with psychophysiological data has a relatively shorter history than measuring emotions with self-reported scales. The former is typically referred to as an objective approach whereas the latter as a subjective one. Methodologically, it is always advisable to employ both approaches in research studies to triangulate empirical findings and thus strengthen the conclusions to be drawn. Nonetheless, it is not uncommon that objective and subjective emotional measures are not significantly correlated [13]. This may be related to the fact that the number of data points captured by different sensors continuously is much larger than that self-reported by participants, whose user experience would be undesirably disrupted when their interaction with a digital artefact was interrupted too often to report on their emotions.

While such a contrast in the number of data points is a measurement issue, the more fundamental conceptual issue is the intricate relationship between the psychological and physiological domains [29]. According to [29], a formal specification of psychological elements as a function of physiological elements is only possible in two types of relations: one-to-one and many-to-one. Whereas one-to-one relations are intuitive and neat to analyse, they are rare and hard to prove, because there can be other psychological elements apart from the one of interest (e.g. arousal) that lead to the same physiological response (e.g.

increase in skin conductance). Hence, many-to-one relations are more probable. However, [29] state that we cannot assure whether a physiological response is a marker of a psychological state or a concomitant (or invariant) of that state, even though associated changes between the two are observed. This line of reasoning can explicate some of the relationships between the two types of measures as reported in the related studies (e.g. [21,65,98]), because participants may respond physiologically to an experience that is not assessed with a specific psychological instrument such as a questionnaire (e.g., Self-Assessment Manikin [17]). Nonetheless, in acknowledging the high complexity of specifying functional mappings between the two domains, the current conceptualizations of their logical relations [29], while entailing further empirical evidence, lay a strong foundation for psychophysiological inferences about behavioural processes.

Furthermore, an ongoing debate in emotion research is whether to conceptualize and measure emotions as distinct states (categories) or relative points along certain dimensions. According to the distinct-state approach, each emotion should be examined as unique [55]. For instance, using Scherer's [100] component model of emotion, fear – one of the so-called basic emotions [36] - is associated with a distinct pattern of cognitive appraisal, experiential quality, physiological response, motor expression and behaviour tendency. The major issue lies with the distinct-state approach is that there are obvious overlaps and resemblances across states. The dimensional approach is the alternative, which involves identifying basic dimensions that account for the similarities and differences among emotional states. This approach was rooted in the work of Wundt from 1912 (cited in [35]), which was further explored by Osgood in the 1950s. Accordingly, three orthogonal dimensions were identified, including (a) evaluative or valence (pleasure – displeasure); (b) potency (dominance or control); (c) activity or arousal or activation. Osgood's three-dimensional model evolved into Semantic Differential Measures of Emotional States [94,95] and Self-Assessment Manikin (SAM) [18]. Subsequently, Russell argued to drop the dimension of dominance, which should be considered as antecedent or consequent of emotion rather than an emotion per se. However, some researchers even argued for the fourth dimension (i.e. predictability; [41]). While the two dimensions of valence and arousal structure proved robust and reliable in general, there remain some terminological confusions as different researchers refer to the same constructs with different terms (cf. the review in [35].

SAM is a widely used pictorial tool for measuring emotions in HCI (Fig. 2). The underlying assumption in using SAM is that individuals are the best source of information on their emotional experiences [70]. SAM consists of pictures of manikins in a scale for each of the dimensions: valence, arousal and dominance. SAM scales range from a smiling, happy figure to a frowning, unhappy figure for valence and from a wide-awake, excited figure to sleepy, relaxed figure for arousal. The dominance dimension in SAM represents control with changes in size of the manikins, where a large figure implies maximum control in the situation. With the argument that dominance is inherently cognitive rather than affective, some researchers exclude this dimension when administering SAM. However, in accordance with the cognitive appraisal theory of emotion [37,38] and the component-model of emotion [100], it is justifiable to include dominance as a constitute dimension of emotion.

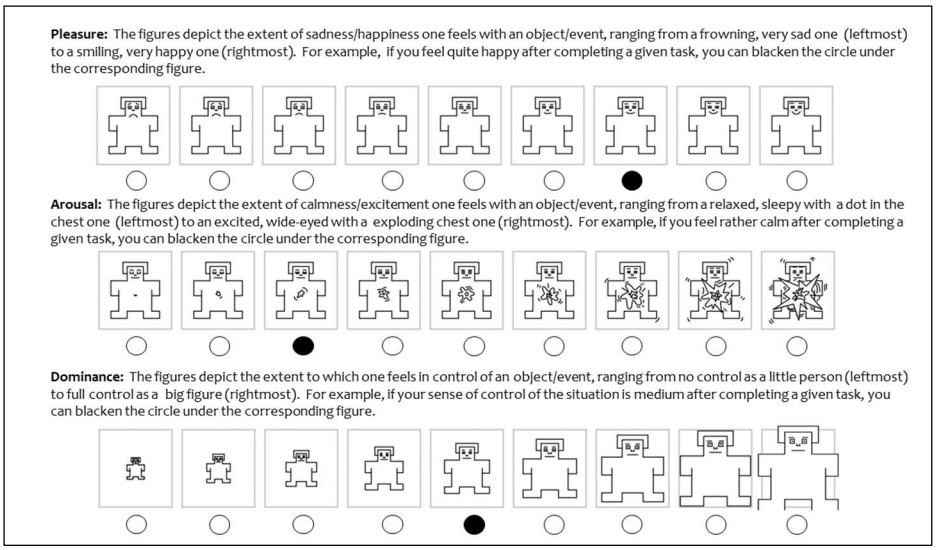

Fig. 2. Self-Assessment Manikin (SAM) with three dimensions - Pleasure, Arousal and Dominance – being measured with 9-point scales (adapted from [18]).

3 METHOD

Two sets of laboratory-based experiments were conducted in two countries, Denmark and the UK, which are designated as Study 1 and Study 2 in their chronological order. Study 2 was essentially a replication of Study 1, following the same research protocol but examining different levels of the two key variables – intervening time and intervening affect. An additional feature in Study 2 was the use of gaze data derived from an eye-tracker as additional visual cues to support cued-recall. Using eye-tracking scanpaths as augmented cues has been proven effective in previous studies (Section 2.1), and they were particularly relevant to Study 2 because of the prolonged 24-hour delay.

Subsequently, we describe the system (an email client seeded with usability problems), physiological sensors, procedures, and materials used in the two studies. Wherever appropriate, we provide separate descriptions for Study 1 and Study 2 (e.g., participants). All the procedures and equipment were pilot tested prior to the actual experimentation.

Nonetheless, a caveat is that our research does *not* aim to demonstrate the impact of CRD on the quality of think-aloud verbal data (Section 2.1). In our study, talking was rendered impractical when the sensors were being attached to participants lest the psychophysiological signals would be corrupted. Retrospective verbalizations could only have been made when participants would view the video playback for the second time after the sensors had been detached. However, such repeated exposure to the same video would likely lead to the *extraneous learning effect* that would not allow us to make any conclusive claim on the quality of verbalizations.

Instead, we aim to evaluate empirically the effectiveness of CRD by studying the extent to which participants experience similar emotions during cued-recall compared to actual interaction, and therefore are able to accurately recall and reflect after the fact. In addition, we aim to demonstrate the flexibility of CRD which can be delayed by up to 24 hours without any consequential impact on emotional recall. This can help mitigate the fatigue effect when participants are asked to carry out CRD immediately after actual interaction, especially a long evaluation session.

3.1 System

As this research work focused on emotional experiences, we needed a system capable of inducing emotional responses in users. Previous studies have shown that negative stimuli elicit stronger responses than positive stimuli [24]. To this end we developed an email client seeded with usability problems. The rationale is that people in general are familiar with an email client and have expectations how it typically works. A non-usable email client with awkward behaviours can lead to mismatched or disconfirmed expectations [12], eliciting frustration and surprise in its users.

3.1.1 Email Client

The system was designed with a set of functionalities typical of an email client, such as a list of contact persons, file attachment, making drafts and mailing options. Using the email client, participants had to complete 11 tasks, 7 of which included seeded usability problems and the remaining 4 were designed following best practices according to established principles obtained from [10,88] (Table 1).

To prevent the ordering effect, the sequence of the tasks was randomized with the exception of tasks T1 and T2, which were always fixed in order as the first two tasks so that the participants could familiarize themselves with the basic functionality of the email client. Fig. 3 shows a screenshot of the developed email client. The bottom right shows a small window with the task instructions along with a red and a green button. The red button was selected if participants could not solve the task and wanted to skip it. The green button was chosen when participants believed they had solved the task.

Fig. 3. The email client seeded with usability problems.

Table 1. List of eleven tasks; those marked with * (T5 to T11) are seeded with usability problems.

Task	Name	Description
T1	Send an email	Send an email with any text to two contacts
T2	Reply to an email	Reply to an email
T3	Save a draft	Create an email and save it as a draft
T4	Write and delete an email	Write an email containing any text, save it a draft and delete it
T5*	Add attachment	Add an attachment to an email *Seeded usability problem: When adding an attachment to an email, the program freezes for 2 seconds during the first three attempts.
T6*	Add contact	Add a new contact to the contacts catalogue *Seeded usability problem: Upon pressing the "Add Contact" button, it fails to respond the first three times
T7*	Send a Draft	Find a draft, either by creating an email and drafting it or selecting a pre-created draft, and send it *Seeded usability problem: In attempting to open a draft to send, it throws an exception, effectively blocking the access
T8*	Create a draft	Create a draft with the body: "I got £3.50, £5, and £2 from them." *Seeded usability problem: When attempting to write a text containing special characters, the keyboard layout changed to a different language setting, making the characters unavailable
T9*	Write an email	Create an email with the body: "Hi, my name is x and I am participating in a usability test" *Seeded usability problem: While typing an email, the caret randomly altered its location, making it difficult to write sentences without typing errors
T10*	Remove contact	Remove a specific contact from the contacts catalogue *Seeded usability problem: When attempting to remove a contact from the contact list, the contact was not removed and the list turned black
T11*	Write an email 2	Write an email with the body text: "Hello, I am having a birthday party 10 days from now, and this is your invitation." *Seeded usability problem: In attempting to write a new mail, it results in a simulation of the "Microsoft Windows Not Responding" window

3.2 Hardware for Physiological Measures

To measure experienced emotions in real time we relied on two physiological sensors: Galvanic Skin Response (GSR) and Heart Rate (HR), which are commonly used in the research on measuring emotions (see e.g. [9,24,68,71,87]). GSR (or electrodermal activity, EDA) has been one of the most widely used response systems in the history of psychophysiology [33], given its ease of measurement and sensitivity to psychological states and processes. Experimental treatments that can induce changes in GSR typically induce changes in HR as well; measuring both is a usual practice. While we considered other types of physiological measures, such as facial expression with EMG and brain activity with EEG, factors such as intrusiveness (e.g. electrodes on face) and high costs (e.g. fMRI) were hindering. In fact, one key advantage of GSR and HR measures is the intuitiveness and affordability of the instruments involved.

For GSR we used a Mindplace ThoughtStream [76], which measures the electrical resistance on the skin surface. This sensor returns a skin resistance value measured in MOhms and has a measuring frequency of

50ms. For HR we used an Arduino Mega 2560 with a Pulse-Sensor [69], that specifically measured Beats Per Minute (BPM) with a frequency of 20ms.

Participants worked on a laptop running Windows 10, using an external mouse and keyboard to avoid static electricity and heat from the laptop. Such factors were discovered to affect sensors during pilot testing.

Additionally, in Study 2 a desktop Tobii T-120 eye-tracker was used to capture participants' gaze data, which were displayed as enriched visual cues to facilitate recall. For instance, the scanpaths shown in Fig. 4 were automatically generated by the eye-tracker to be part of the video playback. We did not analyse any eye-tracking data as emotional responses.

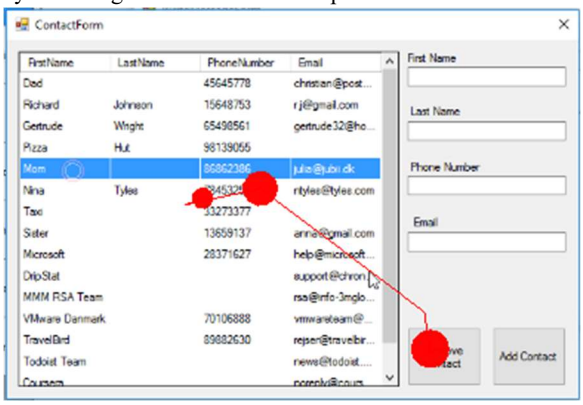

Fig. 4. Example illustrating a scanpath.

3.3 Experimental Conditions

Study 1 and Study 2 investigated the two main variables with overlapping as well as distinct levels (see Table 2 for an overview). Note that there is no intention to merge the data from the two studies, considering different contextual and personal factors that might influence participants' responses.

3.3.1 Intervening Time

In our research, we examine the level of correlation between emotions experienced during the actual interaction and the cued recall session with different lengths of time gap between these evaluation phases. To establish a baseline group, we set a condition with no delay between the actual interaction and the cued-recall (0 minute intervening time, see Table 2). Previous studies describe 20-30 minutes as the tipping point after which memory decay becomes too severe for free recall, cf. [31,106]. Hence, we introduced a delay of 30 minutes. We additionally included longer intervening delays, including a group with 60-minute delay (Study 1) and another with a 24-hour delay (Study 2). This was done in order to observe the effects of extending beyond the tipping point.

Table 2. An overview of the levels of the two key variables in the two studies

	Intervening Time	Intervening Affect
Study 1	0-minute (n=12)	N/A
	30-minute (n=12)	Negative valence / high arousal Positive valence / high arousal Neutral valence / low arousal
	60-minute (n=12)	Negative valence / high arousal Positive valence / high arousal Neutral valence / low arousal
Study 2	0-minute (n=13)	N/A
	60-minute (n=13)	Neutral valence / high arousal Neutral valance / low arousal
	24-hour (n=19)	Neutral valence / high arousal Neutral valance / low arousal

3.3.2 Intervening Affect

During the time that passes between the actual interaction and cued- recall session, participants may be exposed to different events that put them in an affective state that potentially influences their perception of the cues to be presented. Studies have shown that emotions influence how environments are perceived [108,110]. We aimed to control for the emotional experiences that participants would experience during the intervening time. For instance, a participant during the 60-minute intervening time played some video games and became excited. The excitement might influence the extent to which he could recall his emotional response when viewing the video in the cued recall session. We regulated such an influence through affect priming in which we induced affective states using the International Affective Picture System (IAPS) [17]. Studies have shown that priming images affects subsequent assessments where, e.g. the time needed to evaluate a target stimuli is significantly shorter when the prime and target are affectively congruent [50]. Avero and Calvo also note that affect priming with images is "a relatively robust phenomenon" [5]. The IAPS used for this study consists of approximately 1200 images with varying motifs, each with an associated value of valence and arousal. This type of stimuli was chosen as it is well documented and used extensively to induce emotional reactions [18,63]. Five clusters of stimuli (cf. Table 2), with each consisting of 15 images, were selected. Fig. 5 illustrates their distributions according to the two-dimensional model of arousal and valence. The images were presented to

the participants when they returned to the lab for the cued recall session before viewing the video recording.

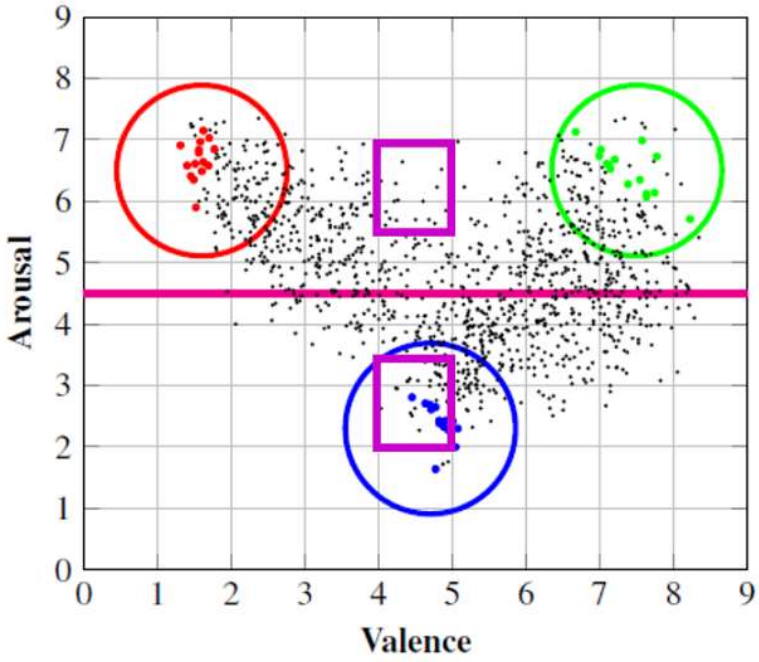

Fig. 5. A plot of the IAPS images distributed on the arousal/valence dimensions. Circles illustrate IAPS images used in Study 1 (Red circle: positive-high arousal; Green circle: negative-high arousal; Blue circle: neutral-low arousal. Rectangles illustrate IAPS images used in Study 2 (Upper rectangle: neutral valence with high arousal; Lower rectangle: neutral valence with low arousal).

Arguably we could measure the affective state without affect priming and presenting the IAPS images, however, the spread of affective states could be very broad in contrast to the selected states induced. This would limit our abilities to explain and discuss our observations.

3.4 Participants

In the following we outline participant details for Study 1 and Study 2.

- **Study 1:** 39 participants, all computer science students, signed up for the experiment on a voluntary basis. Due to sensor failure we had to discard data sets from three participants resulting in a total of 36 participants for the study (17 female / 19 male), mean age 22.85 years (SD=2.68).
- **Study 2:** 61 participants signed up for the study. Due to sensor failure, the physiological data of the first 16 participants were corrupted and discarded. Nonetheless, as each participant followed properly the experimental procedure and completed the SAM accordingly (Section 2.4; Fig. 2), we analyzed all 61 sets of SAM data and the valid 45 sets of physiological data (22 female / 23 male), mean age 27.4 years (SD: 3.12). Participants were primarily students majoring in different subjects, such as informatics, engineering, and social science.

In both studies, participants took a Big Five personality test [8,56]. Results showed no significant differences among the groups of participants. None of them had prior knowledge about the purpose of the study or the design of the email client, but were debriefed after the session.

3.5 Procedure

In the following we outline the three phases of our research protocol (Fig. 6).

3.5.1 Phase 1 - Initial interaction with the email client

As the participants arrived at the usability lab, they were informed about the agenda of the experimental session, but at a level generic enough to not reveal its real purpose. We told participants that they were to try out a new email client and that we aimed to assess its robustness and the user experience. Upon approval of the agenda, participants signed a consent form and completed a questionnaire with demographic data such as name, age and current occupation. Then, they were asked to wear the sensors and informed about how they functioned and what they measured. Next, participants were asked to solve the set of 11 tasks with the email client (see Table 1), including typical ones such as adding/removing a contact, sending/writing/deleting an email. They had to press a green or red button, confirming that the task was completed or that they were unable to complete it (see Fig. 3, bottom right). To lower the risk of faulty sensor readings, participants were instructed to limit movement during interaction, and to use only one hand when interacting with the keyboard and mouse. Sensors were attached to participants' non-dominant hand. The test started with a 3 minute resting period to establish a baseline for the sensors.

During the test, the screen and participants' interaction with the email client were recorded using screen capture software. After the end of the last task, all recordings were stopped and participants in the 0-minute intervening time condition moved directly to Phase 3 (cued-recall). All other participants had the sensors removed after which they were free to choose whether to stay in one of the waiting rooms nearby or leave the lab premises (most relevant for the 60 minute and 24 hour conditions). Participants would return to the lab at the agreed time.

3.5.2 Phase 2 – Intervening Affect

Once participants returned to the lab at an agreed upon time, they were led into one of the empty waiting rooms. There they were asked to sit and rest for three minutes before working on a task as part of the affect priming. The task started with displayed emotional stimuli in the form of a series of 15 IAPS images. Each picture was shown for 20 seconds. Participants who received positive affect priming would view IAPS images of positive valence only (i.e. the set marked by a red circle in Fig. 5), so and so forth.

3.5.3 Phase 3 - Cued-recall with sensors

Participants were asked to watch the video recording of their initial interactions with the email client from Phase 1. They were again asked to remain as motionless as possible and relax for the first 3 minutes to obtain a baseline measure of the physiological data. Participants that had been through the intervening affect in Phase 2 were re-equipped with sensors before the video was started.

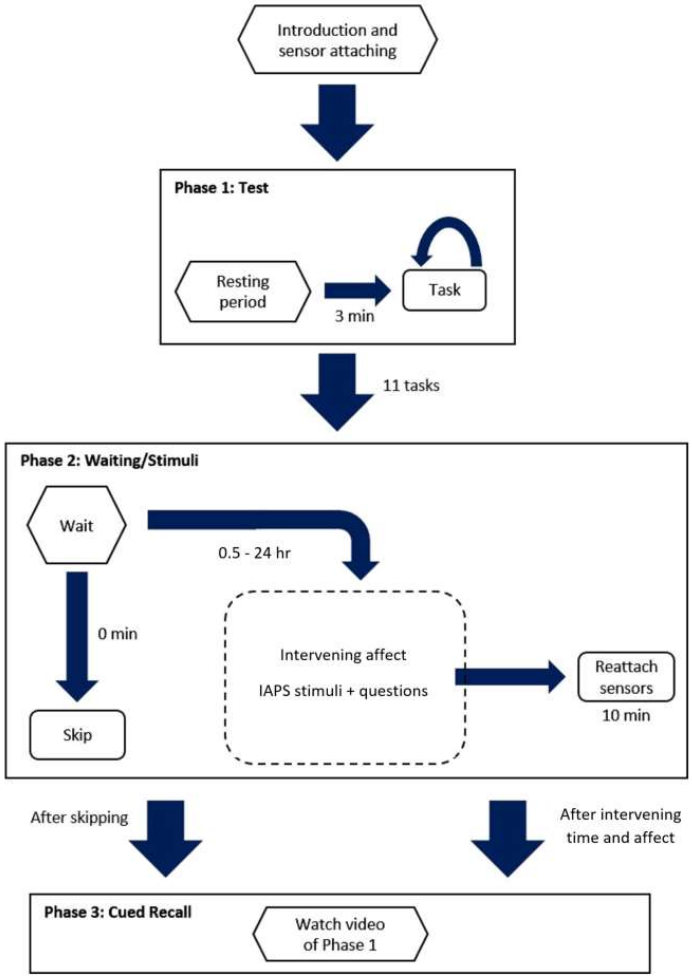

Fig. 6. The research protocol consisting of 3 main phases. Hexagons indicate start in each phase.

3.6 SAM

To complement the objective psycho-physiological data, self-reported subjective measures of emotional responses were collected. However, the main metrics we use to answer the research questions are mainly based on the real-time physiological data obtained in Phases 1 and 3 Fig. 6. These provide a considerable higher number of data points to calculate correlations. We therefore refer to Appendix C to view the SAM data results and analyses.

The instrument SAM (Self-Assessment Manikin) [18] was employed to complement the real-time physiological data, given its established psychometric properties. A 9-point scale for each of the three dimensions – Valence/Pleasure, Activation/Arousal, and Control/Dominance – was used (Section 2.4). As a pictorial emotional measurement tool, SAM was found to be applicable cross-culturally [28,78]. Furthermore, as the measure was taken after each task, to minimize cognitive load, the instrument used needed to be simple. Note, however, the post-task SAM data for Phase 1 and Phase 3 were only collected in Study 2.

3.7 Data Processing

The sample distribution in our data was non-uniform, and data from the actual interaction and cued-recall did not always align fully. To account for this during analysis, the data from each sensor was processed in 4 steps:

1. Synchronizing data from cued-recall and test based on screen capture footage
2. Artefact removal such as abnormal variations that are physically impossible to obtain
3. Splitting data into tasks
4. Account for missing data through hole filtering

3.7.1 Synchronization

The data from each sensor was synchronized using the screen capture footage of cued-recall. It was synchronized by discounting data from the cued-recall data set, if the data point was collected before the start of the initial resting period. This was done to compensate for delays in showing the video during cued-recall.

3.7.2 Artefact Removal

The removal of artefacts was treated differently for each sensor. For the GSR sensor, a moving median filter with a window size of 25 samples was applied. Artefacts were removed from the HR data by only considering samples where a heartbeat was measured. The GSR and HR data were pre-processed by replacing outliers with values corresponding to either the 1st or the 99th percentile of the data.

3.7.3 Missing Data Filtering

A sensor failing to record data manifests itself as missing data in either the actual interaction part or the cued-recall part of a data set. To account for this, when a period of missing data was present after synchronization, data from the same period in the other part of the data set was removed. This is illustrated in Fig. 7.

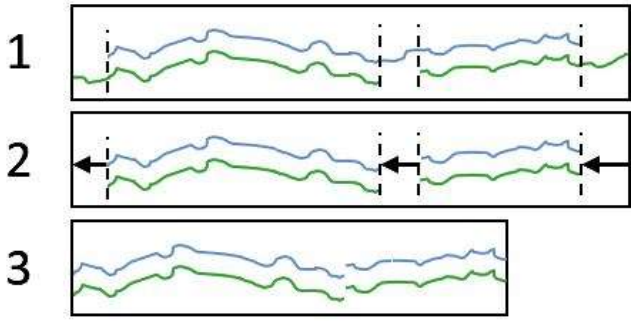

Fig. 7. The blue and green lines are data from actual interaction and cued-recall, respectively. The holes are first identified, then all data within the holes are removed. Finally, the result is kept as the filtered data set.

3.7.4 Time Shifting

During cued recall, as mentioned in Section 2.3, participants may experience either (a) partial cues making them anticipate what is about to happen in the clip or (b) remembering the episode once the cue is specific and complete, i.e. not anticipating what happened. Therefore, we should not expect graphs of physiological signals from original exposure and cued recall to be directly comparable on the exact same temporal axis. There is a need for calibrating the temporal axes, that is, shifting these axes in order to obtain an aligned basis for comparison. In this study, we shifted the relative positions of the axes in increments of 0.01 seconds up to a total of ±20 seconds. We shifted the graph obtained through cued-recall in relation to the graph from actual interaction. For instance, a shift level of -0.01 means that the axis from cued recall was shifted leftward in relation to the axis from actual interaction. A positive shift mean rightward shifting of the cued recall graph. For each ±0.01 shift we calculated correlation. Once correlations for all 0.01 increments within

the ±20 seconds interval were calculated, we located the maximum (highest) correlation level along with the corresponding level of shifting. This algorithm and the calculation of correlations were all made in Python.

In the results section that follows, we present the level of correlation obtained between the actual interaction and the cued-recall session for each of our experimental conditions. These correlations represent emotional reactions measured through the GSR and HR sensors. For each correlation result, we outline two versions: 1) Correlations based on using an optimal level of shifting and 2) correlations based on no shifting.

The levels of shifting that would lead to optimal correlations were comparable across both studies and all our experimental conditions. Optimal shift level in terms of the GSR sensor data revealed a mean of -0.39 seconds (SD= 4.51), i.e. the sensor data from the cued-recall session was shifted with less than half a second advance compared to the data obtained during the actual interaction. Note that this is the overall mean value across all 12 conditions and both studies. A one-way ANOVA (evaluated for normally distributed data through visual inspection of Q-Q plots, .95 level) revealed no significant differences in this respect $F(11, 66)=0.64$, $p=0.78$.

Considering the HR sensor data we found a mean shift level of 0.047 seconds (SD = 4.09) meaning the cued-recall graph was correlated with a slight retard compared to the HR data obtained during the actual interaction. Similar to the GSR data, an ANOVA test (also assessed for the normality assumption) revealed no significant differences between any conditions in the two studies $F(11, 66)=0.44$, $p=0.92$.

3.8 Hypotheses

Based on the literature review (Section 2), which informed the experimental design (Section 3.1-3.7), the following hypotheses are formulated, H1-H3 are applicable to both Study 1 and Study 2 (as described in Section 3.2: GSR = galvanic skin response; HR = heart rate), H4 for Study 1 with stimuli of different valence (in appendix C we also list H5-H7 regarding post-task SAM data).

- **H1**: The GSR and HR measures taken during the actual interaction with a system are significantly correlated with those taken during the cued recall.
- **H2**: Correlations between the GSR and HR measures taken during the actual interaction and those during the cued recall decrease significantly with the increase of the length of intervening time.
- **H3**: Correlations between the GSR and HR measures taken during the actual interaction and those taken during the cued recall decrease significantly with the increase of the arousal level of the stimuli.
- **H4**: Correlations between the GSR and HR measures taken during the actual interaction and those taken during the cued recall decrease significantly with the increase of the negativity of the stimuli.

4 RESULTS

4.1 Effect of Time on Correlations – Sensor Data

This section outlines effects of the independent variable of time on the correlations of GSR and HR sensor data between actual interaction and the cued-recall debriefing sessions.

4.1.1 Effect of time – GSR Study 1

Fig. 8 shows a range of boxplots distributed according to using optimal graph shifting as mentioned in section 3.7.4 (grey) versus no shifting (white) for each of the three intervening time conditions (no waiting time, 0.5 hour wait or 1 hour wait). The boxplots express the level of correlation across all participants and tasks for the given time conditions within Study 1. As shown in Table 2, participants in the 0.5 and 1 hour delay conditions experienced different intervening affect. As an example, a subset of participants in the 0.5 hour wait condition experienced “negative valence / high arousal” while others waited 0.5 hours but experienced “positive valence / high arousal” or “neutral valence / low arousal”. The same applies for the 1 hour waiting time condition. For the purpose of analyzing the effect of time delay, we pooled all the stimuli

conditions under their corresponding time conditions as per Table 2. Thus, data from all participants in the 0.5 hour waiting condition were pooled together to reflect the effect of this particular waiting time on sensor correlations between actual interaction and cued-recall. The same goes for participants in the 1 hour condition.

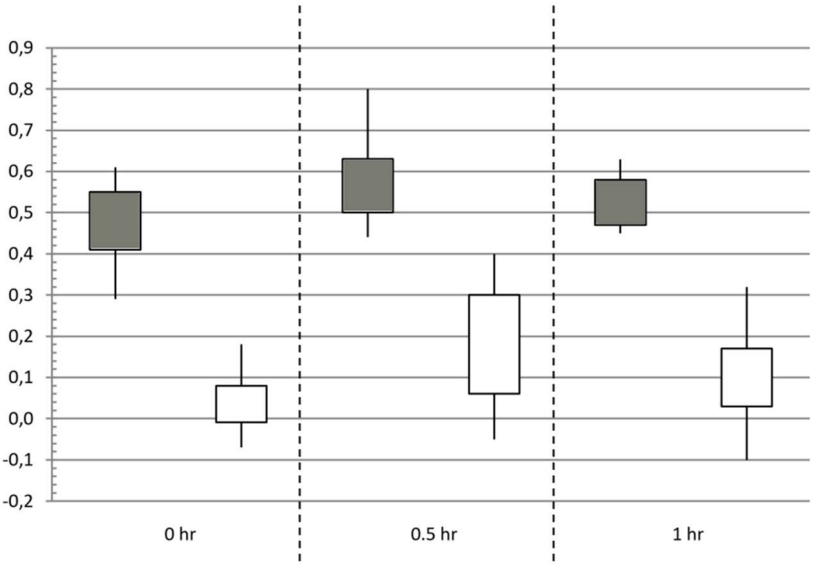

Fig. 8. Study 1 time effects. Pearson correlations of GSR data between actual interaction and cued-recall are shown. Boxplots distributed according to using optimal graph shifting (grey) versus no shifting (white) for each of the three conditions (waiting time of 0 hours, 0.5 hour or 1 hour). N=12 for each condition.

Table 3 outlines the mean level of correlation (Pearson) distributed according to optimal graph shifting versus no shifting for each of the three time conditions in Study 1. The mean level of correlation based on shifting varies between .469 and .542, which are statistically significant ($N_{\text{mean no. of GSR datapoints}}$=2161, p<.01) and is higher compared to not using shifting (.03 - .17). We note that N=2161 refers to the mean number of data points gathered by the GSR sensor for each participant. A one-way ANOVA (tested for normally distributed data using visual inspection of Q-Q plots, .95 level) indicates a highly significant difference between the six data sets reflected in Fig. 8. F(5, 478)=27.29, p<0.000. A Tukey HSD pairwise comparison (.95 level) of all shifted versus non-shifted correlation levels indicates significant differences (p=0). No significant differences were found between any of the shifted correlation levels (p=[.998;1]). This pattern is comparable to the cases of not applying shifting, i.e. no significant difference between these correlations either (p=[.996 - .998]).

Table 1. Mean level of correlation distributed according to optimal graph shifting versus no shifting for each of the three conditions (waiting time of 0 hours, 0.5 hour or 1 hour). Pearson correlations of GSR data between actual interaction and cued-recall are shown.

	Time Condition	Mean	SD
Optimal shift	No wait	.469	.308
	0.5 hour wait	.542	.383
	1 hour wait	.529	.338
No shift	No wait	.032	.419
	0.5 hour wait	.17	.474
	1 hour wait	.104	.434

In the following we revisit hypotheses H1 and H2 which relate to the correlation data obtained through the GSR sensor and the effect of time delay between actual interaction and cued-recall. H1 states that GSR measures taken during the actual interaction with a system and during the cued recall are significantly correlated. The above findings for Study 1 show correlation levels between actual interaction and cued-recall to be around the .5 level or above. This was also the case when at the different intervening time delays (0, 0.5 and 1 hour). Thus, in case of Study 1, we accept H1. H2 states that correlations in sensor data decrease significantly as a delays get longer. By isolating the factor of intervening time delay we observed the tendency for correlation levels to increase from the shortest intervening time of 0 hours to the second intervening time level of 0.5 hour. From the second intervening time level to the third (and longest) of 1 hour we saw a slight decrease in correlation level. Yet, we did not find any significant differences between correlation levels caused by the factor of intervening time. Thus, we reject H2 based on the data obtained in Study 1.

4.1.2 Effect of time – GSR Study 2

The boxplots shown in Fig. 9 represent the effect of intervening time within Study 2 on GSR data correlations between actual interaction and cued-recall. Correlation data based on shifting versus no shifting is also represented (grey vs. white boxplots respectively). The independent variable of wait time spans a longer duration than in the previous study (0 hours, 1 hour and 24 hours). Here we also pooled participants experiencing different intervening stimuli as described in section 4.1.1.

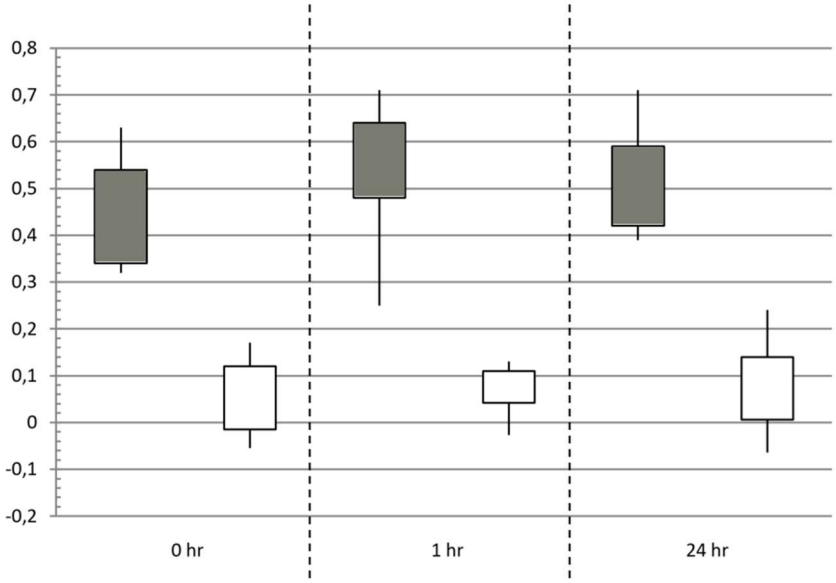

Fig. 1. Study 2 time effects. Boxplots distributed according to using optimal graph shifting (grey) versus no shifting (white) for each of the three conditions (waiting time of 0 hours, 1 hour or 24 hours). N=13 for the no wait and 1 hour wait conditions, N=19 for the 24 hour condition. Pearson correlations of GSR data between actual interaction and cued-recall are shown.

Table 4 shows an overview of the mean correlations within each time condition in Study 2. Using shifting, correlations vary between .445-.535, which are statistically significant ($N_{\text{mean no. of GSR datapoints}}$=1.200, $p<.01$). In the case of not using shifting the correlations vary between .021 - .076. A one-way ANOVA (tested for normally distributed data using visual inspection of Q-Q plots, .95 level) indicates a highly significant difference between the six data sets reflected in Fig. 9 $F(5,592)=31.01$, $p<0.000$. A Tukey HSD pairwise comparison (.95 level) shows no significant differences between all shifted correlation levels (p=[.973;1]), which also applies when comparing between the non-shifted cases (p=[.996 - .999]). Comparing all pairs of shifted and non-shifted correlation levels indicates significant differences (p=0).

Table 2. Mean level of correlation distributed according to optimal graph shifting versus no shifting for each of the three conditions (waiting time of 0 hours, 1 hour or 24 hours). Pearson correlations of GSR data between actual interaction and cued-recall are shown.

	Time Condition	Mean	SD
Optimal shift	No wait	.445	.398
	1 hour wait	.535	.33
	24 hours wait	.512	.401
No shift	No wait	.021	.495
	1 hour wait	.069	.462
	24 hours wait	.076	.498

In terms of hypotheses H1 and H2 we see similar findings pertaining to the GSR sensor data compared to Study 1. In Study 2 we found correlation levels between actual interaction and cued-recall to be around the .5 level or above, also when considering the different intervening time delays (0, 1 and 24 hours). For Study 2 we also accept H1 while rejecting H2.

4.1.3 Effect of time – HR Study 1

The boxplots in Fig. 10 outline the level of correlation between HR data obtained during interaction and during cued-recall within Study 1. These reflect the time conditions of 0 wait, 0.5 hours wait and 1 hour wait. Data from participants experiencing different intervening stimuli were pooled together as described in section 4.1.1.

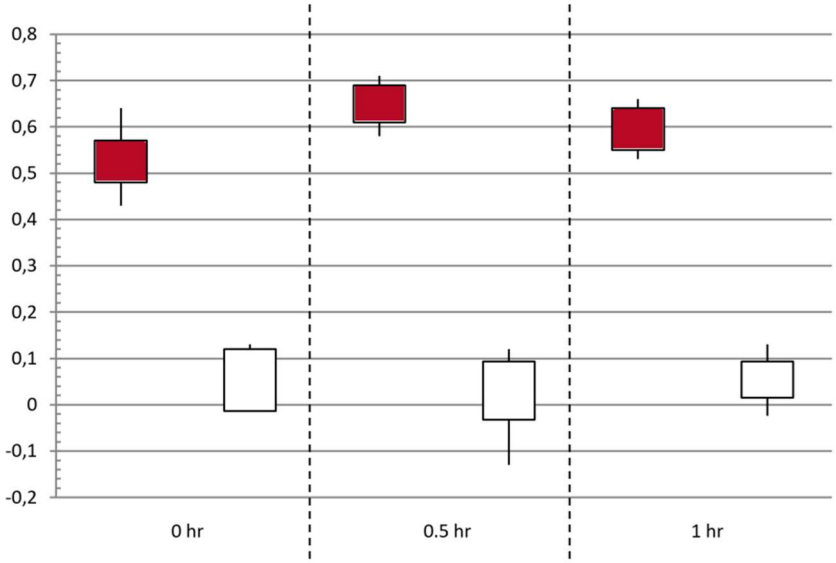

Fig. 2. Study 1 time effects. Boxplots distributed according to using optimal graph shifting (red) versus no shifting (white) for each of the three conditions (waiting time of 0 hours, 0.5 hour or 1 hour). N=12 for each condition. Pearson correlations of HR data between actual interaction and cued-recall are shown.

Table 5 gives an overview of the mean correlation levels. When using shifting the mean varies between .5-.63, which are statistically significant ($N_{\text{mean no. of HR datapoints}}$=5208, p<.01). A one-way ANOVA (tested for normally distributed data using visual inspection of Q-Q plots, .95 level) indicates a highly significant difference between the six data sets reflected in Fig. 10 F(5, 476)=58.36, p<0.000. A pairwise comparison (Tukey HSD, .95 level) reveals no significant differences between any of the shifted correlation levels (p=[.356;.999]), which is similar to the cases not applying shifting (p=[.996-.999]). Comparing all pairs of shifted and non-shifted correlation levels indicates significant differences (p=0).

Table 3. Mean level of correlation distributed according to optimal graph shifting versus no shifting for each of the three conditions (waiting time of 0 hours, 0.5 hour or 1 hour). Pearson correlations of HR data between actual interaction and cued-recall are shown

	Time Condition	Mean	SD
Optimal shift	No wait	.506	.239
	0.5 hour wait	.63	.232
	1 hour wait	.555	.264
No shift	No wait	.022	.391
	0.5 hour wait	.052	.445
	1 hour wait	.075	.37

Correlations of HR data between actual interaction and cued-recall in Study 1 are slightly higher than those for the GSR sensor. In Study 1 we found correlation levels between actual interaction and cued-recall to be around the .5-.6 level when considering the different intervening time delays (0, 0.5 and 1 hour). Therefore we also accept H1 in terms of HR data correlations and again reject H2 due to the non-significant differences in correlation levels between delays.

4.1.4 Effect of time – HR Study 2

The boxplots shown in Fig. 11 represent the effect of time stimuli from Study 2 (0 hours, 1 hour and 24 hours) on HR correlations. Similar to the analyses above, we pooled data from participants experiencing different intervening stimuli during the waiting time.

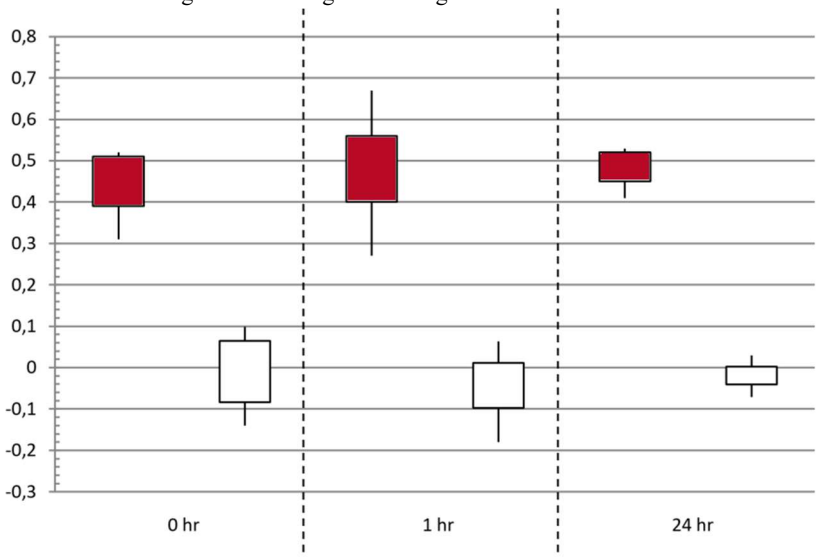

Fig. 3. Study 2 time effects. Boxplots distributed according to using optimal graph shifting (red) versus no shifting (white) for each of the three conditions (waiting time of 0 hours, 1 hour or 24 hours). N=13 for the no wait and 1 hour wait conditions, N=19 for the 24 hour condition. Pearson correlations of HR data between actual interaction and cued-recall are shown.

Table 6 outlines correlation means. When using shifting the correlations vary between .47-.49, which are statistically significant ($N_{\text{mean no. of HR datapoints}}$=5208, $p<.01$). A one-way ANOVA (tested for normally distributed data using visual inspection of Q-Q plots, .95 level) indicates a highly significant difference between the six data sets reflected in Fig. 11 $F(5, 590)=97.26$, $p<0.000$. A Tukey HSD comparison test (.95 level) indicates no significant differences between any of the shifted correlations (p=[.999 - .1]). This also applies when comparing between the non-shifted cases (p=[.999-1]). Comparing all pairs of shifted and non-shifted correlation levels indicates significant differences (p=0).

Table 4. Mean level of correlation distributed according to optimal graph shifting versus no shifting for each of the three conditions (waiting time of 0 hours, 1 hour or 24 hours). Pearson correlations of HR data between actual interaction and cued-recall are shown.

	Time Condition	Mean	SD
Optimal shift	No wait	.471	.239
	1 hour wait	.476	.234
	24 hours wait	.493	.236
No shift	No wait	-.014	.342
	1 hour wait	-.051	.306
	24 hours wait	-.021	.317

Considering Study 2, the correlations of HR data between actual interaction and cued-recall were slightly lower than those for the GSR sensor. HR correlations are just below .05 no matter which of the time delay conditions (0, 1 and 24 hours) are taken into account. Thus, for Study 2 and the HR data, we accept H1 but reject H2 as there were no significant differences in correlation levels between delays.

4.2 Effect of Stimuli on Correlations – Sensor Data

This section outlines effects of the independent variable of stimuli on the correlations of GSR and HR sensor data between actual interaction and the cued-recall debriefing sessions.

4.2.1 Effect of stimuli – GSR Study 1

Fig. 12 shows the effects on correlations when introducing different valence stimuli for Study 1. Correlations are based on GSR data obtained during actual interaction and cued-recall. A subset of participants in both the 0.5 and 1 hour waiting time conditions experienced positive stimuli as the intervening affect, while others experienced negative or neutral valence stimuli. To analyse the effect of such intervening stimuli, we pooled correlation data from all participants that experienced e.g. positive valence, hereby combining data from 0.5 and 1 hour waiting time conditions. From left to right the boxplots show the positive stimuli condition. This positive stimuli leads to a mean GSR correlation between actual interaction and cued-recall of .545 using shifting, which is statistically significant ($N_{\text{mean no. of GSR datapoints}}$=2161, p<.01) and .076 not using shifting. Similar levels of GSR correlations are seen in case of introducing neutral and negative stimuli, which are also statistically significant in cases of shifted values ($N_{\text{mean no. of GSR datapoints}}$=2161, p<.01). Table 7 outlines the means and standard deviations.

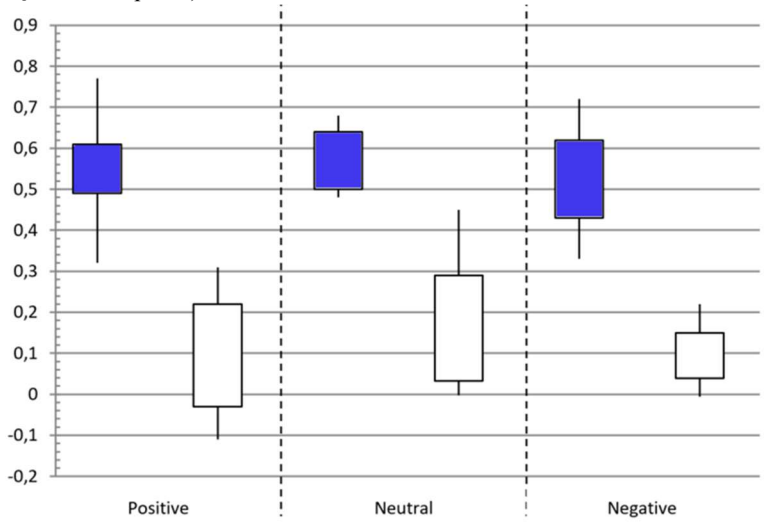

Fig. 4. Study 1 stimuli effects. Boxplots distributed according to using optimal graph shifting (blue) versus no shifting (white) for each of the three conditions (positive, neutral and negative stimuli). N=6 for the

positive stimuli condition, N=13 for the neutral stimuli condition, N=5 for the negative stimuli condition. Pearson correlations of GSR data between actual interaction and cued-recall are shown.

A one-way ANOVA (tested for normally distributed data using visual inspection of Q-Q plots, .95 level) indicates a highly significant difference between the six data sets reflected in Fig. 12 $F(5,318)=15.62$, $p<0.000$. A Tukey HSD pairwise comparison reveals no significant differences between shifted correlation levels (p=[.999;1], .95 level). This pattern is comparable to the cases of not applying shifting, no significant difference between these correlation levels either (p=[.603-.999], .95 level). Comparing all pairs of shifted and non-shifted correlation levels indicates significant differences (p=0, .95 level).

Table 5. Mean level of correlation distributed according to optimal graph shifting versus no shifting for each of the four conditions (no stimuli, positive, neutral and negative stimuli). Pearson correlations of GSR data between actual interaction and cued-recall are shown.

	Time Condition	Mean	SD
Optimal Shift	Positive	.545	.268
	Neutral	.539	.409
	Negative	.516	.322
No shift	Positive	.076	.416
	Neutral	.176	.502
	Negative	.105	.355

Hypothesis 4 states that correlations between the GSR measures taken during actual interaction and during cued recall decrease significantly with the increase of the negativity of the stimuli. The above results from Study 1 show a slightly lower correlation level when introducing negative valence compared to positive valence and neutral valence. Yet, this difference is not significant and hence we reject H4.

4.2.2 Effect of stimuli – GSR Study 2

Fig. 13 shows the effects of different arousal stimuli on the correlation between actual and cued-recall GSR data. Data was obtained during Study 2 and pooled as per the description in section 4.2.1 and instead based on the waiting times of 1 hour and 24 hours relevant for this specific study. In case of high arousal stimuli we see correlations of .55 when using shifting (blue boxplot), which are statistically significant ($N_{\text{mean no. of GSR datapoints}}=2161$, $p<.01$). Correlation level is .075 when not using shifting (white boxplot). These levels are similar to those of introducing low arousal, which are also significant when using shifting ($N_{\text{mean no. of GSR datapoints}}=2161$, $p<.01$). Table 8 outlines the means and standard deviations.

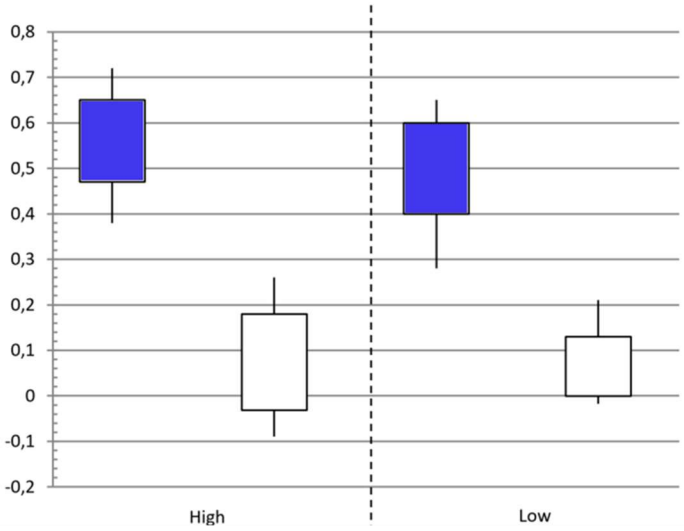

Fig. 5. Study 2 stimuli effects. Boxplots distributed according to using optimal graph shifting (blue) versus no shifting (white) for each of the two conditions (high arousal and low arousal). N=17 for the high arousal stimuli, N=15 for the low arousal condition. Pearson correlations of GSR data between actual interaction and cued-recall are shown.

A one-way ANOVA (tested for normally distributed data using visual inspection of Q-Q plots, .95 level) indicates a highly significant difference between the four data sets reflected in Fig. 13 $F(3, 422)=38.55$, $p<0.000$. A Tukey HSD pairwise comparison reveals no significant differences between shifted correlation levels (p=[.886;.999], .95 level). This pattern is comparable to the cases of not applying shifting, no significant difference between these correlation levels either (p=[.999-1], .95 level). Comparing all pairs of shifted and non-shifted correlation levels indicates significant differences (p=0, .95 level).

Table 6. Mean level of correlation distributed according to optimal graph shifting versus no shifting for each of the two conditions (high arousal and low arousal). Pearson correlations of GSR data between actual interaction and cued-recall are shown.

	Time Condition	Mean	SD
Optimal shift	High arousal	.55	.371
	Low arousal	.491	.375
No shift	High arousal	.075	.518
	Low arousal	.071	.445

In case of Study 2 we introduced stimuli that differed in terms of arousal levels (high vs. low). This relates to hypothesis H3 stating that correlations between the GSR measures taken during actual interaction and during cued recall increase significantly with the increase of the arousal level of the stimuli. In Study 2 we found that the correlation level of the high arousal condition to be slightly higher than for the low arousal condition, although the differences were non-significant. Thus, we reject H3.

4.2.3 Effect of stimuli – HR Study 1

Fig. 14 provides an overview of the effects on correlations when introducing different valence stimuli in case of Study 1. Correlations are based on HR measures during actual interaction and cued-recall and data is pooled as described earlier. The boxplots show that the positive stimuli leads to a mean correlation of .677 using shifting and .059 without shifting. In terms of shifting, correlation level is statistically significant ($N_{\text{mean no. of HR datapoints}}=5208$, $p<.01$). This is similar to the mean level of correlation seen in case of introducing neutral and negative stimuli.

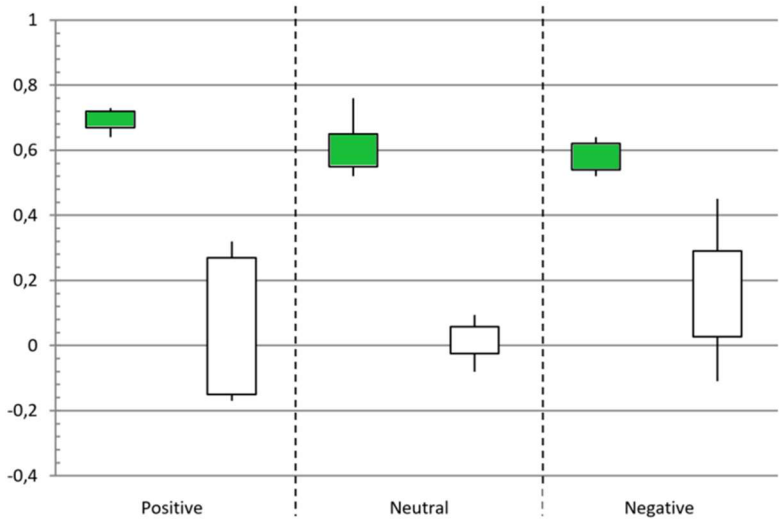

Fig. 6. Study 1 stimuli effects. Boxplots distributed according to using optimal graph shifting (green) versus no shifting (white) for each of the three conditions (positive, neutral and negative stimuli). N=6 for the positive stimuli condition, N=13 for the neutral stimuli condition, N=5 for the negative stimuli condition. Pearson correlations of HR data between actual interaction and cued-recall are shown.

A one-way ANOVA (tested for normally distributed data using visual inspection of Q-Q plots, .95 level) indicates a highly significant difference between the six data sets reflected in Fig. 14 $F(5,318)=41.35$, $p<0.000$. A Tukey HSD pairwise comparison reveals no significant differences between shifted correlation levels (p=[.292;1], .95 level). This is similar to the case of comparing all pairs without using shifting (p=[.728-.1], .95 level). Comparing all pairs of shifted and non-shifted correlation levels indicates significant differences (p=0, .95 level).

Table 7. Mean level of correlation distributed according to optimal graph shifting versus no shifting for each of the three conditions (positive, neutral and negative stimuli). Pearson correlations of HR data between actual interaction and cued-recall are shown.

	Time Condition	Mean	SD
Optimal shift	Positive	.677	.23
	Neutral	.577	.239
	Negative	.537	.285
No shift	Positive	.059	.419
	Neutral	.028	.38
	Negative	.162	.463

We now return to hypothesis 4 on the decrease in correlation levels as a result of negativity of induced stimuli. In terms of the above HR data, we observed that the negative stimuli condition lead to a lower correlation level compared to the positive and neutral stimuli conditions. The difference, however, was not significant and we reject H4 once more as we did in case of the GSR data.

4.2.4 Effect of stimuli – HR Study 2

Fig. 15 gives an overview of the effects of different arousal stimuli from Study 2 on the correlation between actual and cued-recall HR data. Data was pooled as described previously. Table 10 outlines the means and standard deviations. In case of the high arousal stimuli we see correlations of .505 when using shifting (leftmost green boxplot) and -.031 when not using shifting (leftmost white boxplot). Correlation level when

using shifting is statistically significant ($N_{\text{mean no. of HR datapoints}}$=5208, p<.01). These levels are similar to those of introducing low arousal stimuli during the waiting periods.

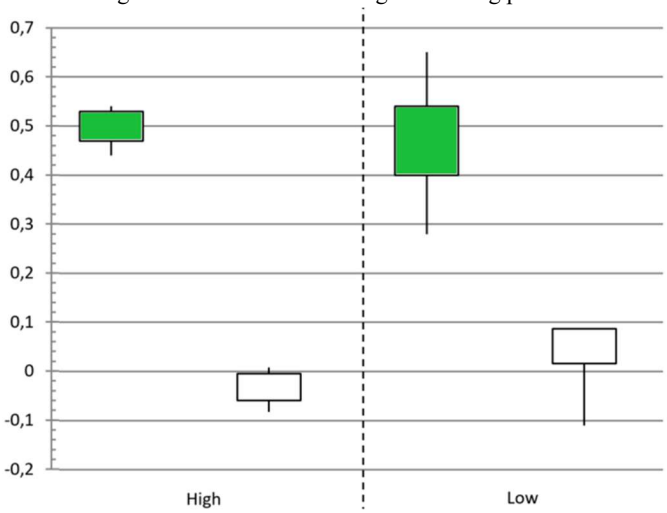

Fig. 7. Study 2 stimuli effects. Boxplots distributed according to using optimal graph shifting (green) versus no shifting (white) for each of the two conditions (high arousal and low arousal). N=17 for the high arousal stimuli, N=15 for the low arousal condition. Pearson correlations of HR data between actual interaction and cued-recall are shown.

A one-way ANOVA (tested for normally distributed data using visual inspection of Q-Q plots, .95 level) indicates a highly significant difference between the four data sets reflected in Fig. 15 F(3, 420)=124.8, p<0.000. A pairwise comparison test (Tukey HSD) shows no significant differences between shifted correlation levels (p=[.994;.1], .95 level), which is similar when comparing all pairs of non-shifted correlations (p=[.999-1], .95 level). Comparing all pairs of shifted and non-shifted correlation levels indicates significant differences (p=0, .95 level).

Table 8. Mean level of correlation distributed according to optimal graph shifting versus no shifting for each of the four conditions (no stimuli, high arousal and low arousal). Pearson correlations of HR data between actual interaction and cued-recall are shown.

	Time Condition	Mean	SD
Optimal shift	High arousal	.505	.241
	Low arousal	.465	.229
No shift	High arousal	-.031	.304
	Low arousal	-.035	.323

For hypothesis H3, which state that correlations of HR measures increase as arousal level increases. In Study 2 we found that correlation level of the high arousal condition to be higher in comparison to the low arousal condition. This difference was not significant and H3 is rejected.

5 DISCUSSION

Cued recall debriefing (CRD) is a research method for gaining insights into participants' thoughts and feelings in connection to a system after they have interacted with it. This is to address the issue of flow interruption when participants are required to think aloud concurrently or they are stopped intermittently to describe their felt experiences. There can be different reasons to defer CRD to a later time, not immediately following the completion of actual interactions with a system. Participants may be cognitively and

emotionally exhausted after working out some challenging tasks with a complex system. A salient reason is to avoid the adverse effect of fatigue on participants, especially when the duration of actual interaction is long, say an hour.

Nonetheless, as a retrospective approach, CRD is susceptible to memory decay, leading to missing or inaccurate recall of events and associated emotions. Such omission and inaccuracy can be more severe when the time gap between the actual interaction and CRD session is non-trivial, say 30 minutes or longer [31,106], because other activities irrelevant to the interaction may interfere with the memory of interaction experiences.

The main research question we aimed to answer with our empirical results is how the two variables – the length of time delay and the type of external stimuli immediately preceding CRD – influence the consistency between emotional experiences elicited in actual interaction and those in viewing the video of the interactions. Such consistency is an indicator to what extent participants can re-immerse into the events retrospectively, and we have assessed it in terms of our research hypotheses, which we revisit here with regard to our findings.

5.1 Revisiting the Hypotheses

In this section, we revisit each of the four hypotheses presented in Section 3.8 by summarizing the related key findings.

5.1.1 H1: The GSR and HR measures taken during the actual interaction with a system are significantly correlated with those taken during the cued recall

H1 was confirmed. For both Study 1 and Study 2, results under different conditions show that there are statistically significant correlations between the two sets of GSR measures (i.e. actual interaction and cued-recall) and between the two sets of HR measures.

However, the average correlations under different conditions are generally higher in Study 1 than those in Study 2, though the order of magnitude is consistent within the respective study, as shown in Table 11 (extracted from the results in Section 4.2).

Table 9. Summary of shifted correlations for GSR and HR measures per factor (intervening time, intervening affect) for the two studies.

	Intervening Time		**Intervening Affect**	
	GSR	HR	GSR	HR
Study 1	0.513	0.555	0.533	0.597
Study 2	0.497	0.493	0.521	0.485

Several plausible explanations for the observed discrepancies can be identified. First, in terms of intervening time, Study 2 included a 24-hour group, which presumably should have lower recall accuracy. Second, stimuli of neutral valence used in Study 2 as compared with those of positive and negative valence could have less impact on recall. Third, participants in Study 2 were more heterogeneous in terms of educational background, culture, and age as compared with a cohort of computer science students in Study 1. The gender ratio was well balanced in both studies.

We also chose to gather subjective data through SAM ratings in order to further validate our findings. Overall we only found a smaller subset of cases in which subjective ratings differed between actual interaction and cued-recall (see Table C1 in appendix C). Isolating the effect of the *intervening time*, we found that only two of eleven tasks (T4 “Write and delete email” and T5 “Add attachment”) showed significant differences in subjective ratings during actual interaction and cued-recall. These observations suggest that the length of intervening time did not have any significant impact on the extent to which the participants could recall their emotions. Similar findings apply in terms of the *intervening affect*, where only one of eleven tasks (T2 “Reply to email”) shows a significant difference in ratings between actual interaction and cued-recall. This suggests that the intervening affect played no role in influencing the participants’ emotional recall. This can be attributed to the successful re-immersion in the tasks with the augmented cues, which could in principle eliminate the effect, if any, of the emotion induced by the IAPS images. Finally, we

also examined the relationship between the objective physiological data and subjective self-report data (see Appendix C, Table C5 for actual interaction and Table C6 for cued-recall). None of the mean GSR data for actual interaction or cued-recall were significantly correlated with the SAM ratings. In contrast, significant correlations were found between the mean HR data and SAM ratings for two tasks (T2, T5) in the case of actual interaction and for four tasks (T1, T8, T9, T10) in the case of cued-recall. A plausible explanation is the contrasting number of data points between GSR/HR and SAM. For example, the number of GSR/HR data points for participant P03 of Study 2 is in the order of thousands per task, ranging from 1725 to 4931. Collapsing a spread of values into the single mean value (in order to compare to SAM ratings) might be problematic as it cancels out highs and lows of emotional changes. Also, while previous studies have indeed found correlations between objective physiological data and subjective ratings of emotions, particularly between measures of GSR and arousal ratings [30,61,109], others have found such correlations to be more elusive. Choi et al. [30] found significant positive correlation between objective HR data and subjective pleasure ratings but negative correlations between HR data and dominance ratings. Similarly, Albanese and colleagues found correlations between objective and subjective data in some situations, but not in others [1]. In their study it was found that HR measures correlate with subjective ratings when participants were at rest, but not so when exercising.

5.1.2 H2: Correlations between the GSR and HR measures taken during the actual interaction and those taken during the cued recall decrease significantly with the increase of the length of intervening time

H2 was rejected. Surprisingly, the length of intervening time had no significant effect on the level of correlation in Study 1 or Study 2. Another unexpected observation is that for both the GSR and HR measures, the 0-hour group in Study 1 and Study 2 had the lowest correlation, albeit the differences were not statistically significant. A possible explanation is that the participants might reflect on the experimental activities during the intervening period, because wearing those specific sensors and interacting with the email client with seeded usability problems were not mundane tasks. Their reflection could reinforce their memory of the events in the experimental session [3], which did not erode with time as assumed.

5.1.3 H3: Correlations between the GSR and HR measures taken during the actual interaction and those taken during the cued recall decrease significantly with the increase of the arousal level of the stimuli

H3 was rejected based on the findings of Study 2. Two sets of images of neutral valence with low and high levels of arousal were presented to the participants with the assumption that the latter would induce stronger emotions, interfering with the recall of past experiences. However, no expected effect on performance in terms of lower correlation levels was observed. Similar to our analysis of the unexpected findings for H3, the role of frequent exposure to the related images can be a contributing factor for the no-significant-difference phenomenon.

Nonetheless, the visual stimuli might still work in "rinsing" so-called "emotional noises" that participants would bring to the CRD session (e.g., responses of other emotion-inducing activities in which participants have been engaged during the waiting time). However, to ascertain the regulating function of such stimuli, a control study is required where participants are not presented any pre-selected stimuli when the intervening time is over and shortly before a CRD session.

5.1.4 H4: Correlations between the GSR and HR measures taken during the actual interaction and those taken during the cued recall decrease significantly with the increase of the negativity of the stimuli

H4 was rejected based on the related findings of Study 1. In fact, the correlations for the negative stimuli were the lowest (cf. the positive and neutral ones; Table 7 and Table 9) for both GSR and HR, albeit the differences were statistically insignificant. The assumption of negativity dominance [92] might not always be the case in an HCI context. According to the notion of negativity bias, the images with negative contents might stimulate the participants with potent unpleasant emotions that would bias their recall of the actual interaction. In contrast, the images with positive contents might exert the beneficial effect, enhancing the participants' attention [43,66] to the cued recall exercise. Consequently, the performance of the latter could

have been better than that of the former in terms of higher levels of correlations between the two sets of psychophysiological measures taken in the actual interaction and CRD session. However, the predicted trend could not be statistically demonstrated.

This surprising finding can be explicated by the argument that the impact of the IAPS images may be compromised after their extensive use [32] in a number of research studies in various domains. The participants might be exposed to a subset of the images or similar entities beforehand in other contexts. Another observation is the increasing use of image-sharing social media [84], especially among young adults like the participants in Study 1; this may result in their overexposure to negative images and thus insensitivity towards such contents.

While we were aware of the availability of GAPED – a new set of images as an alternative to IAPS – created by Scherer and colleagues [32], we opted for IAPS because of its wider scope of validation in terms of cultural contexts. Nonetheless, it is reasonable to anticipate that GAPED can be comparably powerful in the coming years due to the expanding related validation work.

Another explanation could be that affect priming only has a temporary effect from the beginning of the cued-recall session and arguably does not cancel out all intermediary emotions experienced during the time delay. Nevertheless we did not observe any significant differences across the different intervening stimuli, hereby suggesting that CRD is robust in terms enabling participants to re-immerse into past experiences independent of the time delay (at least up to 24 hours) and intervening experiences.

5.2 Implications

Omodei and McLennan originally argued that successful re-immersion through video cues would enable study participants to re-experience emotions similar to when the original event took place [83]. While some recent work in the field of HCI has been conducted to verify this claim [9,22,25], our findings provide further indication of the validity of using cued-recall debriefing to retrospectively uncover emotional reactions of study participants. This work extends previous studies by systematically showing that factors of intervening time and intervening affect does not seem to significantly impact the re-immersion that takes place during cued-recall debriefing based on videos showing participants' interactions with a system.

A widely applied method such as experience sampling was proposed in the 1980's by Larson and Csikszentmihaly [64] as a way to let study participants report on their experiences during the day by, for instance, prompting at random or fixed moments [101]. However, while reducing the bias of the memory-experience gap, this also burdened participants and interfered with their daily activities [101]. As a response to this, Kahneman suggested the Day Reconstruction Method [58] in which participants reflect on their daily experiences towards the end of the day, which has also been used in HCI studies (cf. [62]). The Day Reconstruction Method, however, is based on free recall, hereby being susceptible to the memory-experience gap, which is critical to consider when aiming to understand emotional reactions given their fleeting nature [24,100]. CRD leans on the advantages of both experience sampling and day reconstruction as cues gathered throughout the day can be used to report experiences at the end of the day. Given this backdrop, we elaborate on the implications of these findings for CRD in the wild.

Our contribution indicates that CRD, due to its stability in re-immersing over time and across various intervening stimuli, has the potential of complementing established ethnographic approaches used to study phenomena in the wild. This should be considered as an alternative to experience sampling and day reconstruction. As Scollon argues, it increases response rates if study participants can elaborate on their experiences at a later, more convenient time than is the case in typical experience sampling studies [101], and CRD seems to be robust for up to 24 hours after the fact. Rogers and Marshall [81] suggest the use of physiological data when conducting studies in the wild as they enable continuous data gathering on emotional experiences. The insights gained about the necessity of an intervening time period for a CRD session and of a "reset procedure" (or emotional rinsing) as the prelude of such a delayed CRD session can have considerable implications for research practice. These insights challenge the assumption that more consistent recall of thoughts and emotions can occur if CRD takes place immediately after the actual interaction when the memory is still fresh than if CRD is delayed. Thus, our work provides a critical step

further towards enabling HCI researchers to use CRD with physiological sensors for studying emotional reactions experienced in naturalistic settings.

An examples use-case may be to measure emotional reactions over the course of a day through GSR or HR sensor. This can be done without interfering with daily activities as opposed to experience sampling and would give insights on e.g. high and low arousal episodes occurring during a day. Such measures can be coupled with automatic recordings of video, photos or other contextual data to provide cues that enable valid reflections of emotional experiences. Thus, besides the quantitative physiological data, study participants can describe their experiences qualitatively using cued-recall, hereby providing subjective insights into their experiences. Indeed, studies utilizing CRD are emerging in HCI with study participants wearing physiological sensors like in our case in order to enable reflections of the past, not only through video cues [9,22,24] but also through other types of cues based on screenshots and contextual cues such as photos, location data, nearby Bluetooth-enabled devices, usage logs etc. [26,59,93,105]. Our study is, to the best of our knowledge, the first to go beyond early exploration of CRD as we systematically assess its validity with regards to the factors of intervening time and affect.

A caveat to consider though, is that participants in our study were asked to view ~20 minutes of data showing their entire interaction with an email client. This will not be feasible to do for data captured over the course of an entire day. Studies within HCI and Lifelogging have discussed this issue, which we touch upon in the following. Lifelogging studies focus on using wearable technology to make people reflect on and describe episodes from their daily lives [26]. Such studies have used image data for some time with the purpose supporting study participants in describing daily events. The SenseCam developed by Microsoft is an example of a widely applied camera for such studies (cf. [4,11,53,59,60,89,102]). SenseCam captures an image every 30 seconds, yet, while providing much data, the amount can be overwhelming for participants, who likely may not wish to browse through 20,000+ images taken over the course of a day. Therefore, some filtering is necessary in order to select a manageable set of data that is then presented to study participants at the end of the day. SenseCam can do this through light and facial recognition sensors and a recent study have utilized a GSR sensor for selecting and presenting images based on the daily events that led to the highest levels of arousal [26].

Studies from HCI have also applied a manual approach to data filtering by letting study participants interpret physiological data on their own in order to extract and describe the most important episodes that happened during a day. Such manual filtering was based on showing raw GSR data in one study [23] while another applied visual transformation into abstract shapes and colours, although with varying preference [105]. Thus, the vast amount of cues captured throughout an entire day can be filtered in terms of extracting and presenting the most critical cues, and this can even be done by the study participants themselves [23,105]. Our study shows the potential of participants being able to re-immerse themselves into past experiences up to 24 hours after the fact, which implies that cued-recall can be considered a valid complement to established ethnographic data collection methods utilized to study phenomena in the wild. Furthermore, using physiological sensors in conjunction with CRD has the potential of enabling quantitative and qualitative insights into people's emotional experiences at a more fine-grained scale than seen in studies based on experience sampling and day reconstruction.

5.3 Limitations

A caveat is that we did not identify a time threshold (i.e. the intervening time duration) or a stimulus type (i.e. the intervening affect with a specific combination of valence and arousal level) to produce the highest possible quality of CRD, which entails different experimental setups. Future work should entail identification of sweet spots in terms of these factors, which for instance could be to uncover more precisely at which point the intervening time duration leads to a decline in CRD's ability to re-immerse into past experiences. Furthermore, the effect of the intervening affect could have been gauged by measuring a participant's affective state immediately after she has viewed the IAPS images. The challenge, however, is to identify a right instrument that is valid, usable but short enough to ensure that the induced affective state remains. Nonetheless, this can lead to a paradoxical question whether such a measuring process would alter

the state being measured or even become yet another form of affect priming. We aim to explore this intriguing issue in our future work.

Also, a constraint of our studies is that in the first viewing of the video (Phase 3 in Fig. 6), participants could not perform a running commentary because any talking can interfere with sensor measurements. Hence, we cannot draw any conclusion about the qualitative insights gained through cued-recall. However, the original study conducted by Omodei and McLennan [83] report on CRD enabling *"direct, vivid, and immediate feedback about what an individual was actually doing at each point on the course and what thoughts, feelings, choices, plans, and decisions directed these actions"*. This is also corroborated in [26] where it is reported that participant through CRD could *"vividly recall details about their past experiences"*. Nonetheless, we deem that it is relevant, as our future work, to conduct empirical studies to compare systematically the quality of retrospective think-aloud (TA) with cues to that without cues (cf. Section 2.1). In both conditions, physiological data (e.g. HR, GSR) for tracking emotional responses will be taken to substantiate the findings of this present work. Note, however, the feasibility of the planned work relies much on the availability of mobile wearable sensors with desirable characteristics, including their stability/sensitivity to bodily movements associated with TA (controlling noisy data), usability (minimising physical comfort), and affordability.

It is extremely resource-demanding for handling a vast body of psychophysiological data. Within-subject experimental design is difficult, if not impossible, to implement for this research work whereas between-subject one requires a good number of participants. These and other practical concerns have not allowed us to incorporate in our research studies the corresponding control groups, i.e. the groups without any intervening time are not presented any pre-selected stimuli to examine the relationship between the quality of CRD and the variables of interest.

6 CONCLUSION

Over the course of two studies, we examined the robustness of Cued-Recall Debriefing (CRD) as a retrospective method for assessing user experience. In our studies of CRD robustness we analysed the effect of varying the duration of intervening time between actual interaction and CRD sessions. We also induced experiences of varying valence and arousal combination towards the end of the intervening time period to see how these impact CRD experiences. The main dependent variable is the level of correlation in physiological sensor data between actual interaction and CRD. For both studies we applied GSR and HR sensors and showed significant levels of correlation between actual interaction and CRD, regardless of the size of time delay (30 minutes, 60 minutes or 24 hours). We also found significant correlations across all valence experiences (positive, neutral and negative) and two levels of arousal (high and low). That level of correlation, however, required the datasets from actual interaction and CRD to be shifted with a few seconds (~.4 seconds). Results in terms of subjective SAM ratings are more mixed, although with the main tendency of insignificant differences in ratings between actual interaction and cued-recall. Overall, we argue for the robustness of CRD, which can be applied in diverse settings in today's research landscape. Finally, as psychophysiological sensors have become more wearable, further work is needed to assess their uses in the wild outside the confinements of the lab.

ACKNOWLEDGMENTS

We would like to thank Pamela Andrade, Michael Fuglsang, Henrik Haxholm, Vincent Hocquemiller and Benjamin Hubert for their invaluable assistance in facilitating the experiments and data gathering.

REFERENCES

<BIBL>

1. Marco Albanese, Michaelis Neofytou, Taoufik Ouarrak, Steffen Schneider, and Wolfgang Schöls. 2016. Evaluation of heart rate measurements in clinical studies: a prospective cohort study in patients with

heart disease. *European Journal of Clinical Pharmacology* 72, 7: 789–795. https://doi.org/10.1007/s00228-016-2046-9

2. Obead Alhadreti and Pam Mayhew. 2018. Rethinking Thinking Aloud: A Comparison of Three Think-Aloud Protocols. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (CHI '18), 44:1--44:12. https://doi.org/10.1145/3173574.3173618

3. John R. Anderson and Gordon H. Bower. 1979. *Human associative memory*. Lawrence Erlbaum Associates, Hillsdale, NJ.

4. Mattias Arvola, Johan Blomkvist, and Fredrik Wahlman. 2017. Lifelogging in User Experience Research: Supporting Recall and Improving Data Richness. *The Design Journal* 20, sup1: S3954–S3965. https://doi.org/10.1080/14606925.2017.1352898

5. Pedro Avero and Manuel G. Calvo. 2006. Affective priming with pictures of emotional scenes: The role of perceptual similarity and category relatedness. *Spanish Journal of Psychology* 9, 1: 10–18. https://doi.org/10.1017/S1138741600005928

6. Javier A. Bargas-Avila and Kasper Hornbæk. 2011. Old wine in new bottles or novel challenges? A critical analysis of empirical studies of User Experience. 2689–2698. https://doi.org/10.1145/1978942.1979336

7. R.F. Baumeister, E. Bratslavsky, C. Finkenauer, and K.D. Vohs. 2001. Bad is stronger than good. *Review of General Psychology* 5, 4: 323–370. https://doi.org/10.1037/1089-2680.5.4.323

8. V Benet-Martinez and O P John. 1998. Los Cinco Grandes across cultures and ethnic groups: multitrait multimethod analyses of the Big Five in Spanish and English. *Journal of personality and social psychology* 75, 3: 729–750.

9. Todd Bentley, Lorraine Johnston, and Karola von Baggo. 2005. Evaluation Using Cued-recall Debrief to Elicit Information About a User's Affective Experiences. In *Proc. OzCHI* (OZCHI '05), 1–10. Retrieved from http://dl.acm.org/citation.cfm?id=1108368.1108403

10. David. Benyon. *Designing interactive systems : a comprehensive guide to HCI, UX and interaction design*. Retrieved May 2, 2019 from https://catalogue.pearsoned.co.uk/educator/product/Designing-Interactive-Systems-A-comprehensive-guide-to-HCI-UX-and-interaction-design-3E/9781447920113.page

11. Emma Berry, Narinder Kapur, Lyndsay Williams, Steve Hodges, Peter Watson, Gavin Smyth, James Srinivasan, Reg Smith, Barbara Wilson, and Ken Wood. 2007. The use of a wearable camera, SenseCam, as a

pictorial diary to improve autobiographical memory in a patient with limbic encephalitis: A preliminary report. *Neuropsychological Rehabilitation* 17, 4–5: 582–601. https://doi.org/10.1080/09602010601029780

12. Anol Bhattacherjee. 2001. Understanding information systems continuance: An expectation-confirmation model. *MIS Quarterly: Management Information Systems* 25, 3: 351–370. https://doi.org/10.2307/3250921

13. Kirsten Boehner, Rogério DePaula, Paul Dourish, and Phoebe Sengers. 2007. How emotion is made and measured. *International Journal of Human-Computer Studies* 65, 4: 275–291. https://doi.org/10.1016/J.IJHCS.2006.11.016

14. T. Boren and J. Ramey. 2000. Thinking aloud: reconciling theory and practice. *IEEE Transactions on Professional Communication* 43, 3: 261–278. https://doi.org/10.1109/47.867942

15. Gordon H Bower and Joseph P Forgas. 2000. Affect, memory, and social cognition. *Cognition and emotion.*, 87–168.

16. V.A. Bowers and H.L. Snyder. 1990. Concurrent versus retrospective verbal protocols for comparing window usability. In *Human Factors Society 34th Meeting*, 1270 – 1274.

17. M. Bradley. 2015. Media Core. Retrieved August 1, 2016 from http://csea.phhp.ufl.edu/media.html

18. Margaret M. Bradley and Peter J. Lang. 1994. Measuring emotion: The self-assessment manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry* 25, 1: 49–59. https://doi.org/http://dx.doi.org/10.1016/0005-7916(94)90063-9

19. Jason J. Braithwaite and Derrick G. Watson. 2015. *Issues surrounding the normalization and standardisation of skin conductance responses (SCRs)*. Birmingham.

20. Jennifer L Branch. 2000. Investigating the Information-Seeking Processes of Adolescents: The Value of Using Think Alouds and Think Afters. *Library & Information Science Research* 22, 4: 371–392. https://doi.org/10.1016/S0740-8188(00)00051-7

21. Casey L. Brown, Natalia Van Doren, Brett Q. Ford, Iris B. Mauss, Jocelyn W. Sze, and Robert W. Levenson. 2020. Coherence between subjective experience and physiology in emotion: Individual differences and implications for well-being. *Emotion* 20, 5: 818–829. https://doi.org/10.1037/emo0000579

22. A. Bruun, E.L.-C. Law, M. Heintz, and P.S. Eriksen. 2016. Asserting real-Time emotions through cued-recall: Is it valid? In *ACM International Conference Proceeding Series*. https://doi.org/10.1145/2971485.2971516

23. Anders Bruun. 2018. It's Not Complicated: A Study of Non-specialists Analyzing GSR Sensor Data to Detect UX Related Events. In *Proceedings of the 10th Nordic Conference on Human-Computer Interaction* (NordiCHI '18), 170–183. https://doi.org/10.1145/3240167.3240183

24. Anders Bruun and Simon Ahm. 2015. Mind the Gap! Comparing Retrospective and Concurrent Ratings of Emotion in User Experience Evaluation. In *15th IFIP TC13 Conference on Human-Computer Interaction (INTERACT)*, 237–254. https://doi.org/http://dx.doi.org/10.1007/978-3-319-22701-6_17

25. Anders Bruun, E.L.-C. Effie Lai-Chong Law, Matthias Heintz, and L.H.A. Lana H A Alkly. 2016. Understanding the Relationship Between Frustration and the Severity of Usability Problems: What Can

Psychophysiological Data (Not) Tell Us? In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (CHI '16), 3975–3987. https://doi.org/10.1145/2858036.2858511

26. Anders Bruun and Martin Lynge Stentoft. 2019. Lifelogging in the Wild: Participant Experiences of Using Lifelogging as a Research Tool. In *Proceedings of the 17th IFIP TC13 Conference on Human-Computer Interaction (INTERACT)*, 431–451. https://doi.org/10.1007/978-3-030-29387-1_24

27. Tony W Buchanan. 2007. Retrieval of Emotional Memories. *Psychological bulletin* 133, 5: 761–779. https://doi.org/10.1037/0033-2909.133.5.761

28. Teah-Marie Bynion and Matthew T Feldner. 2017. Self-Assessment Manikin. In *Encyclopedia of Personality and Individual Differences*, Virgil Zeigler-Hill and Todd K Shackelford (eds.). Springer International Publishing, Cham, 1–3. https://doi.org/10.1007/978-3-319-28099-8_77-1

29. John T. Cacioppo, Louis G. Tassinary, and Gary G. Berntson. 2009. Psychophysiological Science: Interdisciplinary Approaches to Classic Questions About the Mind. In *Handbook of Psychophysiology*. Cambridge University Press, 1–16. https://doi.org/10.1017/cbo9780511546396.001

30. Kwang Ho Choi, Junbeom Kim, O. Sang Kwon, Min Ji Kim, Yeon Hee Ryu, and Ji Eun Park. 2017. Is heart rate variability (HRV) an adequate tool for evaluating human emotions? – A focus on the use of the International Affective Picture System (IAPS). *Psychiatry Research* 251: 192–196. https://doi.org/10.1016/j.psychres.2017.02.025

31. M Csikszentmihalyi and R Larson. 1987. Validity and reliability of the Experience-Sampling Method. *The Journal of nervous and mental disease* 175, 9: 526–536.

32. Elise S Dan-Glauser and Klaus R Scherer. 2011. The Geneva affective picture database (GAPED): a new 730-picture database focusing on valence and normative significance. *Behavior research methods* 43, 2: 468–477. https://doi.org/10.3758/s13428-011-0064-1

33. Michael E. Dawson, Anne M. Schell, Diane L. Filion, and Gary G. Berntson. 2009. The Electrodermal System. In *Handbook of Psychophysiology*. Cambridge University Press, 157–181. https://doi.org/10.1017/cbo9780511546396.007

34. Nicola Eger, Linden J Ball, Robert Stevens, and Jon Dodd. 2007. Cueing Retrospective Verbal Reports in Usability Testing Through Eye-movement Replay. In *Proceedings of the 21st British HCI Group*

Annual Conference on People and Computers: HCI...But Not As We Know It - Volume 1 (BCS-HCI '07), 129–137. Retrieved from http://dl.acm.org/citation.cfm?id=1531294.1531312

35. Panteleimon Ekkekakis. 2013. *The Measurement of Affect, Mood, and Emotion A Guide for Health-Behavioral Research*. Cambridge University Press.

36. Paul Ekman. 1992. An argument for basic emotions. *Cognition and Emotion* 6, 3–4: 169–200. https://doi.org/10.1080/02699939208411068

37. Phoebe C. Ellsworth. 2013. Appraisal Theory: Old and New Questions. *Emotion Review* 5, 2: 125–131. https://doi.org/10.1177/1754073912463617

38. Phoebe C Ellsworth. 1994. Levels of thought and levels of emotion. In *The nature of emotion: Fundamental questions*, Paul Ekman and Davidson R.J. (eds.). Oxford University Press, New York, NY, US, 192–196.

39. K Anders Ericsson and Herbert A Simon. 1980. Verbal reports as data. *Psychological Review* 87, 3: 215–251. https://doi.org/10.1037/0033-295X.87.3.215

40. Karl Anders Ericsson and Herbert Alexander Simon. 1993. *Protocol analysis: Verbal reports as data, Rev. ed.* The MIT Press, Cambridge, MA, US.

41. Johnny R.J. Fontaine, Klaus R. Scherer, Etienne B. Roesch, and Phoebe C. Ellsworth. 2007. The World of Emotions is not Two-Dimensional. *Psychological Science* 18, 12: 1050–1057. https://doi.org/10.1111/j.1467-9280.2007.02024.x

42. Jodi Forlizzi and Katja Battarbee. 2004. Understanding Experience in Interactive Systems. In *Proc. DIS* (DIS '04), 261–268. https://doi.org/10.1145/1013115.1013152

43. Barbara L Fredrickson and Christine Branigan. 2005. Positive emotions broaden the scope of attention and thought-action repertoires. *Cognition & emotion* 19, 3: 313–332. https://doi.org/10.1080/02699930441000238

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

45. Tamara van Gog, Fred Paas, Jeroen J. G. van Merriënboer, Puk Witte, Jeroen J G van Merrienboer, and Puk Witte. 2005. Uncovering the problem-solving process: cued retrospective reporting versus concurrent and retrospective reporting. *Journal of experimental psychology. Applied* 11, 4: 237–244. https://doi.org/10.1037/1076-898X.11.4.237

46. Zhiwei Guan, Shirley Lee, Elisabeth Cuddihy, and Judith Ramey. 2006. The validity of the stimulated retrospective think-aloud method as measured by eye tracking. In *Proceedings of the SIGCHI conference on Human Factors in computing systems - CHI '06*, 1253. https://doi.org/10.1145/1124772.1124961

47. Maaike J Van Den Haak, Menno De Jong, Peter Jan Schellens, Menno De, Jong & Peter, Jan Schellens, Maaike J Van Den Haak, Menno D T De Jong, and Peter Jan Schellens. 2003. Retrospective vs.

concurrent think-aloud protocols: Testing the usability of an online library catalogue. *Behaviour & Information Technology* 22, 5: 339–351. https://doi.org/10.1080/0044929031000

48. John Paulin Hansen. 1991. The use of eye mark recordings to support verbal retrospection in software testing. *Acta Psychologica* 76, 1: 31–49. https://doi.org/10.1016/0001-6918(91)90052-2

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

50. Dirk Hermans, Adriaan Spruyt, Jan De Houwer, and Paul Eelen. 2003. Affective priming with subliminally presented pictures. *Canadian Journal of Experimental Psychology* 57, 2: 97–114. https://doi.org/10.1037/h0087416

51. Morten Hertzum, Pia Borlund, and Kristina B. Kristoffersen. 2015. What Do Thinking-Aloud Participants Say? A Comparison of Moderated and Unmoderated Usability Sessions. *International Journal of Human-Computer Interaction* 31, 9: 557–570. https://doi.org/10.1080/10447318.2015.1065691

52. Morten Hertzum, Kristin D. Hansen, and Hans H.K. Andersen. 2009. Scrutinising usability evaluation: does thinking aloud affect behaviour and mental workload? *Behaviour & Information Technology* 28, 2: 165–181. https://doi.org/10.1080/01449290701773842

53. Steve Hodges, Lyndsay Williams, Emma Berry, Shahram Izadi, James Srinivasan, Alex Butler, Gavin Smyth, Narinder Kapur, and Ken Wood. 2006. SenseCam: A Retrospective Memory Aid BT - UbiComp 2006: Ubiquitous Computing. 177–193.

54. Aulikki Hyrskykari, Saila Ovaska, Pävi Majaranta, Kari-Jouko Räihä, and Merja Lehtinen. 2008. Gaze Path Stimulation in Retrospective Think-Aloud. *Journal of Eye Movement Research* 2, 4. https://doi.org/10.16910/jemr.2.4.5

55. Caroll E Izard. 1993. Organizational and motivational functions of discrete emotions. In *Handbook of emotions.* Guilford Press, New York, NY, US, 631–641.

56. Oliver P John, Laura P Naumann, and Christopher J Soto. 2008. Paradigm Shift to the Integrative Big Five Trait Taxonomy: History, Measurement, and Conceptual Issues. In *Handbook of Personality: Theory and Research*, Oliver P John, Richard W Robbins and Lawrence A Pervin (eds.). Guilford, New York, 114–156. Retrieved from http://www.ocf.berkeley.edu/~johnlab/bigfive.htm

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

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

59. Liadh Kelly and Gareth J. F. Jones. 2010. An exploration of the utility of GSR in locating events from personal lifelogs for reflection. *Self*: 82–85. Retrieved September 21, 2018 from https://www.semanticscholar.org/paper/An-Exploration-of-the-Utility-of-GSR-in-Locating-Kelly/12e73eb02d357ee0085465d0cff6aa322e19e515

60. Jacqueline Kerr, Simon J. Marshall, Suneeta Godbole, Jacqueline Chen, Amanda Legge, Aiden R. Doherty, Paul Kelly, Melody Oliver, Hannah M. Badland, and Charlie Foster. 2013. Using the SenseCam to

Improve Classifications of Sedentary Behavior in Free-Living Settings. *American Journal of Preventive Medicine* 44, 3: 290–296. https://doi.org/10.1016/J.AMEPRE.2012.11.004

61. Sander Koelstra, Christian Mühl, Mohammad Soleymani, Jong Seok Lee, Ashkan Yazdani, Touradj Ebrahimi, Thierry Pun, Anton Nijholt, and Ioannis Patras. 2012. DEAP: A database for emotion analysis; Using physiological signals. 3, 1: 18–31. https://doi.org/10.1109/T-AFFC.2011.15

62. Sari Kujala and Talya Miron-Shatz. 2013. Emotions, Experiences and Usability in Real-life Mobile Phone Use. In *Proc. CHI* (CHI '13), 1061–1070. https://doi.org/10.1145/2470654.2466135

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

64. Reed Larson and Mihaly Csikszentmihalyi. 1983. The Experience Sampling Method. *New Directions for Methodology of Social & Behavioral Science* 15: 41–56.

65. Lucian Leahu, Steve Schwenk, and Phoebe Sengers. 2008. Subjective objectivity: Negotiating emotional meaning. In *Proceedings of the Conference on Designing Interactive Systems: Processes, Practices, Methods, and Techniques, DIS*, 425–434. https://doi.org/10.1145/1394445.1394491

66. Angela Y. Lee and Brian Sternthal. 1999. The Effects of Positive Mood on Memory. *Journal of Consumer Research* 26, 2: 115–127. Retrieved from http://www.jstor.org/stable/10.1086/209554

67. Clayton Lewis. 1982. *Using the "thinking-aloud" method in cognitive interface design*. IBM T.J. Watson Research Center, Yorktown Heights N.Y. Retrieved June 19, 2019 from https://www.worldcat.org/title/using-the-thinking-aloud-method-in-cognitive-interface-design/oclc/38214134

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

69. World Famous Electronics Llc. 2015. Pulse Sensor. Retrieved August 1, 2016 from http://pulsesensor.com

70. Sascha Mahlke. 2008. *Visual aesthetics and the user experience*.

71. Sascha Mahlke, Michael Minge, and Manfred Thüring. 2006. Measuring Multiple Components of Emotions in Interactive Contexts. In *CHI EA* (CHI EA '06), 1061–1066. https://doi.org/10.1145/1125451.1125653

72. Regan L. Mandryk and M. Stella Atkins. 2007. A Fuzzy Physiological Approach for Continuously Modeling Emotion During Interaction with Play Technologies. *Int. J. Hum.-Comput. Stud.* 65, 4: 329–347. https://doi.org/10.1016/j.ijhcs.2006.11.011

73. Regan L Mandryk, M Stella Atkins, and Kori M Inkpen. 2006. A Continuous and Objective Evaluation of Emotional Experience with Interactive Play Environments. In *Proceedings of the SIGCHI*

Conference on Human Factors in Computing Systems (CHI '06), 1027–1036. https://doi.org/10.1145/1124772.1124926

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

75. Sharon McDonald, Tingting Zhao, and Helen M. Edwards. 2013. Dual Verbal Elicitation: The Complementary Use of Concurrent and Retrospective Reporting Within a Usability Test. *International Journal of Human-Computer Interaction* 29, 10: 647–660. https://doi.org/10.1080/10447318.2012.758529

76. MindPlace. 2014. Mindplace Thoughtstream. Retrieved August 1, 2016 from http://www.mindplace.com/

77. Talya Miron-Shatz, Arthur Stone, and Daniel Kahneman. 2009. Memories of yesterday's emotions: does the valence of experience affect the memory-experience gap? *Emotion* 9, 6: 885–891. https://doi.org/10.1037/a0017823

78. Jon D Morris. 1995. Observations: SAM: The self-assessment manikin: An efficient cross-cultural measurement of emotional response. *Journal of Advertising Research* 35, 6: 63–68.

79. Brendan D. Murray, Alisha C. Holland, and Elisabeth A. Kensinger. 2013. Episodic Memory and Emotion. In *Handbook of Cognition and Emotion*, M.D. Robinson, E.R. Watkins and E Harmon-Jones (eds.). Guilford Press, 156–175.

80. Jakob Nielsen. 1993. *Usability engineering*. Academic Press. Retrieved June 19, 2019 from https://www.sciencedirect.com/book/9780125184069/usability-engineering

81. K N Ochsner. 2000. Are affective events richly recollected or simply familiar? The experience and process of recognizing feelings past. *Journal of experimental psychology. General* 129, 2: 242–261.

82. Kenneth R Ohnemus and David W Biers. 1993. Retrospective versus Concurrent Thinking-Out-Loud in Usability Testing. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 37, 17: 1127–1131. https://doi.org/10.1177/154193129303701701

83. Mary M. Omodei and Jim McLennan. 1994. Studying complex decision making in natural settings: Using a head-mounted video camera to study competitive orienteering. *Perceptual and Motor Skills* 79, 3, Pt 2: 1411–1425. https://doi.org/10.2466/pms.1994.79.3f.1411

84. Andrew Perrin. 2015. *Social Media Usage: 2005-2015*. Retrieved from http://www.pewinternet.org/2015/10/08/social-networking-usage-2005-2015/

85. Linda W.P. Peute, Nicolette F. de Keizer, and Monique W.M. Jaspers. 2015. The value of Retrospective and Concurrent Think Aloud in formative usability testing of a physician data query tool. *Journal of Biomedical Informatics* 55: 1–10. https://doi.org/10.1016/J.JBI.2015.02.006

86. Pierre Philippot, Celine Baeyens, and Celine Douilliez. 2006. Specifying emotional information: Regulation of emotional intensity via executive processes. *Emotion (Washington, D.C.)* 6, 4: 560–571. https://doi.org/10.1037/1528-3542.6.4.560

87. Rosalind W. Picard, Szymon Fedor, and Yadid Ayzenberg. 2016. Multiple Arousal Theory and Daily-Life Electrodermal Activity Asymmetry. *Emotion Review* 8, 1: 62–75. https://doi.org/10.1177/1754073914565517

88. Jenny Preece, Yvonne Rogers, and Helen Sharp. *Interaction design : beyond human-computer interaction*. Retrieved May 2, 2019 from https://www.wiley.com/en-us/Interaction+Design%3A+Beyond+Human+Computer+Interaction%2C+4th+Edition-p-9781119020752

89. Jenny S Radesky, Caroline J Kistin, Barry Zuckerman, Katie Nitzberg, Jamie Gross, Margot Kaplan-Sanoff, Marilyn Augustyn, and Michael Silverstein. 2014. Patterns of mobile device use by caregivers and

children during meals in fast food restaurants. *Pediatrics* 133, 4: e843-9. https://doi.org/10.1542/peds.2013-3703

90. Donald A Redelmeier and Daniel Kahneman. 1996. Patients' memories of painful medical treatments: real-time and retrospective evaluations of two minimally invasive procedures. *Pain* 66, 1: 3–8. https://doi.org/http://dx.doi.org/10.1016/0304-3959(96)02994-6

91. Donald A Redelmeier, Joel Katz, and Daniel Kahneman. 2003. Memories of colonoscopy: a randomized trial. *Pain* 104, 1–2: 187–194. https://doi.org/http://dx.doi.org/10.1016/S0304-3959(03)00003-4

92. Paul Rozin and Edward B Royzman. 2001. Negativity Bias, Negativity Dominance, and Contagion. *Personality and Social Psychology Review* 5, 4: 296–320. https://doi.org/10.1207/S15327957PSPR0504_2

93. D M Russell and M Oren. 2009. Retrospective Cued Recall: A Method for Accurately Recalling Previous User Behaviors. In *System Sciences, 2009. HICSS '09. 42nd Hawaii International Conference on*, 1–9. https://doi.org/10.1109/HICSS.2009.370

94. J A Russell and A Mehrabian. 1974. Distinguishing anger and anxiety in terms of emotional response factors. *Journal of consulting and clinical psychology* 42, 1: 79–83. Retrieved June 19, 2019 from http://www.ncbi.nlm.nih.gov/pubmed/4814102

95. James A Russell and Albert Mehrabian. 1977. Evidence for a three-factor theory of emotions. *Journal of Research in Personality* 11, 3: 273–294. https://doi.org/10.1016/0092-6566(77)90037-X

96. J. Edward Russo, Eric J. Johnson, and Debra L. Stephens. 1989. The validity of verbal protocols. *Memory & Cognition* 17, 6: 759–769. https://doi.org/10.3758/BF03202637

97. J Edward Russo. 1979. A software system for the collection of retrospective protocols prompted by eye fixations. *Behavior Research Methods & Instrumentation* 11, 2: 177–179. https://doi.org/10.3758/BF03205643

98. Sinué Salgado and Osman Skjold Kingo. 2019. How is physiological arousal related to self-reported measures of emotional intensity and valence of events and their autobiographical memories? *Consciousness and Cognition* 75. https://doi.org/10.1016/j.concog.2019.102811

99. L. Salmerón, J. Naumann, V. García, and I. Fajardo. 2017. Scanning and deep processing of information in hypertext: an eye tracking and cued retrospective think-aloud study. *Journal of Computer Assisted Learning* 33, 3: 222–233. https://doi.org/10.1111/jcal.12152

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

101. Christie N. Scollon, Chu Kim-Prieto, and Ed Diener. 2003. Experience Sampling: Promises and Pitfalls, Strengths and Weaknesses. *Journal of Happiness Studies* 4, 1: 5–34. https://doi.org/10.1023/A:1023605205115

102. Abigail J. Sellen, Andrew Fogg, Mike Aitken, Steve Hodges, Carsten Rother, and Ken Wood. 2007. Do life-logging technologies support memory for the past? In *Proceedings of the SIGCHI conference on Human factors in computing systems - CHI '07*, 81. https://doi.org/10.1145/1240624.1240636

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

104. Adam P R Smith, Richard N A Henson, Michael D Rugg, and Raymond J Dolan. 2005. Modulation of retrieval processing reflects accuracy of emotional source memory. *Learning & Memory* 12, 5: 472–479. https://doi.org/10.1101/lm.84305

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

106. Arthur A. Stone, Joseph E. Schwartz, John M. Neale, Ssaul Shiffman, Christine A. Marco, Mary Hickcox, Jean Paty, Laura S. Porter, and Laura J. Cruise. 1998. A comparison of coping assessed by ecological

momentary assessment and retrospective recall. *Journal of personality and social psychology* 74, 6: 1670–1680. https://doi.org/10.1037//0022-3514.74.6.1670

107. K Lynn Taylor and Jean-Paul Dionne. 2000. Accessing problem-solving strategy knowledge: The complementary use of concurrent verbal protocols and retrospective debriefing. *Journal of Educational Psychology* 92, 3: 413–425. https://doi.org/10.1037/0022-0663.92.3.413

108. Jonathan R Zadra and Gerald L Clore. 2011. Emotion and perception: the role of affective information. *Wiley interdisciplinary reviews. Cognitive science* 2, 6: 676–685. https://doi.org/10.1002/wcs.147

109. Feng Zhou. 2014. Emotion Prediction from Physiological Signals: A Comparison Study Between Visual and Auditory Elicitors. *Interact. Comput.* 26, 3: 285–302. https://doi.org/10.1093/iwc/iwt039

110. Philippe Zimmermann. 2008. Beyond Usability – Measuring Aspects of User Experience.

</BIBL>

Appendix A: A survey of the related work on comparing concurrent think aloud (CTA) and retrospective think-aloud (RTA) [1] Table 12.

Table 10. Comparison studies on concurrent think-aloud (CTA)and retrospective think-aloud (RTA) in the context of usability testing.

Author	Year	System	Comparison	Key Findings
Bowers & Snyder [16]	1990	Microsoft Word	CTA vs. RTA	CTA: Descriptive comments RTA: Explanatory comments/ Enhancement suggestions
Hansen [48]	1991	Text editor	Cued RTA: plain video playback vs. gaze-paths video playback	RTA gaze-paths: more problem-focused verbal comments
Ohnemus & Biers [82]	1993	Databased management system	CTA vs. Immediate RTA vs. 24-hour-delayed RTA	RTA > CTA: feedback for design CTA ~ RTA: Task performance RTA immediate ~ RTA delayed: value of verbal data
Branch [20]	2000	Multimedia encyclopedia	CTA vs. RTA uncued	CTA and RTA provide complementary information
Van den Haak, De Jong & Schellens [47]	2003	Online library	CTA vs. RTA	CTA ~ RTA: Total number of UPs; Severity of UPs; Experience with TA; Task performance
van Gog, Paas, van Merriënboer, & Witte [45]	2005	Computer-simulated electronic circuit	CTA vs. RTA uncued vs. RTA gaze-paths video playback	CTA > RTA gaze-paths > RTA uncued: amount of problem-solving process information (action, how and why)
Guan, Lee, Cuddihy & Ramey [46]	2006	Problem-solving tasks	RTA plain video playback, validated with gaze-paths, not used as cues	RTA: Low risk of fabrication; Reliable despite omission; Validity unaffected by task complexity

[1] Those related studies with primary tasks *not* involving interactive technology should be analysed separately. For instance, Russo et al 1989 conducted a comparison study on problem-solving tasks with four TA conditions: CTA, RTA with one's own solution (*response-cued*) , RTA with the original problem (*stimulus-cued*), RTA with eye-fixations data (*prompted*). Their results showed that CTA led to lower task performance (reactivity) and RTA had the issues of omission and fabrication (nonveridicality). Kuusela and Paul (2000) conducted a comparison study on decision-making tasks on insurance offers between CTA and RTA (free-recall, un-cued) and concluded that CTA outperformed RTA in terms of number of verbal protocol segments.

Eger, Ball, Stevens & Dodd [34]	2007	Search Engine	CTA vs. RTA plain video playback vs. RTA gaze-paths video playback	RTA gaze-paths > RTA plain > CTA: More UPs; Useful for evaluating more complex system
Hyrskykari et al [54]	2008	Car brokerage website	CTA vs. RTA gaze-paths video playback	RTA > CTA: More data of higher quality; cognitive > behavioural
Peute, de Keizer & Jaspers [85]	2015	Digital health tech	CTA vs. RTA	CTA < RTA: Lower task performance for complex tasks; longer time, more errors, lower completion rate. CTA > RTA: More UPs of minor cosmetic nature. RTA > CTA: Verbalisations explanatory, more useful for redesign, more insights into complex UPs
McDonald, Zhao & Edwards [75]	2013	University website	CTA vs. RTA stimulus-cued (the website evaluated)	CTA>RTA: Number of UPs RTA> CTA: Better understanding about causes and impacts of UPs
Alhadreti & Mayhew [2]	2018	University library website	CTA vs. Hybrid (CTA+RTA) vs. RTA plain video playback	CTA ~ Hybrid > RTA: more UPs, minor problems on layout; cost-effective (time)

Appendix B: Data normalization algorithm

Here below is a concise description of the algorithm for data normalization, which is based on Braithwaite & Watson [19].

a) Read recorded data from HR and GSR files and split them into the different tasks (and resting period). Splitting of Recall data is based on "Video playback started" event in ExternalEvents. Values less than 0 are ignored, as those are clearly sensor faults.

b) Outliers (less than Q1 - 1.5 * IQR and larger than Q3 + 1.5 * IQR) are identified and removed for each task data sets (including resting period).

c) Minimum value during resting period is identified and all values in tasks less than this value are removed from the data sets.

d) Overall maximum is identified, by checking for the largest value in all the tasks.

e) Currently the Z-score values are calculated with the formula "ZScore = (SCR - mean) / stdev" based on the cleaned datasets (outliers and values less than resting period minimum are removed). Specifically, mormalized values are calculated for each value, by subtracting minimumDuringResting and dividing by the personal range (maximumValueOverAllTasks - minimumValueDuringResting).

Appendix C: SAM Data

C.1.1 SAM Data Collection Method

The Self-Assessment Manikin (SAM) [18] was employed in Study 2 (phases 1 and 3) to complement the real-time physiological data. We used a 9-point scale for the three dimensions – Valence/Pleasure, Activation/Arousal, and Control/Dominance, see section 2.4.

The two independent variables of Study 2 were the length of time with three levels (0-hour, 1-hour and 24-hour) and the stimuli of neutral valence with two levels of arousal (High, Low). Each participant was asked to complete SAM after each task in Phase 1 and Phase 3 (Fig. 6 in the main text).

- **Actual-Task{i} (i = 1 to 11):** In the actual interaction phase, after each of 11 tasks has been performed;
- **Recall-Task{i} (i = 1 to 11):** In the cued-recall phase, after the video of each of 11 tasks has been viewed;

C.1.2 SAM Hypotheses

The variable SAM-delta is operationalized as subtracting a Recall SAM rating from its corresponding Actual SAM Rating; this is applied for each of the three dimensions (Pleasure, Arousal and Dominance). The following SAM-related hypotheses (H5-H6) are formulated:

- **H5:** There are no significant differences in the SAM ratings between the Actual and Recall tasks for all the eleven tasks.
- **H6:** There are significant differences in the SAM-delta ratings **(a)** among the three groups with different lengths of intervening time (0-hour, 1-hour and 24-hour); **(b)** between the two stimuli groups (High arousal vs. Low arousal).

In Study 2, both objective physiological data (GSR, HR) and self-report subjective data were collected. As discussed in Section 2.4, it is uncommon that these two sets of data are not significantly correlated [11]. The related hypothesis is formulated as follows:

- **H7**: There are no significant correlations between the objective physiological data and self-report SAM data.

C1.3 SAM Results: Effect of Time and Stimuli - Post-task SAM Ratings

Significant differences in SAM ratings between actual and recall tasks

As the data are not normally distributed (Shapiro Wilk test, $p < .05$), non-parametric two-related sample (within-group) Wilcoxon Signed Ranks tests were used to evaluate whether the participants, for each of the eleven tasks, had different ratings for each of the three SAM dimensions after actual interaction and after cued-recall. Table C1 summarizes the findings.

Table C1. A summary of the results of significant differences in the SAM ratings *between actual interaction and cued-recall*. When a significant difference ($p<.05$) for a dimension is detected, the corresponding symbol (P = Pleasure, A = Arousal, D = Dominance) together with *Z* value of Wilcoxon Signed Ranks test is entered in the respective cell. The red bordered columns, T4 and T10, show the highest and lowest number of significant differences under different conditions, respectively.

	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11
All (N=61)	P 3.18 D 2.37	–	P 2.19 D 2.73	D 2.92	–	P 1.96	P 2.3 D 2.56	–	–	–	–
Time: 0-h Stimuli-No (N=18)	P 2,57 D 1.99	–	–	–	D 2.05	–	P 2.2	–	–	–	–
Time: 1-h (N=21)	P 2.35	–	–	–	–	–	–	–	–	–	–
Time: 24-h (N=22)	–	D 2.22	D 2.37	D 3.13	–	–	–	–	–	–	A 2.63
Stimuli-High (N=23)	P 2.31	–	–	D 3.11	–	–	–	–	–	–	–
Stimuli-Low (N =20)	–	P 2.42	–	P 2.27	–	–	–	–	–	–	–

Some interesting patterns emerged. When considering all conditions, the number of significant differences in any of the three dimensions (P, A or D) of the SAM ratings between the actual interaction and cued-recall sessions was low (cf. 17 out of 66 cells in Table C1) with a notable contrast between T4 (4 out of 6 conditions) and T10 (none). In fact, T10 was a simpler task involving one action (remove an email) as compared three actions for T4 (write -> save -> delete an email). Note that T10 was seeded with usability problems while T4 was not.

Table C2. Descriptive statistics of the SAM ratings for the two tasks over all participants. For T4, the difference in the Dominance rating between actual interaction and cued-recall was significant whereas none of the dimensions was significant for T10.

N = 61 (All participants)	T4				T10			
	Actual Interaction		Cued-Recall		Actual Interaction		Cued-Recall	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Pleasure	6.38	1.54	6.11	1.58	5.97	1.57	5.55	1.81
Arousal	3.66	1.85	3.39	1.74	4.13	1.85	3.90	1.85
Dominance	*7.36*	1.78	*6.75*	1.91	6.41	1.99	6.05	2.07

The descriptive statistics (Table C2) show that in both actual interaction and cued-recall, T10 had lower Pleasure and Dominance ratings than T4. In contrast, T10 had higher Arousal ratings than T4. These suggest that the usability problems seeded in T10 might undermine positive feeling and sense of control while increasing the activation level, and that the respective levels remained irrespective of the intervening time and stimuli. In fact, 12 out of 17 significant differences were tied to T1, T2, T3 and T4, which were *not* seeded with usability problems. We argued that the usability problems seeded in the other tasks (T5-T11) led to stronger emotional responses in the participants, who could recall them more consistently with the given cues. Overall, H5 was rejected.

Significant differences in SAM-delta ratings across the conditions

We looked into the extent to which the participants change their ratings from the actual interaction to cued-recall sessions. The range of changes can be from -8 to +8, given the 9-point Likert scale of SAM we used. To render visualisation less complicated (Fig. C1), we grouped the changes into five sub-ranges: Decreased by -5 to lower bound (-6, -7 or -8); Decreased by -1 to -4; Unchanged (0); Increased by 1 to 4; Increased by 5 to upper bound (6, 7 or 8).

In general, Pleasure and Arousal showed similar patterns. For most of the tasks, the percentages of Unchanged and Decreased by -1 to -4 were comparable and slightly higher than Increased by 1 to 4. The percentages of the extreme changes at both ends were relatively small (less than 5%). However, Dominance showed a different pattern: the percentage of Decreased by -1 to -4 was higher than that Unchanged for seven tasks. The percentage of the extreme change Increased by 5 to 7 was higher than in the case of Pleasure and Arousal. Among the eleven tasks, T7 ('Send a Draft' with seeded usability problem) showed a unique pattern for all three ratings: The percentage of Increased by 1 to 4 was higher than that of the other subranges. As shown in Table C1 (the first row), the increases in the Pleasure and Dominance ratings were statistically significant. The seeded usability problem of T7 caused the blocked access to draft emails, leading to unpleasant feeling and perceived loss of control in the participants during actual interaction (MeanPleasure = 3.74, SD =1.58; MeanDominance = 3.69, SD=2.02), but the intensity of these negative emotional responses reduced over time in cued-recall (MeanPleasure = 4.21, SD=1.87; MeanDominance = 4.48, SD=2.39).

Fig. C1. Changes in the SAM ratings (a) Pleasure (top), Arousal (middle) and (c) Dominance (bottom) across the eleven tasks and over all cases (N = 61).

To evaluate H6a and H6b, for each of the eleven tasks, we computed *delta ratings* for each of the three SAM dimensions for each participant. In other words, each participant had 33 delta ratings (i.e. 3 dimension x 11 tasks). As the delta scores were not normally distributed (Shapiro Wilk test, $p < .05$), non-parametric k-

independent (between-group) Kruskal-Wallis tests were used to evaluate whether there were significant differences among the three intervening time groups (0-hour, N = 18; 1-hour, N = 21; 24-hour, N = 22).

Results showed that among the 33 SAM-delta ratings, only two were significant, namely Pleasure for T4 (H = 5.96, df = 2, p = 0.05) and Dominance for T5 (H=7.21, df =2 , p = 0.027) (for descriptive statistics, see Table C3). Note that T4 is 'write and delete email' without seeded usability problems whereas T5 is 'add attachment' with seeded usability problems (Table 1 in the main text). Mann-Whitney tests were used to evaluate the delta Pleasure ratings for T4. The 0-hour group differed significantly from the 1-hour (U = 118, Z = 2.059, p = 0.039) and from the 24-hour groups (U = 119.5, Z = 2.23, p = 0.026), indicating that 0-hour group tended to increase their Pleasure ratings (M = 0.44, SD = 1.25) whereas the 1-hour and 24-hour tended to decrease their ratings (M=0.57, SD = 2.23; M= 0.55, SD = 1.26) in the Recall phase. Similarly, in case of Dominance for T5, results of the Mann-Whitney tests showed a highly significant difference between 0-hour and 24-hour (U = 100, Z = 2.681, p<.001), with the 0-hour group decreasing their Dominance rating during cued-recall (M = 1.5, SD = 2.75) and the 24-hour group increasing it (M=0.86, SD = 2.32).

Table C3. Descriptive statistics for the Pleasure-delta ratings for T4 and Dominance-delta ratings for T5 where significant differences among the three intervening time groups were found.

Pleasure-Delta for T4	Mean	SD	Medium	Range
0-hour (N=18)	-0.44	1.25	0	(-4) to 1
1-hour (N=21)	0.57	2.23	1	(-4) to 7
24-hour (N=22)	0.55	1.26	0.5	(-2) to 3
Dominance-Delta for T5	Mean	SD	Medium	Range
0-hour (N=18)	1.5	2.75	1	(-2) to 8
1-hour (N=21)	0.1	2.57	1	(-6) to 4
24-hour (N=22)	-0.86	2.32	-1	(-6) to 4

Similarly, Mann-Whitney tests were applied to evaluate whether there were significant differences between the two intervening affect groups (i.e. High- and Low-arousal-stimuli). Results showed that there was only one significant difference in the Pleasure-delta for T2 (U = 152, Z = 1.953, p = 0.05) (for descriptive statistics, see Table C4), indicating that the High-arousal-stimuli group tended to increase their Pleasure rating (M = 0.04, SD = 1.80), though slightly, whereas the Low-arousal-stimuli group tended to decrease it (M=1.00, SD =1.59).

Table C4. Descriptive statistics for the Pleasure-delta ratings for T2 where significant difference between the two intervening affect groups was found.

Pleasure-Delta for T2	Mean	SD	Medium	Range
Low-Arousal (N=23)	-0.04	1.79	0	(-4) to 4
High-Arousal (N=20)	1	1.59	1	(-2) to 4

Correlation between SAM ratings and GSR/HR data

To verify the hypothesis on the relationship between the objective physiological data and subjective self-report data, bivariate Pearson's correlation tests between the two sets of data were computed, given the normal distribution (Shapiro Wilk test, p>0.05). Prior to that, the data normalization process of GSR/HR data [19] was carried out to control individual differences, and then the mean GSR/HR per task was calculated.

Results are shown in Table C5 (actual interaction) and Table C6(cued-recall). None of the mean GSR data for actual interaction or cued-recall were significantly correlated with the SAM ratings. In contrast, significant correlations were found between the mean HR data and SAM ratings for two tasks (T2, T5) in the case of actual interaction and for four tasks (T1, T8, T9, T10) in the case of cued-recall.

To examine whether there would be the peak-end effect as observed in [19], we computed the correlations between the maximum GSR/HR measures (i.e. the peak value) of a specific task and the

corresponding post-task SAM ratings; the same for the last GSR/HR measures of that task (i.e. the end value). For GSR of actual interaction, T7 had significant correlations for the peak and end value, and T8 for the peak value. For GSR of cued-recall, none of the tasks had any significant correlation for the peak and only T7 for the end. A slightly different pattern was observed for HR: for actual interaction, two tasks (T8 and T9) for the peak and two tasks (T6 and T7) for the end were significant; for recall, only one task (T8) for the peak and four tasks (T2, T3, T8 and T9) for the end. Overall, the Peak-End effect was not strongly supported by the findings.

Table C5. Correlations between SAM ratings and GSR/HR data per task (T) for actual interaction. *N = 45, ns.= non-significant.* In case of significant correlation (*p<.05)*, the corresponding symbol (P=Pleasure, A=Arousal, D=Dominance) and *r* value is entered in the respective cell.

	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11
GSR (mean)	ns	ns	ns	ns	ns	Ns	ns	ns	ns	ns	ns
GSR (peak)	ns	ns	ns	ns	ns	Ns	P(.42) D(.38)	A(.35)	ns	ns	ns
GSR (end)	ns	ns	ns	ns	Ns	Ns	P(.36) D(.37)	ns	ns	ns	ns
HR (mean)	ns	A(.34)	ns	ns	A(.31) D(.32)	Ns	ns	ns	ns	ns	ns
HR (peak)	ns	ns	ns	ns	ns	ns	ns	P(.32)	D(.33)	ns	ns
HR (end)	ns	ns	ns	ns	ns	P(.32)	D(.31)	ns	ns	ns	ns

Table C6. Correlations between SAM ratings and GSR/HR data per task (T) for cued-recall. *N = 45, ns.= non-significant.* In case of significant correlation (*p<.05)*, the corresponding symbol (P=Pleasure, A=Arousal, D=Dominance) and *r* value is entered in the respective cell.

	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11
GSR(mean)	ns	ns	Ns	ns	ns	ns	ns	ns	ns	ns	Ns
GSR (peak)	ns	ns	Ns	ns	ns	ns	ns	ns	ns	ns	Ns
GSR (end)	ns	ns	Ns	ns	ns	ns	A(.31)	ns	ns	ns	Ns
HR (mean)	ns	ns	Ns	ns	ns	ns	ns	P(.31)	P(.31)	P(.41) A(.37)	Ns
HR (peak)	ns	ns	Ns	ns	ns	ns	ns	D(.33)	ns	ns	Ns
HR (end)	ns	A(.48)	P(.31)	ns	ns	ns	ns	P(.31)	D(.37)	ns	Ns

<enddoc>