HHS Public Access

Author manuscript

Proc ACM Interact Mob Wearable Ubiquitous Technol. Author manuscript; available in PMC 2021 June 22.

Published in final edited form as:

Proc ACM Interact Mob Wearable Ubiquitous Technol. 2019 December ; 3(4): 1–27. doi:10.1145/3369806.

ReVibe: A Context-assisted Evening Recall Approach to Improve Self-report Adherence

MASHFIQUI RABBI[*],
Harvard University, 1 Oxford Street, Cambridge, MA, 02134, USA

KATHERINE LI,
University of Michigan, Ann Arbor, MI, USA

H. YANNA YAN,
University of Michigan, Ann Arbor, MI, USA

KELLY HALL,
Yale University, Ann Arbor, MI, USA

PREDRAG KLASNJA,
University of Michigan, Ann Arbor, MI, USA

SUSAN MURPHY
Harvard University, Cambridge, MA, USA

Abstract

Besides passive sensing, ecological momentary assessments (EMAs) are one of the primary methods to collect in-the-moment data in ubiquitous computing and mobile health. While EMAs have the advantage of low recall bias, a disadvantage is that they frequently interrupt the user and thus long-term adherence is generally poor. In this paper, we propose a less-disruptive self-reporting method, "assisted recall," in which in the evening individuals are asked to answer questions concerning a moment from earlier in the day assisted by contextual information such as location, physical activity, and ambient sounds collected around the moment to be recalled. Such contextual information is automatically collected from phone sensor data, so that self-reporting does not require devices other than a smartphone. We hypothesized that providing assistance based on such automatically collected contextual information would increase recall accuracy (i.e., if recall responses for a moment match the EMA responses at the same moment) as compared to no assistance, and we hypothesized that the overall completion rate of evening recalls (assisted or not) would be higher than for in-the-moment EMAs. We conducted a two-week study (N=54) where participants completed recalls and EMAs each day. We found that providing assistance via contextual information increased recall accuracy by 5.6% ($p = 0.032$) and the overall recall completion rate was on average 27.8% ($p < 0.001$) higher than that of EMAs.

[*]Mashfiqui Rabbi is the corresponding author mrabbi@fas.harvard.edu.

Additional Key Words and Phrases:

Context-aware computing; self-report adherence; engagement; mobile health; real-world study; experience sampling; ESM; EMA; recall; episodic memory; interruption

1 INTRODUCTION

While regular engagement in data collection is an integral part of many mobile health (mHealth) applications, it is very difficult to achieve [3, 43, 69]. A recent industry market research report shows that 74% people stop engaging with mHealth apps after only 10 uses [43, 69]. Passive sensing can improve adherence to data collection, but many subjective experiences cannot yet be passively sensed accurately (e.g., perceived stress, loneliness, helplessness) [2, 16, 56]. These subjective experiences have to be self-reported by higher-burden methods like self-reports. The question is then, how can one improve engagement in self-reporting? In this paper, we propose a novel self-reporting method, to collect the same data as an in-the-moment Ecological Momentary Assessment (EMA), but with less user burden.

Ecological Momentary Assessment (EMAs) is a popular yet high-burden method to capture in-situ subjective experiences. Since people report their feelings/experiences at the moment they are asked, EMA drastically reduces recall bias. However, a key limitation of EMA is that it interrupts a user's daily work flow, which can be burdensome and cause app abandonment. A common strategy to offset EMA burden is to provide financial incentives [35, 45, 77, 82]. But financial incentives are not always scalable, particularly when the regular self-reports are to be used as part of a just-in-time intervention or in settings in which data collection is desired over long periods of time. Strategies that do not require large amounts of money are needed to improve regular engagement in self-reporting.

One strategy to reduce financial incentives for EMA is to lower its burden[21]. The idea here is that if self-reporting burden can be lowered then people will need lower incentives to self-report [21, 52]. Some researchers have already tried to lower burden to improve EMA adherence. However, the current low-burden approaches make three compromises in order to lower burden: (i) shortening the EMA questionnaires, often to just one question [1, 17, 28, 32, 51, 70, 85], thus reducing the reliability and range of data that can be collected; (ii) asking EMAs only at interruptible moments [17, 60, 70, 85], thus limiting the range of situations that can be studied; and(iii) requiring the individual to wear or carry an extra non-smartphone-based device, such as a smartwatch, that is easier to access in-the-moment than a smartphone [17, 60, 70, 75, 85], thus increasing the complexity and cost of conducting EMA studies. It remains an open question whether one can reduce the burden of EMAs without making these compromises.

Recalling earlier moments from the day in the evening, or *evening recall*, is a less burdensome alternative to EMA that does not require the above compromises. Evening recall is arguably less burdensome because the individual can schedule the recall for a convenient window of time in the evening at which the individual is likely interruptible [38]. Furthermore, evening recalls do not require an additional device to lower-burden and we can

ask questions about both interruptible and uninterruptible moments from earlier in the day. As individuals are more cognitively available at the time of the evening recall, evening recall does not require that questionnaires are shortened nearly as drastically as they are for an EMA [1, 17, 28, 32, 51, 70, 85]. Thus evening recall has great potential to improve self-report adherence because it is less burdensome, does not require the use of additional devices, nor extremely short questionnaires that are completed at only interruptible moments.

However, a downside of evening recall is recall bias—i.e., certain experiences may not be remembered accurately or in their entirety [36, 59]. Fortunately, the literature in episodic memory suggests that contextual data surrounding a moment can improve the recall of memories of that moment [10, 36, 38, 59]. Episodic memory uses a particular moment's contextual details as an index for memories of that moment [36]. The day reconstruction method (DRM) by Kahneman et al. is a famous example of collecting recall-based self-reports where contextual information is provided to improve episodic memory. However, the context in DRM is provided manually by participants where, in the first part of the questionnaire, participants self-report what they were doing as episodes which are between 15 minutes to 2 hours long. In the second part, participants use the episode list from the first part to self-report their mood, where they were, who they were with, etc. [37, 38]. Providing such episode information manually is highly burdensome, which limits DRMs utility for daily self-reporting [37]. Luckily, phone sensors can automatically pick up many details of such episodes (e.g., changes in location, ambiance, physical activity) and participants do not have to be burdened to manually recall episodes. We thus hypothesize that when an app provides automatically captured contextual details from phone sensors as part of an evening recall, the individual will be able to use the information to more accurately recollect the moment that took place in that context[1].

To reduce self-report burden while keeping recall bias low, we developed a mobile phone application called ReVibe. ReVibe stands for "Remembering/Reviving the Vibe" of a specific moment. ReVibe automatically records contextual information, such as location, physical activity (e.g., walking, running, driving, etc.), and ambient sounds (e.g., people talking near by), from a user's phone. Then at a user-specified evening time, ReVibe prompts the user to answer questions concerning various moments earlier in the day. The recall is assisted with contextual information collected during those moments. For example., at 8PM, ReVibe can ask a user to recall stress at 3:15PM, providing the user with the information about her context (location, physical activity, and ambient sound) around 3.15PM to assist the requested recall.

To evaluate ReVibe, we conducted a 14-day study (N=54) to address two scientific questions. The first question was whether providing contextual information improves evening recall accuracy. To assess the effect of contextual information on recall accuracy, each evening recall was randomized to provide or not provide contextual information: location, physical activity (e.g., walking, sitting) and ambiance (e.g., people talking nearby)

[1]For rest of this paper, we will refer to context as automatically captured physical activity, location, ambient sound data from phone sensors.

about the moment being recalled. To measure evening recall accuracy, we collected EMA data for the same moment that the evening recall also asked about. EMA responses thus acted as “ground truth” to which evening recall could be compared. [38]. Recall accuracy was then measured by comparing evening recall answers to the corresponding EMA answers. The second scientific question was whether adherence rates to evening recalls (assisted by context or not) were greater than adherence rates to EMA. Finally, we conducted exploratory and qualitative analyses to further understand how assisted recall functions.

The specific contributions of this paper are the following:

- Design and development of the ReVibe application which enables assisted evening recall with contextual information and provides a experimental platform to examine whether evening recall affects adherence and whether contextual information affects recall accuracy (section 3.1, 3.2.2).
- In an evaluation experiment on 54 participants for 14 days, we found that assisted evening recall holds much promise for reducing burden as compared to EMA yet provides more accurate answers than unassisted evening recall. In particular, the experimental results support the finding that providing contextual information with evening recall increases the accuracy of evening recalls. The results also show that the completion rate is higher for evening recalls than for EMAs (section 4.1, 4.2).
- In-depth quantitative and qualitative analyses that point to additional benefits of using evening recall to improve self-report adherence and using contextual information to improve recall accuracy (section 4.4).

2 MOTIVATION AND RESEARCH QUESTIONS

Improving recall with contextual information:

Recalling earlier moments from the day in the evening can be a low-burden alternative to EMAs, but evening recall of earlier events can have recall bias due to difficulties in recovering memories in their entirety. Luckily, theories of episodic memory suggest that we can improve recall using contextual information. A proper description of episodic memory is beyond the scope of the paper, but we describe a few key ideas below and include a detailed literature review in section 5. Episodic memory, as a distinctive information processing system, receives and stores information about events and organizes the events based on their temporal-spatial relations [73]. For instance, episodic memory stores that I met a friend at Starbucks last night. Studies have also found certain memory aids are helpful to facilitate recall [74]. Giving people cues, such as contextual information by which they originally encoded the target events will help them remember these events. Tulving and others have used contextual information (such as location, companions, date and time, and thoughts when a target event occurs) as cues to facilitate recall of a target event. Other studies [14, 25, 63, 64, 79] have found that context information concerning the activity itself is most effective for memory recollection, followed by location, people involved, and thoughts at that moment. In these studies, time was found to be the least effective retrieval cue.

Since modern mobile apps can easily capture contextual information such as activity, location, and ambient sound, we investigate whether contextual information about daily moments captured by mobile phones can improve the recall of those daily moments. To the best of our knowledge, such a question has not been tested in real-world settings. Previous studies used controlled experiments with predefined tasks (e.g., math task, campus tours) [58] and predefined contexts (e.g., lab, library, etc.) [49], to test how context influences recall. These controlled experiments captured context cues manually. Moreover their use of raw audio [39] and photos [68] raises privacy issues. A review of these controlled experiments are included in section 5. In our work, we use a uncontrolled setting where contexts are specific to the users as they went on with their daily lives. Furthermore, the contextual cues are captured unobtrusively and automatically using privacy sensitive passive sensing (more in section 3.1.2). Our specific research question regarding context-supported recall is the following:

> ***RQ1*** Is evening recall of a daily moment more accurate when passively collected and privacy sensitive contextual information from the moment is provided during recall than when such contextual information is not provided?

Improving self-report adherence by lowering burden:

User engagement with mHealth apps is generally low [69] and adding EMAs, which are high-burden and interruptive, to these apps can further reduce use. A user's ability to answer EMAs in the moment may be low because they may be unavailable [60] or cognitively busy [23, 52], because of work [2] or meetings [81]. Furthermore, EMAs can adversely affect a user's work flow [7, 24, 29]. Past studies show that it often takes up to 15 minutes to recover from task switching [33], and if users are experiencing flow [18] when they are interrupted, the task-switching can have negative effects (e.g., annoyance, frustration, stress) [5, 6, 22].

The question then is: how can we improve adherence of self-reports of daily moments? Here we turn to theories of persuasion. The Elaboration Likelihood Model (ELM) [52] and BJ Fogg's Behavior Model (FBM) [21] state two necessary components of persuasiveness of a task: motivation to perform the task and the ability to complete the task. Motivating self-report with different incentives such as money and gamification is well explored. [76]. A less-explored direction is to improve self-report adherence by lowering burden. Past efforts make compromises by reducing the number of questions in the EMA [1, 17, 28, 32, 51, 70, 85], requiring an extra device other than a smartphone [1, 28, 32, 51], or asking EMAs at times when people are more interruptible [17, 60, 70, 75, 85]. A complete review of the existing low-burden EMA methods is in Section 5. In this paper, we propose a novel recall-based method that is low-burden without the compromises of prior low-burden EMA approaches.

Evening recall of earlier moments from the day is a low-burden alternative to EMAs that does not require drastically shortening questionnaires, using extra devices, or gathering EMAs only at interruptible moments. Evening recall may be low burden because people are often at home in the evening, and likely to be more available [60] or less stressed out [2, 38, 81]. Burden can be further reduced by letting people choose the time they are asked to do the recall; past work shows that people are more receptive to interruptions if they can control

their timing [46]. Having people self-report at a time in the evening that they select as interruptible can reduce their negative reaction to EMA. Furthermore, since evening recalls are likely already low burden, they do not require additional compromises to reduce burden like past approaches; e.g., an investigator does not have to shorten the already short self-report questionnaire to lower burden. An investigator also does not have to ask at interruptible moments to reduce burden; e.g., an evening recall prompt can ask what was your stress level when you were having a meeting around 11AM earlier today. Thus, evening recalls have several promising features which may reduce burden and improve adherence. We formally test these hypotheses with the following research question.

RQ2 Is the completion rate of evening recalls about earlier daily moments higher than that of EMAs?

3 METHODS

In this section, we detail our method to answer the research questions RQ1 and RQ2. We first give an overview of the ReVibe application. We then describe a study design that uses ReVibe to answer our two research questions.

3.1 The ReVibe Application

3.1.1 Overview: ReVibe is an Android mobile app to recall the details of daily moments in the evening, whereby evening recall is augmented with contextual information to improve recall accuracy. Contextual information is automatically captured from the phone sensors. Figure 2 shows examples of answering evening recalls in ReVibe. These examples are captured on the lead author of this paper[2]. In Figure 2a, after 8PM, the lead author was asked to recall the moment at 5.54PM earlier on the same day. Figure 2b shows several pieces of contextual information from 15 minutes before to 15 minutes after 5.54PM. ReVibe showed the location on a map, his or her physical activity and ambience (e.g., if anyone is talking nearby). Figure 2c shows another example of contextual information for another recall moment at 11.03AM; in this example, the lead author was commuting to a mall from home around 11.03AM on a weekend. The contextual information about 11:03AM showed movement (combination walking and vehicle rides) and ambient noise nearby (other people talking in the bus).

ReVibe is also a platform to conduct experiments and answer scientific inquiries concerning evening recall. Investigators can: (i) sample a set of time points during the day (ii) for each sampled time point, randomize whether to ask an EMA and whether to ask an evening recall, (iii) if an evening recall is scheduled for a time point, investigators can decide whether to provide contextual information or not. The two randomizations will be combined to answer ***RQ1*** and ***RQ2*** as we will describe in section 3.2.

3.1.2 System architecture of the ReVibe app: ReVibe is comprised of two modules: a context sensing module and a self-report generation module.

[2]Note that the lead author was not a participant in the user study described later. We are not showing contextual from actual study participants for privacy reasons.

Context sensing module: The current version of ReVibe captures three kinds of context streams: (1) the participant's location (2) the participant's physical activity, such as sitting, walking, traveling in a vehicle, etc., and (3) the ambiance, such as whether there was human speech nearby. Note that these three sensor streams are available as Android APIs (physical activity and location [26]) or open source libraries (ambiance recognition) [54, 81].

ReVibe continuously records physical activity and location as one minute summaries [26]. ReVibe takes activity predictions from Google's activity recognition API (e.g., sitting, walking, running, traveling in a vehicle, etc.) [30] and computes a representative activity type every minute: the activity type that happens the highest number of times with highest confidence is chosen as the representative activity for the minute. Location sensing is performed if the representative activity type in the last minutes is not stationary. Location sensing is excluded for stationary minutes since location sensing is battery intensive and participant's location likely did not change when he or she is stationary. ReVibe captures locations using the Google's Fused location API at the highest accuracy setting [31].

The ambience detection stream continuously collects audio from the phone microphone and determines ambiance (silence, noise, or talking) in a privacy-sensitive way. First, if the root mean square value of the audio data (i.e., loudness) is below a certain threshold, the ambiance is classified as silence [44]. If the ambiance is not silent, then a classifier determines whether human speech is present or not. The human speech classifier, originally developed by Basu [8], uses hidden Markov models that infer human speech through auto-correlation and spectral entropy based features. This ambience detection technique is widely used and has been deployed in several field studies [54, 80, 81, 83] and is available as an add-on library in several open source mobile sensing frameworks [20, 40, 67]. Note that the classifier runs on the phone and all the raw audio is discarded after the ambiance is determined. No actual spoken words are recorded, making the ambiance classification process privacy sensitive. As with activity recognition, ReVibe computes a representative ambiance level every minute. The representative ambiance level is the dominant ambiance type (i.e., silence, noise or talking) within the minute.

Self-reporting module: The self-reporting module of ReVibe has two parts: a scheduler to sample a set of time points and a module to generate the surveys. ReVibe can generate two types of surveys: the first is a EMA survey for each sampled time point and with a predefined set of questions. The second type of survey is the evening recall survey, where participants are asked to recall their experience at a specific time earlier in the day. These evening recall surveys can also include contextual information about the time. Specifically, 30 minutes of contextual information around the sampled time point is pulled from the context sensing streams (15 minutes before and 15 minutes after the sampled time). The location information is shown on a map. The D3 visualization library is used to visualize minute-by-minute summaries of physical activity and ambiance levels [12]. Finally, in regard to generating the surveys, ReVibe includes a flexible survey generation tool that can take a JSON formatted survey template and generate surveys that include binary, Likert scale, multiple choice, and free form responses.

In ReVibe, a scheduler runs at 12AM to select a set of time points and randomize whether or not an EMA or an evening recall is requested for a selected time point. The project investigator determines the number of time points and when these time points will be scheduled. The investigator can also sample the time points within a given range (e.g., a random time between x and $x + r$). Once a set of time point are sampled, the scheduler randomizes whether an EMA or an evening recall will be asked for a time point. ReVibe includes EMAs because EMAs are gold standard method for in-the-moment self-reporting, and the accuracy and adherence of evening recall are assessed against EMAs [38] (more on EMAs in section 3.2 and 4). If a time point is randomized for an EMA, ReVibe's scheduler sends a push notification at that time point, and if the notification is clicked then ReVibe uses the survey generator module to create an EMA. ReVibe's scheduler also sends another push notification at a user-specified time in the evening. If this notification is clicked, ReVibe uses a listview to present a number of time points for evening recall. These time points are randomly selected from the time points scheduled at 12AM earlier in the day. Once a time point is clicked on the listview, ReVibe generates an evening recall where a participant is asked to recall the corresponding time point. ReVibe also allows investigators to randomize whether or not to include contextual information from the time-point to assist recall.

3.2 Study Protocol

We conducted a 14-day study to answer the two research questions from section 2. Again, these two questions are: (i) whether contextual information improves the accuracy of recall and (ii) whether people have higher adherence to self-report when it's administered via evening recall surveys compared to EMA. Since EMAs are the current gold standard for in-the-moment self-reporting, we used them as a benchmark against which to compare our evening recall based approach. We measured the accuracy of evening recall by comparing answers to the recalls with EMAs conducted for the same daily moments. We also compared the adherence rate of recall to the number of responses to EMAs. At the end of 14 days, we conducted usability surveys and gathered responses to open ended questions on how evening recalls compared to EMAs. Below we give a detailed protocol of this study.

3.2.1 Recruitment protocol and eligibility—We used a variety of methods to invite potential participants. First, an email was sent from the University of Michigan's Registrar to a random sample of 2,000 students. Invitations to participate were also posted on Facebook. Interested parties were asked to fill out a screening questionnaire. The screening questionnaire collected information about each respondent's demographic, level of expertise in using smart phone apps, and availability for an intake interview. An important inclusion criterion was owning an Android smartphone, because ReVibe only runs on Android OS. At the intake interview, recruiters provided a brief overview of the study and gathered informed consents. Participants then installed the ReVibe application on their phones. The recruiters worked with the participants to ensure that the app was installed correctly and answered any questions the participants had. In the intake interview, participants also set a time when they normally wake up in the morning and when in the evening they could complete recall surveys. This study was approved by the Institutional Review Board of the University of

Michigan (HUM00116942). The study was financially compensated. Participants earned 50 cents for each completed EMA and evening recall.

3.2.2 Study design and experimental manipulations—To answer ***RQ1*** and ***RQ2*** from section 2, we ran a 14-day study. We choose a 14-day study duration because several prior EMA studies are of this length [76]. Below, we first describe he questionnaire we used for self-reports. We then discuss how we randomized EMAs and recalls to answer how evening recall improve adherence and whether contextual information improve recall accuracy.

Table 1 contains a seven-question questionnaire that we used for both recalls and EMAs. These seven questions were designed in consultation with a professor in the Department of Psychiatry at University of Michigan, and they are similar to those a researcher would ask in an EMA mental health study (i.e., a mixture of sensitive and cognitively demanding questions). Note that the fifth question is about sensitive topics; we maintained the anonymity of the answer to this question by asking a randomly selected question from Table 2. We kept only the response without recording which question from Table 2 was asked. The question order is randomized for EMA and recall to remove any order effects [41]. Finally, in the recalls, we gave an option "cannot remember," which was not included in the EMAs, since participants were asked to do EMAs in the moment at the selected time points.

In the 14-day study, ReVibe was used to sample 4 time points per day, two of which were randomized for EMA. During the other two time points, participants received a push notification that simply said "remember this moment." ReVibe also sampled two of the 4 time points for evening recall. The time points that were both randomized to EMA and selected for recall were used to assess recall accuracy (i.e., ***RQ1***). A recall was accurate if the answers to the evening recall matched the answers to the EMA. Furthermore, the quantity of EMA responses were compared with the quantity of evening recall responses to assess the adherence of evening recalls (i.e., ***RQ2***). Specific details of the 14-day study follow.

For each day of the 14-day study, we selected four time points with the goal to ask EMA or evening recalls at those moments. We refer to these four time points as "decision points" [48]. We uniformly spread these decision points over 12 hours each day (one decision every three hours) in order to capture the variability of day-to-day life [38]. The four decision points were chosen as follows: at the start of the study, participants set a time when they wake up in the morning. Starting from this wake-up time, four decision points were selected as follows: (i) 12 hours after the wake-up time was divided into four 3-hour consecutive non-overlapping blocks. One decision point was chosen at a random time within each 3-hour block, and (ii) no two decision points were within one hour of each other. This gap ensured there was enough time between two successive self-reported moments and we are not gathering data about the same moment.

Once the four decision points are chosen, they were randomized for different types of self-reports. At the beginning of the day, two of the four decision points were randomly selected to send a push notification that asked the participant to fill out an EMA. These EMAs asked

the questions from Table 1. During the other two time points, participants were sent a push notification that just asked them to "remember this moment." For both EMA and "remember this moment," the notification automatically timed out after 1 hour and was removed from the notification tray. Finally, at the end of every study day, two of the four decision points were randomly chosen for recall. Note, no specific preference was made on whether the same decision point was an EMA. In each recall, participant were asked to recall their experience during the selected decision points via the same questions from Table 1 used for EMA, with an additional "cannot remember" option provided. The participant specified an evening time when he or she wanted to be prompted to recall the two decision points. A manipulated variable during the recall was the availability of certain contextual information. One of the two decision points selected for evening recall was randomly chosen and was amended with information about the moment's location and ambiance (e.g., people talking nearby) and the participant's physical activity (e.g., walking, sitting) at that moment.

Figure 3 shows a visualization of how the different randomizations are spread out across different days for a participant. Let t_W and t_r respectively be the wake-up time and recall time set by the participant. The circles represent different decision points. For each of the 3-hour block from t_w, there is one decision point; i.e., for each 3-hour block of $[t_w, t_w + 3h)$, $[t_w + 3h, t_w + 6h)$, $[t_w + 6h, t_w + 9h)$, *and* $[t_w + 9h, t_w + 12h$, there is one decision point. A filled circle represents a decision point randomized for asking for an EMA and an empty circle represents a decision point randomized to not ask a EMA. At time t_r, two of the four decision points are selected for recall. One of these decision points may have contextual information provided. We added a * to denote the decision points for which contextual information is provided. Note, the randomizations will be different for another user—i.e., the times of the decision points, and which decision points are selected for EMAs and evening recall would be different due to randomization.

3.2.3 Exit protocol—After the 14-day study, we conducted a web-based exit survey where we measured burden and asked open ended questions. We used the three dimensions from the user burden scale by Suh et al [65], namely (i) interruption of daily work flow or social situations (ii) mental load to complete questionnaire, and (iii) privacy burden of the research questions. We also asked open ended questions about the usefulness of different contextual cues, differences in participants' experiences in answering EMAs and recalls, and how the ReVibe app could be improved.

3.3 Analysis Plan

Now that we have discussed the study design and the data we captured, we state analysis plan to answer the research question ***RQ1*** and ***RQ2***. We also specify a variety of additional quantitative and qualitative analyses that provide further clarification of the analyses of ***RQ1*** and ***RQ2***.

***RQ1* analysis:** This analysis focuses on how recall *accuracy* changes when contextual information is provided. The accuracy of recalls is assessed with EMAs, as has been done in earlier studies [10, 38, 49]. The research hypothesis precisely is the following:

H1: For time points with both EMA and recall, the responses to the questions in recalls are more likely to match the corresponding question responses in EMA when contextual information is provided during recall than when no contextual information is provided.

For this hypothesis, we use stress (Q1), mood (Q2), sensitive (Q5), and mindfulness (Q6) question responses (see Table 1) because they are always asked and their answers can be unambiguously compared because they are not open-ended questions. We evaluated a binary outcome; 1=if EMA and recall response to a question are the same for a time point, 0=otherwise. This binary outcome is evaluated separately for each of the four questions (i.e., Q1, Q2, Q5, Q6), and a dataset was created, where for each question, there is one row in the dataset. In other words, the total number of rows in the analyzed dataset is 387*4=1568, which is four times the 387 decision points that we included for H1's analysis. Configured this way, the dataset has a three-level structure, where we have decision points nested within a person and questions nested within a decision point. We excluded the third and fourth questions from Table 1 because they are open-ended and, even if we used manual coding to determine matches, we would need to quantify the uncertainty of inter-coder reliability and factor that into the analysis, none of which is straightforward. We also excluded the seventh question, because it is only available when the answer to the sixth question is "yes." Finally, we define availability of contextual information during recall as a binary intervention, where intervention =1 if recall has context and =0 if recall is without context.

To estimate the effect of context on recall accuracy, we use generalized estimating equations (GEE) with working independent correlation matrix and robust standard error [11, 66]. The robust standard error adjusts for correlated outcomes, in this case, multiple questions answered by the same participant. Here, we arranged the data in such a way that the data points from each participant occupy contiguous rows. When the data set is organized in this way, GEE will only use between-subject differences to estimate the standard errors, and we do not need to separately adjust for the nested structure of questions within a single decision point. Finally, we include two covariates to reduce noise [55, 71, 72]: these covariates are "time gap between recall and EMA" and "day in the study." We included the time gap between recall and EMA because we expect that the chance that people will not be able to recall rises as the time gap increases [36]. We include day of the study to see if recall quality drops as people spend more time in the study. We code "day in the study" as 0,1,2,…, 13 for the 14-day study and "time gap between recall and EMA" in hours; both of these covariates are continuous. The analysis only uses time points when both EMA and recall responses are available.

***RQ2* analysis:** This analysis deals with the quantity of self-reported data using the evening-recall-based method in comparison to using EMAs. This research question addresses the issue of low engagement with mHealth applications, because a positive answer would mean we can collect more self-reported data using recall-based methods. The research hypothesis precisely is the following:

H2: The completion rate is higher for recalls than for EMAs.

Our analysis plan to test this hypothesis is the following. We used a binary outcome variable, where 1=when a response for EMA or recall is completed after a notification and 0=otherwise. Here we compare the adherence rate to EMA with adherence rate to recall. Self-report type is coded as binary where 0 represents EMA and 1 represents recall. We use two covariates to reduce noise [55, 71, 72]: “day in the study” and an interaction between “day in the study” and intervention, or self-input type. We included “day in the study” because earlier studies have shown that self-report rates typically drop over time [19]. We code “day in the study” as 0,1,2,…, 13 for the 14-day study. We added the interaction between intervention and “day in the study” as a covariate because adherence rates may change differently for EMA and recall. Finally, we use a generalized estimating equation with an independent correlation matrix and robust standard error [11, 66]. The robust standard error adjusts for correlated outcomes, that is the multiple binary outcomes on each participant [11].

Additional quantitative and qualitative analyses: We conducted a series of quantitative and qualitative analyses to investigate how evening recall affects self-report adherence and how contextual information influences recall accuracy. In particular, we conducted descriptive analyses of the quantitative exit survey data and thematic analyses of the qualitative data [13]. Recall that the qualitative exit-survey questions asked about the differences in participants’ experiences of answering EMAs and recalls, the usefulness of different contextual cues, and how the ReVibe app could be improved. The quantitative exit survey question asked participants to rate the usefulness of different contexts (i.e., location, physical activity, ambient sound) for supporting recall on a 7 point Likert scale (1=not useful all, 7=very useful). Additionally we tested whether evening recalls were indeed less burdensome than EMA. Note, evening recall being low-burden is a driving hypothesis behind this paper. As such, we analyzed potential differences between user burden ratings from the exit survey for EMA and evening recall. The survey asked three questions from the user burden scale by Suh et al. [65], namely (i) interruption of daily work flow or social situations (ii) mental load to complete questionnaire, and (iii) privacy burden of the questionnaire. Since these responses are in ordinal scale, we used a non-parametric Mann-Whitney-U test to ascertain statistical significance.

4 RESULTS

4.1 Participant Sample and Dataset

We recruited 56 participants (23 males and 33 females) from the University of Michigan campus. The participants were undergraduate and graduate students. The mean age of participants was 19.7 (μ= 19.7, σ = 2.7). All participants were proficient in using smartphones: when asked about their expertise to use mobile apps using a seven point Likert scale, where 1=rarely use mobile app, 7=very comfortable with mobile apps, participants provided an average rating of 6.36 (μ= 6.36, σ = 0.4). 27 out of 56 participants reported keeping a journal to log their daily lives. When we asked about why they were interested in the study 33 participants said they were interested in self-monitoring, 54 participants reported the study to be interesting and 55 participants said they were interested in the financial incentives.

We excluded two participants from the study because they were using Huawei phones, which ran a modified Android OS that regularly killed ReVibe's background process responsible for scheduling notifications and recording contextual data. Thus data from 54 participants (22 male, 32 female) was used in these analyses. Of these 54 participants, one participant switched to an iPhone on the eighth day of the study (our ReVibe application only runs on Android) and thus only the first seven days of data are included. The remaining 53 participants used the ReVibe application for 14 days. Once the 14-day study concluded, we asked participants to complete an exit survey. All 54 participants completed the exit survey and collected study compensation. The analysis for H2 use decision points that are scheduled for notification for an EMA or an evening recall. This condition is satisfied by 1317 EMA decision points and 1432 evening recalls decision points, and we include them for the analysis. Note, ideally the phone should notify 1568 times for both EMA and evening recalls, but the phone sometimes failed to schedule notifications because ReVibe's background process was killed or suspended by the Android OS. The analysis for H1 uses only data from time points at which there was an EMA and an evening recall for this time point. In total for all participants 616 time points were randomized and notified for both EMA and recall, out of which 387 were completed (i.e., 62.9% response rate). Thus the analysis for H1 uses 387 decision points.

4.2 Analysis for H1: Effect of Contextual Information on Recall Accuracy

Table 3 shows the results for the **H1** analysis. The results indicate that when contextual information is provided recalls are more likely to match EMA responses for the same question. The coefficient on the *context* variable is significant (p=0.034). The 'log odds ratio' for the context coefficient is 0.27 with a standard error of 0.12. This suggests that when contextual information is provided the probability of getting a match between EMA and recall increases by 5.6%. In addition, the *time gap*, which is the time difference between EMA and recall in hours, is significant (p=0.016). Furthermore, the coefficient on the time gap variable is negative, which means that if the time gap between EMA and recall is longer then the probability of a match between EMA and recall is lower.

4.3 Analysis for H2: Adherence Comparison between EMA and Evening Recall

Remember that for the H2 analysis, we are interested in the relative completion rates of EMAs and evening recalls; here, the outcome of interest is whether a self-report (EMA/recall) is completed and the experimental manipulation is whether the self-report is an EMA or a recall. In this analysis, if the phone schedules a notification for recall or EMA then the point is included in the analysis. We included 1317 decision points for EMAs and 1432 decision points for evening recalls for analysis.

Table 4 shows the results of the **H2** analysis. The results indicate that when the self-report type is a recall, it is more likely to be completed than when the self-report is an EMA (p<0.0001). The log odds ratio for the self-report coefficient is 1.1 with a standard error of 0.1. This suggests that recalls are 23.6% more likely to be answered than EMAs on the first day of the study. We also found day of study to be significant ($p < 0.001$) and the log odds ratio is negative. However, the interaction between day of study and self-report type is positive, which approaches significance ($p = 0.06$). These results, in combination, mean that

we expect EMA completion to decrease by 0.8% each day. However, recall completion decreases by only 0.2% each day (−0.8+0.6). Therefore, across the 14-day study, EMA completion rate decreased 11.2% whereas recall completion rate dropped only by 2.8%. i.e., by the 14-th day of the study, recall completion rate would be 32% higher than EMAs.[3]

4.4 Additional Quantitative and Qualitative Results

In this subsection, we describe additional quantitative and qualitative results that point to further benefits of using evening recall to improve self-report adherence and benefits of using contextual information to improve recall accuracy. Note again, unlike the analyses of H1 and H2, these additional analyses are not adequately powered. So, the results should interpreted as patterns that triangulate and provide additional insights to the findings of the analyses of H1/H2. We enumerate these additional results below:

1. **Evening recalls are less-burdensome than EMAs:** In both qualitative and qualitative analyses, we found that evening recalls are less burdensome than EMAs. We quantitatively measured user burden by including a sub-scale from the user burden scale by Suh et al.[65] in the exit survey, where we asked about interruption burden, mental effort, and privacy burden in a seven item Likert scale (1-Not at all, 7-very much). Figure 6 shows the distributions of various types of user burden. We found the interruption burden of recall ($\mu = 2.2$, $\sigma = 1.3$) to be lower than that of EMA ($\mu = 4.4$, $\sigma = 1.5$), and the difference is significant ($W = 461, p < 10e^{-5}, d = 1.06$). The mental effort burden of recall ($\mu = 2.2$, $\sigma = 1.4$), was similar to EMA ($\mu = 1.9$, $\sigma = 1.2$) and the difference was not significant ($W = 1724, p = 0.17, d = 0.18$). However, privacy burden(for recall burden) ($\mu = 1.7$, $\sigma = 1.03$ was lower than that of EMA ($\mu = 2.4$, $\sigma = 1.6$), and the difference was significant ($W = 1153, p = 0.021, d = 0.35$). Therefore, EMA imposed more interruption and privacy burden than evening recall.

Qualitative results echoed that evening recall was less burdensome than EMAs. Several participants (26%=14/54) mentioned that EMA interrupted their daily work flow; some were also annoyed by the random timing of EMAs and their requirement of immediate attention. One participant said *"I found the daily moment prompts to be unenjoyable just because they would literally come at the most random parts of day. There are times where having to stop and reflect on your mood and if you had any intrusive thoughts is detrimental to your day and your productivity."* On the other hand, several participants (22%=12/54) saw recall as convenient and more predictable. For instance, one participant said *"It [EMA] is highly irritating as it demands immediate attention. Evening recall was better because I could set it up at a convenient time."*. Another participant said, *" I think it was easier to just respond to the moments at the end of the day because I knew it was coming and I was able to prepare for it more than if it just popped up during the day. "* Some mentioned that due to their busy

[3]Note that, we asked participants to complete two recalls in the evening, and our results for recall completion rate (i.e., RQ2) consider the two recalls as distinct self-report completion. However, in 98.4% of days, participants completed both evening recalls for the day, since two evening recalls are asked at the same time. The same cannot be said for EMAs; i.e., completing one EMA does not improve the odds to complete the other EMA since they are asked separately. Nonetheless, if we consider each discrete engagement in self-reporting after a notification (i.e., we consider an EMA completion after a notification as one discrete engagement and one/two evening report completions after an evening notification prompt as one discrete engagement) the fraction of discrete engagement for self-report per notification is 50.7%,56.1%, 68.1%, 82.6% respectively for the first, second, third and fourth decision points, and 87.7% for evening recall notifications.

schedule they couldn't always answer the EMAs, *"I preferred the evening applications, as I'm usually pretty busy during the day and couldn't always get to the daily events [EMA], but pretty much always could make it to the evening recalls"*.

2. **EMA adherence was lower in the morning compared to afternoon:** We found that EMA adherence was higher for the afternoon moments than morning moments, although no such difference was observed for evening recall. For this analysis, we used the same dataset as **H2**, with the added variable decision time in the day. Note, an earlier decision point in the day will be in the morning and a later decision point will be in the afternoon. We found that the time of the decision point had a significant moderating affect on the probability to answer EMA/recall. Results are shown on Table 5.

The coefficient for decision point has marginal effect of 7.1% ($p < 0.001$), which means EMA adherence was higher in the afternoon compared to in morning. Specifically, EMA adherence increased by 7.1%,14.2%,21.3% for 2nd, 3rd, 4th decision point from the first morning decision point. However, for recall, the coefficient for decision point has marginal effect of 0.2% (=7.1–6.9). i.e., recall adherence did not change much. That is, people generally replied to decision point from morning or afternoon equally when it comes to self-reporting evening recall.

3. **Evening recall promoted self-reflection:** In the qualitative analysis, a surprising finding was that some participants (20.4%=11/55) mentioned they were more *self-reflective* when they were filling out their recalls. One participant mentioned that recalls were *"a better reflection of how I felt overall after the day."* Another participant mentioned evening recall made her more aware of patterns of daily life, thoughts, and behavior; *"evening recall allowed me to think about my day using more effort, which in turn, forced me to reflect upon my day more. Asking for answers [EMAs] throughout the day allowed for surface, minimal amounts of self reflection, while evening recall allowed a more in depth reflection. Evening recall also allowed me to pick up patterns of my daily life, and possibly ways to change my patterns of thought and behavior."* Another participant mentioned evening recall was an opportunity to analyze whether s/he was having any negative thoughts. For EMAs, on the other hand, a few participants(11.1%=6/54) reported that they were paying less attention and just filled out the EMA in the moment. For instance, one participant said *"I would bust out the survey when prompted in 30 seconds or so, just using my stream of thought."*

4. **Contextual information during recall reduced recency effect:** Recalling past events suffers from the recency effect—i.e., more recent memories are recalled at higher accuracy than less recent memories [36]. We saw a similar effect in section 4.2, where recall accuracy dropped as the time gap between EMA and recall increased. However, we also found evidence that recall accuracy may have dropped less with increased time-gap when contextual information was provided during recall ($p = 0.09$). For this analysis, we used the same dataset as H1. The analysis used the same outcome and statistical method as the analysis for H1. To the model we added predictors for day-in-the-study, time gap between EMA and recall, and an interaction term between this time-gap and availability of context. Table 6 shows the results. The coefficient for time-gap has marginal effect of −1.3% ($p =$

0.004), which indicate recall accuracy dropped by 1.3% with each hour of time gap between EMA and evening recall. However, when contextual information is provided, recall accuracy improved by 0.6% with each hour of time gap between EMA and evening recall ($p = 0.09$). In real terms, this result suggests that the recency effect in recall accuracy was smaller when contextual information was provided.

5. **The absolute value of recall accuracy was high:** So far, we only discussed relative increase of recall accuracy when contextual information was provided. However, the absolute value of recall accuracy is important because a lower accuracy would mean evening recall cannot be a useful measure of in-the-moment stress, mood, etc. Fortunately, the absolute value of recall accuracy was quite high. For this comparison, we used EMA as ground truth similar to Kahneman et al. [38]. We compared the absolute difference between EMA and evening recall responses for the same moment: e.g., if self-reported stress levels are 4 out of 5 in EMA and 3 out of 5 in evening recall then the absolute difference is $|4 - 3| = 1$. Thus, a lower absolute difference would mean EMAs and recall closely match with each other and vice versa[4]. If the difference is small then evening recalls are good approximations of EMAs.

Table 7 shows the results of the questions from Table 1 and 2 that we included for the analysis for H1. As one can see, for Likert scale questions, the mean and standard deviation are low when contextual information are provided during recall (middle column in Table 7). The mean is also less than 1, which means on average the absolute difference between EMA and evening recall is less than one point. This result is further reinforced by the right most column of Table 7, which shows that the absolute difference between EMA and evening recall is ≤ 1, 86.1–93.6% of the times depending on various question types. Finally, for binary scale, we compare exact matches between EMA and recall for the same moment. We found EMA and evening recalls match 85.8–87.7% depending on various questions.

6. **Location was rated as the most useful contextual information:** In the exit survey, participants rated usefulness of location, physical activity and ambient sound in improving recall accuracy. We used a 7 point Likert scale (1=not useful at all, 7=extremely useful). Location was rated as the most useful context ($\mu = 4.6$, $\sigma = 2.1$) followed by physical activity ($\mu = 3.1$, $\sigma = 1.7$) and ambient sound ($\mu = 2.6$, $\sigma = 1.7$). Note that we have no quantitative evidence on how much each type of context affected recall accuracy. We address this issue in more detail in the limitations section (Section 7).

5 RELATED WORK

Before we provide a discussion of the results, we give a detailed overview of the related works of ReVibe to further situate the discussion and impact of the results.

[4]Note, the absolute difference cannot be less than 0 or greater than 5 since maximum of the likert scale is 5

Lowering Burden of EMAs

Here we review previously developed methods to lower EMA burden, which can be grouped into two general approaches. We also show that none of these approaches has a clear superiority to ReVibe.

The first approach to lowering burden of EMAs used additional non-smart-phone based devices; e.g., off-the-shelf wearables (e.g., a smartwatch or Google Glass), or devices created specifically for EMAs. These devices are either stand-alone or have to be synced with a smartphone. They rely on two mechanisms to lower EMA burden: (i) being easier to access than pulling out a phone out of the pocket, (ii) asking questions that are answerable with a single interaction (e.g., one tap). Hernandez et al. [28] were the first to investigate EMA adherence on phones, smartwatches, and Google Glass. They asked either Likert scale or 2D grid (e.g., Russell's affect grid) questions that were answerable by single taps. In an evaluation study, EMA adherence on smartwatch and Google Glass was 13% higher than on phones. But this difference was not statistically significant. Other "extra-device" approaches lowered burden further by asking only one question at a time. Heed is a small disc-shaped device with a touch sensitive surface [51], which can be placed in the environment like a smart home device (e.g., Amazon Echo or Google Home). Users can tap on Heed's surface to self-report in a Likert scale[5]. Keppi is a portable palm-sized cylindrical device with a pressure sensitive surface [1]. Users can self-report different level of a measure (e.g., pain) by gripping Keppi at different intensities. μ-EMA is another low-burden EMA method that uses a Moto 360 smartwatch to ask one question at a time with time gaps between questions [32, 53]. The early results of μ-EMA show higher adherence than phone EMA has [32, 53]. But phones and watches had the same adherence rate when multiple EMA questions were asked back-to-back on the watch without a delay [53]. This result suggests that μ-EMA's high adherence is due to asking one question at a time rather than to the watch as a form factor.

The "extra-device" approach has a few disadvantages. First, users must own and carry an extra device, and these devices are not as ubiquitously adopted as smartphones. Second, in smartwatch based approaches, the small surface area of watches can cause occlusion or the fat finger problem [9, 62]. This small surface area limits the amount of text that can be displayed and precludes the typing of free-form responses. Several solutions to the fat-finger problem have been proposed: zooming [50], adding extra buttons [50], making the sides or the back of the device touch enabled [4], adding extra degrees of freedom for interaction [84]. However, these alternate forms of interactions have not yet been widely used for EMA studies. Third, devices like Heed or Keppi that ask a single question at a time do not allow investigators to capture multidimensional data on a particular moment to answer scientific questions like how stress is related to mood or loneliness [81]. Phones, on the other hand, can do EMAs that require users to answer a number of questions in order to complete the self-report. An μ-EMA user can be asked to provide the same information gathered in a

[5]Note that, Heed also allowed participants to self-report on their smartphones. Authors of Heed reported that smartphone was preferred for self-reporting on-the-go, but participants found the small disc-shaped device more convenient for self-reporting when it was available [51]. Since on-the-go self-reporting on a smartphone is same as EMAs and this paper focuses on lowering burden options for EMAs, we only discuss the Heed system's small disc-shaped device.

phone EMA by answering single question prompts in succession. But users are less likely to answer several μ-EMA prompts than a single μ-EMA prompt [53].

One advantage of extra-device-based EMA methods is they gather information in the moment. In-the-moment data is critical for applications that require immediate action based on survey responses. However, the goal of many health intervention apps is to predict adverse health conditions ahead of time [34, 47]. ReVibe's recall approach is appropriate for such proactive intervention applications.

The second approach to lowering burden of EMA is opportunistic methods to capture EMAs on mobile phones. One method is to predict interruptible moments and ask EMAs in those moments [60]. This is a well-explored area [75]. However, most interruption prediction models are offline because it is hard to integrate and sync multiple sensors to predict interruptible moment. It is unclear what kind of adherence rates one would get when an interruption prediction model selects the best times for EMAs. Another opportunistic method of capturing EMA is to embed EMAs in activities that users are already doing on their phone. For instance, Truong et al [70] and Zhang et al. [85] modified the swipe gesture to unlock devices for EMAs. Choe et al. [17] used lock screen widgets to capture sleep data. However, most modern phones have disabled lock-screen widgets and replaced swipe-to-unlock gestures with face ID or fingerprint ID. Nonetheless, the idea of opportunistic interaction to capture of EMA is generalizible. EMAs can be integrated with other common phone interactions. e.g., the Youtube mobile app asks users to do surveys before playing videos. Other opportunistic ways to capture EMA could be to ask questions when users open the phone or interact with a certain app. However, two limitations of these opportunistic interaction approaches is that they ask only one EMA question at a time [17, 70, 85] and their completion times are not uniformly distributed across time. Thus opportunistic EMA data could be biased to moments when users are interruptible and less stressed out [2, 28, 38, 61].

Use of Context to Improve Recall of Memory

The role of context in constructing and recollecting human memory is well documented in psychology [36]. Episodic memory is a part of explicit long-term memory, which is responsible for storing and retrieving event related memories. e.g., my last birthday party or I took a bus in the morning to work[6]. The predominant theory of episodic memory holds that our episodic memory is organized by contexts. Context is a complex high-dimensional entity, and context changes quickly as new events happen in our lives. Environmental factors (location, ambiance, etc.) are shown to be a part of context that encodes memory, but there are internal or external factors that also contribute to context-based memory encoding. Furthermore, since context is high-dimensional, it can change significantly within a short period. Thus, recollecting past memory of events can be difficult because context likely shifted as new things happened in life. However, if some memory cues are issued, they can help us to get back to past context when the event occurred then we would be able to remember past moments vividly.

[6]The other part of explicit long-term memory is semantic memory, which helps us recall meaning of words. Non-explicit, i.e, implicit long-term memory includes memories that we cannot clearly articulate, e.g., how to ride a bicycle

Free recall and cued recall are two common ways memory retrieval is studied in cognitive psychology [36]. Most of these studies use word memory tests, where researchers utter a list of words and participants recall the words that have been said. In free recall scenarios, no assistance is provided during recall and participants are free to recall any words they have heard. For cued recall, some assistance is provided during the recall; e.g., adjacent words, category of words, or the contexts of recall being same as when researchers uttered the sequence of works. Many experiments showed that cued recall produces better memory recollection (i.e., number of words recalled) than free recall. More precisely, free recall has a recency effect, where participants were better able to recall the words they had heard most recently than words that were uttered earlier. However, in cued recall, if hints were given participants recalled words not among those they heard most recently. In the ReVibe experiment (RQ1), the above can be translated in the following way: the control condition is a free recall, whereas the experimental condition is a cued recall. Furthermore, in the control free-recall condition, people should be able to recall events later in the day, but in the cued-recall scenario people should be able to recall moments even in the middle or earlier parts of the day.

Prior to our study, no experiment has manipulated real-world contexts captured from phones to assess their effect on cued recall. For instance, some daily diary studies used context to improve daily diary entries on the implicit assumption that contextual information can improve recall. Carter et al. used photos, audio, and location taken at various moments of the day to assist more detailed daily diary entries [15]. Rabbi et al. ran a food logging study where participants took photos of their meals during the day and labeled those photos at the end of the day [57]. Laerhoven et al. [78] ran a study where a wrist-worn sensor with accelerometer and tilt-sensors was used to continuously monitor activities throughout the day and participants used the sensor tracked information to recall what they were doing at different times of the previous day. While these studies assume that contextual information improves recall, they did not systematically manipulate context to assess its effect on recall accuracy. There are studies that systematically manipulated context to assess its effect on recall. But these studies are in controlled setting. Rahman et al. [58] ran a controlled study where participants completed a pre-defined math task in different locations. Contextual information (GPS, raw audio) about those places was captured when participants completed the math task. Later participants were asked to recall their stress with and without contextual information. Niforatos et al. [49] ran another controlled study where participants took a predefined walk on campus and contextual information was captured using pictures, taken either by participants or by passive sensors (Narrative Clip). One week later, participants were asked to recall those moments and the authors compared how passive and participant-taken pictures influenced recall. Our study, on the other hand, was in uncontrolled settings, where contextual information is captured as users were going about their daily activities. Furthermore, our contextual information is privacy sensitive, which make ReVibe readily usable for self-report studies. Both Rahman et al. [58] and Niforatos et al. [49] used raw audio and pictures, which raise privacy concerns [39, 68].

6 DISCUSSION AND FUTURE WORK

In this work, we created the ReVibe app, which uses a low-burden evening recall method to improve adherence of regular self-reports. Since evening recall can be less accurate due to recall bias, we used contextual information to improve recall accuracy. Another feature we included in ReVibe is that it contains an experimental platform to answer scientific questions about the accuracy of and adherence to evening recalls. We used ReVibe's experimentation capabilities to run a 14-day study on 54 participants, where we investigated two scientific questions:(i) how much the accuracy of evening recall improves when contextual information is provided, and (ii) how evening recall adherence compares to EMA adherence.

We found that contextual information improved accuracy of evening recall by 5.6% (p = 0.034). While the effect is not large, it suggests that even relatively simple contextual information that can be readily gathered in an unobtrusive, automated, and privacy-preserving way can help increase recall accuracy. While a 5.6% accuracy increase may not be meaningful in all settings, in settings where recall accuracy is important but lower adherence to EMA, especially long-term, is problematic, even small increases in accuracy in a method that has a high level of adherence could be significant. Contextual information during recall had other benefits as well. The absolute value of recall accuracy was high, and the difference between EMA and recall response was lower with a smaller standard deviation when contextual information was provided during recall (Table 7). Recency effect on recall accuracy was also less when contextual information was provided (section 4.4, Finding 4). Furthermore, recall accuracy with contextual information did not degrade as the study progressed (section 4.2), which is a positive finding since response quality in behavior studies often drops over time due to repeated exposure [27]. Together, these findings suggest that contextually-supported evening recall is a promising strategy for collecting high-quality data about individuals' lived experience.

We also found that adherence to evening recalls was on average 27.8% higher than the adherence to EMAs ($p < 0.001$) (section 4.3). This is a statistically significant result with a large effect size. Our exploratory and formative analyses also overwhelmingly supported this result. In the user burden scale, we found interruption and privacy burden were both significantly higher for EMAs than for evening recall (section 4.4, Finding 1). In the open-ended user feedback (section 4.4, finding 1), participants found the EMAs to interrupt their workflow and they were annoyed by their random timing. On the other hand, many participants found the more predictable recalls convenient.

Additional exploratory and formative analyses in section 4.4 also pointed to further nuances of how evening recall may influence self-report adherence. One important finding was that EMA adherence dropped significantly over time, while no such adherence drop was observed for evening recall adherence (section 4.3). Furthermore, EMA adherence in the earlier part of the day were lower than EMA adherence later in the day (likely because students are busier in the morning). Again, no such within-day drop was observed for recalling of various moments across the day during the evening recall (section 4.4, Finding 2). Finally, a surprising finding was that, in the open-ended exit survey, participant reported being more **self-reflective** during evening recall than during EMAs (section 4.4, Finding 3).

Participants were most likely freer in the evening and had more time to think and reflect on their data. This is an important finding because self-reflection may increase people's motivation, which may contribute to the adherence to evening recall [16, 21, 52].

Nonetheless, our findings also point to a few limitations of the evening recalls. One issue is the higher level of "self-reflection" during evening recall; i.e., EMAs captured more immediate in-the-moment experience whereas evening recalls produced a cognitively-processed version of what happened earlier on the day. In future work, we plan to explore how this difference in recall varies for different types of experiences, but in some settings these differences in what is reported may be meaningful and will need to be kept in mind when deciding on a data-collection method. Another issue with evening recall was that even with included contextual information it was still not as accurate as the EMA. While it's possible that the evening recall will never match EMA in absolute recall accuracy, we suspect that the gap can be narrowed much more substantially than we were able to do in ReVibe. The recall accuracy reported in this paper is just the start and it can be improved in future versions. One idea could be to create more semantically meaningful context which can provoke better recall. Another idea is to use "serial recall" from episodic memory literature which argues that people recall better when they are incrementally asked about memories from the most recent to the less and less recent [36, 38]. We will explore these directions in future work.

In summary, evening recalls have better adherence properties than EMAs, and their accuracy increased with contextual information. However, recall accuracy can be further improved with better design of contextual information, and our future work will focus on designing better representations of contexts that support better recall of memory.

7 LIMITATION

One limitation of our paper is we did a complete case analysis for H1 in section 4.2; i.e., we only considered decision points with both EMA and evening recall [42]. Thus, we do not know the accuracy of evening recalls for decision points when participants were not asked an EMA or were asked an EMA but participants did not respond. However, these missing EMA data can be filled with missing data techniques like multiple imputation [42]. Missing data imputation, nonetheless, is a non-trivial process and they have not been extensively studied for mobile health. Our future work will consider multiple imputation of missing EMA data and do a revised analysis of the imputed data set.

Another limitation of the current work is that we did not test which context is more helpful to improve recall accuracy. i.e., whether location is a more useful context for recall than physical activity. We did not do so for two reasons. First, the differences in effects of various contexts are likely to be small, so detecting these differences would require a large sample. Second, our current version of context—i.e., location, physical activity, ambient sound—is only the first version of the context to improve recall. Now that we have found evening recall adherence is significantly higher than that of EMAs, we have more incentive to develop better contextual representations for memory recollection. This better contextual information can be a complex function of location, movement, and other information.

Another potential limitation of our work is that participants were pinged four times a day and two of these pings were randomly chosen for EMAs. The extra two non-EMA pings, which we call “remember the moments,” were simple push notifications to participants to look at their surroundings so that they can recall those moments better during the evening recall. Remember the moments are simple text notifications that show up in the notification tray with a vibration and participants do not need to interact with or open these notifications. While participants can ignore the “remember the moments” prompts, if they notice their surroundings they may recall those moments more accurately according to cued recall theory [36]. These two “remember the moments” pings may cause additional burden and hamper EMA adherence. But, our qualitative data from exit surveys do not indicate that these two extra pings caused additional burden. In future work, we will explicitly verify the benefit and burden of remember the moments prompts.

Another limitation of our current work is that we did not explicitly check the quality of the EMA responses. Recent work in Ubicomp raised concerns about the accuracy of EMA responses [60, 76, 77]. This research used contextual information (e.g., location) to validate whether participants are responding authentically. But in our case, contextual information is used to improve recall. So we could not use context to validate survey response and improve recall at the same time. Future work will investigate how to check the quality of EMA answers with the added constraint of contextual information being used to improve recall.

ACKNOWLEDGMENTS

This work has been supported by the Michigan Institute for Data Science (PI: SM), NIDA P50 DA039838 (PI: Linda Collins), NIAAA R01 AA023187 (PI: SM), NHLBI/NIA R01 HL125440 (PI: PK), NIBIB U54EB020404 (PI: Santosh Kumar), NIH/NCI U01CA229437 (PI: Inbal Nahum-Shani). We thank the members of Statistical Reinforcement Learning group at Harvard University and Dynamic Decision-Making Lab at University of Michigan for their input throughout the SARA development process.

REFERENCES

[1]. Adams Alexander T, Murnane Elizabeth L, Adams Phil, Elfenbein Michael, Chang Pamara F, Sannon Shruti, Gay Geri, and Choudhury Tanzeem. 2018. Keppi: A Tangible User Interface for Self-Reporting Pain. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. ACM, 502.

[2]. Adams Phil, Rabbi Mashfiqui, Rahman Tauhidur, Matthews Mark, Voida Amy, Gay Geri, Choudhury Tanzeem, and Voida Stephen. 2014. Towards personal stress informatics: Comparing minimally invasive techniques for measuring daily stress in the wild. In Proceedings of the 8th International Conference on Pervasive Computing Technologies for Healthcare. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), 72–79.

[3]. Flurry Analytics. 2012. App Engagement: The Matrix Reloaded,. http://flurrymobile.tumblr.com/post/113379517625/app-engagement-the-matrix-reloaded

[4]. Ashbrook Daniel, Lyons Kent, and Starner Thad. 2008. An investigation into round touchscreen wristwatch interaction.. In Mobile HCI. 311–314.

[5]. Bailey Brian P and Iqbal Shamsi T. 2008. Understanding changes in mental workload during execution of goal-directed tasks and its application for interruption management. ACM Transactions on Computer-Human Interaction (TOCHI) 14, 4 (2008), 21.

[6]. Bailey Brian P and Konstan Joseph A. 2006. On the need for attention-aware systems: Measuring effects of interruption on task performance, error rate, and affective state. Computers in human behavior 22, 4 (2006), 685–708.

[7]. Bailey Brian P, Konstan Joseph A, and Carlis John V. 2001. The Effects of Interruptions on Task Performance, Annoyance, and Anxiety in the User Interface.. In Interact, Vol. 1. 593–601.

[8]. Basu Sumit. 2002. Conversational scene analysis. Ph.D. Dissertation. MaSSachuSettS InStitute of Technology.

[9]. Baudisch Patrick and Chu Gerry. 2009. Back-of-device interaction allows creating very small touch devices. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM, 1923–1932.

[10]. Belli Robert F. 1998. The structure of autobiographical memory and the event history calendar: Potential improvements in the quality of retrospective reports in surveys. Memory 6, 4 (1998), 383–406. [PubMed: 9829098]

[11]. Boruvka Audrey, Almirall Daniel, Witkiewitz Katie, and Murphy Susan A. 2018. Assessing time-varying causal effect moderation in mobile health. J. Amer. Statist. Assoc 113, 523 (2018), 1112–1121.

[12]. Bostock Michael, Ogievetsky Vadim, and Heer Jeffrey. 2011. D^3 data-driven documents. IEEE transactions on visualization and computer graphics 17, 12 (2011), 2301–2309. [PubMed: 22034350]

[13]. Braun Virginia and Clarke Victoria. 2006. Using thematic analysis in psychology. Qualitative research in psychology 3, 2 (2006), 77–101.

[14]. Brewer WF. 1988. Memory for randomly sampled autobiographical events. (1988), 21–90.

[15]. Carter Scott and Mankoff Jennifer. 2005. When participants do the capturing: the role of media in diary studies. In Proceedings of the SIGCHI conference on Human factors in computing systems. ACM, 899–908.

[16]. Choe Eun Kyoung, Abdullah Saeed, Rabbi Mashfiqui, Thomaz Edison, Epstein Daniel A, Cordeiro Felicia, Kay Matthew, Abowd Gregory D, Choudhury Tanzeem, Fogarty James, et al. 2017. Semi-automated tracking: a balanced approach for self-monitoring applications. IEEE Pervasive Computing 16, 1 (2017), 74–84.

[17]. Choe Eun Kyoung, Lee Bongshin, Kay Matthew, Pratt Wanda, and Kientz Julie A. 2015. SleepTight: low-burden, self-monitoring technology for capturing and reflecting on sleep behaviors. In Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing. ACM, 121–132.

[18]. Csikszentmihalyi Mihaly. 1997. Finding flow: The psychology of engagement with everyday life. Basic Books.

[19]. Eysenbach Gunther. 2005. The law of attrition. Journal of medical Internet research 7, 1 (2005), e11. [PubMed: 15829473]

[20]. Ferreira Denzil, Kostakos Vassilis, and Dey Anind K. 2015. AWARE: mobile context instrumentation framework. Frontiers in ICT 2 (2015), 6.

[21]. Fogg BJ. 2009. A behavior model for persuasive design. In Proceedings of the 4th international Conference on Persuasive Technology. ACM,40.

[22]. Galluch Pamela. 2009. Interrupting the workplace: Examining stressors in an information technology context. (2009).

[23]. Gilbert Daniel T, Pelham Brett W, and Krull Douglas S. 1988. On cognitive busyness: When person perceivers meet persons perceived. Journal of personality and social psychology 54, 5 (1988), 733.

[24]. Gillie Tony and Broadbent Donald. 1989. What makes interruptions disruptive? A study of length, similarity, and complexity. Psychological research 50, 4 (1989), 243–250.

[25]. Godden Duncan R and Baddeley Alan D. 1975. Context-dependent memory in two natural environments: On land and underwater. British Journal of psychology 66, 3 (1975), 325–331.

[26]. Google Play Service. 2014. http://developer.android.com/google/play-services/index.html. [Online; accessed 1 April 2014].

[27]. Groves Philip M and Thompson Richard F. 1970. Habituation: a dual-process theory. Psychological review 77, 5 (1970), 419. [PubMed: 4319167]

[28]. Hernandez Javier, McDuff Daniel, Infante Christian, Maes Pattie, Quigley Karen, and Picard Rosalind. 2016. Wearable ESM: differences in the experience sampling method across wearable

devices. In Proceedings of the 18th International Conference on Human-Computer Interaction with Mobile Devices and Services. ACM, 195–205.

[29]. Cutrell Edward Czerwinski Mary Horvitz Eric. 2001. Notification, disruption, and memory: Effects of messaging interruptions on memory and performance. In Human-computer Interaction: INTERACT'01: IFIP TC. 13 International Conference on Human-Comupter Interaction, 9th-13th July 2001, Tokyo, Japan. IOS Press, 263.

[30]. Google Inc. 2017. Activity Recognition. https://developers.google.com/android/reference/com/google/android/gms/location/ActivityRecognitionClient

[31]. Google Inc. 2017. Fused Location. https://developers.google.com/android/reference/com/google/android/gms/location/FusedLocationProviderClient

[32]. Intille Stephen, Haynes Caitlin, Maniar Dharam, Ponnada Aditya, and Manjourides Justin. 2016. μEMA: Microinteraction-based ecological momentary assessment (EMA) using a smartwatch. In Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing. ACM, 1124–1128.

[33]. Iqbal Shamsi T and Horvitz Eric. 2007. Disruption and recovery of computing tasks: field study, analysis, and directions. In CHI, Vol. 7. 677–686.

[34]. Jaques Natasha, Taylor Sara, Sano Akane, Picard Rosalind, et al. 2017. Predicting tomorrowâĂŹs mood, health, and stress level using personalized multitask learning and domain adaptation. In IJCAI 2017 Workshop on Artificial Intelligence in Affective Computing. 17–33.

[35]. Johnson Daniel, Deterding Sebastian, Kuhn Kerri-Ann, Staneva Aleksandra, Stoyanov Stoyan, and Hides Leanne. 2016. Gamification for health and wellbeing: A systematic review of the literature. Internet Interventions 6 (2016), 89–106. [PubMed: 30135818]

[36]. Kahana Michael Jacob. 2012. Foundations of human memory. OUP USA.

[37]. Kahneman Daniel, Krueger Alan B, Schkade David A, Schwarz Norbert, and Stone Arthur A. 2004. The day reconstruction method (DRM). Instrument documentation.

[38]. Kahneman Daniel, Krueger Alan B, Schkade David A, Schwarz Norbert, and Stone Arthur A. 2004. A survey method for characterizing daily life experience: The day reconstruction method. Science 306, 5702 (2004), 1776–1780. [PubMed: 15576620]

[39]. Klasnja Predrag, Consolvo Sunny, Choudhury Tanzeem, Beckwith Richard, and Hightower Jeffrey. 2009. Exploring privacy concerns about personal sensing. In International Conference on Pervasive Computing. Springer, 176–183.

[40]. Mobile Data 2 Knowledge. 2017. mCerebrum. https://github.com/MD2Korg/mCerebrum

[41]. Lavrakas Paul J. 2008. Encyclopedia of survey research methods. Sage Publications.

[42]. Little Roderick JA and Rubin Donald B. 2019. Statistical analysis with missing data. Vol. 793. Wiley.

[43]. Localytics. [n. d.]. 24% of Users Abandon an App After One Use. http://info.localytics.com/blog/24-of-users-abandon-an-app-after-one-use

[44]. Lu Hong, Yang Jun, Liu Zhigang, Lane Nicholas D, Choudhury Tanzeem, and Campbell Andrew T. 2010. The Jigsaw continuous sensing engine for mobile phone applications. In Proceedings of the 8th ACM Conference on Embedded Networked Sensor Systems. ACM, 71–84.

[45]. Lynn Peter. 2001. The impact of incentives on response rates to personal interview surveys: Role and perceptions of interviewers. International Journal of Public Opinion Research (2001).

[46]. McFarlane Daniel. 2002. Comparison of four primary methods for coordinating the interruption of people in human-computer interaction. Human-Computer Interaction 17, 1 (2002), 63–139.

[47]. Menictas Marianne, Rabbi Mashfiqui, Klasnja Predrag, and Murphy Susan. 2019. Artificial intelligence decision-making in mobile health. The Biochemist 41, 5 (2019), 20–24. [PubMed: 33828355]

[48]. Nahum-Shani Inbal, Smith Shawna N, Tewari Ambuj, Witkiewitz Katie, Collins Linda M, Spring Bonnie, and Murphy S. 2014. Just in time adaptive interventions (jitais): An organizing framework for ongoing health behavior support. Methodology Center technical report 14–126 (2014).

[49]. Niforatos Evangelos, Cinel Caterina, Mack Cathleen Cortis, Langheinrich Marc, and Ward Geoff. 2017. Can Less be More?: Contrasting Limited, Unlimited, and Automatic Picture Capture for

Augmenting Memory Recall. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies 1, 2 (2017), 21.

[50]. Oney Stephen, Harrison Chris, Ogan Amy, and Wiese Jason. 2013. ZoomBoard: a diminutive qwerty soft keyboard using iterative zooming for ultra-small devices. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM, 2799–2802.

[51]. Paruthi Gaurav, Raj Shriti, Gupta Ankita, Huang Chuan-Che, Chang Yung-Ju, and Newman Mark W. 2017. HEED: situated and distributed interactive devices for self-reporting. In Proceedings of the 2017 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2017 ACM International Symposium on Wearable Computers. ACM, 181–184.

[52]. Petty Richard E and Cacioppo J. 1986. Elaboration likelihood model. Handbook of theories of social psychology. London, England: Sage (1986).

[53]. Ponnada Aditya, Haynes Caitlin, Maniar Dharam, Manjourides Justin, and Intille Stephen. 2017. Microinteraction Ecological Momentary Assessment Response Rates: Effect of Microinteractions or the Smartwatch? Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies 1, 3 (2017), 92.

[54]. Rabbi Mashfiqui, Ali Shahid, Choudhury Tanzeem, and Berke Ethan. 2011. Passive and In-Situ assessment of mental and physical well-being using mobile sensors. In Proc. 13th ACM Intï£¡l Conf. Ubiquitous Computing. 385–394.

[55]. Rabbi Mashfiqui, Aung Min SH, Gay Geri, Reid M Cary, and Choudhury Tanzeem. 2018. Feasibility and Acceptability of Mobile Phone–Based Auto-Personalized Physical Activity Recommendations for Chronic Pain Self-Management: Pilot Study on Adults. Journal of medical Internet research 20, 10 (2018), e10147. [PubMed: 30368433]

[56]. Rabbi Mashfiqui, Meredith Philyaw Kotov Rebecca Cunningham, Bonar Erin E, Nahum-Shani Inbal, Klasnja Predrag, Walton Maureen, and Murphy Susan. 2018. Toward increasing engagement in substance use data collection: development of the Substance Abuse Research Assistant app and protocol for a microrandomized trial using adolescents and emerging adults. JMIR research protocols 7, 7 (2018), e166. [PubMed: 30021714]

[57]. Rabbi Mashfiqui, Pfammatter Angela, Zhang Mi, Spring Bonnie, and Choudhury Tanzeem. 2015. Automated Personalized Feedback for Physical Activity and Dietary Behavior Change With Mobile Phones: A Randomized Controlled Trial on Adults. JMIR mHealth uHealth 3, 2 (14 5 2015), e42. 10.2196/mhealth.4160 [PubMed: 25977197]

[58]. Rahman Tauhidur, Zhang Mi, Voida Stephen, and Choudhury Tanzeem. 2014. Towards accurate non-intrusive recollection of stress levels using mobile sensing and contextual recall. In Proceedings of the 8th International Conference on Pervasive Computing Technologies for Healthcare. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), 166–169.

[59]. Robinson Michael D and Clore Gerald L. 2002. Belief and feeling: evidence for an accessibility model of emotional self-report. Psychological bulletin 128, 6 (2002), 934. [PubMed: 12405138]

[60]. Sarker Hillol, Sharmin Moushumi, Ali Amin Ahsan, Rahman Md Mahbubur, Bari Rummana, Hossain Syed Monowar, and Kumar Santosh. 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. ACM, 909–920.

[61]. Shiffman Saul, Stone Arthur A, and Hufford Michael R. 2008. Ecological momentary assessment. Annu. Rev. Clin. Psychol 4 (2008), 1–32. [PubMed: 18509902]

[62]. Siek Katie A, Rogers Yvonne, and Connelly Kay H. 2005. Fat finger worries: how older and younger users physically interact with PDAs. In IFIP Conference on Human-Computer Interaction. Springer, 267–280.

[63]. Smith Steven M, Glenberg Arthur, and Bjork Robert A. 1978. Environmental context and human memory. Memory & Cognition 6, 4 (1978), 342–353.

[64]. Smith Steven M and Vela Edward. 2001. Environmental context-dependent memory: A review and meta-analysis. Psychonomic bulletin & review 8, 2 (2001), 203–220. [PubMed: 11495110]

[65]. Suh Hyewon, Shahriaree Nina, Hekler Eric B, and Kientz Julie A. 2016. Developing and Validating the User Burden Scale: A Tool for Assessing User Burden in Computing Systems. In

Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems. ACM, 3988–3999.

[66]. Pepe Margaret Sullivan and Anderson Garnet L. 1994. A cautionary note on inference for marginal regression models with longitudinal data and general correlated response data. Communications in statistics-simulation and computation 23, 4 (1994), 939–951.

[67]. Tangmunarunkit Hongsuda, Hsieh Cheng-Kang, Longstaff Brent, Nolen S, Jenkins John, Ketcham Cameron, Selsky Joshua, Alquaddoomi Faisal, George Dony, Kang Jinha, et al. 2015. Ohmage: A general and extensible end-to-end participatory sensing platform. ACM Transactions on Intelligent Systems and Technology (TIST) 6, 3 (2015), 38.

[68]. Thomaz Edison, Parnami Aman, Essa Irfan, and Abowd Gregory D. 2013. Feasibility of identifying eating moments from first-person images leveraging human computation. In Proceedings of the 4th International SenseCam & Pervasive Imaging Conference. ACM, 26–33.

[69]. Torous John, Nicholas Jennifer, Larsen Mark E, Firth Joseph, and Christensen Helen. 2018. Clinical review of user engagement with mental health smartphone apps: evidence, theory and improvements. Evidence-based mental health 21, 3 (2018), 116–119. [PubMed: 29871870]

[70]. Truong Khai N, Shihipar Thariq, and Wigdor Daniel J. 2014. Slide to X: unlocking the potential of smartphone unlocking. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM, 3635–3644.

[71]. Tsiatis Anastasios. 2007. Semiparametric theory and missing data. Springer Science & Business Media.

[72]. Tsiatis Anastasios A, Davidian Marie, Zhang Min, and Lu Xiaomin. 2008. Covariate adjustment for two-sample treatment comparisons in randomized clinical trials: a principled yet flexible approach. Statistics in medicine 27, 23 (2008), 4658–4677. [PubMed: 17960577]

[73]. Tulving Endel et al. 1972. Episodic and semantic memory. Organization of memory 1 (1972), 381–403.

[74]. Tulving Endel and Thomson Donald M. 1973. Encoding specificity and retrieval processes in episodic memory. Psychological review 80, 5 (1973), 352.

[75]. Turner Liam D, Allen Stuart M, and Whitaker Roger M. 2015. Interruptibility prediction for ubiquitous systems: conventions and new directions from a growing field. In Proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing. ACM, 801–812.

[76]. Van Berkel Niels, Ferreira Denzil, and Kostakos Vassilis. 2018. The experience sampling method on mobile devices. ACM Computing Surveys (CSUR) 50, 6 (2018), 93.

[77]. Van Berkel Niels, Goncalves Jorge, Hosio Simo, and Kostakos Vassilis. 2017. Gamification of mobile experience sampling improves data quality and quantity. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies 1, 3 (2017), 107.

[78]. Van Laerhoven Kristof, Kilian David, and Schiele Bernt. 2008. Using rhythm awareness in long-term activity recognition. In 2008 12th IEEE International Symposium on Wearable Computers. IEEE, 63–66.

[79]. Wagenaar Willem A. 1986. My memory: A study of autobiographical memory over six years. Cognitive psychology 18, 2 (1986), 225–252.

[80]. Wang Rui, Aung Min SH, Abdullah Saeed, Brian Rachel, Campbell Andrew T, Choudhury Tanzeem, Hauser Marta, Kane John, Merrill Michael, Scherer Emily A, et al. 2016. CrossCheck: toward passive sensing and detection of mental health changes in people with schizophrenia. In Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing. ACM, 886–897.

[81]. Wang Rui, Chen Fanglin, Chen Zhenyu, Li Tianxing, Harari Gabriella, Tignor Stefanie, Zhou Xia, Ben-Zeev Dror, and Campbell Andrew T. 2014. StudentLife: assessing mental health, academic performance and behavioral trends of college students using smartphones. In Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing. ACM, 3–14.

[82]. Werbach Kevin and Hunter Dan. 2012. For the win: How game thinking can revolutionize your business. Wharton Digital Press.

[83]. Wyatt Danny, Choudhury Tanzeem, and Bilmes Jeff A. 2008. Learning Hidden Curved Exponential Family Models to Infer Face-to-Face Interaction Networks from Situated Speech Data.. In AAAI. 732–738.

[84]. Xiao Robert, Laput Gierad, and Harrison Chris. 2014. Expanding the input expressivity of smartwatches with mechanical pan, twist, tilt and click. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM, 193–196.

[85]. Zhang Xiaoyi, Pina Laura R, and Fogarty James. 2016. Examining unlock journaling with diaries and reminders for in situ self-report in health and wellness. In Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems. ACM, 5658–5664.

CCS Concepts:

- **Human-centered computing** → **Empirical studies in ubiquitous and mobile computing**; *Ubiquitous and mobile computing systems and tools*;
- **Applied computing** → *Health informatics*.

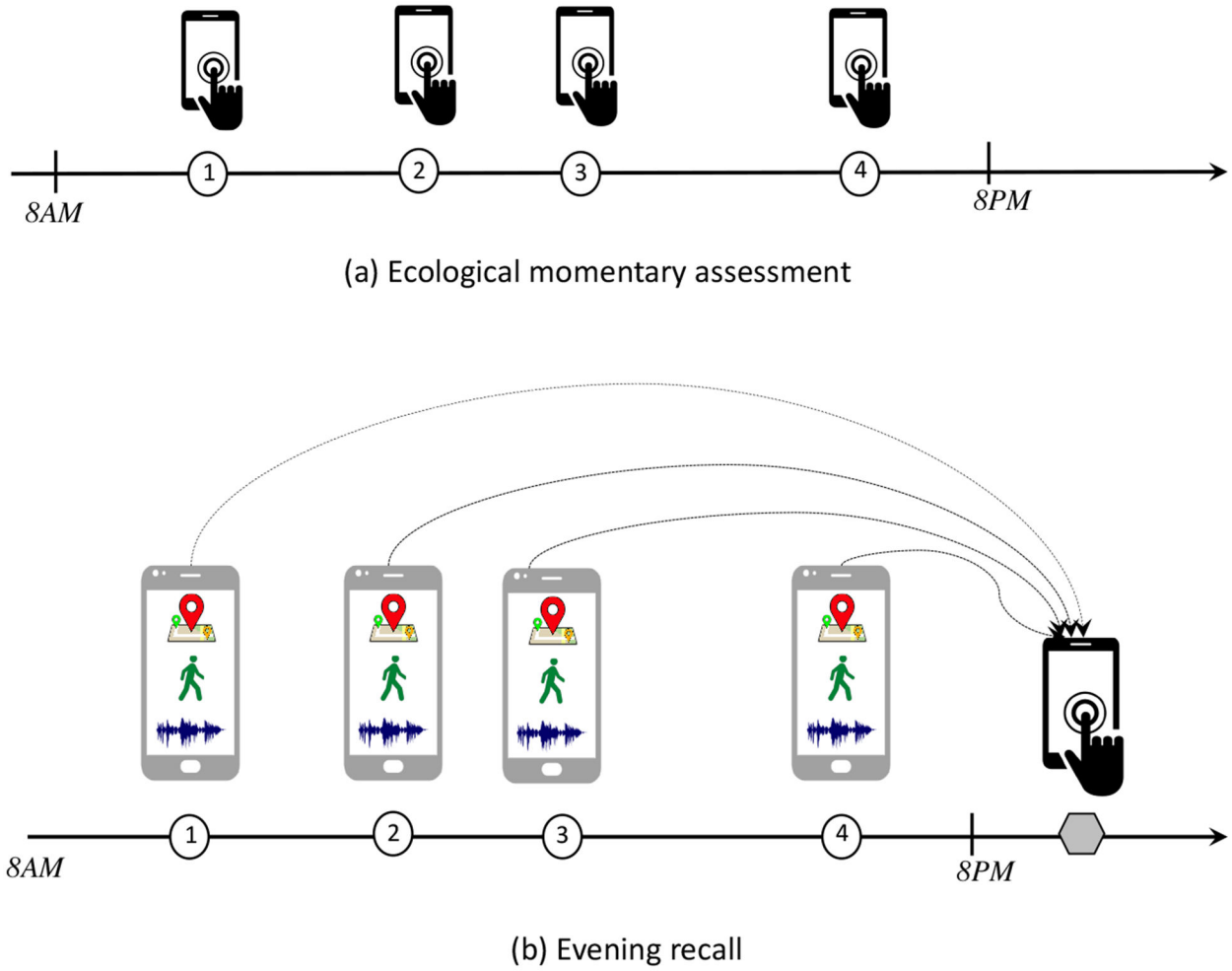

Fig. 1.

Conceptual diagram of EMA and evening recall. EMA interrupts multiple times a day for self-report. Evening recall captures contextual information passively through out the day and asks self-report in the evening.

Fig. 2.

Examples of recalling in the ReVibe application from the lead author's phone. (a) Answering recall question about at an earlier moment in the day (b) contextual information around the moment in 'a' (c) contextual information for another recall moment when a participant is commuting to work.

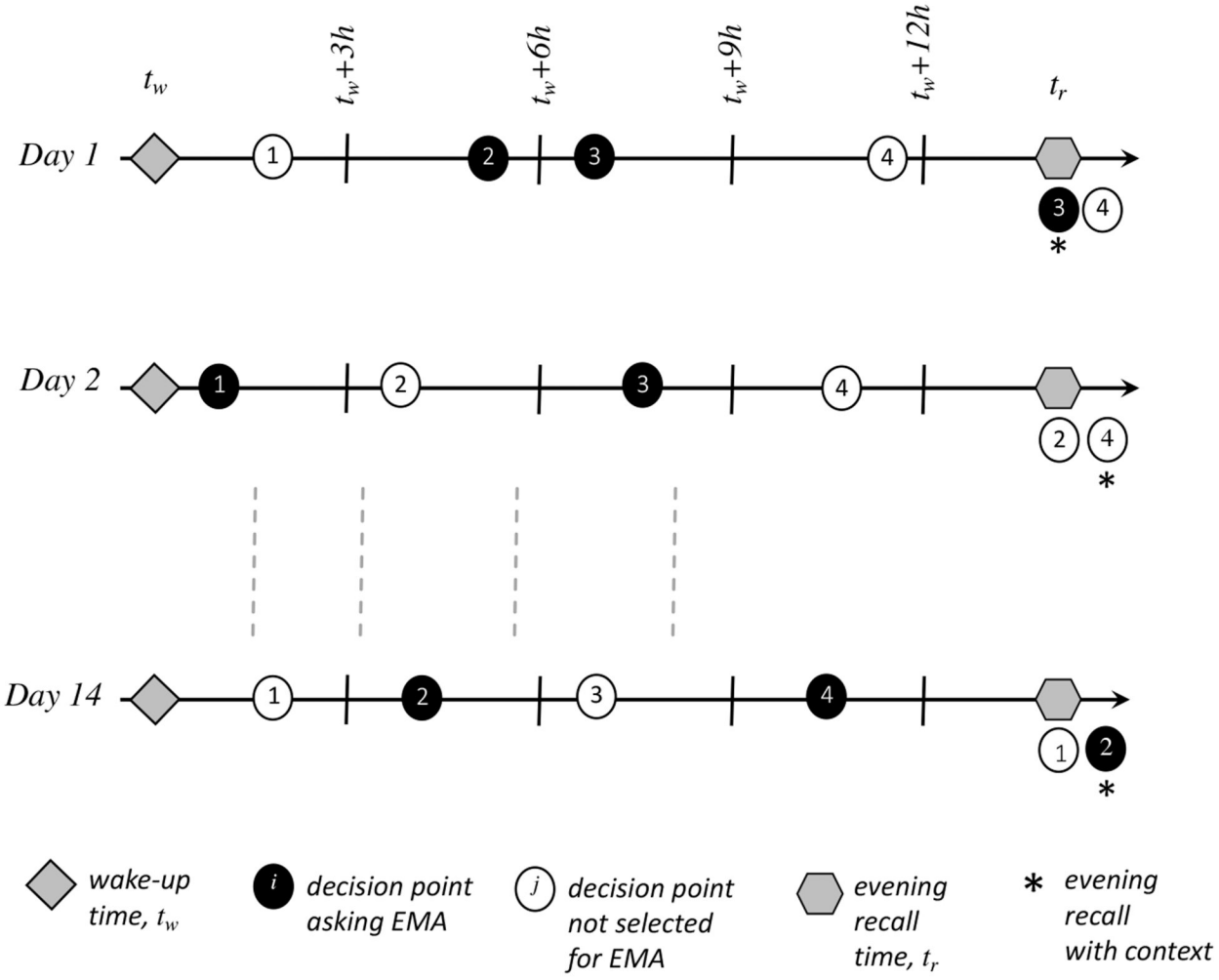

Fig. 3.

Example of randomizations in the ReVibe evaluation study for one user. For a different user, the randomizations of decision points, EMAs and evening recalls could be different.

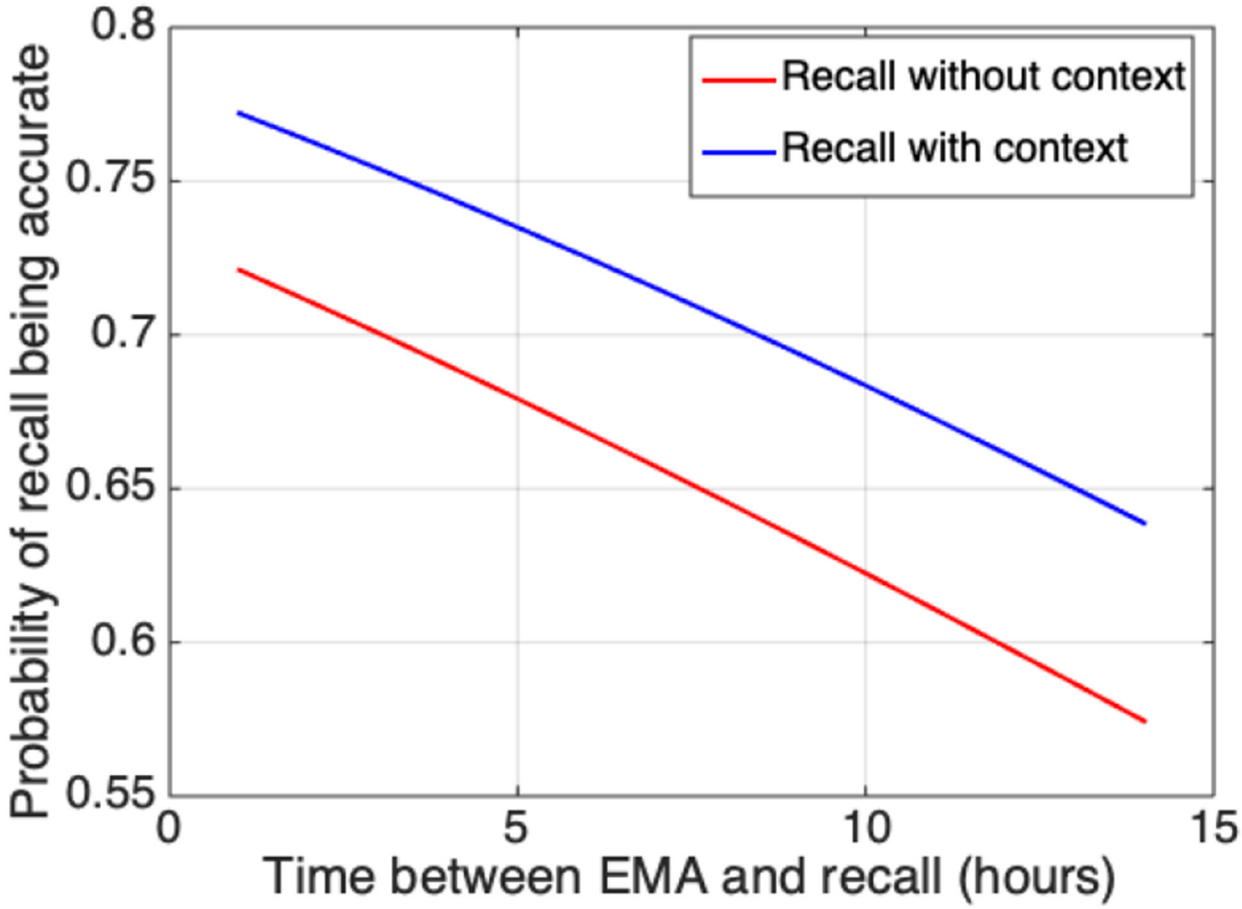

Fig. 4.
Fitted probability of recall accuracy for the model in Table 3

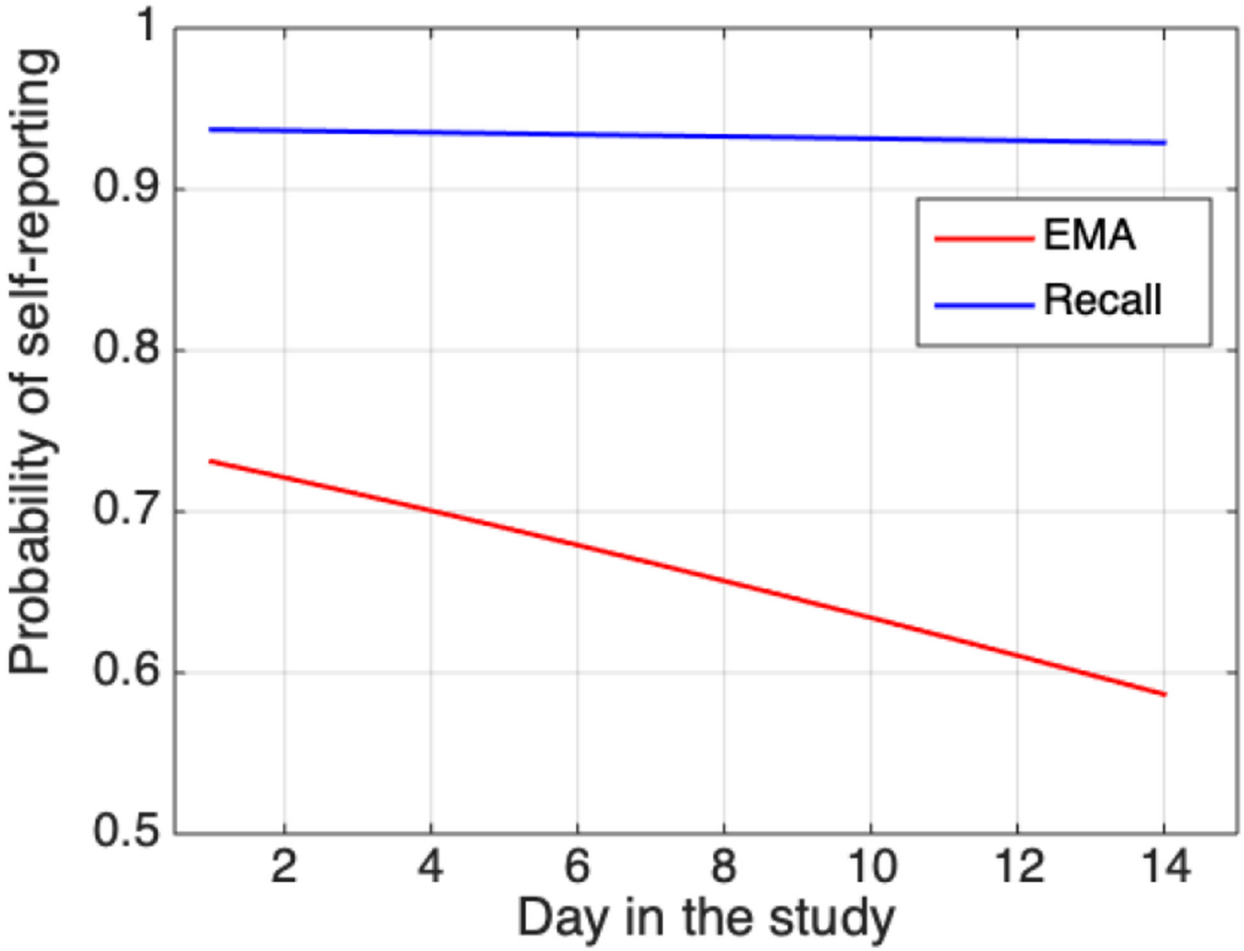

Fig. 5.
Fitted probability of self-reporting for the model in Table 4

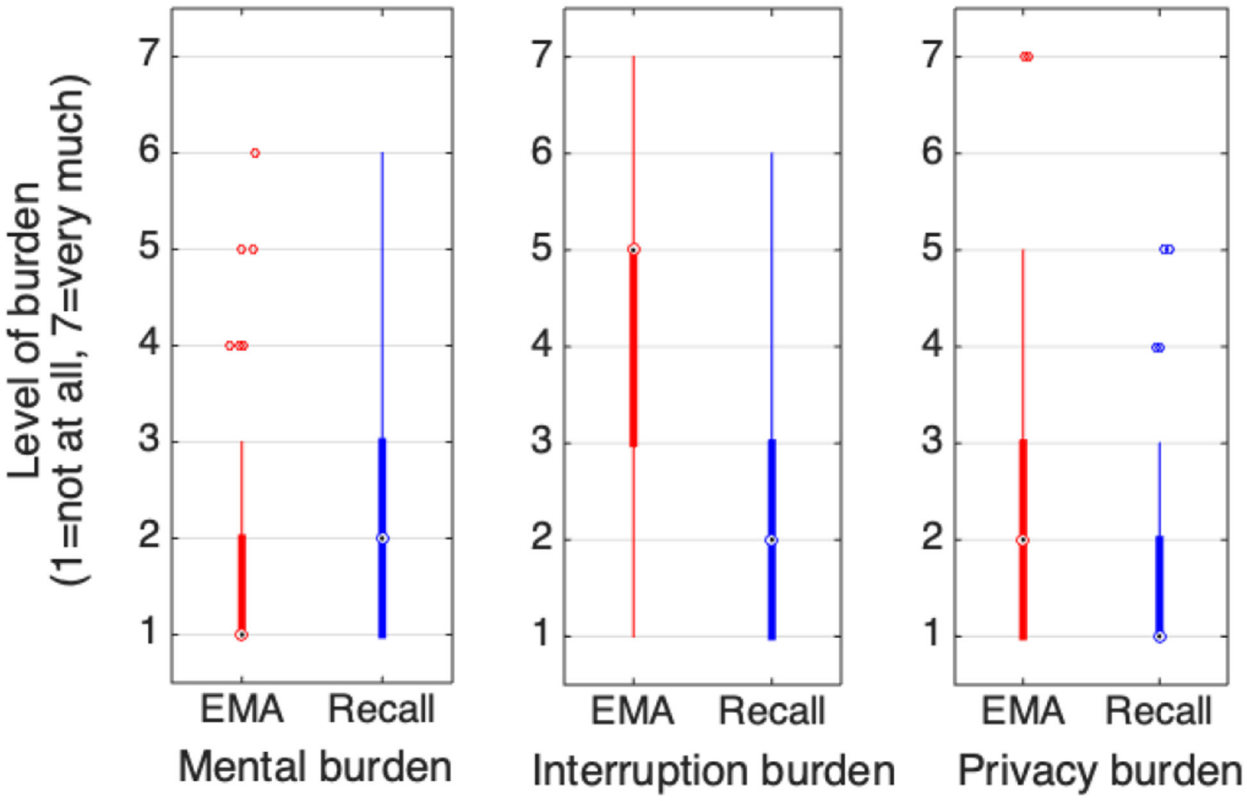

Fig. 6.

Box plots for self-reported user burden. Likert scale (1-Not at all, 7-very much). (left) mental effort (middle) interruption (right) privacy.

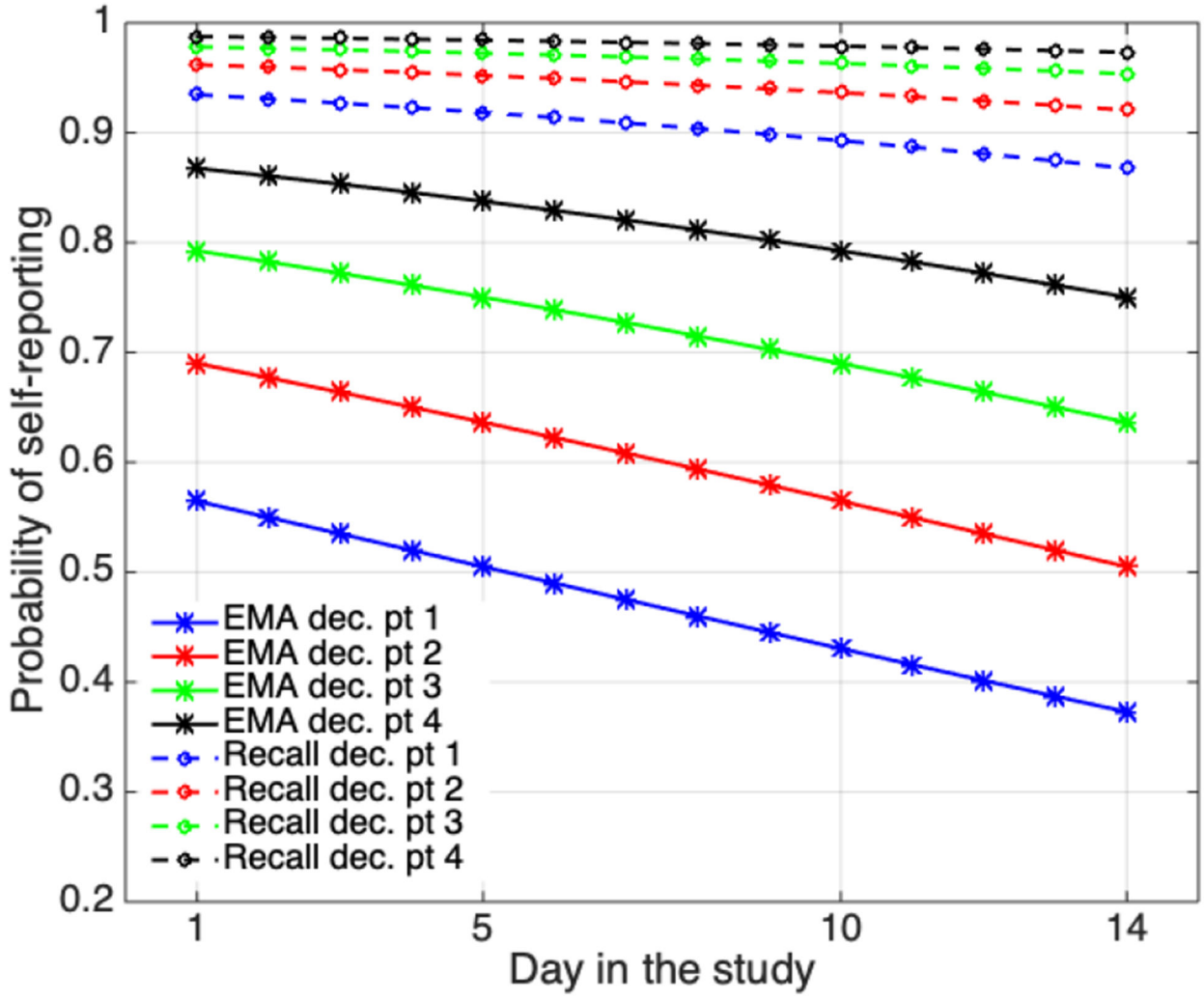

Fig. 7.

Fitted probability of self-report adherence for the model in Table 5. Dotted lines represent recalls, and dotted lines represent EMAs

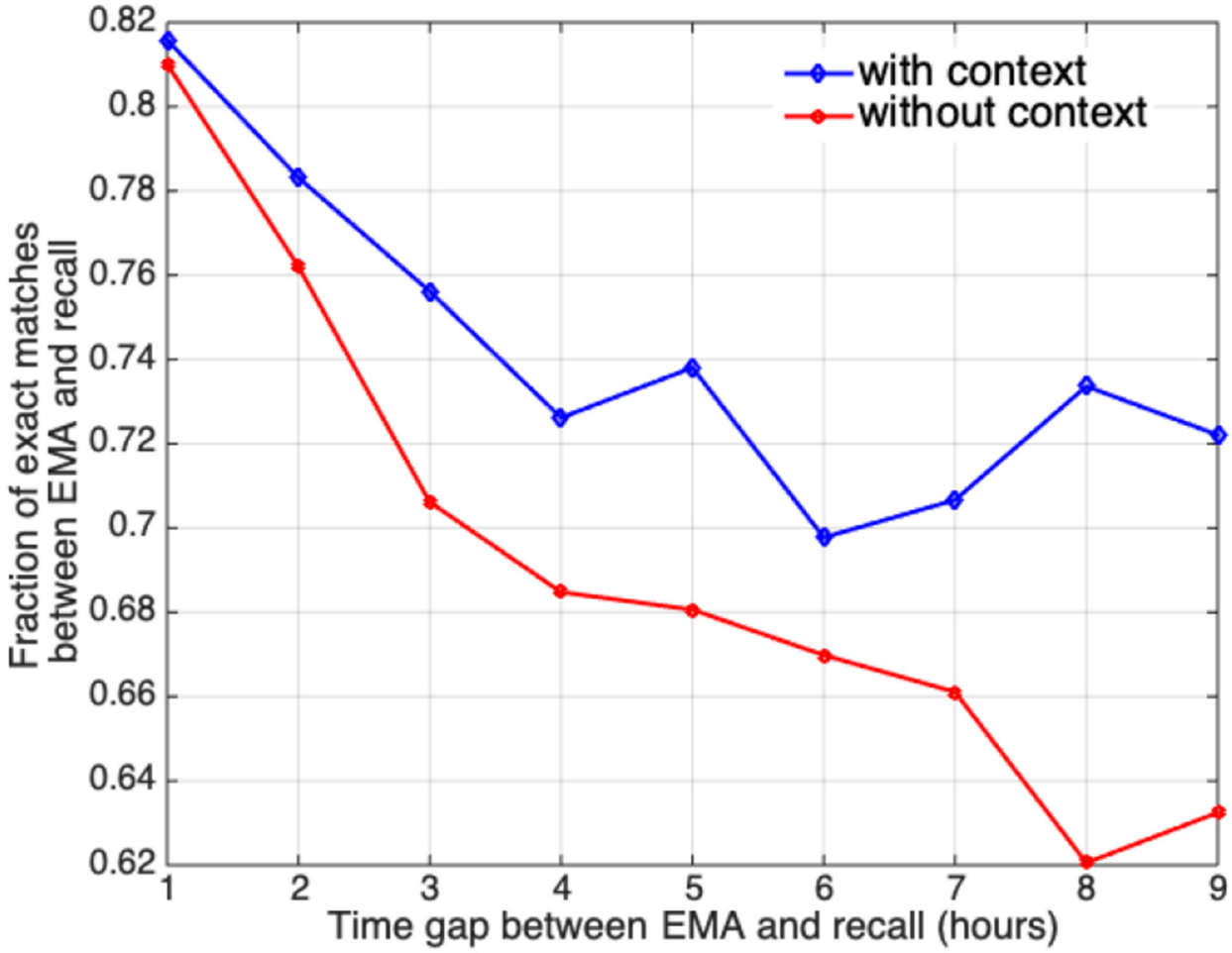

Fig. 8.

Visualization of fraction of exact matches between EMA and recall against time-gap between EMA and recall as measured in hours.

Table 1.

List of questions asked in EMA and Recall

Questions in the EMA and Recall
Q1. Rate your overall stress level at this moment
• *Likert scale. 1–5. Not at all to very much*
Q2. Rate your overall mood at this moment
• *Likert scale. 1–5. Sad to smile*
Q3. What are you doing now?
• *E.g., Watching TV, Doing chores, Studying, Going to school, Going to shopping*
Q4. Who are you with right now?
• *Freeform question*
Q5. Random selection of one of the 7 sensitive questions from Table 2
Q6. Is there something occupying your mind that you can't get out of your mind?
• *Yes or No*
Q7. How much difficulty are you having in getting it out of your mind? [This question will be asked if the answer to question 6 is yes]
• *Likert scale. 1–5. Not at all to very much*

Table 2.

Replacement questions for the sensitive question, which is 5th on the list, in Table 1

Sensitive questions for question 5 in Table 1
Q5.1. Did you take any medication in the last 4 hours that makes you feel good?
• Yes or No
Q5.2. Have you been critical of yourself in the last 4 hours?
• Yes or No
Q5.3. Rate how much you currently feel inadequate
• Likert scale. 1–5. Not at all to very much
Q5.4. Did you lie to someone important in the last 4 hours?
• Yes or No
Q5.5. Rate how strong any racially prejudiced thoughts were in the last 4 hours.
• Likert scale. 1–5. Not at all to very much
Q5.6. Did you lie unnecessarily in the last 4 hours?
• Yes or No
Q5.7. How strong are your sexual thoughts about one or more people currently around you?
• Likert scale. 1–5. Not at all to very much

Table 3.

Summary of results for **H1**analysis. Context variable is 1 if contextual information is given during recall and 0 else. Time gap is time difference between EMA and recall in hours.

Predictor	β	SE β	p	**marginal effect (%)**
Intercept	0.96	0.16	<0.001	-
Intervention:Context	0.27	0.12	0.032	5.6%
Day in study	0.01	0.02	0.56	0.2%
Time gap	−0.05	0.02	0.016	−1%

Table 4.

Summary of results for **H2**analysis. Self-report type variable is 0 if self-report is an EMA and 1 if self-report is an evening recall

Predictor	***β***	**SE *β***	***p***	**marginal effect (%)**
Intercept	1.00	0.15	<0.001	-
Self-report type	1.7	0.25	<0.001	23.6%
Day in study	−0.06	0.018	<0.001	−0.8%
Day in study×Self-report type	0.04	0.024	0.06	0.6%

Table 5.

Summary of results for decision point as a moderator for H2. Self-report type variable is 0 if self-report is an EMA and 1 if self-report is a recall. Decision point is coded as 0,1,2,3 for 1st, 2nd, 3rd, 4th decision point respectively

Predictor	β	SE β	p	marginal effect (%)
Intercept	0.26	0.17	0.12	-
Self-report type	2.4	0.3	<0.001	31.6%
Day in study	−0.06	0.019	<0.001	−0.86%
Decision point	0.54	0.059	<0.001	7.1%
Day in study×Self-report type	0.03	0.024	0.07	0.72%
Decision point×Self-report type	−0.53	0.11	< 0.001	−6.9%

Table 6.

Summary of results of how contextual information change recency effect of evening recall. Time-gap is number of hours between EMA and evening recall, Context variable is 1 when context is provided during evening recall and 0 otherwise. Day is study is 0 for day 1, and 13 for day 14 of the study.

Predictor	β	SE β	p	marginal effect (%)
Intercept	1.09	0.16	<0.001	-
Day in study	0.01	0.02	0.4	0.2%
Time gap	−0.06	0.02	0.004	−1.3%
Time gap×Context	0.03	0.01	0.09	0.6%

Table 7.

Various descriptive statistics of the absolute difference between EMA and evening recall responses for the same moment. Note for the sensitive question, we merged the results of similar question types because we did not capture which question we asked.

Question	Inter-vention	Absolute difference mean (sd)	Exact match (%)	Match within 1-point (%)
Stress (Q1)[a]	context	0.6 (0.9)	61.2%	**86.1**%
	no context	0.7 (1.1)	57.1%	84.5%
Mood (Q2)[a]	context	0.5 (0.9)	63.0%	**91.4**%
	no context	0.7 (1.2)	55.0%	89.0%
Sensitive[a]	context	0.2 (0.6)	83.0%	**93.6**%
(Q5.3,Q5.5,Q5.7)	no context	0.5 (1.0)	70.1%	91.5%
Sensitive[b]	context	-	**85.8**%	-
(Q5.1,Q5.2,Q5.4,Q5.6)	no context	-	85.7%	-
Mindfulness[b]	context	-	**87.7**%	-
(Q6)	no context	-	79.0%	-

[a]Likert scale question

[b]Binary yes/no question