Prompto: Investigating Receptivity to Prompts Based on Cognitive Load from Memory Training Conversational Agent

SAMANTHA W. T. CHAN, Augmented Human Lab
SHARDUL SAPKOTA, Augmented Human Lab
REBECCA MATHEWS, Augmented Human Lab
HAIMO ZHANG, Augmented Human Lab
SURANGA NANAYAKKARA, Augmented Human Lab

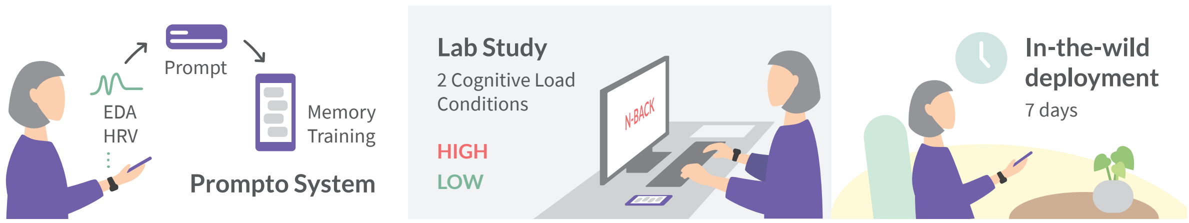

Fig. 1. Overview of the Prompto system which measured user's electrodermal activity (EDA) and heart-rate variability (HRV) to trigger prompts for memory training sessions (left), the setup of our lab study which issued prompts during high and low cognitive load conditions (middle), and our evaluations which included Prompto's in-the-wild deployment for 7 days (right).

Prospective memory lapses, which involve forgetting to perform intended actions, affect independent living in older adults. Although memory training using smartphone applications could address them, users are sometimes unaware of available times for training or forget about it, presenting a need for proactive prompts. Existing applications mostly provide time-based prompts and prompts based on users' cognitive contexts remain an under-explored area. We developed Prompto, a conversational memory coach that detects physiological signals to suggest training sessions when users are relaxed and potentially more receptive. Our study with 21 older adults showed that users were more receptive to prompts and memory training under low cognitive load than under high cognitive load. Interviews and an in-the-wild deployment of Prompto indicated that majority of users appreciated the concept, found it helpful and were likely to respond to its prompts. We contribute towards developing technologies with cognitive context-aware prompting based on users' physiological readings.

CCS Concepts: • **Human-centered computing** → **User studies**; **Natural language interfaces**.

Additional Key Words and Phrases: receptivity, cognitive load, conversational agent, physiological sensing, memory, older adults, context-aware notifications, conversational agent

Authors' addresses: Samantha W. T. Chan, samantha@ahlab.org, Augmented Human Lab; Shardul Sapkota, shardul@ahlab.org, Augmented Human Lab; Rebecca Mathews, rebecca@ahlab.org, Augmented Human Lab; Haimo Zhang, haimo@ahlab.org, Augmented Human Lab; Suranga Nanayakkara, suranga@ahlab.org, Augmented Human Lab.

2474-9567/2020/12-ART121 $15.00
https://doi.org/10.1145/3432190

ACM Reference Format:
Samantha W. T. Chan, Shardul Sapkota, Rebecca Mathews, Haimo Zhang, and Suranga Nanayakkara. 2020. Prompto: Investigating Receptivity to Prompts Based on Cognitive Load from Memory Training Conversational Agent. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 4, 4, Article 121 (December 2020), 23 pages. https://doi.org/10.1145/3432190

1 INTRODUCTION

Prospective memory (PM) relates to remembering to perform intended actions [27], such as remembering to take medication. It remains a major factor for independent living in older adults [40, 41], yet PM lapses are the most common type of lapses reported in everyday forgetting [21, 75]. Digital reminder systems and virtual assistants on smartphones, like Google Assistant [59] and Siri [46], serve as external PM aids [8]. PM lapses can also be addressed by maintaining or improving cognitive functions through smartphone memory training applications (apps) [84]. Users might miss opportunities for training sessions when they are unaware of available times or forget to do so. Thus, many existing memory training apps allow users to set time-based reminders for the sessions to promote regular training [22, 55, 61].

However, according to Sarker et al. [82], a well-timed prompt may not be effective if the user is not cognitively available to engage in the intervention. Inspired by their work, we focused on exploring proactive prompts based on users' cognitive contexts and physiological states. We developed Prompto, a conversational agent which guides users in applying an effective memory technique, the "when-then" technique, which has been known to strengthen PM in older adults [18, 102]. Prior work indicated that digital memory training with this method could allow users to be more on-time in performing PM tasks [15]. We extend this work by presenting the technique in a dialogue-based, natural-language interface. Using users' electrodermal activity (EDA) and heart rate variability (HRV) readings, Prompto estimates their cognitive load, and initiates dialogue to suggest moments for training when they are more relaxed and potentially more receptive to undergoing training sessions (Figure 1-left). Natural-language interaction and physiological understanding of users and could create effective conversational memory tutors which relate to and understand us better. Our work contributes with:

- an investigation of the effects of cognitive load on older adults' receptivity to prompts (Figure 1-middle) showing that users were significantly more responsive to prompts and engaged in more memory training in low cognitive load compared to high cognitive load;
- the design and implementation of Prompto, a conversational agent which prompts for memory training sessions based on physiological readings during low cognitive load;
- an analysis of user attitudes towards Prompto and usage behaviour in an in-the-wild deployment (Figure 1-right) revealing that a majority of users appreciated the concept, found Prompto helpful and were likely to respond to such prompts (in-the-wild response rate at 67.9%, lab study response rate at 90.5%).

2 RELATED WORK

2.1 When-Then Technique

The "when-then" technique [33, 34], also called implementation intentions, was mainly taught by researchers in the lab setting and was shown to enhance PM performance [18]. The technique has two steps: 1) to formulate and verbalise a "when-then" sentence [10], for example, "When I leave home after breakfast tomorrow, then I will bring the documents.", and 2) to visualise yourself performing the action. It helps users to form stronger associations between situation cues (event, time or location) and intended actions [35], to increase the likelihood of performing their PM tasks. Our system guides and works with users to apply the technique on their own tasks. Since using the "when-then" technique demands high attention and planning [63], Prompto was developed to initiate practice sessions when users are more cognitively available.

2.2 Memory Training Systems

Digital memory training can be classified into process-based training and strategy-oriented training [9, 31, 41]. In process-based training, memory exercises are repeated with increasing difficulty depending on the user's performance. Many digital cognitive training programmes use the process-based training approach for microlearning-styled memory training on computers and smartphone apps [22, 55, 61, 84]. SMART [68] improved working memory function in older adults, BrainHQ [22, 87] was shown to strengthen auditory memory, and Cognifit [61, 85] was found to enhance executive function and working memory in older adults. Specific to PM, Virtual Week [77], a computerised board game, used tasks associated with daily living to train PM. The virtual game made by Lin et al. [57] used tasks related to fishing and commerce to exercise PM.

Digital strategy-oriented training systems make mnemonic strategies accessible on-the-go and enable users to train on the use of these strategies. The memory palace project [37], Physical Loci [70], NeverMind [78] and HoloMoL [97] enabled users to use the method of loci, in which individuals mentally associate physical objects (loci) in a familiar location to information they want to remember, and enable users to virtually visualise these loci. SuperMemo [95] and Anki [26] facilitate spaced repetition, a method shown to improve remembering in adults with dementia [12], by using digital flash cards where new knowledge is learned and recalled over increasingly longer periods. The most relevant work was QuizBot [79], which facilitated spaced retrieval of factual knowledge through a conversational interface. It was more engaging and enabled higher recall than traditional flashcard apps.

Our work combines strategy-oriented training and process-based training for PM, training the use of the "when-then" technique and encouraging users to practice it through microlearning sessions. BrainHQ, CogniFit and Anki allow users to set time-based prompts for starting training sessions. However, there remains a lack of research on physiological-based prompts in memory training systems.

2.3 Conversational Agents for Memory and Learning

Conversational agents have previously been used for reminiscence where users recall and talk about their past memories [13, 90, 99]. Virtual assistants [46, 59], as well as embodied conversational agents like MANA [74] and Billie [96], support PM by managing and reminding users of their PM tasks. Since our work was aimed at teaching the use of a memory technique, it was closely related to pedagogical agents and intelligent tutors. Hirch et al. [42] discussed the design of social robots as memory coaches to help the elderly as memory training tools. They suggested that memory exercises should be conducted in a comfortable and concentrated learning environment, and should be trained with personal memories. Sansen et al. [81] also proposed a concept for a cognitive and physical robot coach. The concepts have yet to be developed into testable prototypes. Affective intelligent tutors, such as AutoTutor [25] and ALEKS [24], use users' physiological signals during teaching to adapt the systems' delivery and content of learning accordingly. We explored a different application of physiological measures - to estimate attention and cognitive load before the learning begins.

2.4 Context-Aware Notifications

App notifications on mobile devices are perceived negatively by users when triggered at inappropriate times. For example, when they are occupied with an on-going task as these notifications lead to a negative impact on not only their performance on the on-going task [32, 53] but also their emotional well-being [64, 101]. On the other hand, as shown by Iqbal and Horvitz [52], users are also sometimes amenable to mobile interruptions if they perceive the awareness coming from the notification as important. Thus, with a need to appropriately deliver notifications to users, recent studies have focused on using contextual information to understand how users interact with notifications. These studies have broadly yielded in works that investigate 1) interruptibility and attention management [1, 20, 44, 64, 73, 100], and 2) receptivity and engagement prediction [28, 62, 72, 91] with notifications in ubiquitous computing.

Research on interruptibility and attention management is concerned with investigating the use of behavioral and device usage patterns to attract user's attention via notifications. For instance, Ho and Intille [44] found through the use of wireless accelerometers that messages delivered at activity transitions were more positively received than messages delivered at random times. Yuan et al. [100] used device related features like notification mode, device screen status, internet connectivity, and features related to people's personality traits to show that users were more willing to be interrupted when in a positive mood. This result was also affirmed by physiological investigations and experience surveys from Sarker et al. [82] who found that users were more attentive to notifications when they were happy or energetic than when they were stressed.

Research on receptivity and engagement prediction, on the other hand, is concerned with whether the users engage with the notifications. As Turner et al. [91] have identified, being receptive or engaged - which indicates that a user consumed the content of the notification, is different from being reachable – which indicates that a user's attention was attracted. Pielot et al. [72] used a machine learning approach with data from experience sampling and found that users were more likely to consume the notification content when at home or travelling than when at work. Dingler et al. [23] developed QuickLearn, a vocabulary trainer, which introduced system-initiated microlearning sessions to users using notifications during idle moments. They found that users were more receptive to sessions when they were mobile and idle. Mehrotra et al. [65] found that users might be more receptive when idle but take more time to respond to such notifications compared to while the user is busy. Users were more attentive to notifications when busy but were likely to consider them disruptive and dismiss them. Their results also showed that disruptive notifications (54%) tend to be accepted, mainly due to the content being useful.

These prior works have shown that users' cognitive and affective contexts (mood, stress, emotions, attention, engagement, idling etc.) are factors in interruptibility management and receptivity prediction. Therefore, since physiological signals can be used to infer such contexts, we employed a physiological sensing approach for detecting opportune moments for interruption. Furthermore, these prior works usually require users' intermittent attention for manual data input which add to their burden, annoyance, and frustration – resulting in possible abandonment of the application by the users [94]. We thus build on these previous works, target idle and relaxed moments, and investigate how receptive users would be to memory training based on only physiological signals.

2.5 Physiological Measures

Our bodies' physiological activities, such as heart rate and perspiration, are regulated by our autonomic nervous system. Electrodermal activity (EDA) refers to electrical changes on the skin, generally measured through variations in skin conductivity due to reaction of sweat glands to external stimuli. Any variation in skin conductance of more than 0.01 microSiemens (μS) has been considered a Skin Conductance Response (SCR) [4]. EDA has been used as an index of cognitive load [17, 67, 86], where phasic features and sum of SCR values (accumulative SCR) were recommended features. Bahrainian et al. found that a memorable segment of a conversation is preceded by a fast decrease in EDA signal, after the local minimum [2]. Another measure that we identified was heart rate variability (HRV). It reflects changes in the time intervals between consecutive heartbeats called inter-beat-intervals (IBI) [83]. IBI is derived from photoplethysmography (PPG) signals which measure blood volume pulse. The root mean square of successive differences between normal heartbeats (RMSSD) has been a commonly-used time-domain estimate of HRV [11, 83]. HRV was found to be a good indicator to assess cognitive load [36] and a decreased HRV has been linked with sustained cognitive load and attention [43, 89]. Our work utilised EDA and HRV measures to estimate cognitive load.

3 PROMPTO

3.1 Design

Prompto was developed as a result of three design phases over 20 months. We describe the user studies, findings and insights from each phase.

3.1.1 Phase 1. As a starting point, we wanted to understand our users' needs and context regarding memory and memory interventions. A focus group with 8 older adults (*M*=68 yrs, *SD*=8, 4 female) was held to discuss on and gain insights into: 1) the memory lapses faced in daily life and 2) the solutions used to address them. The most relevant findings and observations were:

- Out of the types of memory lapses faced in everyday life (attention, episodic, prospective and semantic), they most frequently faced PM lapses. This finding was consistent with previous studies [21, 75].
- Writing to-do lists, using calendars and reminders were among the commonly-used solutions. A few mentioned memory training methods such as doing crossword puzzles and Sudoku.
- The use of Lumosity [55], a memory training app, by one participant sparked interest in other participants to try it for themselves. Their interest was because the app gave daily brain exercises, they saw it as a way to keep their brains active, and most of them had smartphones to use the app.

PM lapses remain a critical issue. Despite current solutions, the issue persists and commonly-used solutions only support memory but may not enhance or maintain memory functions. Memory training apps and methods have been used to fill this gap and keep the mind active. However, the cognitive exercises and games presented in many memory training methods are abstract and may be hard for users to apply their learning to their own daily PM tasks.

3.1.2 Phase 2. To address these needs and problems, we developed a prototype (Prototype 1) which facilitated practice of the "when-then" technique with naturalistic PM tasks that were common for older adults. Daily practice tasks were given. Users typed in details of each PM task in text fields on Prototype 1 to formulate sentences in the "when-then" format. Prototype 1 then showed them a timer and asked them to visualise themselves performing the PM task until the timer stopped. A 12-day in-the-wild testing of Prototype 1 and interviews were conducted with 10 users (*M*=70 yrs, *SD*=5, 3 male), two of the users were also participants from the focus group in Phase 1. Further details on Phase 1 and 2 are available [15]. The relevant findings are summarised below:

- Users were more on-time at completing PM tasks and felt that their PM improved.
- Users sometimes forgot their training sessions with the practice tasks for the day.
- Seven out of ten users indicated in the interviews that they would like to try out a conversational interface.
- Users also mentioned in the interviews about setting reminders for the sessions and choosing idle or leisure times to begin them, such as during commute or when watching TV.

There were benefits in using Prototype 1 but we exposed a potential problem that users might miss training sessions, resulting in irregular practice. Previous studies advocate that regular and frequent practice in memory training are needed to achieve sufficient neurobiological improvement [31, 77]. Thus, there needed to be a better way to engage users and promote regular training.

3.1.3 Phase 3. As such, we developed our second prototype (Prototype 2 - Prompto) which used the same memory technique but presented the training through a conversational interface and it incorporated physiological interaction to initiate training sessions when users were relaxed or idle. Since conversational agents use natural language and have been positively received by older adults [6, 76, 93], older adults might relate better to a conversational interface and it might motivate regular application of the “when-then” technique. Learning through a conversational interface would also be similar to how it has been verbally taught [18] and could support the experience of verbalising the "when-then" sentence (first step of memory technique). The growing interest by the Human-Computer Interactions community in tapping into user's internal state and context-aware notifications inspired our focus on exploring physiological interaction with Prompto.

Think-Aloud Testing with 3 users (*M*=63 yrs, *SD*=7, 1 male) was conducted to improve the conversation design. User testing sessions with 2 participants (55 and 57 yrs, 1 male, 30 minutes) were held to get initial reactions on the prototype [16]. They were asked to wear a physiological sensing device, then go through a baseline period and stressor activity. After that they were asked stay seated and relaxed, Prompto detected their relaxed state and they received prompts to start a training session. The findings from these two tests were as follows:

- Users liked being guided on how to use the technique through examples in the conversation but they disliked wordy responses.
- User reactions were positive in which they appreciated the idea of being prompted to do memory training during idle moments.

In this paper, we present the continuation of the work in Phase 3 to further investigate the use of Prompto with more users and how it might affect user receptivity in both a lab and in-the-wild environments [14]. Conversational responses were made more concise and long responses were broken into separate ones when necessary. Other major usability issues were resolved before the study in Section 4.

3.2 Implementation

In this subsection, we describe how Prompto was implemented in terms of hardware, conversation design and software (Figure 2).

3.2.1 Hardware. To support the interaction flow, we presented the conversational agent on a smartphone, which has been shown to be the main device for conversational agent use [98]. We implemented Prompto using a

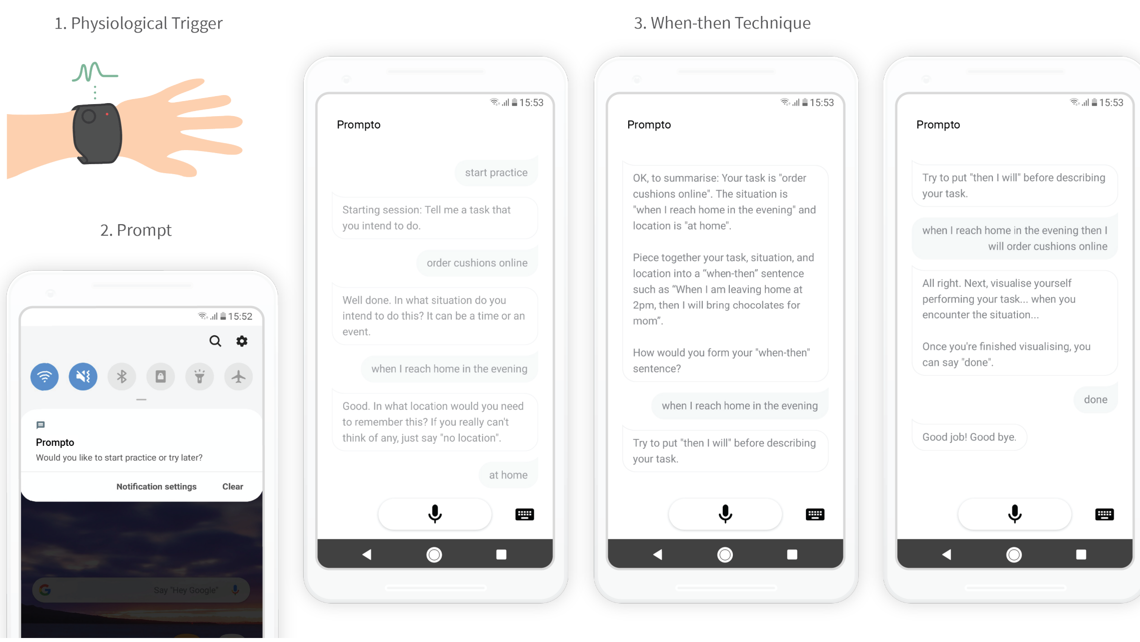

Fig. 2. Interaction flow and screenshots of Prompto: 1. physiological readings are monitored for appropriate user states, which triggers prompts from Prompto, 2. prompt is displayed as a notification and an audible notification sound, 3. user goes through memory training session for verbalisation and visualisation steps of the "when-then" technique.

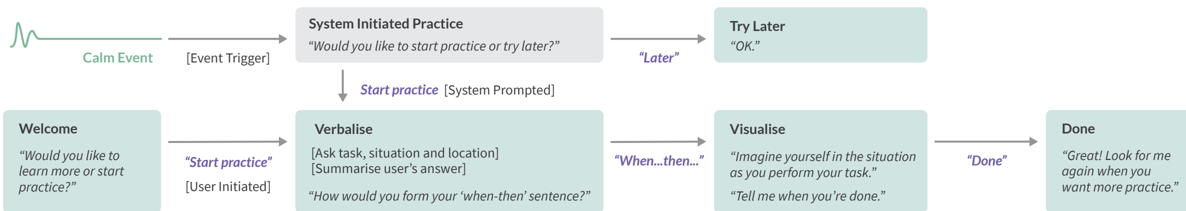

Fig. 3. Conversational flow for memory training session which could be system prompted via an event trigger from a detected calm event (top row) or user-initiated (bottom row). The boxes represent the intents for the Dialogflow agent and quoted text within boxes are example responses by the agent when the intent is triggered. The quoted text beside the arrows represent example user utterances which lead to the respective intents.

commercially-available physiological sensing device, the E4 wristband (E4) [48], and an Android phone (Samsung Galaxy A8, Android version 9). The E4 had the required EDA and PPG sensors, and it streamed the readings to an app on the phone via Bluetooth.

3.2.2 Conversational Interface and Structure. The app ran the conversational interface and communicated with a DialogFlow [58] agent which managed the dialogue between Prompto and the user. The app had a voice interface and a text-based chat interface for users who were uncomfortable with speech interaction or require silent interaction especially in public spaces. Having both voice and text-based interfaces has been suggested to overcome privacy and preference limitations of purely voice interfaces [69].

In DialogFlow terminologies, "intents" represented user intentions that were matched based on users' speech. Prompto starts with the "Welcome" intent that presents two choices: 1) to learn more about the system and technique or 2) to start practice. Figure 3 depicts the conversation flow for one session following the choice to start memory training. Users could initiate a practice session at any time, whereas proactive initiation occurred when the DialogFlow agent received an event request from the app (a feature in the DialogFlow API), invoking the "System Initiated Practice" intent. Upon receiving this intent, Prompto showed a visual notification and played an audible notification sound before speaking and informed the user that it noticed the user being more "calm", prompting to start practice or try again later. Once users chose to start practice, Prompto guided the user through the verbalisation step where the user was prompted to think about a task they intended to do, and related situation cues of event, time or location. Prompto provided a summary of user responses, then encouraged them to verbalise the "when-then" sentence and corrected the sentence when needed. The user was then led to the second step, visualisation, to visualise oneself performing the task with the situation cues in mind. After that, the session was complete and the user could have more practice, learn more or leave the conservation.

3.2.3 Software. The app also ran a cognitive load detector for lowered cognitive load that derived EDA and HRV-related features from incoming E4 readings. Baseline readings for calibration (taken when user is seated and relaxed) could be measured during the first 5 minutes upon receiving EDA and PPG data. We used EDA readings (in μS) and RMSSD (in *ms*) from PPG readings over 1-minute intervals. We implemented threshold detection for RMSSD of more than 30% of the average baseline value (increased HRV relates to lowered cognitive load) and EDA value within ±10% of average baseline value (near baseline EDA relates to lowered cognitive load). These intervals and thresholds were developed according to formulae and findings from previous work [17, 36, 83, 86]. The values had to be within thresholds for 1 minute, after which, an event request was sent to the DialogFlow agent which activated a prompt from the agent to suggest starting a practice session.

4 RECEPTIVITY STUDY

To investigate if there was value in triggering prompts for memory training during the low cognitive load state (relaxed or idle moments). We ran a within-subjects study with two conditions: prompting during high cognitive load condition (HCL) and prompting during low cognitive load condition (LCL). The prompts were issued independently of the cognitive load detector in the previous section, therefore its accuracy was irrelevant to this study. We used the n-back task for HCL, a previously established task to induce cognitive load [30, 36, 71]. For LCL, participants were told to have a break and relax while they watched a nature video, which was akin to relaxing and watching TV at home [29].

We compared users' receptivity to the prompts between conditions, analysing how the prompts were handled, response time and subjective ratings on how appropriate the timings of the prompts were. Prior work used ratings on how "disruptive" prompts were [65], we utilised the term "appropriateness" as an inverse to "disruptiveness". We also examined the subjective usefulness of Prompto. Finally, semi-structured interviews were held to gain better understanding of users' reactions to the system.

4.1 Participants

We recruited 21 participants (M=66 yrs, SD=6, 6 male) who were in the age range of 50 to 80, familiar with apps and computers, fluent or native English speakers, had normal or corrected-to-normal hearing and vision, and had no history of neurological diseases or epilepsy. Participants were recruited through email invitation and word-of-mouth. Ethical approval was obtained before the studies. Participants gave their consent to use their data and were reimbursed in local currency equivalent to USD $15.

4.2 Procedure

At the start of the session, we connected the E4 wristband to Prompto and assisted the participants to wear the E4 on their non-dominant wrist to reduce movement artefacts. This was to collect ground truth recordings of EDA and HRV data to objectively validate how participants' physiological responses were under LCL and HCL. During the first 10 minutes, as shown in Figure 4, they were introduced to Prompto and were asked to test it. The next 5 minutes was a baseline period where they were asked to sit and relax. This was followed by 5 minutes of completing practice rounds of the n-back task. In a round of the n-back task (auditory 2-back), participants heard a sequence of letters (trials) and were to press a key on the keyboard if they thought the current letter matched the one from 2 letters back (2 trials ago). Each round lasted for 48 seconds with 24 trials (with 4 possible matches), presented at 2 seconds per trial and received full score if they correctly identified matches without errors. They were allowed a 10-minute break before resuming the session and were instructed to write 4 to 5 tasks for the upcoming week, specifying situation and location accordingly. This task list was kept on the table for their reference in the next part.

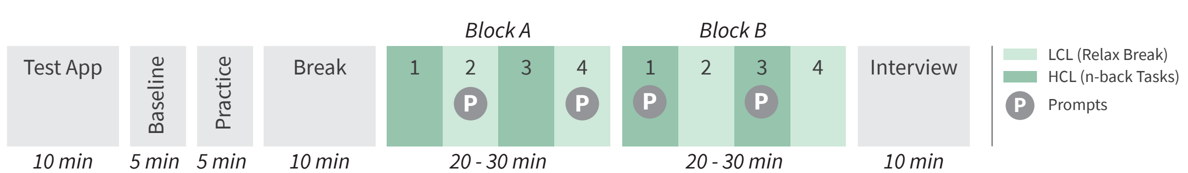

Fig. 4. Procedure for the Receptivity Study: participants went through two counterbalanced blocks of tasks and questionnaires with four alternating segments of relax breaks and n-back tasks. Prompts were given during LCL in Block A, and were given during HCL in Block B.

In the main study period, they underwent two 30-minute blocks of tasks (Blocks A and B) and were informed that they would be testing two different apps (one in each block). The order of the blocks were counterbalanced. Under each block, there were four segments in which participants alternated between doing 5-minute segments of n-back tasks (HCL, segments 1 and 3), where they were told to complete as many n-back rounds as they could, and 5-minute segments of a relax break while they watched a nature video (LCL, segments 2 and 4). In Block A segments 2 and 4 (prompting under LCL condition), prompts were presented while users were in the relax break. In Block B segments 1 and 3 (prompting under HCL condition), prompts were given while they were doing the n-back rounds. Participants self-paced the rounds and could pause them at anytime, therefore had chances to handle incoming prompts during HCL. For these segments with prompting, Prompto randomly issued two prompts that were spaced at least 3 minutes apart from each other to ensure participants had time to finish handling the first prompt. At the end of each segment, participants rated in questionnaires how appropriate the timings of the prompts were, which gave them at least a 2-minute break before the next segment/condition, such cool-down time has been used in previous work [17].

Prompts were given in both audio and visual modalities, where an audible notification sound was played before Prompto asked participants if they wanted to "start a memory practice session or try later" and the same text bubble appeared on the chat interface. At the start of each block, participants were told that the prompts might happen during any segment, and could be handled in three ways: 1) Say "Start Practice" to Prompto and complete a memory training session, 2) tell it to "Try Later" or 3) "Ignore" it. They were asked to only use voice input (by pressing the on-screen microphone button) for this study as we wanted to keep the memory training sessions as short as possible to reduce fatigue. Speech input was shown to be faster than text [80] while our preliminary testing indicated that memory training sessions using speech input on Prompto were usually under 3 minutes. Participants could refer to their written task list during training sessions to speed up entry.

Participants were given an optional break before starting a 10-minute semi-structured interview. The session was approximately 2 hours long.

4.3 Apparatus and Setup

Although participants were told that they were testing two different apps, we actually installed the same app on two phones (Samsung Galaxy A8, Android version 9) and presented one phone during each block. Participants sat in front of a computer screen with a keyboard, the phone and task list on the desk (Figure 1-middle). The screen showed the n-back tasks through Brain Workshop [45] and the nature video [66]. Participants did not take coffee, tea or smoke at least an hour before the study. Ambient conditions of room temperature, lighting and noise were controlled across sessions.

4.4 Results

4.4.1 Validation of Induced Cognitive Load. To validate that participants had higher cognitive load in that condition compared to the low cognitive load condition, we report EDA and HRV measures that objectively estimate workload, as well as subjective workload scores.

Objective Cognitive Load: We conducted continuous decomposition analysis on EDA data using Ledalab [4] in Matlab R2018b. SCRs were defined as any variation in skin conductance of more than 0.01 μS. We report the number of SCRs and sum of SCR amplitudes (accumulative SCR). We also analysed the HRV data using Kubios HRV 3.3 [88] in Matlab R2018b and report RMSSD values. We used EDA and HRV from segments that did not have prompts to avoid confounding factors as it has been shown that prompts cause SCRs [29] and interaction with the agent could also cause higher cognitive load. We did not find any significant difference in the number of SCRs (p=.71), accumulative SCR (p=.89), and RMSSD (p=.61) readings at baseline and during LCL. From this, we confirmed that relax breaks (LCL) induced roughly the same cognitive load as during baseline.

Since the normality assumption was not met according to the Shapiro-Wilk normality test (p<.05) for the number of SCRs and accumulative SCR, we conducted Wilcoxon signed rank tests to compare the differences. There was significantly higher number of SCRs (Z=203, p<.001) during HCL (*median*=73.5) compared to during LCL (*median*=49.2). There was a significant difference between the sum of SCR amplitudes (Z=206, p<.001) in HCL (*median*=3.65μS) than in LCL (*median*=1.34μS). For RMSSD values, a t-test showed a significant difference ($t_{35}=-3.37$, p<.001) between values during HCL (M=46.7, SD=11.9) and LCL (M=89.2, SD=10.4). Therefore, the two conditions indeed induced physiologically measurable cognitive load differences.

Subjective Workload: Participants were asked to fill in a NASA-TLX [38] form after each segment and each subjective workload rating represented a participant's average raw NASA-TLX score from four segments for low workload and four for high workload. A Wilcoxon signed rank test between the subjective workload ratings for LCL and HCL confirmed a significant difference in the scores (Z=231, p<.0001, $median_{low}$=16.4, $median_{high}$=65.3). Hence, participants reported experiencing higher mental load while doing n-back rounds (HCL) and lower mental load during relax breaks (LCL).

4.4.2 Receptivity to Prompts. Participants were asked to confirm if they noticed each prompt at the end of each segment where prompts were given. Two prompts (1 in HCL and 1 in LCL) were not perceived by one participant, it was likely that the prompts were not noticeable enough for her. As every prompt was noticed by participants, except these two, we took the two prompts as anomalies and they were not used in the analysis.

Handling of Prompts: The prompts were handled in three ways: 1) User said "Start Practice" (SP) to Prompto and completed that training session, 2) told it to "Try Later" (TL) or 3) "Ignored" (I) it. We grouped them into two pairs of outcomes, one pair compared outcomes of "Started Practice" ("SP", i.e., user started training session) and "No Practice" ("NP", i.e., user handled prompts with TL or I), the other pair compared outcomes of "Responded" ("R", i.e., user handled with SP or TL) and "Ignored" ("I", i.e., user ignored the prompt). We report the average count for these outcomes across conditions and conducted t-tests to compare the differences as depicted in Figure 5.

We found that participants started significantly more training sessions under LCL compared to under HCL ($Average\ Count_{LCL}$=2.52, $Average\ Count_{HCL}$=1.14, SD_{LCL}=1.29, SD_{HCL}=1.01, $t_{38}=-3.86$, p<.001), as shown in Figure 5a. There was a significantly higher "No Practice" count under HCL than under LCL (*Average*

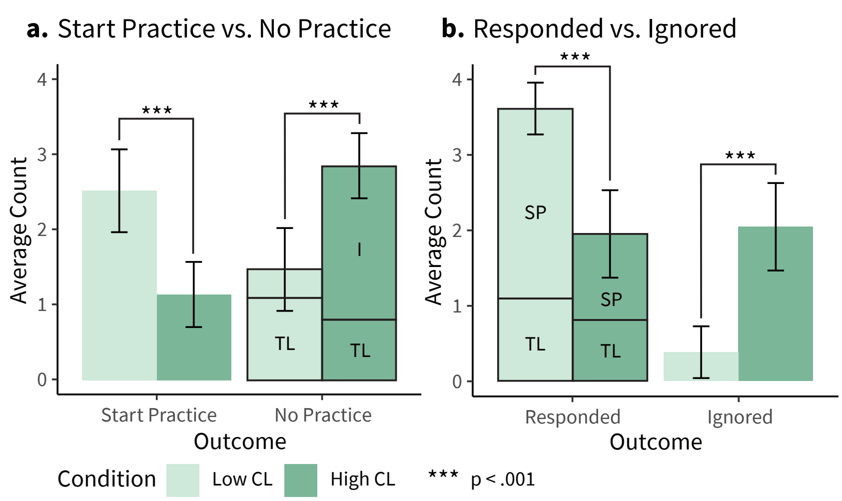

Fig. 5. Average count comparing outcomes: a. Start Practice and No Practice, and b. Responded and Ignored.

$Count_{LCL}$=1.48, $Average\ Count_{HCL}$=2.86, SD_{LCL}=1.29, SD_{HCL}=1.01, t_{38}=3.86, p<.001). When comparing the outcome pair of "Responded" and "Ignored" as shown in Figure 5b, participants were significantly more responsive to prompts under LCL than under HCL ($Average\ Count_{LCL}$=3.62, $Average\ Count_{HCL}$=1.95, SD_{LCL}=.80, SD_{HCL}=1.36, t_{32}=−4.84, p<.001). They were also significantly more likely to ignore prompts under HCL than under LCL ($Average\ Count_{LCL}$=0.38, $Average\ Count_{HCL}$=2.05, SD_{LCL}=.80, SD_{HCL}=1.36, t_{32}=4.84, p<.001). Hence, participants were likely to respond to prompts and to start training sessions during LCL. They were equally likely to respond or ignore to prompts during HCL and were likely to dismiss prompts if responding.

Response Time: We measured the response time as the time difference between when the prompt was issued and when the participants responded with either SP or TL. Since the normality assumptions were not met (Shapiro-Wilk normality test, p<.05), we conducted Mann-Whitney U tests to compare the differences in response times across conditions and across the two possible responses of SP and TL. We found no significant difference in the overall response times between LCL and HCL as shown in Figure 6a (U=377, p=.5, $median_{low}$=10s, $median_{high}$=10s). Between the conditions (Figure 6b), there were no significant differences in response times for SP (U=151, p=.96, $median_{low}$=10s, $median_{high}$=10.25s) and TL (U=47, p=.23, $median_{low}$=9.25s, $median_{high}$=8s). Thus, regardless of during LCL or HCL, participants took approximately the same amount of time to respond to the prompts.

Subjective Appropriateness of Prompt Timing: Participants were asked to rate how appropriate the timing of each perceived prompt was in each condition. Ratings were on a scale of 1 (not appropriate at all) to 9 (very appropriate). We took the average of the ratings of all participants for each condition, which we report as the average subjective appropriateness. Since the normality assumption was satisfied according to the Shapiro-Wilk normality test (p>.05), we used a paired t-tests to compare the differences in the ratings. The average subjective appropriateness rating for prompts in LCL (M=7.04, SD=1.65) was significantly higher (t_{20}=−4.7, p<.001) than

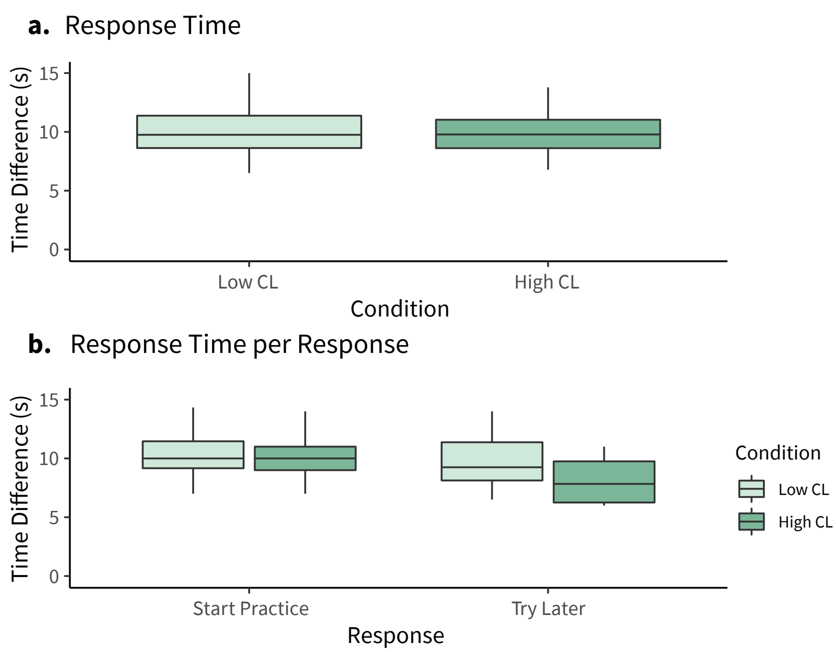

Fig. 6. a. Response times under LCL and HCL, and b. response time across conditions for "Start Practice" and "Try Later" responses.

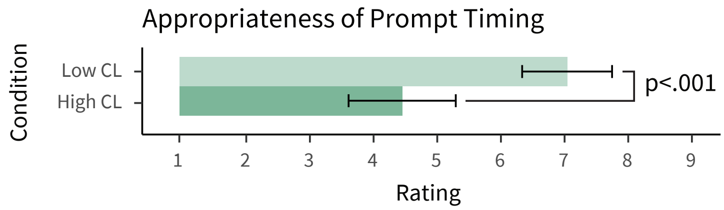

Fig. 7. Average subjective appropriateness of prompt timing ratings between conditions

in HCL (M=4.45, SD=1.97) as shown in Figure 7. Therefore, participants felt that the timings of prompts received during LCL were more appropriate than when received under HCL.

4.4.3 Usefulness Ratings. Participants rated on how useful they felt the apps from each condition were, on a scale of 1 (not useful at all) to 9 (very useful). Participants mainly rated the usefulness of the technique rather than the prompts. A Wilcoxon signed rank test showed that the usefulness rating for the app in LCL (*median*=7) was significantly greater (Z=7.5, p<.01) than that for HCL (*median*=6). Although the apps were the same and had the same functionality, users felt that the app presented in LCL was more useful than the one presented in HCL.

4.4.4 Semi-Structured Interviews. Our interviews sought to gain understanding of user behaviours and requirements. It had four guiding questions regarding the reasons for prompt handling, feedback on Prompto's concept, possibility of future usage and suggestions for improvement. We transcribed the interview responses and present the findings categorised according to the four questions. Using the thematic analysis method [7], two researchers independently coded the data to extract main ideas from the responses and generate initial key themes. After which, they reviewed the themes, codes and responses, and lastly, defined the final key themes as described within the categories below.

Reasons for Prompt Handling: 15 participants (71.4%) expressed that they started practice during LCL because they felt more available (*"Because I'm just relaxing and I've got time to attend to the phone."*, P14). 4 participants (19%) commented on ignoring the prompts during HCL due to being annoyed, frustrated or interrupted by them as they wanted to complete the n-back rounds (*"It was very annoying when it prompted when I was trying to do something. I did notice a difference when I was relaxed."*, P12). We iterate that the rounds could be paused by participants at anytime they chose. 3 participants (14.2%) decided to start practice after some of the prompts during HCL, because focus was already lost due to the prompts (P8), as means of escape from n-back rounds (P15) and because it was easier (P9). We anticipated this behaviour and noted that these were rare occurrences. Overall, more participants felt available to engage in training sessions during LCL.

After receiving answers for this part, we revealed to the participants the purpose of the study and the idea of Prompto prompting during times of low cognitive load (relaxing or idling times).

Feedback on Concept: 11 participants (52.3%) embraced the idea of Prompto prompting for memory training sessions while they have lower cognitive load, when they are more relaxed or during idle moments and thought that it made sense. (*"I quite like the idea of biofeedback, informing the timing of when I do certain things. I quite like that, it's sort of a friendly helper."*, P19). 5 of these participants would like the system to detect and prompt during their preferred times of mornings (P3, P9), early evenings (P5) or idle moments (P6, P18).

In contrast, 4 participants (19%) mentioned that they preferred not to be disturbed during the relaxed moments (*"I didn't like being interrupted when I was relaxing. It's like someone coming to annoy me when I'm watching TV."*, P21). Interestingly, two of these participants indicated preferring to be prompted when busy (*"I'm used to working under pressure. My grandchildren interrupt me constantly when I'm doing something so in my mind I actually found [that] the second one [the app during HCL] did not intrude."*, P9). Hence, a slight majority of participants liked the concept prompting for training sessions during relaxed or idle moments but we do need to acknowledge that a small minority did not prefer it.

Possibility of Future Usage: 13 participants (61.9%) did foresee themselves using the system in the near future, reasons include finding the system useful, and feeling that they did face memory issues and needed memory training (*"I'm not getting any younger, memory is not getting any better."*, P5). 6 participants (28.6%) felt that the system was not applicable for them now but in the future (after 3 to 5 years) or it was more suitable for others (*"Maybe in the future when I'm no longer working."*, P4). Two participants felt that they were unlikely to use it because they were already using external aids to help them remember (*"I think I've got some quite strong techniques already. I do a lot of lists, I remind myself the day before."*, P19). However, one of these participants (P19) did state her interest to use the memory technique *"I would experiment with it [the technique]"*. Thus, most participants foresaw themselves using Prompto for the value of maintaining or improving their memory through the technique.

Suggestions for Improvement: Six participants (28.6%) did not mind wearing a wristband, such as the E4, when using the system on a regular basis. This might be because most of these six people regularly wore their own watches or Fitbit [51] devices and were therefore used to wearing devices on their wrists. On the other hand, 7 participants (33.3%) disliked wearing wristbands or watches because they do not wear watches in general (P1, P5), avoided wearing metallic objects (P8, P13) or disliked the form factor of the E4 itself and commented that it was *"ugly"*, *"big"* and *"clunky"* to wear (P3, P14, P21). Participants suggested that the wearable device or the wristband needed to be slim (P3, P8, P14, P21), discreet (P8, P16), multi-functional with watch and sensing features (P6, P14), light-weight (P8), durable (P16) and attractive (P3). A few recommended different form factors, such as a necklace (P8, P16), ring (P3), clip on the pocket (P1) or on the bra (P16). Therefore, participants were either against or neutral towards wearing wrist-worn sensors and future designs should consider meeting these elicited user requirements or using different form factors.

4 participants were familiar with using virtual assistants, such as Siri [46] (P1, P15) and Alexa [47] (P10, P11). 3 participants (14.3%), including P10 and P11, felt that the system's speech recognition should improve. Other improvements were that the conversation be made more fun and have animations (P21), have customisable voice and prompt sound (P3, P19), personalised messages (*"it's quite nice when it uses your name"*, P3) and encouraging messages (*"One thing that I do like about the Fitbit is that when I reach 10,000 steps it would say 'well done', that's quite nice if you put messages like that."*, P3). Although the conversational interface was usable, a small group of participants felt that there was room for improvement. We note that these suggested improvements were minor and would not affect functionality of the system.

5 participants thought that the system should have several alternatives available other than the "when-then" technique, for example, have *"techniques that you could progress your way through"* (P15), specific memory applications like *"learning a language"* (P19) or one that *"tests on tasks that have been entered before"* (P12). P18 and P19 saw a separate use for the physiological state detector part of the system where it could be used to understand themselves better for mindfulness and wellness purposes. Thus, a few participants saw value and potential in extending Prompto's capabilities to other applications beyond memory training.

5 IN-THE-WILD DEPLOYMENT

From the Receptivity Study (Section 4), we see that there are benefits to prompting during low cognitive load. In this deployment, we wanted to see if these benefits could still hold and to identify the additional challenges and requirements to realistically prompt based on physiological triggers in-the-wild.

5.1 Users

Participants from the Receptivity Study were invited via email to test Prompto in-the-wild. Seven community-dwelling users volunteered to take part (M=67.4 yrs, SD=4.5, 2 male). Two of them worked full-time, two of them worked on an ad hoc or part-time basis and the other three were retired but active. Four of them lived with their spouses, two lived alone and one lived with family.

5.2 Implementation and Pilot Test

Prompto was implemented with the same hardware and conversation design as described in Section 3.2. It still issued prompts when low cognitive load was estimated and users could start conversations with it at anytime.

The prompting system and thresholds were tweaked based on analysing data from the Receptivity Study and a 1-week pilot field test. The test was done with the seven users in which they used Prompto and wore the wristband each day. Prompto did not issue any prompts and recorded their EDA and PPG readings. They were asked to converse with it whenever they wanted. This test provided us with an estimate of how often users would use Prompto without prompts and enabled us to make final adjustments to the thresholds for each user.

Raw EDA readings (4 Hz) from the E4 wristband were directly used by Prompto. The wristband automatically processed PPG signals to provide inter-beat-intervals (IBI) values at 1 Hz. We implemented a function in Prompto which calculated the root mean square of successive differences between IBI values (RMSSD) [11] from the IBI data based on the following equation:

$$RMSSD = \sqrt{\frac{1}{N-1}\sum_{n=1}^{N-1}(IBI_{n+1} - IBI_n)^2} \quad (1)$$

The same equation was used by Kubios HRV [54, 88]. For every IBI reading received (about every second), the RMSSD was calculated from the previous 1 minute of IBI data (the 60 most recent IBI values, N = 60). Since a low EDA and high RMSSD indicated a low cognitive load state, the baseline EDA (EDA_{base}) and RMSSD ($RMSSD_{base}$) values for each user were identified from the logged data in the 1-week test by separately finding the lowest EDA reading and highest RMSSD value where there was no physical movement (accelerometer data was smooth) [50]. To detect a lowered cognitive load state, Prompto computed the average EDA (EDA_{avg}) and RMSSD ($RMSSD_{avg}$) over 1 minute intervals. If EDA_{avg} was greater than 60% of EDA_{base} and less than 140% of EDA_{base}, and $RMSSD_{avg}$ was greater than 80% of $RMSSD_{base}$, a lowered cognitive load state was detected and a prompt was issued. Both EDA and RMSSD criteria had to be met in the same 1-minute window. A demonstration application of the prompting system is available online[1].

For the main deployment, we set a limit for Prompto to issue at most 2 prompts a day and prompts were restricted to be at least 5 hours apart. This was set so as to follow suggestions by previous studies which gave on average 1.5 treatments per day and advised not to give too many treatments per day [56]. The volume of the notification sound was made customisable and can be put to silent or vibration-only. In the chat view, Prompto's speech output could be interrupted at any time by pressing anywhere on the chat screen. It also automatically listened for speech input from user, thus users did not need to press the on-screen microphone button to speak to it. We implemented an iOS version of Prompto as majority of users had iOS phones. Although three users mentioned in the interviews (from previous section) about their dislike for wristbands and one user preferred not to be disturbed during relax moments, we did not modify the system to fit these preferences but confirmed that they were still willing to test the system.

[1] https://github.com/cwtsam/prompto-demo

5.3 Procedure

Prior to deployment, we met each user in a 15-minute session where we guided them to install the app on their phones and passed a E4 wristband to them. They were asked to use Prompto and wear the E4 on their non-dominant wrist for 7 days (could be non-consecutive). During these days, Prompto issued prompts according to the users' detected cognitive load. They were informed that they could handle the prompts in the same three ways as they did in the Receptivity Study: "Start Practice", "Try Later" or "Ignore". The conversations and responses to prompts were recorded via the app. For each day of usage, participants were to rate the appropriateness of timings for each prompt and give written feedback to answer three guiding questions: "What worked/didn't work?", "What did you like?" and "What do you wish to see?". They did not use other memory training apps and did not wear the wristband while sleeping or bathing.

5.4 Findings

We report the findings of the deployment in two parts: The first part describes the results of the usage behaviour, responses to prompts and subjective ratings of appropriateness of prompt timings. The second part describes the key themes from our thematic analysis [7] of the written feedback. Feedback responses were coded by a researcher to generate themes, another researcher reviewed the themes and generated the final analysis.

5.4.1 Responses to Prompts and Subjective Appropriateness of Prompt Timings: A total of 28 prompts were issued, 19 of them (67.9%) were responded to: 13 with "Start Practice" (46.4%) and 6 with "Try Later" (21.4%), while 9 prompts were "Ignored" (32.1%). Thus, an average of 4 prompts (SD=2.4) were issued per user and an average of less than 1 prompt was issued per day. This might be due to the users not often being in low cognitive load state or due to the limitations of cognitive load detection system. Users took an average of 8.3 minutes to respond. They also gave an average appropriateness of timings (1 to 9 scale) rating of 6.2 for the prompts. We recorded 41 conversations in total. Assuming that 13 of the conversations were started due to the prompts, this meant that there were possibly 28 conversations that were user-initiated (users did use Prompto without being prompted). The total number of conversations (41) was slightly higher than the number of conversations (38) during the pilot (user-initiated with no prompts) but this meant there were fewer user-initiated conversations in the deployment (28) compared to the pilot. This might be due to users' excitement in trying something new in the pilot (a kind of novelty effect) but they used Prompto less after the first week. It could also be because they were aware that they would receive prompts and sometimes waited for a prompt to start a conversation. Although the response rate (67.9%) and ratings for appropriateness of prompt timings (6.2) during deployment were lower than the response rate (3.62 out of 4, 90.5%) and ratings (7.04) for the LCL condition in the Receptivity Study, the majority of prompts still received responses and there remained a higher chance that the response would be to start a practice session with Prompto.

5.4.2 Feedback: Having used Prompto first-hand and in-the-wild, three users commented that they liked the physiological-based prompting and felt it was helpful for them. P3 explained in her feedback that she liked it especially when she received prompts between times when she usually had breakfast and reading in the morning. Their comments were consistent with their speculative opinions from their interviews in the Receptivity Study.

A key challenge in prompting based physiological signals is addressing user acceptance of the sensing device. All of the users felt that wearing the E4 wristband and connecting it to the app was simple. However, similar to the findings from the Receptivity Study interviews, there remained mixed comments regarding the E4. P10 felt that it was quite unobtrusive, commenting, *"I forget that it's even there"*, whereas P13 thought that it was *"cumbersome"* and *"bulky"*.

Another challenge is ensuring that the users' lifestyles and habits are minimally affected when using the system. Prompto required constant Bluetooth communication between the E4 and the app to receive physiological data. However, because of this, majority of the users (4 users) remarked that the app was battery hungry. They were cautious of using Prompto, especially during busy days, to prevent from being negatively affected if their phone's battery ran out. P21 worked in an office, she was not in habit of taking her phone with her whenever she went to the restroom and would leave it on her desk. Thus, the E4 would disconnect when she left for the restroom and she found it a hassle to reconnect it upon her return.

Three users wished to see more functionality in Prompto then just going through the memory technique and wished to have more control over the prompting. P11 and P12 suggested that Prompto could also record and keep track of their tasks like a to-do list. P10 noted that *"...my [physiological] signals should not be the only thing to trigger [prompts]"*, and that Prompto should also know the current activity he was doing.

6 DISCUSSION

6.1 Receptivity

6.1.1 Effect on Prompt Responses: During LCL, participants were more responsive to prompts and engaged in more training sessions compared to during HCL. The lower percentage of responses to prompts in the deployment compared to the study was likely because a few prompts were not noticed and other unforeseen external factors (social, etc.) prevented users from responding in the in-the-wild setting. Participants were unlikely to start training sessions due to prompting during HCL (about 25% chance). This was consistent with a previous study that showed that stress reduced the probability of being available [82]. As seen in Figure 5b for HCL condition, the average count of responses and ignores were approximately the same value of 2. Thus, users were almost equally likely to respond to or ignore prompts during HCL (50% to 50% chance). Furthermore, for the HCL condition, users still rated app's content as useful (at 6 out of the scale of 9) and rated "appropriateness" of prompt timings was low (high "disruptiveness"). These results seem to agree with the previous finding that 54% of disruptive notifications could be answered if the content was useful [65].

6.1.2 Effect on Response Time. The marginally faster (but not significant) response times under HCL compared to LCL was expected as we note from the study by Mehrotra et al. [65] that an increase in complexity of an ongoing task may have users respond faster to prompts and be more attentive to them. We postulate that with more participants or more prompts and repetition, we might see a significant difference in response times. As our study was held in a lab setting, participants might have felt a more immediate need to answer prompts compared to an in-the-wild setting. The response times during our in-the-wild deployment (8.3 minutes) were understandably much longer than response times in our study (10 seconds). Nonetheless, they were more comparable to timings shown by Mehrotra et al. (3 to 7 minutes).

6.1.3 Effect on Subjective Usefulness. Perceived usefulness of Prompto was heavily tied with the usefulness of the memory technique and should have been similar across conditions, yet participants rated Prompto as more useful during LCL than during HCL. This might be because they initiated more training sessions during LCL, and thus had more frequent exposure to the technique, allowing them to understand it better and realise its usefulness when applying it to their own tasks.

6.1.4 Availability. Availability and rated appropriateness of prompt timings are subjective and highly context-based. Users felt more available and rated prompts in LCL to be more appropriately timed. These ratings were lower in the deployment. We agree with what P10 mentioned in the deployment that cognitive and physiological states should not be the only context used. As gathered from our interviews, we also need to take into account users' preferences on when are suitable timings for prompting. We note that although multiple contexts have to be used for a more robust and accurate system [82], our work first focused on examining the physiological context on its own, which could then be added to relevant multi-context models of receptivity [19].

6.2 Implications

6.2.1 Prompting Based on Cognitive Context. Prior work suggest that idle times, which we equate as times of LCL, are the most appropriate moments for interruption and users might be willing to accept more prompts [1, 65]. We confirm this quantitatively and qualitatively, and for the first time with prompting from conversational agents. Other pedagogical agents and prompting systems beyond the memory training context could benefit from adopting prompting mechanisms which trigger at these times. We found that prompting under HCL would still have a 50% chance of response from the user. Research work on stress interventions and virtual therapists, which often require prompting during HCL, could still benefit from such a mechanism.

6.2.2 Combining Context Inputs. Prompting using physiological context is challenging and would require additional context inputs to be realistically implemented. Accounting for user contexts of time and location through a scheduling system that looks at the user's calendar [60] and Global Positioning System (GPS) sensor on the phone might work. However, underlying privacy issues have to first be addressed. Having a do-not-disturb mode could be useful for accommodating and respecting users' preferences of prompt timings. A drawback might be that users would still need to remember to toggle to that mode. Time-based prompts still remain the most common and simplest to implement on mobile settings. Thus, future systems could look at issuing "physio-modulated time-based prompts", in which users would set a timed prompt (for example, every day at 9 a.m.) and the system would only trigger the prompt if it finds a low cognitive load event within a time window of the set timing (if ±15 minute window, prompt might be between 8:45 a.m. to 9:15 a.m.).

6.2.3 Physical Form and Placement of Physiological Sensors. From our findings, wearing wrist-worn E4 was easy to use but not ideal for older adults as it involved user burden [82] of wearing and charging, and was not unobtrusive enough [5]. We did observe participants having normal watches, Apple Watches and Fitbits. We believe that it would be possible to have easy integration with what they currently have, although most users agreed that a smartphone implementation was good enough. This would negate the need to wear additional and expensive devices like the E4. Furthermore, while they stated that the placement on wrist was not ideal, it was hard to pinpoint a form and placement that was acceptable by most older adults. Having an all-in-one sensor device in watches, necklaces and clip-on attachments to clothes, as suggested by participants, might be the most viable and practical remedy. Developing a ear-worn device similar to hearing aids might make it more discreet.

6.3 Limitations

6.3.1 Generalisability and Participant Selection. Since participants were older adults, it is difficult to generalise some of our findings for a broader population. Nonetheless, we unveiled some findings and implications which other researchers on receptivity, conversational agents and physiological sensing could utilise in their studies with various age groups. There is also a possibility that the positive receptivity seen with Prompto could be a result of a participant selection bias as we recruited participants who had initial interest in memory-related interventions. Collecting further information from participants who decided not to join the study on their reasons in doing so would be informative for developing next iterations of the system.

6.3.2 Sensor Placement. Quality EDA and PPG readings are strongly reliant on the locations of sensor placement on the body and need robust skin contact to be attained. Finger and palmar surface measurements for EDA are known to be optimal and recommended [92]. Wrist placement of the EDA sensor in the E4 might be sub-optimal compared hand placements but it is a more practical and comfortable option for in-the-wild settings where users are likely to use and move their hands. Designers of the E4 at Empatica Inc. have noted this as well and have designed the EDA sensor to be suited and more sensitive to readings at the wrist [50]. Alternative sensor placements like on necklaces and clothes, as suggested in Section 6.2.3, might result in intermittent skin contact and measurement, thus require further design and experimentation if they were to be implemented.

6.3.3 Motion Artefacts. Despite users being shown how to wear the E4 device and being instructed to place it on their non-dominant wrist to minimise motion artefacts, there was bound to be unavoidable noise in EDA and PPG measurements due to user's motion and it is also unknown how well users placed the sensors. This might prevent the measurements from averaging within the threshold range and explain the lower-than-expected prompting rate during the in-the-wild deployment. We relied on the E4's own filtering algorithm to discard noisy, movement-related IBI data [49] but further reliably testing is needed. Future iterations of the Prompto app could include a noise-filtering algorithm for the physiological measurements to address the limitation.

6.3.4 Effects of Age on Autonomic Responses. Age-related decline in autonomic responses of EDA [3] and time-domain HRV [83] as people advance in age are known to result in lowered skin conductance responses, EDA levels and RMSSD values. We were aware that this may affect that the cognitive load detection performance of Prompto. Thus, we attempted to overcome this by customising the thresholds. A machine-learning based model could be used for detection in the future.

6.4 Future Work

6.4.1 Cognitive Load Detection. We chose to use the E4 as it had the sensors needed. Other physiological sensing devices could be considered. Detection using eye-based features [30], conversational cues [25] and respiration [39] are three promising methods for estimating cognitive load which could be investigated in future versions.

6.4.2 Speech Recognition. Android (Google) and Apple speech recognition were not satisfactory in performance as users' speech were often wrongly recognised. We agree that supporting both voice and text-based input in Prompto was advantageous compared to voice-only input [69]. For future iterations, we intend to use other speech-to-text engines known for high recognition accuracy, such as Baidu Deep Speech [80].

6.4.3 Content. Due to nature of "when-then" technique, there was no sense of progression and rounds of conversation felt boring after a while, especially since the technique is quickly learnt [15]. As a step towards providing an alternative selection of memory methods to introduce variety and progression, we have added a mode on Prompto to facilitate a different memory technique, the spaced retrieval method, in a form of a dialogue-based quiz that is similar to QuizBot [79]. Initial reactions from older adults indicated that such a mode may not be useful for them as they felt that it was only for trivia and entertainment. Students might be a better user group for it.

7 CONCLUSION

Prospective memory training through conversational agents could benefit older adults but still requires regular practice sessions. We designed Prompto to deliver such training while using physiological readings to infer users' cognitive contexts, prompting them to engage in sessions during low cognitive load and when they are cognitively available. Our investigation on the effects of cognitive load on older adults' receptivity to prompts showed that users were more responsive and engaged in more sessions under low cognitive load than under high cognitive load. From our interviews and in-the-wild deployment, Prompto was found to be useful and helpful, and users were likely to respond to its prompts in both in the lab and in-the-wild. Our discussions implied that prompting at low and high cognitive load could be beneficial depending on the application, a combination of physiological context with other context inputs (e.g., time) as well as a close consideration for form and placement might be needed for future developments. We believe our work would encourage further research on technologies and conversational agents with prompting and notification systems which take into account users' cognitive contexts.

ACKNOWLEDGMENTS

We would like to thank Assoc. Prof. Lynette Tippett for her advice, Vipula Dissanayake for his assistance in development and Yvonne Chua for her assistance in designing the illustrations. This work was supported by *Assistive Augmentation* research grant under the Entrepreneurial Universities (EU) initiative of New Zealand.

REFERENCES

[1] Christoph Anderson, Isabel Hübener, Ann-Kathrin Seipp, Sandra Ohly, Klaus David, and Veljko Pejovic. 2018. A Survey of Attention Management Systems in Ubiquitous Computing Environments. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 2, Article 58 (July 2018), 27 pages. https://doi.org/10.1145/3214261

[2] Seyed Ali Bahrainian and Fabio Crestani. 2017. Towards the Next Generation of Personal Assistants: Systems That Know When You Forget. In *Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval* (Amsterdam, The Netherlands) *(ICTIR '17)*. Association for Computing Machinery, New York, NY, USA, 169–176. https://doi.org/10.1145/3121050.3121071

[3] Marta Barontini, Julio O. Lázzari, Gloria Levin, Inés Armando, and Susana J. Basso. 1997. Age-related changes in sympathetic activity: biochemical measurements and target organ responses. *Archives of Gerontology and Geriatrics* 25, 2 (1997), 175–186. https://doi.org/10.1016/S0167-4943(97)00008-3

[4] Mathias Benedek and Christian Kaernbach. 2010. A continuous measure of phasic electrodermal activity. *Journal of Neuroscience Methods* 190, 1 (2010), 80–91. https://doi.org/10.1016/j.jneumeth.2010.04.028

[5] J. H. M. Bergmann and A. H. McGregor. 2011. Body-worn sensor design: what do patients and clinicians want? *Annals of Biomedical Engineering* 39, 9 (2011), 2299–2312. https://doi.org/10.1007/s10439-011-0339-9

[6] Timothy W. Bickmore, Lisa Caruso, Kerri Clough-Gorr, and Tim Heeren. 2005. "It's just like you talk to a friend" relational agents for older adults. *Interacting with Computers* 17, 6 (2005), 711–735. https://doi.org/10.1016/j.intcom.2005.09.002

[7] Virginia Braun, Victoria Clarke, Nikki Hayfield, and Gareth Terry. 2018. *Thematic Analysis*. Springer Singapore, Singapore, 1–18. https://doi.org/10.1007/978-981-10-2779-6_103-1

[8] R. N. Brewer, M. R. Morris, and S. E. Lindley. 2017. How to Remember What to Remember: Exploring Possibilities for Digital Reminder Systems. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 3, Article 38 (Sept. 2017), 20 pages. https://doi.org/10.1145/3130903

[9] Sarah Susanne Brom and Matthias Kliegel. 2014. Improving everyday prospective memory performance in older adults: Comparing cognitive process and strategy training. *Psychology and Aging* 29, 3 (2014), 744–755. https://doi.org/10.1037/a0037181

[10] Christina Burkard, Lucien Rochat, Anaëlle Blum, Joëlle Emmenegger, Anne-Claude Juillerat Van der Linden, and Martial Van der Linden. 2014. A daily-life-oriented intervention to improve prospective memory and goal-directed behaviour in ageing: A pilot study. *Neuropsychological Rehabilitation* 24, 2 (2014), 266–295. https://doi.org/10.1080/09602011.2014.887023

[11] A. J. M. M. Camm, Marek Malik, J. T. G. B. Bigger, Günter Breithardt, Sergio Cerutti, R. Cohen, Philippe Coumel, E. Fallen, H. Kennedy, R. E. Kleiger, et al. 1996. Heart rate variability. Task Force of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology. *Circulation* 93, 5 (1996), 1043–1065. https://doi.org/10.1161/01.CIR.93.5.1043

[12] Cameron Camp, Michael Bird, Catherine Cherry, et al. 2000. Retrieval strategies as a rehabilitation aid for cognitive loss in pathological ageing. In *Cognitive rehabilitation in old age*. Oxford University Press. http://hdl.handle.net/1885/89084

[13] Joana Campos and Ana Paiva. 2010. May: My memories are yours. In *International Conference on Intelligent Virtual Agents*. Springer, 406–412. https://doi.org/10.1007/978-3-642-15892-6_44

[14] Samantha W. T. Chan. 2020. Biosignal-Sensitive Memory Improvement and Support Systems. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI EA '20)*. Association for Computing Machinery, New York, NY, USA, 1–7. https://doi.org/10.1145/3334480.3375031

[15] Samantha W. T. Chan, Thisum Buddhika, Haimo Zhang, and Suranga Nanayakkara. 2019. ProspecFit: In Situ Evaluation of Digital Prospective Memory Training for Older Adults. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 3, 3, Article 77 (Sept. 2019), 20 pages. https://doi.org/10.1145/3351235

[16] Samantha W. T. Chan, Haimo Zhang, and Suranga Nanayakkara. 2019. Prospero: A Personal Wearable Memory Coach. In *Proceedings of the 10th Augmented Human International Conference 2019* (Reims, France) *(AH2019)*. Association for Computing Machinery, New York, NY, USA, Article 26, 5 pages. https://doi.org/10.1145/3311823.3311870

[17] Fang Chen, Jianlong Zhou, Yang Wang, Kun Yu, Syed Z. Arshad, Ahmad Khawaji, and Dan Conway. 2016. *Stress and Cognitive Load*. Springer International Publishing, Cham, 185–194. https://doi.org/10.1007/978-3-319-31700-7_12

[18] Xing-jie Chen, Ya Wang, Lu-lu Liu, Ji-fang Cui, Ming-yuan Gan, David HK Shum, and Raymond C. K. Chan. 2015. The effect of implementation intention on prospective memory: a systematic and meta-analytic review. *Psychiatry research* 226, 1 (2015), 14–22. https://doi.org/10.1016/j.psychres.2015.01.011

[19] Woohyeok Choi, Sangkeun Park, Duyeon Kim, Youn-kyung Lim, and Uichin Lee. 2019. Multi-Stage Receptivity Model for Mobile Just-In-Time Health Intervention. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 3, 2, Article 39 (June 2019), 26 pages.

https://doi.org/10.1145/3328910
[20] Minsoo Choy, Daehoon Kim, Jae-Gil Lee, Heeyoung Kim, and Hiroshi Motoda. 2016. Looking Back on the Current Day: Interruptibility Prediction Using Daily Behavioral Features. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (Heidelberg, Germany) *(UbiComp '16)*. Association for Computing Machinery, New York, NY, USA, 1004–1015. https://doi.org/10.1145/2971648.2971649
[21] Sarah Clinch and Cecilia Mascolo. 2018. Learning from Our Mistakes: Identifying Opportunities for Technology Intervention against Everyday Cognitive Failure. *IEEE Pervasive Computing* 17, 2 (2018), 22–33. https://doi.org/10.1109/MPRV.2018.022511240
[22] Posit Science Corporation. 2002. BrainHQ. Retrieved 2019-07-03 from https://www.brainhq.com/
[23] Tilman Dingler, Dominik Weber, Martin Pielot, Jennifer Cooper, Chung-Cheng Chang, and Niels Henze. 2017. Language Learning On-the-Go: Opportune Moments and Design of Mobile Microlearning Sessions. In *Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and Services* (Vienna, Austria) *(MobileHCI '17)*. Association for Computing Machinery, New York, NY, USA, Article 28, 12 pages. https://doi.org/10.1145/3098279.3098565
[24] Sidney D'Mello. 2014. Emotional rollercoasters: day differences in affect incidence during learning. In *The Twenty-Seventh International Flairs Conference*. https://www.aaai.org/ocs/index.php/FLAIRS/FLAIRS14/paper/view/7777
[25] Sidney D'Mello and Art Graesser. 2013. AutoTutor and Affective Autotutor: Learning by Talking with Cognitively and Emotionally Intelligent Computers That Talk Back. *ACM Trans. Interact. Intell. Syst.* 2, 4, Article 23 (Jan. 2013), 39 pages. https://doi.org/10.1145/2395123.2395128
[26] Damien Elmes. 2015. Anki. Retrieved 2019-07-24 from http://ankisrs.net
[27] Michael W. Eysenck. 2012. *Fundamentals of Cognition*. Psychology Press. https://books.google.com/books?id=CVeacQAACAAJ
[28] Joel E. Fischer, Nick Yee, Victoria Bellotti, Nathan Good, Steve Benford, and Chris Greenhalgh. 2010. Effects of Content and Time of Delivery on Receptivity to Mobile Interruptions. In *Proceedings of the 12th International Conference on Human Computer Interaction with Mobile Devices and Services* (Lisbon, Portugal) *(MobileHCI '10)*. Association for Computing Machinery, New York, NY, USA, 103–112. https://doi.org/10.1145/1851600.1851620
[29] Pascal E. Fortin, Elisabeth Sulmont, and Jeremy Cooperstock. 2019. Detecting Perception of Smartphone Notifications Using Skin Conductance Responses. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–9. https://doi.org/10.1145/3290605.3300420
[30] Lex Fridman, Bryan Reimer, Bruce Mehler, and William T. Freeman. 2018. Cognitive Load Estimation in the Wild. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–9. https://doi.org/10.1145/3173574.3174226
[31] Nicola J. Gates, Perminder S. Sachdev, Maria A. Fiatarone Singh, and Michael Valenzuela. 2011. Cognitive and memory training in adults at risk of dementia: a systematic review. *BMC geriatrics* 11, 1 (2011), 55. https://doi.org/10.1186/1471-2318-11-55
[32] Tony Gillie and Donald Broadbent. 1989. What makes interruptions disruptive? A study of length, similarity, and complexity. *Psychological research* 50, 4 (1989), 243–250. https://doi.org/10.1007/BF00309260
[33] Peter M. Gollwitzer. 1993. Goal achievement: The role of intentions. *European review of social psychology* 4, 1 (1993), 141–185. https://doi.org/10.1080/14792779343000059
[34] Peter M. Gollwitzer. 1999. Implementation intentions: strong effects of simple plans. *American psychologist* 54, 7 (1999), 493. https://doi.org/10.1037/0003-066X.54.7.493
[35] Peter M. Gollwitzer and Paschal Sheeran. 2006. Implementation intentions and goal achievement: A meta-analysis of effects and processes. *Advances in experimental social psychology* 38 (2006), 69–119. https://doi.org/10.1016/S0065-2601(06)38002-1
[36] Eija Haapalainen, SeungJun Kim, Jodi F. Forlizzi, and Anind K. Dey. 2010. Psycho-Physiological Measures for Assessing Cognitive Load. In *Proceedings of the 12th ACM International Conference on Ubiquitous Computing* (Copenhagen, Denmark) *(UbiComp '10)*. Association for Computing Machinery, New York, NY, USA, 301–310. https://doi.org/10.1145/1864349.1864395
[37] Joshua Harman. 2001. Creating a Memory Palace Using a Computer. In *CHI '01 Extended Abstracts on Human Factors in Computing Systems* (Seattle, Washington) *(CHI EA '01)*. Association for Computing Machinery, New York, NY, USA, 407–408. https://doi.org/10.1145/634067.634306
[38] Sandra G. Hart. 2006. NASA-task load index (NASA-TLX); 20 years later. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 50. Sage publications Sage CA: Los Angeles, CA, 904–908. https://doi.org/10.1177/154193120605000909
[39] Jennifer A. Healey and Rosalind W. Picard. 2005. Detecting stress during real-world driving tasks using physiological sensors. *IEEE Transactions on Intelligent Transportation Systems* 6, 2 (2005), 156–166. https://doi.org/10.1109/TITS.2005.848368
[40] Alexandra Hering, Matthias Kliegel, Peter G. Rendell, Fergus I. M. Craik, and Nathan S. Rose. 2018. Prospective memory is a key predictor of functional independence in older adults. *Journal of the International Neuropsychological Society* 24, 6 (2018), 640–645. https://doi.org/10.1017/S1355617718000152
[41] Alexandra Hering, Peter G. Rendell, Nathan S. Rose, Katharina M. Schnitzspahn, and Matthias Kliegel. 2014. Prospective memory training in older adults and its relevance for successful aging. *Psychological Research* 78, 6 (2014), 892–904. https://doi.org/10.1007/s00426-014-0566-4

[42] Linda Hirsch, Anton Björsell, Mikael Laaksoharju, and Mohammad Obaid. 2017. Investigating Design Implications Towards a Social Robot as a Memory Trainer. In *Proceedings of the 5th International Conference on Human Agent Interaction* (Bielefeld, Germany) *(HAI '17)*. Association for Computing Machinery, New York, NY, USA, 5–10. https://doi.org/10.1145/3125739.3125755

[43] Nis Hjortskov, Dag Rissén, Anne Katrine Blangsted, Nils Fallentin, Ulf Lundberg, and Karen Søgaard. 2004. The effect of mental stress on heart rate variability and blood pressure during computer work. *European Journal of Applied Physiology* 92, 1-2 (2004), 84–89. https://doi.org/10.1007/s00421-004-1055-z

[44] Joyce Ho and Stephen S. Intille. 2005. Using Context-Aware Computing to Reduce the Perceived Burden of Interruptions from Mobile Devices. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Portland, Oregon, USA) *(CHI '05)*. Association for Computing Machinery, New York, NY, USA, 909–918. https://doi.org/10.1145/1054972.1055100

[45] Paul Hoskinson. 2008. Brain Workshop. Retrieved 2019-08-24 from http://brainworkshop.sourceforge.net/

[46] Apple Inc. 2011. Siri. Retrieved 2019-07-03 from https://www.apple.com/siri/

[47] Amazon.com Inc. 2014. Alexa. Retrieved 2019-08-29 from https://www.amazon.com/b?node=17934671011

[48] Empatica Inc. 2013. E4 Wristband. Retrieved 2019-08-09 from https://www.empatica.com/research/e4/

[49] Empatica Inc. 2020. E4 data - IBI expected signal. Retrieved 2020-08-12 from https://support.empatica.com/hc/en-us/articles/360030058011-E4-data-IBI-expected-signal

[50] Empatica Inc. 2020. What should I know to use EDA data in my experiment? Retrieved 2020-08-12 from https://support.empatica.com/hc/en-us/articles/203621955

[51] Fitbit Inc. 2007. Fitbit. Retrieved 2019-08-29 from https://www.fitbit.com

[52] Shamsi T. Iqbal and Eric Horvitz. 2010. Notifications and Awareness: A Field Study of Alert Usage and Preferences. In *Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work* (Savannah, Georgia, USA) *(CSCW '10)*. Association for Computing Machinery, New York, NY, USA, 27–30. https://doi.org/10.1145/1718918.1718926

[53] John G. Kreifeldt and Mary E. McCarthy. 1981. Interruption as a test of the user-computer interface. (1981). https://ntrs.nasa.gov/citations/19820005848

[54] Kubios. 2020. About HRV - Kubios HRV. Retrieved 2020-08-12 from https://www.kubios.com/about-hrv/

[55] Lumos Labs. 2007. Lumosity. Retrieved 2019-07-03 from https://www.lumosity.com/

[56] Peng Liao, Walter Dempsey, Hillol Sarker, Syed Monowar Hossain, Mustafa al'Absi, Predrag Klasnja, and Susan Murphy. 2018. Just-in-Time but Not Too Much: Determining Treatment Timing in Mobile Health. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 4, Article 179 (Dec. 2018), 21 pages. https://doi.org/10.1145/3287057

[57] Han Lin, Jinghua Hou, Han Yu, Zhiqi Shen, and Chunyan Miao. 2015. An Agent-Based Game Platform for Exercising People's Prospective Memory. In *2015 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*, Vol. 3. 235–236. https://doi.org/10.1109/WI-IAT.2015.42

[58] Google LLC. 2010. Dialogflow. Retrieved 2019-08-09 from https://dialogflow.com/

[59] Google LLC. 2016. Google Assistant. Retrieved 2019-07-03 from https://assistant.google.com/

[60] Tom Lovett, Eamonn O'Neill, James Irwin, and David Pollington. 2010. The Calendar as a Sensor: Analysis and Improvement Using Data Fusion with Social Networks and Location. In *Proceedings of the 12th ACM International Conference on Ubiquitous Computing* (Copenhagen, Denmark) *(UbiComp '10)*. Association for Computing Machinery, New York, NY, USA, 3–12. https://doi.org/10.1145/1864349.1864352

[61] CogniFit Ltd. 1999. CogniFit. Retrieved 2019-07-04 from https://www.cognifit.com/

[62] Akhil Mathur, Nicholas D. Lane, and Fahim Kawsar. 2016. Engagement-Aware Computing: Modelling User Engagement from Mobile Contexts. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (Heidelberg, Germany) *(UbiComp '16)*. Association for Computing Machinery, New York, NY, USA, 622–633. https://doi.org/10.1145/2971648.2971760

[63] Mark A. McDaniel, Daniel C. Howard, and Karin M. Butler. 2008. Implementation intentions facilitate prospective memory under high attention demands. *Memory & Cognition* 36, 4 (2008), 716–724. https://doi.org/10.3758/MC.36.4.716

[64] Abhinav Mehrotra and Mirco Musolesi. 2017. Intelligent Notification Systems: A Survey of the State of the Art and Research Challenges. *CoRR* abs/1711.10171 (2017). arXiv:1711.10171 http://arxiv.org/abs/1711.10171

[65] Abhinav Mehrotra, Veljko Pejovic, Jo Vermeulen, Robert Hendley, and Mirco Musolesi. 2016. My Phone and Me: Understanding People's Receptivity to Mobile Notifications. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 1021–1032. https://doi.org/10.1145/2858036.2858566

[66] LoungeV Films Relaxing Music and Nature Sounds. 2008. Relaxing Music with Amazing Nature Scenery HD Video 1080p - 6 Hours. Retrieved 2019-08-24 from https://www.youtube.com/watch?v=TdpBRZ0dZhw

[67] Nargess Nourbakhsh, Yang Wang, and Fang Chen. 2013. GSR and Blink Features for Cognitive Load Classification. In *Human-Computer Interaction – INTERACT 2013*, Paula Kotzé, Gary Marsden, Gitte Lindgaard, Janet Wesson, and Marco Winckler (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 159–166. https://doi.org/10.1007/978-3-642-40483-2_11

[68] Seo Jin Oh, Sungmin Seo, Ji Hyun Lee, Myeong Ju Song, and Min-Sup Shin. 2018. Effects of smartphone-based memory training for older adults with subjective memory complaints: a randomized controlled trial. *Aging & Mental Health* 22, 4 (2018), 526–534. https://doi.org/10.1080/13607863.2016.1274373 PMID: 28071929.

[69] Cathy Pearl. 2016. *Designing Voice User Interfaces: Principles of Conversational Experiences*. "O'Reilly Media, Inc.". https://dl.acm.org/citation.cfm?id=3086829

[70] Simon T. Perrault, Eric Lecolinet, Yoann Pascal Bourse, Shengdong Zhao, and Yves Guiard. 2015. Physical Loci: Leveraging Spatial, Object and Semantic Memory for Command Selection. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. ACM, New York, NY, USA, 299–308. https://doi.org/10.1145/2702123.2702126

[71] Bastian Pfleging, Drea K. Fekety, Albrecht Schmidt, and Andrew L. Kun. 2016. A Model Relating Pupil Diameter to Mental Workload and Lighting Conditions. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 5776–5788. https://doi.org/10.1145/2858036.2858117

[72] Martin Pielot, Bruno Cardoso, Kleomenis Katevas, Joan Serrà, Aleksandar Matic, and Nuria Oliver. 2017. Beyond Interruptibility: Predicting Opportune Moments to Engage Mobile Phone Users. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 3, Article 91 (Sept. 2017), 25 pages. https://doi.org/10.1145/3130956

[73] Martin Pielot, Rodrigo de Oliveira, Haewoon Kwak, and Nuria Oliver. 2014. Didn't You See My Message? Predicting Attentiveness to Mobile Instant Messages. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Toronto, Ontario, Canada) *(CHI '14)*. Association for Computing Machinery, New York, NY, USA, 3319–3328. https://doi.org/10.1145/2556288.2556973

[74] David M. W. Powers, Martin H. Luerssen, Trent W. Lewis, Richard E. Leibbrandt, Marissa Milne, John Pashalis, and Kenneth Treharne. 2010. MANA for the Ageing. In *Proceedings of the 2010 Workshop on Companionable Dialogue Systems*. Association for Computational Linguistics, 7–12. https://www.aclweb.org/anthology/W10-2702.pdf

[75] Laura Ramos, Elise van den Hoven, and Laurie Miller. 2016. Designing for the Other 'Hereafter': When Older Adults Remember about Forgetting. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 721–732. https://doi.org/10.1145/2858036.2858162

[76] Lazlo Ring, Lin Shi, Kathleen Totzke, and Timothy Bickmore. 2015. Social support agents for older adults: longitudinal affective computing in the home. *Journal on Multimodal User Interfaces* 9, 1 (2015), 79–88. https://doi.org/10.1007/s12193-014-0157-0

[77] Nathan S. Rose, Peter G. Rendell, Alexandra Hering, Matthias Kliegel, Gavin M. Bidelman, and Fergus I. M. Craik. 2015. Cognitive and neural plasticity in older adults'prospective memory following training with the Virtual Week computer game. *Frontiers in Human Neuroscience* 9 (2015). https://doi.org/10.3389/fnhum.2015.00592

[78] Oscar Rosello, Marc Exposito, and Pattie Maes. 2016. NeverMind: Using Augmented Reality for Memorization. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology (UIST '16 Adjunct)*. ACM, New York, NY, USA, 215–216. https://doi.org/10.1145/2984751.2984776

[79] Sherry Ruan, Liwei Jiang, Justin Xu, Bryce Joe-Kun Tham, Zhengneng Qiu, Yeshuang Zhu, Elizabeth L. Murnane, Emma Brunskill, and James A. Landay. 2019. QuizBot: A Dialogue-Based Adaptive Learning System for Factual Knowledge. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3290605.3300587

[80] Sherry Ruan, Jacob O. Wobbrock, Kenny Liou, Andrew Ng, and James Landay. 2016. Speech is 3x faster than typing for english and mandarin text entry on mobile devices. *arXiv preprint arXiv:1608.07323* (2016). https://hci.stanford.edu/research/speech/paper/speech_paper.pdf

[81] Hugues Sansen, Gérard Chollet, Cornelius Glackin, Kristiina Jokinen, and Badii A. Torres. 2016. The Roberta IRONSIDE Project A Cognitive and Physical Robot Coach for Dependent Persons. *Handicap 2016, Paris* (2016). http://www.iieta.org/sites/default/files/Journals/MMC/MMC_C/2016.77.2_15.pdf

[82] Hillol Sarker, Moushumi Sharmin, Amin Ahsan Ali, Md. Mahbubur Rahman, Rummana Bari, Syed Monowar Hossain, and Santosh Kumar. 2014. Assessing the Availability of Users to Engage in Just-in-Time Intervention in the Natural Environment. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (Seattle, Washington) *(UbiComp '14)*. Association for Computing Machinery, New York, NY, USA, 909–920. https://doi.org/10.1145/2632048.2636082

[83] Fred Shaffer and J. P. Ginsberg. 2017. An overview of heart rate variability metrics and norms. *Frontiers in Public Health* 5 (2017), 258. https://www.frontiersin.org/article/10.3389/fpubh.2017.00258

[84] Tejal M. Shah, Michael Weinborn, Giuseppe Verdile, Hamid R. Sohrabi, and Ralph N. Martins. 2017. Enhancing Cognitive Functioning in Healthly Older Adults: a Systematic Review of the Clinical Significance of Commercially Available Computerized Cognitive Training in Preventing Cognitive Decline. *Neuropsychology Review* 27, 1 (2017), 62–80. https://doi.org/10.1007/s11065-016-9338-9

[85] Evelyn Shatil, Jaroslava Mikulecka, Francesco Bellotti, and Vladimír Bureš. 2014. Novel television-based cognitive training improves working memory and executive function. *PLoS One* 9, 7 (2014), e101472. https://doi.org/10.1371/journal.pone.0101472

[86] Yu Shi, Natalie Ruiz, Ronnie Taib, Eric Choi, and Fang Chen. 2007. Galvanic Skin Response (GSR) as an Index of Cognitive Load. In *CHI '07 Extended Abstracts on Human Factors in Computing Systems* (San Jose, CA, USA) *(CHI EA '07)*. Association for Computing Machinery, New York, NY, USA, 2651–2656. https://doi.org/10.1145/1240866.1241057

[87] Glenn E. Smith, Patricia Housen, Kristine Yaffe, Ronald Ruff, Robert F. Kennison, Henry W. Mahncke, and Elizabeth M. Zelinski. 2009. A Cognitive Training Program Based on Principles of Brain Plasticity: Results from the Improvement in Memory with Plasticity-based Adaptive Cognitive Training (IMPACT) Study. *Journal of the American Geriatrics Society* 57, 4 (2009), 594–603. https://doi.org/10.1111/j.1532-5415.2008.02167.x

[88] Mika P. Tarvainen, Juha-Pekka Niskanen, Jukka A. Lipponen, Perttu O. Ranta-Aho, and Pasi A. Karjalainen. 2014. Kubios HRV–heart rate variability analysis software. *Computer methods and programs in biomedicine* 113, 1 (2014), 210–220. https://doi.org/10.1016/j.cmpb.2013.07.024

[89] Julian F. Thayer, Anita L. Hansen, Evelyn Saus-Rose, and Bjorn H. Johnsen. 2009. Heart rate variability, prefrontal neural function, and cognitive performance. *Annals of Behavioral Medicine* 37, 2 (2009), 141–153. https://doi.org/10.1007/s12160-009-9101-z

[90] Myrthe Tielman, Marieke van Meggelen, Mark A. Neerincx, and Willem-Paul Brinkman. 2015. An Ontology-Based Question System for a Virtual Coach Assisting in Trauma Recollection. In *International Conference on Intelligent Virtual Agents*. Springer Cham, 17–27. https://doi.org/10.1007/978-3-319-21996-7_2

[91] Liam D. Turner, Stuart M. Allen, and Roger M. Whitaker. 2017. Reachable but not receptive: Enhancing smartphone interruptibility prediction by modelling the extent of user engagement with notifications. *Pervasive and Mobile Computing* 40 (2017), 480–494. https://doi.org/10.1016/j.pmcj.2017.01.011

[92] Marieke van Dooren, J. J. G. (Gert-Jan) de Vries, and Joris H. Janssen. 2012. Emotional sweating across the body: Comparing 16 different skin conductance measurement locations. *Physiology & Behavior* 106, 2 (2012), 298–304. https://doi.org/10.1016/j.physbeh.2012.01.020

[93] Laura Pfeifer Vardoulakis, Lazlo Ring, Barbara Barry, Candace L. Sidner, and Timothy Bickmore. 2012. Designing Relational Agents as Long Term Social Companions for Older Adults. In *International Conference on Intelligent Virtual Agents*. Springer, 289–302. https://doi.org/10.1007/978-3-642-33197-8_30

[94] Aku Visuri, Niels van Berkel, Chu Luo, Jorge Goncalves, Denzil Ferreira, and Vassilis Kostakos. 2017. Predicting Interruptibility for Manual Data Collection: A Cluster-Based User Model. In *Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and Services* (Vienna, Austria) *(MobileHCI '17)*. Association for Computing Machinery, New York, NY, USA, Article 12, 14 pages. https://doi.org/10.1145/3098279.3098532

[95] Piotr A. Wozniak and Edward J. Gorzelanczyk. 1994. Optimization of repetition spacing in the practice of learning. *Acta Neurobiologiae Experimentalis* 54 (1994), 59–59. https://ane.pl/pdf/5409.pdf

[96] Ramin Yaghoubzadeh, Marcel Kramer, Karola Pitsch, and Stefan Kopp. 2013. Virtual Agents as Daily Assistants for Elderly or Cognitively Impaired People. In *International workshop on intelligent virtual agents*. Springer, 79–91. https://doi.org/10.1007/978-3-642-40415-3_7

[97] Yuki Yamada, Keisuke Irie, Kota Gushima, Fumiko Ishizawa, Mohammed Al Sada, and Tatsuo Nakajima. 2017. HoloMoL: Human Memory Augmentation with Mixed-reality Technologies. In *Proceedings of the 21st International Academic Mindtrek Conference (AcademicMindtrek '17)*. ACM, New York, NY, USA, 235–238. https://doi.org/10.1145/3131085.3131097

[98] Xi Yang, Marco Aurisicchio, and Weston Baxter. 2019. Understanding Affective Experiences with Conversational Agents. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300772

[99] Akihito Yoshii, Helena Malmivirta, Mika Luimula, Paula Pitkäkangas, and Tatsuo Nakajima. 2015. Designing a Map-Based Application and a Conversational Agent for Addressing Memory Problems. In *International Conference on Human-Computer Interaction*. Springer, 340–345. https://doi.org/10.1007/978-3-319-21380-4_58

[100] Fengpeng Yuan, Xianyi Gao, and Janne Lindqvist. 2017. How Busy Are You? Predicting the Interruptibility Intensity of Mobile Users. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 5346–5360. https://doi.org/10.1145/3025453.3025946

[101] Fred R. H. Zijlstra, Robert A. Roe, Anna B. Leonora, and Irene Krediet. 1999. Temporal factors in mental work: Effects of interrupted activities. *Journal of Occupational and Organizational Psychology* 72, 2 (1999), 163–185. https://doi.org/10.1348/096317999166581

[102] Thomas D. Zimmermann and Beat Meier. 2010. The effect of implementation intentions on prospective memory performance across the lifespan. *Applied Cognitive Psychology* 24, 5 (2010), 645–658. https://doi.org/10.1002/acp.1576