Context-Informed Scheduling and Analysis: Improving Accuracy of Mobile Self-Reports

Niels van Berkel
Jorge Goncalves
Peter Koval
first.last@unimelb.edu.au
The University of Melbourne, Australia

Simo Hosio
first.last@oulu.fi
University of Oulu, Finland

Tilman Dingler
first.last@unimelb.edu.au
The University of Melbourne, Australia

Denzil Ferreira
first.last@oulu.fi
University of Oulu, Finland

Vassilis Kostakos
first.last@unimelb.edu.au
The University of Melbourne, Australia

ABSTRACT

Mobile self-reports are a popular technique to collect participant labelled data in the wild. While literature has focused on increasing participant compliance to self-report questionnaires, relatively little work has assessed response accuracy. In this paper, we investigate how participant context can affect response accuracy and help identify strategies to improve the accuracy of mobile self-report data. In a 3-week study we collect over 2,500 questionnaires containing both verifiable and non-verifiable questions. We find that response accuracy is higher for questionnaires that arrive when the phone is not in ongoing or very recent use. Furthermore, our results show that long completion times are an indicator of a lower accuracy. Using contextual mechanisms readily available on smartphones, we are able to explain up to 13% of the variance in participant accuracy. We offer actionable recommendations to assist researchers in their future deployments of mobile self-report studies.

CCS CONCEPTS

• **Human-centered computing** → *Empirical studies in HCI*; *Ubiquitous and mobile computing design and evaluation methods*;

KEYWORDS

Data quality, Ecological Momentary Assessment, EMA, Experience Sampling Method, ESM, smartphones, cognition, context, questionnaires, working memory.

ACM Reference Format:
Niels van Berkel, Jorge Goncalves, Peter Koval, Simo Hosio, Tilman Dingler, Denzil Ferreira, and Vassilis Kostakos. 2019. Context-Informed Scheduling and Analysis: Improving Accuracy of Mobile Self-Reports. In *CHI Conference on Human Factors in Computing Systems Proceedings (CHI 2019), May 4–9, 2019, Glasgow, Scotland UK*. ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3290605.3300281

1 INTRODUCTION

The Experience Sampling Method (ESM) [33], also known as Ecological Momentary Assessment [7, 50], is widely used to obtain *in situ* data on a variety of research topics, including affective state [47], activities [15], and technology usage [12, 43]. Participants in ESM studies answer a set of questions, *i.e.* self-reports, throughout the day, typically for 1-3 weeks [16]. As researchers rely on the input of study participants rather than direct observation, attaining a sufficient level of accuracy and frequency of completed self-reports is crucial. Previous literature has focused on answer frequency, for example by timing questionnaires based on participant context [45] or by increasing motivation [8, 24]. However, assessing the accuracy of ESM responses has received little attention in the literature, despite the considerable implications for study results. Researchers tend to assume that all ESM responses are accurate, without further validating this assumption. For this reason, self-reports of low accuracy can be detrimental in biasing or contaminating study data. With an increased uptake of the ESM in both HCI and the wider scientific community [7], enhancing the accuracy of participant-labelled data is increasingly important.

While it is well-known that cognitive performance differs across individuals [13], human cognition also varies from moment-to-moment. Previous work shows that a person's context can influence cognitive performance [48], including time of day [48, 56], mental state [5, 14], and smartphone usage [25, 31]. In addition, literature on designing ESM studies warns that the study's methodological configuration, *e.g.*, number of daily questionnaires [12], study duration [7, 51], can affect participant motivation. However, the effect of such (contextual) factors on the accuracy of self-reports has not been previously quantified.

CHI 2019, May 4–9, 2019, Glasgow, Scotland UK

ACM ISBN 978-1-4503-5970-2/19/05…$15.00
https://doi.org/10.1145/3290605.3300281

In this paper, we systematically investigate the contextual factors that influence response accuracy in ESM studies. We asked participants to answer objective and verifiable questions *in situ*. Our verifiable questions consisted of a working memory, a recall, and an arithmetic task. In addition, we asked participants to report on their affective state – a common question type in ESM studies [47]. Furthermore, we unobtrusively gathered data on participant interaction with their mobile devices. From these data, we are able to empirically identify the effect of mobile context on ESM response accuracy. Our results show that participant accuracy is higher when the number of recent phone interactions is limited and the phone is not in active use. We also show that accuracy degrades after two weeks of study duration, and that long question completion times are an indicator of reduced accuracy. Based on these results, we offer actionable recommendations that allow researchers to assess and improve the quality of ESM data.

2 RELATED WORK

The ESM was originally introduced by Csikszentmihalyi & Larson in the late 1970's [16]. In Experience Sampling, participants are actively reminded to complete a questionnaire - often multiple times throughout the day. Use of this methodology has seen an increased uptake in the HCI community as well as related disciplines, such as Psychology and Behavioural Medicine [7]. Response rate, defined as the number of completed questionnaires divided by number of presented questionnaires [7], has long been used as an important indicator of the completeness of the measured phenomenon. As such, a variety of previous work has explored ways to increase participant response rate [8, 12, 24, 45]. In this study, however, we aim to gain insights into the accuracy of participant responses in ESM studies.

Participant Accuracy in Experience Sampling

Klasnja et al. [30] studied participants' ability to accurately recall the duration of two physical routine activities (sitting and walking) across six different scheduling strategies (including variations on a randomised, time-based, and journal-based (*i.e.*, proactive self-reports) schedule). Recall error reduced as the number of daily questionnaires increased (and participants thus reported on shorter time spans). Klasnja et al. conclude that a fixed schedule (interval-contingent) leads to the highest accuracy in participant recall, and advise 5 to 8 notifications per day as an optimal balance between accuracy and annoyance [30].

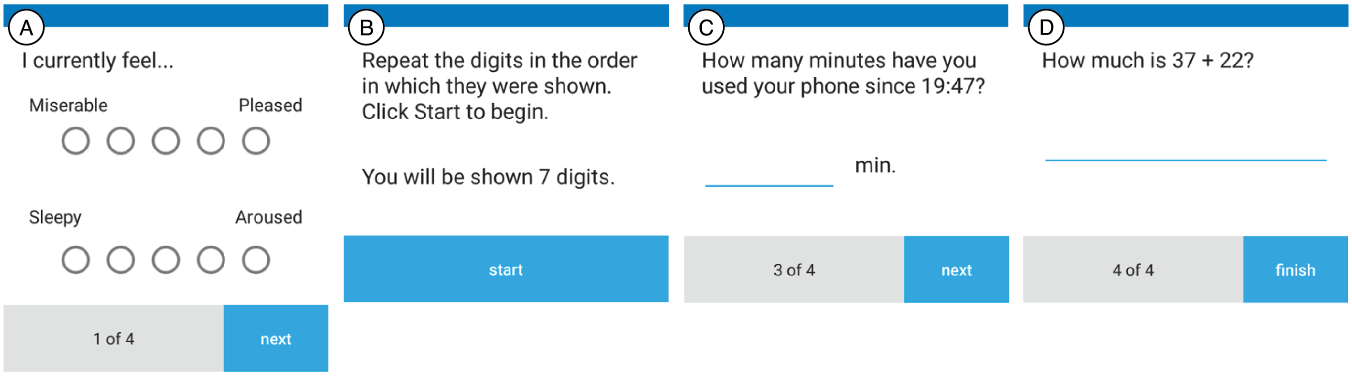
However, the use of fixed schedules has also been criticised. When using a fixed schedule, participants can anticipate the next questionnaire and potentially adjust their behaviour accordingly [49]. Furthermore, an interval-contingent configuration is likely to repeatedly assess the same event (*e.g.*, start of a lecture) [57]. Investigating an alternative input technique for ESM questionnaires, Truong et al. [52] present a mechanism to answer questions during device unlock. Participants answered, *inter alia*, math questions as a ground-truth task, with 81% of these questions answered correctly. However, the authors did not report on any inferences made on the effect of different contexts on participant accuracy.

A popular and straightforward technique to identify and subsequently remove low-accuracy responses is the removal of responses with suspiciously fast completion times, as for example suggested by McCabe et al. [36] (0.5 seconds) or Van Berkel et al. [7] (two standard deviations below the mean). However, these suggestions are arbitrary in nature and have not been empirically verified. Other studies have suggested the removal of participants with an overall low response rate, as the collected data paints an incomplete picture of the collected parameters. This is again, however, an arbitrary practice, with different cut-off values being used in different studies (*e.g.*, less than 50% of tasks completed [60]).

Working Memory

The ESM seeks to reduce reliance on human memory by asking participants to reflect on a short period of time [26]. By reflecting on current or recent events rather than on an entire day, ecological validity is retained and accuracy is less reliant on the participant's ability to accurately reconstruct past events [15, 26]. Although this effectively reduces reliance on participant's long term memory, working memory remains an important factor in answering ESM questionnaires. Working memory is described as "*a brain system that provides temporary storage and manipulation of the information necessary for such complex cognitive tasks as language comprehension, learning, and reasoning.*" [2]. As such, working memory is a crucial component in the completion of ongoing tasks: "*cognitive tasks can be completed only with sufficient ability to hold information as it is processed.*" [14].

Furthermore, working memory capacity has been shown to fluctuate within-persons at a range of timescales. At the micro timescale of seconds, working memory capacity may be reduced when experiencing acute stress [5]. At a meso timescale of hours, working memory capacity has been shown to vary according to circadian processes, with the highest performance at times of peak alertness (*i.e.*, morning for older adults and evening for younger adults [48, 56]). Finally, at the macro level of days, working memory performance gradually increases from Monday to Wednesday after which performance declines and stabilises from Thursday to Sunday [32]. Recent work utilising the ESM has seen an increase in task complexity [8, 12, 43] and requiring different cognitive abilities. Thus, in the current study we examine how moment-to-moment fluctuations in working memory relate to ESM response accuracy.



I currently feel...

Repeat the digits in the order in which they were shown. Click Start to begin.

You will be shown 7 digits.

How many minutes have you used your phone since 19:47?

How much is 37 + 22?

Figure 1: Questions as presented to participants. A) Affect. B) Working memory. C) Recall. D) Addition.

ESM Study Parameters

The maximum number of daily questionnaire notifications is defined as the study's 'inquiry limit'. Although a higher number of answered questionnaires leads to a more complete overview of the participant's day, a higher inquiry limit increases participant burden. The literature describes several ways to reduce participant burden. These include minimising the use of open-ended questions [12], setting a sensible inquiry limit [12], and reducing questionnaire items (*e.g.*, remove items based on current context [7]).

The scheduling of ESM triggers can be grouped into three categories: signal-, interval-, and event-contingent [57]. Each scheduling type can introduce a certain bias; time-based triggers favour frequently occurring contexts, whereas sensor-based triggers "*generate a different view of behaviour than more a complete sampling would provide*" [34]. Wheeler & Reiss recommend a signal-contingent schedule when the goal is to minimise recall bias [57]. Recent developments explore the use of more intelligent scheduling techniques. For example, Rosenthal et al. [45] establish a participant's level of interruptibility based on context. Iqbal & Bailey [27] demonstrate that optimising interruption mid-task reduces resumption lag. Finally, Ho & Intille [23] demonstrate that messages delivered when participants transition to a new activity are better received. In our study, we aim to provide new insights into the accuracy of collected ESM responses. We systematically investigate the effect of participant context and study design through a questionnaire containing verifiable questions. This study aims to support a more informed ESM scheduling in relation to the accuracy of self-reports.

3 METHODOLOGY

To systematically investigate how the accuracy of participants' answers varies as the result of their (mobile) context, we collected human input data through our questionnaire (detailed below) as well as a range of contextual sensor information on the phone. We conducted a study by collecting data via a custom-build smartphone application. Once installed on a participant's phone, the application continuously collected contextual data and presented questionnaires in accordance with the questionnaire schedule (detailed below). Participants were not informed about the correctness of their answers at any point throughout the study.

Questionnaire Items

We describe the individual questionnaire items in detail, and offer an overview in Table 1. Our question selection focuses on verifiable questions given the increased usage of the ESM in studies with a focus on non-intrapsychic participant observations (*i.e.*, participant answers can be verified or compared with one another [8, 43]) and the ability to measure participant accuracy as based on the available ground-truth data.

Affect self-assessment. The first question consists of two Likert-scales in which participants quantify their current affective state (Fig. 1-A). Participants were asked to report their current level of arousal and valence, following Russell's circumplex model of affect [46]. Self-assessment of affective state is widely used in the ESM [47]. As we are unable to assess the accuracy of participants' reported affect, we will consider affect as a contextual variable in our analysis of participant accuracy. The three remaining questions are presented in random order as to diminish order effects. We chose to present the self-assessment of current affect prior to the remaining questions to ensure that participants' affect is not modulated by their performance on the other questions.

Working memory - Digit span. The working memory task is based on a widely used task in experimental psychology, known as the forward digit-span test (Fig. 1-B). Following the display of a sequence of numerical digits, participants are asked to recall the order of these numbers in chronological order. Span tasks have been found to strongly correlate with human performance in problem-solving among other higher order thinking skills [20, 53]. As such, its relation to the ESM is found in studies in which responses are obtained through a combination of a participant's observation and cognitive

Table 1: Summary of questions included in study questionnaire.

	Non-verifiable: **Affect**	**Verifiable:** **Digit span**	**Phone usage**	**Arithmetic**
Construct	Affective state	Working memory	Recall ability	Participant effort
Motivation	Traditionally frequently used measurement construct in self-report studies.	Good representation of participant's ability to complete higher order cognitive tasks, accuracy fluctuates over time.	Recall of experiences is frequently used in ESM studies and has a long history, starting with initial ESM studies.	Use of explicitly verifiable questions has been used to filter fraudulent contributors and increase effort by participants.
Relevant studies	Wide array of studies on subjective well-being [47].	Studies featuring higher order cognitive tasks (*e.g.*, [8, 11, 43]).	Most ESM studies on time or frequency of activities (*e.g.*, [15, 59]).	Used in [52], extensive use of verifiable questions in crowdsourcing.

reasoning. For example, Reeder et al. [43] asked participants to reflect on browser warnings (observation) and their immediate response in browsing behaviour (cognitive reasoning). This type of ESM data collection has seen an increased popularity in our community (*e.g.*, [8, 11, 43]). In this question, each digit is presented for exactly 1 second (following [28]), after which the digit fades away in 0.4 seconds. Digits are randomly sampled without replacement and list length is constant at 7 digits. We chose a length of 7 as a challenging but not impossible, and thereby demoralising, number sequence [28, 58]. Furthermore, Miller [39] established 7 (±2) items as the upper boundary of working memory.

Recall - Phone usage. The questionnaire includes a recall task, asking participants to recall for how long they have used their phone since the time of last notification (Fig. 1-C). Recalled duration of activities has long been used in the ESM, starting with one of the first ESM studies by Csikszentmihalyi et al. [15], in which adolescents reported their daily activities and experience. Recall questions are still common in ESM studies (*e.g.*, [38, 59]). Previous literature shows that the duration of smartphone usage is diverse across users [21], and is challenging to estimate accurately [1]. As our application records the smartphone use of our participants in the background, we are able to compare participants' answers against ground truth. Distribution of complexity is achieved by varying the number of daily notifications (a high number of daily notifications results in a shorter recall period).

Arithmetic - Verifiable question. Finally, our questionnaire included an addition task in which the two numbers are randomly selected between 10 and 99 (Fig. 1-D). This is an explicitly verifiable question [29], a concept borrowed from the crowdsourcing literature [19, 29]. Explicitly verifiable questions have been shown to improve answer quality, as participants are aware their answers are verifiable and can thus be used to identify those providing fraudulent input [29].

Questionnaire Schedule

We collected data for a total of 21 consecutive days, thus including both weekdays and weekends. This is in line with both recent ESM studies [7] and the 2-4 weeks duration as recommended by Stone et al. [51]. Time of questionnaire presentation is randomised over the course of the day, with no questions being asked before 09:00 or after 21:00 to avoid unnecessary participant strain [7, 12]. Upon receiving the notification, participants had 20 minutes to open the notification before it disappears and is marked as 'expired'.

Literature on questionnaire frequency suggests varying practices (*e.g.*, 5-8 per day [30], 10 per day [12], or 8-12 per day [44]), as well as specifying frequency following questionnaire complexity or study duration [7, 51]. Given the lack of consensus on the ideal number of notifications, and the fact that the effect of questionnaire frequency on response accuracy has not been empirically studied, we opted for a varying number of 5-17 of daily notifications. We randomised the distribution of questionnaire across the day, imposing a minimum of 20 minutes between each individual questionnaire.

Statistical Model Construction

We first calculate 10 contextual predictors based on the collected smartphone sensor data, participant smartphone interaction, and other contextual variables. Selection of these variables is informed by previous literature on variance in cognitive performance. Predictors are measured at the time of answering the questionnaire unless otherwise stated. We describe these predictors in detail below:

- **Hour of day**: Following previous work which indicates an effect of time of day on cognitive skills, effect differs based on a person's age [48, 56].
- **Screen state**: Screen turned on or off *upon questionnaire arrival.* Previous work points to inattention during smartphone usage [25].
- **Network type**: 'Wi-Fi' or 'Mobile', proxy for mobility behaviour (*i.e.*, mobile on-the-go, increased likelihood of Wi-Fi when in a location for a prolonged period).

Previous work indicates that a person's environment significantly affects cognitive and affective state [42].

- **Recent notifications**. Number of notifications received *in the preceding 15 minutes*. An increase in smartphone notifications increases inattention [31].
- **Recent phone interactions**. Number of times phone was turned on *in the preceding 15 minutes*, following work on inattention during smartphone usage [25].
- **Study day**: Participant's day of study (1 to 21). Literature on ESM studies points to a decrease in both the number and quality of responses over time [51].
- **Completion time**: Time to complete per individual question. Previous work warns of very short completion times as an indicator of low quality [36].
- **Accuracy arithmetic**: Participant's accuracy in the arithmetic question, used as a verifiable question [29].
- **Arousal**: Self-report, 1 (sleepy) to 5 (aroused).
- **Valence**: Self-report, 1 (miserable) to 5 (happy). Arousal and valence have been shown to affect attention and decision making processes [35].

Recruitment and Procedure

Following ethics approval from our University's ethics committee, we recruited 25 participants using mailing lists of our University (13 female, 12 male, 18-52 years old, M = 26.8, SD = 7.4). Participants were required to have an Android-based smartphone and were recruited from a diverse range of departments (*e.g.*, Arts, Economics, Law) over a 2-week period (rolling enrolment). Our sample consisted of 7 undergraduates, 11 postgraduates, and 7 staff members.

We invited participants to our lab for individual intake sessions. During these sessions, we explained the research goal, obtained consent for the collection of mobile usage data, and went through all questionnaire items. At the end of the study, we conducted another individual debriefing session. During these sessions we carried out semi-structured exit interviews which were directly transcribed by the interviewer, focusing on the participants' self-perceived (changes in) accuracy. Participants received a $15 voucher for their efforts, irrespective of accuracy or number of answers provided.

4 RESULTS

Following a three-week data collection period, a total of 2,539 questionnaires were completed. Participants fully completed an average of 100.8 questionnaires (SD = 37.4). The overall response rate was 50.2%, with 4.2% of notifications actively dismissed by participants. The remaining 45.6% of notifications were automatically dismissed after 20 minutes. This relatively low response rate is most likely the result of a combination of a high number of notifications, a relatively short expiry time, and reimbursements not being linked to the number of completed responses. We now detail how we calculate response accuracy. Following this, we describe the construction of both general and personalised models. Finally, we analyse significant predictors in more detail.

We used the R package *lme4* [3] to perform a linear mixed effects analysis of the relationship between the aforementioned predictors and participant working memory and recall accuracy (*i.e.*, we compute two models). The use of generalised linear mixed-effect models allow us to identify the effect of a set of predictors on an outcome variable (accuracy) while following an arbitrary (*i.e.*, possibly non-normal) distribution. Further, mixed-effect models allow for the analysis of nested data (*i.e.*, questionnaires nested within participants) by modelling variability across upper-level units using random effects. We specify participant as a random effect as to allow for individual differences in our models. For each participant we also construct two individual models, using the same set of predictors but without participant as a random factor. For each individual model we calculate the $\mathbb{R}^2$ value and significance compared to the intercept. Through visual inspection of residual plots we did not find any noteworthy deviations from homoscedasticity or normality.

Response Accuracy

For each response, we calculated accuracy on a scale from 0 to 1. For the working memory task, we compared the participants' answer to the correct number set using the Damerau-Levenshtein distance metric [17]. This metric calculates the minimum edit distance between two sequences, using the operations of insertion, deletion, substitution of a single character, or transposition of two adjacent characters. For this task, the maximum Damerau-Levenshtein distance is 7, and the minimum (completely correct) is 0. We scale the Damerau-Levenshtein distance to the range of 0 to 1 (*i.e.*, a distance of 7 results in an accuracy of 0). Participants obtained an average accuracy of 94.8% (SD = 13.5%).

For the recall task, we calculate accuracy as; $Accuracy = 1 - (\frac{|correct\ answer - given\ answer|}{correct\ answer})$. Therefore, answers which are more than one magnitude away from the correct answer are assigned an accuracy value of zero, with the accuracy increasing as the answer approaches the correct value. Using this technique, we account for the differences in scale between the measured values. To avoid overly stringent evaluation on small numbers, we assigned a value of 1 minute to all participant answers and measured values which were below 3 minutes. Average accuracy was 44.7% (SD = 38.2%).

For the arithmetic task, the answer is either correct (1) or incorrect (0), with an average accuracy of 96.6% (SD = 18.2%).

Working Memory Task

We now report on model construction for the working memory task. Following model selection, a total of four variables remain (Table 2). We perform a likelihood ratio test with the null model [10] and find that our model is statistically significant ($\chi^2(4)$ = 196.15, p <0.001) and explains 13.2% of the variance in accuracy ($\mathbb{R}$ = 0.363, $\mathbb{R}^2$ = 0.132). This means that by using the independent variables listed in Table 2, we can explain 13.2% of fluctuation in participant accuracy for the working memory task. From these variables, completion time and study day have the largest effect on participant accuracy. Both variables affect accuracy in a negative direction; an increase in completion time or a later study day result in a lower accuracy. For the personalised models, the mean adjusted $\mathbb{R}^2$ value is 0.176 (values ranged from 0.008 to 0.607, ±0.167). The $\mathbb{R}^2$ values of the individual models are not significantly higher than the general model, as shown by a two-tailed one-sample t-test, $t(24) = 1.237$, p = 0.229.

Table 2: Effect of predictors on working memory task accuracy.

	Estimate	SE	t value	Pr(>|t|)
(Intercept)	1.193	0.020	61.09	$< 0.001^{***}$
Study day	−0.001	0.001	−2.66	0.008^{**}
Completion time	−0.014	0.001	−14.14	$< 0.001^{***}$
Screen state - On	−0.021	0.014	−1.47	0.156
Network - Wi-Fi	0.034	0.014	2.34	0.029^{*}

Recall Task

The final prediction model contained four predictors (Table 3). The model is statistically significant ($\chi^2(4)$ = 34.576, p < 0.001) and explained 7.7% of variance in recall accuracy ($\mathbb{R}$ = 0.227, $\mathbb{R}^2$ = 0.077). As such, we can explain 7.7% of the variation in participant accuracy for the recall task using the independent variables listed in Table 3. Similar to the working memory model, study day and completion time are the two variables with the largest effect on participant accuracy. For the personalised models, the mean adjusted $\mathbb{R}^2$ value is 0.060 (values ranged from 0.010 to 0.166, ±0.042). The $\mathbb{R}^2$ values of the individual models do not significantly differ from the general model, as shown by a two-tailed one-sample t-test, $t(24) = 1.742$, p = 0.097.

Table 3: Effect of predictors on recall task accuracy.

	Estimate	SE	t value	Pr(>|t|)
(Intercept)	0.578	0.030	19.37	$< 0.001^{***}$
Study day	−0.005	0.001	−4.42	$< 0.001^{***}$
Completion time	−0.007	0.002	−3.63	$< 0.001^{***}$
Screen state - On	−0.088	0.046	−1.91	0.069
Recent interac.	−0.007	0.004	−1.69	0.091

Feature Description

Following model construction, we present a more detailed look at the significant features.

Study day. For both working memory (small estimate) and recall accuracy, study day is a significant predictor. A later study day resulted in a decrease in recall accuracy. Table 4 lists the average accuracy per week for both tasks. While participant accuracy for the working memory remains stable across the three-week period, mean accuracy for the recall question drops considerably in the third week of study.

Table 4: Mean accuracy values per week.

	Working memory	Recall
Week 1	94.8%	46.9%
Week 2	94.3%	46.2%
Week 3	95.3%	39.7%

Completion time. Time taken to complete a question is a significant predictor in both general models. Completion times were 16.46 (±2.80, includes display of numbers) and 5.72 (±3.95) seconds for the working memory and recall questions respectively. Figure 2 shows the completion time for correct and incorrect questions, indicating shorter completion times for correct answers. In contrast with these results, previous literature has focused on short completion times as an indicator of low-quality, and suggest removing answers with a completion time below 0.5 seconds [36]. None of our questions were completed within 0.5 seconds, or even within 1 second. We investigated the effect of extremely short completion times and calculated the average accuracy for the quickest 5% of answers (avg. accuracy in brackets): working memory 98.7% (94.8%), recall 60.2% (44.7%). We then calculated the average accuracy per standard deviation from the mean completion time (Table 5). Both questions show a consistent decrease of accuracy as completion time increases.

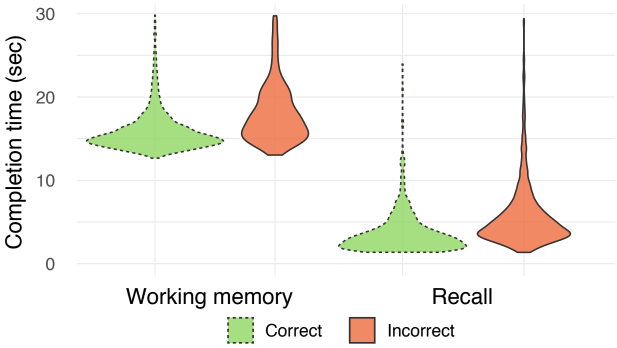


Figure 2: Completion time for correct and incorrect answers.

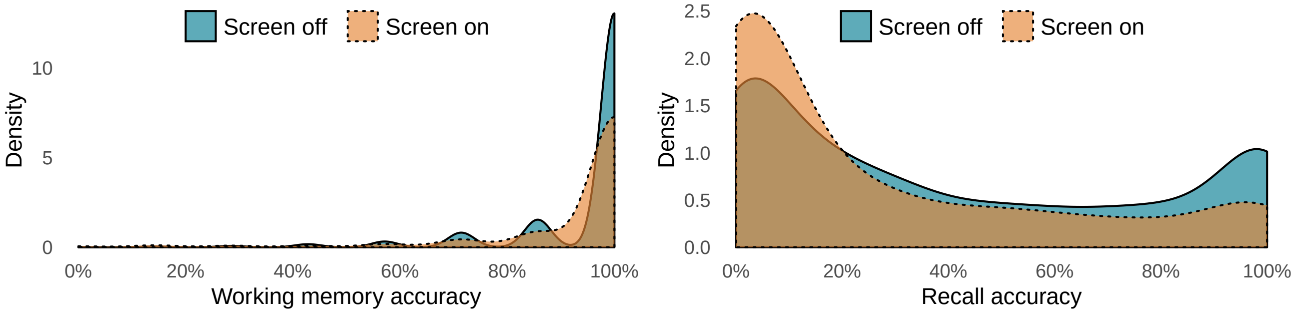


Figure 3: Density plots of the effect of screen state at time of questionnaire notification on accuracy.

Table 5: Question accuracy per standard deviation from the mean completion time.

	∞-2σ	2-1σ	1σ-μ	μ-1σ	1-2σ	2σ-∞
Working memory	98.4%	98.7%	97.3%	94.1%	88.2%	82.9%
Recall	62.4%	53.3%	46.5%	41.0%	39.5%	34.4%

Current phone usage. The variable 'screen state' is a binary variable which describes whether the screen is on or off upon receiving the questionnaire notification. For the majority of questions in our dataset (80.3%), the participants' screen was turned off at the time the questionnaire notification arrived. As shown in Figure 3, the highest accuracy values are obtained when the participant's screen is turned off at notification arrival (*ergo*, not in use).

Network connectivity. The state of the participant's connection (Wi-Fi or mobile) is a significant predictor in the model describing working memory accuracy. The mean differences indicate a limited effect; memory accuracy while connected to Wi-Fi is 96.4% (SE = 0.02), compared to an average accuracy of 94.6% (SE = 0.02) on mobile. Our results show that participants were connected to Wi-Fi in 21.8% of cases.

Recent phone interactions. We find that a larger number of phone interactions in the 15 minutes leading up to the questionnaire is negatively correlated with accuracy for the recall task. We plot the effect of this variable in Figure 4.

Thematic Coding

To develop a richer understanding of participant accuracy, we performed thematic coding on the transcripts of the exit interviews. Thematic coding is used to identify underlying concepts related to the accuracy of our participants. Each interview lasted for roughly 20 minutes and consisted of a set of predefined questions as well as the opportunity for participants to highlight elements which they considered important. We specifically asked our participants what affected their accuracy, including the effect of contextual factors, and whether their accuracy may have changed over time. Two of the authors individually completed two consecutive rounds of inductive coding. Following high level of agreement between the coders, three overarching themes were developed. These three themes highlight a range of effects participants associate with their accuracy.

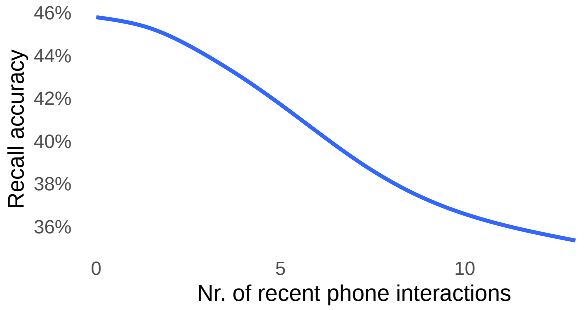


Figure 4: Effect of recent phone interactions (within the last 15 minutes) on recall accuracy.

Mental state. Fifteen participants noted that their current mental state had a considerable effect on the accuracy of their answers. Participants primarily noted tiredness and a low concentration level as examples of negative factors. Fatigue was often reported in combination with time of day (discussed below), but participants also reported experiencing tiredness following lunch or completing a specific task. Although the application did not show accuracy following questionnaire completion, participants seemed to be self-aware of their mistakes. Several participants recalled a scenario in which their performance suffered considerably: "*I remember I was in a grocery store and it [the questionnaire] came up and I thought 'really?!' and I did pretty bad.*" (P16).

Nine participants stated that their phone usage at the time of questionnaire affected their ability or inclination to answer accurately. "*Sometimes I was messaging someone and I'd get a notification. My mind is not focused, I would rush to complete it.*" (P8). For most participants, using the phone while a questionnaire arrives is associated with lower accuracy. Four participants specifically mentioned the completion of questionnaires during a phone call, which they experienced as challenging: "*If I am calling with a friend, and they ask a question while I am answering the question, it is difficult to answer two questions at once.*" (P3). For two participants, however, using the phone when the questionnaire arrived was thought to have a positive effect on accuracy: "*If I was already on my phone I was already 'in the zone'*" (P24).

Finally, three participants mentioned the negative effect of study fatigue. One participant noted that she not only answered fewer notifications but also answered less carefully: "*The questions started becoming a bit repetitive. Both a small drop in quality and quantity of responses*" (P17).

Distraction. A common thread throughout the interviews was the negative effect of distraction on participants' accuracy, as brought up by seventeen participants. Participants noted a wide variety of distractions, including listening to music, walking, direct or indirect interruption by colleagues, or their children. Three participants noted that the memory recall task is most susceptible to distraction. This is not surprising, as participants had to maintain focus on the screen as not to miss any of the numbers. "*My attention was pulled away, and even though that was just for a second, it was already enough*" (P17). In addition, it also required participants to focus and keep the previously displayed numbers in their memory: "*The recalling of seven numbers requires a long attention span*" (P16). Multitasking was also reported as a major culprit by several participants: "*When multitasking I took longer to complete and wanted to get rid of it - so the quality of my answers was probably lower*" (P6), and "*My accuracy was not so good when outside, walking and stuff. I was distracted and looking where I am going*" (P18).

A total of 17 participants commented on the effect of location. From these, seven participants specifically reported feeling less distracted at home, stating that they typically had to pay attention to fewer things. Our interviews also reveal that participants had more time available to answer questionnaires at home. "*When at home I did better, fewer distractions. Sometimes when I am outside I try to rush questionnaire*" (P11). One participant noted how the location affected his strategy in answering the number recall task: "*I read the number sequence out loud [...]. If I couldn't read out loud that caused some more difficulty.*" (P12). A few participants specifically stated that it was not their location but the task at hand which led to their distraction (*e.g.*, doing groceries, in transportation). Social interactions added an extra layer of complexity: "*When I was out with friends I thought it was rude to answer questions, but I did it anyway.*" (P3).

Time of day. Nineteen participants believed that time of day affected their accuracy in some way. Most common was the idea that performance slightly decreased in the early morning and towards the end of the day. For example: "*Late evening probably worse. By that point you're already switched off a little.*" (P12). Other participants simply stated that time of day did not affect their results at all or were more nuanced in the effect of time: "*At the end of the day or early in the morning I did worse, but not that much*" (P5).

5 DISCUSSION

Mehrotra et al. [37] state how future ESM studies should become more interruptibility-aware, with interruptions negatively affecting participants' cognitive state [37]. We extend this idea by quantifying not only participant response state (*i.e.*, provided an answer) but also the accuracy of these responses. Preserving data accuracy is a critical element of ESM studies, yet has received relatively little attention in the literature [6]. As a result, there is a lack of clear guidelines for researchers: "*Cleaning the data–arguably the most difficult component of ESM studies–also needs consideration. The literature lacks a clear discussion and standardization of how to detect errors in PDA data and how to prepare the data for analysis*" [36]. We discuss our findings regarding response accuracy and offer a summary of our recommendations for future ESM studies in Table 6.

Completion time. Our results show that a longer completion time resulted in lower accuracy. While this is to be expected for the working memory question (due the temporal nature of our cognitive system, retaining the order of a sequence becomes increasingly difficult), this effect is also apparent in the recall question. Our thematic coding revealed that long completion times may be an indicator of multitasking. Furthermore, a short completion time did not result in low accuracy but was, in fact, an indicator of high accuracy. This even applied to the 5% quickest completion times. Historically, ESM practitioners have focused on responses with a short completion time as an indicator of low-accuracy [36]. A recent review on the ESM notes that no well-supported cut-off values for completion time are established [7], and as such suggested researchers to determine outliers in their data. Based on our results, we recommend against the removal of answers based on short completion time alone, and instead urge researchers to take note of outliers with long completion times. We recommend to discard answers with completion times two standard deviations above the mean, representing 2.75% of the data. Different dynamics may come into play when answering a question using a different input technique (*e.g.*, multiple-choice, screen unlock input [52]), and these techniques are not considered in our study.

Study duration. Stone et al. [51] were one of the first to quantify recommendations for ESM researchers. In their recommendation on study duration, they warn of a decline in data quality commonly occurring between 2–4 weeks. An analysis from 110 recent ESM studies shows an average study duration of 32 days and median duration of 14 days [7]. While our accuracy results for the working memory task did not change over the study duration, the accuracy of the recall task dropped as the study progressed. We were surprised to find similar accuracy levels for week 1 and 2, but a considerable drop in accuracy in week 3 (Table 4). This is in line with earlier findings on quality of mood reports [4].

Verifiable questions. Kittur et al. [29] stress the importance of including explicitly verifiable questions. These questions can be used to "*signal to users that their answers will be scrutinized, which may play a role in both reducing invalid responses and increasing time-on-task*" [29]. The use of verifiable questions has not been explored extensively for the ESM. Truong et al. [52] include a mathematical question to *"provide a question with ground-truth to test participant effort*" [52]. However, they did not report whether or how this affected the accuracy of participants in accompanying questions. Our results indicate that the included explicitly verifiable question did not provide a significant insight into the accuracy of participants. We argue that this is likely caused by the difference in dynamics between researchers and participants in crowdsourcing and ESM-based studies. Crowdworkers typically work online and do not interact directly with the researcher, whereas ESM participants customarily meet face-to-face with the researcher – establishing some level of rapport [33]. Furthermore, despite the anonymisation of the collected data, participants in ESM studies are likely to feel more directly observed which may reduce impulsive nonsensical data contributions. As such, we observe limited use for including of verifiable questions in studies employing the ESM.

Cognition-Aware Contingency

Research on questionnaire scheduling has focused on improving response rates. This has led to various novel approaches such as considering participant sleeping schedule [12] or level of interruptibility [45, 55]. Alternative input techniques include screen unlock interactions [52] or alert dialogues [54] to further increase response rates and data quantity.

We argue that, while response rate is an important factor in Experience Sampling, the accuracy of answers is at least equally important. Our results show that, based on contextual information, we are able to more accurately assess the accuracy of participants (Tables 2 and 3). This builds on previous work which aimed to assess human accuracy through cognitive tasks [18], manual evaluation of combined self-reports and photographs [61], or majority voting techniques [8]. One commonality between these works is the need for human input. As systems aim to become increasingly cognition-aware, detecting changes in cognition without adding additional task load onto the user is a crucial step. In essence, using cognition-aware scheduling we aim to know whether a participant's answer will be accurate, and therefore valuable, prior to asking the question.

Our results show that answer accuracy is higher for questionnaires which arrive when the phone is not in active use. Participants were more likely to postpone answering the questionnaire until a more convenient time (within the notification expiration time) when they were not using their phone. We argue that a delayed response time does not negatively affect the validity of responses, as participants report on their current state rather than their state at the time of incoming notification. Similarly, we find that when participants frequently use their device in the immediate period leading up to a questionnaire they are less accurate in their responses (Figure 4). By scheduling questionnaires when participants are likely to provide more accurate answers (*e.g.*, when not using their smartphone), researchers can increase the reliability of their study results.

Contextual bias. Taking the above recommendations into account can considerably increase contextual bias. For example, sampling when the phone has not been recently used will limit results to non-phone activities, potentially excluding interactions which take place on the phone, or mental states such as boredom (often leading to smartphone usage [41]). Contextual dissonance is inherent to any sampling strategy [34]. However, this is not to say that no measures can, and should, be taken to reduce the introduced contextual bias. In the case of potential contextual bias, researchers should combine their cognition-aware configuration with one of the well-known time-based contingency configurations (*i.e.*, interval and signal based) [9]. This could, for example, result in a schedule optimising questionnaire arrival in sliding two-hour time-windows. Failing to identify a context in which a participant is more likely to answer accurately, the questionnaire will be presented following a time-based interval. This ensures questionnaires are spread across time and context, while still optimising accuracy.

Implications for Study Design

Researchers conducting ESM studies face multiple study design decisions. This includes five key methodological decisions: notification schedule, inter-notification time, study duration, notification expiry, and inquiry limit [7]. In this study, we focus primarily on the scheduling of notifications, but also quantify the effect of inter-notification time and study duration. We discuss recommendations for future ESM studies based on the results of our study and provide a summary of our recommendations in Table 6. We note that the models on the working memory and recall task explain respectively 13.2% and 7.7% of variance in participant accuracy, whereas the selected individual predictors are (highly) significant (Tables 2 and 3). This indicates that participant

Table 6: Recommendations for optimising participant accuracy in future ESM studies.

Recommendation	Motivation
Limit study duration to two weeks.	Recall accuracy drops after a two-week period. Previous work links this effect to study fatigue [51].
Prioritise sending questionnaires when the phone is not in active use.	Avoid participants rushing to complete the questionnaire to return to their smartphone activity.
Prioritise sending questionnaires when the phone is connected to a Wi-Fi network.	Participants are likely to be in a location with less distractions (*e.g.*, at home rather than in transport).
Do not allow participants to complete questionnaires during a phone call.	Participant likely distracted and mentally occupied when talking with another person.
Combine cognition-aware scheduling with a time-based interval schedule of questionnaires.	Reduce the chance of contextual bias; ensures questionnaires are spread over the day.
Remove responses with completion time two standard deviations above the mean.	The participant was likely distracted while completing the questionnaire.

accuracy varies considerably, resulting in relatively low $\mathbb{R}^2$ values. A relatively high level of unexplainable variability is however to be expected when studying human behaviour and accuracy, as reported by *e.g.* [22, 40].

Our results indicate that contextual smartphone usage affects accuracy. Participants who were actively using their smartphone during an incoming questionnaire notification were found to have a reduced accuracy. Our thematic coding reveals that participants were eager to return to their previous activity (*e.g.*, playing a game, chatting) and likely to experience the questionnaire as inconvenient. We therefore recommend researchers to schedule questionnaires during times in which the smartphone is not in active use. In addition, we find that participant accuracy is slightly increased when connected to a Wi-Fi network – a possible indicator that someone is stationary and/or in a familiar location.

Furthermore, we observe a considerable effect of study duration on participant recall accuracy. As shown in Table 4, recall accuracy dropped considerably in week 3. Based on these results, we advise researchers to limit the duration of studies collecting recall data to 2-weeks. A reduced study duration will likely have a negative effect on the total amount of data collected. This can, however, be compensated by increasing the number of daily questions presented.

Limitations

We discuss some of the limitations that may have affected the presented results. First, to reduce participant strain we limited the number of questions per questionnaire. The included questions were selected to represent a wide range of use cases and for their close resemblance to the skills required when answering ESM questionnaires. However, to assess all cognitive aspects would require a larger questionnaire. Similarly, as the ESM is applied across a wide variety of research domains some question types may not be well represented. Lastly, although we include a wide range of contextual variables it is almost certain that there are more variables at play which affect participant accuracy. As indicated by the thematic coding, location and mobility (*e.g.*, walking, sitting) are likely to be among these variables. The presented work is a first step towards future cognition-aware systems aimed at improving accuracy in self-report studies.

6 CONCLUSION

As researchers utilising self-report methods rely heavily on the answers provided by their participants, measuring and improving response accuracy is an important research avenue. Through a questionnaire assessing participants' performance across a set of questions (recall, working memory, arithmetic) we identify contextual and study-wide factors affecting participant accuracy. Our results show that a short survey completion time did not correlate with a low response accuracy in our questionnaire items. In fact, surveys that took a long time to complete were more likely to be inaccurate. Furthermore, we find that contextual usage factors such as phone usage at the time of questionnaire arrival and the number of recent phone interactions influence participant accuracy. These contextual factors can be used to optimise questionnaire scheduling for improved data accuracy. We offer actionable recommendations to assist researchers in their future deployments of self-report studies. Cognition-aware scheduling provides a novel scheduling technique, which can best be used in combination with traditional ESM scheduling techniques (time or event-based) to reduce contextual bias. With an increased shift of Experience Sampling to the observation of external and verifiable phenomena [8, 11], we expect an increased focus on ensuring accuracy of responses in Experience Sampling.

REFERENCES

[1] Sally Andrews, David A. Ellis, Heather Shaw, and Lukasz Piwek. 2015. Beyond Self-Report: Tools to Compare Estimated and Real-World Smartphone Use. *PLOS ONE* 10, 10 (2015), 1–9. https://doi.org/10.1371/journal.pone.0139004

[2] Alan Baddeley. 1992. Working memory. *Science* 255, 5044 (1992), 556–559.

[3] Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software* 67, 1 (2015), 1–48. https://doi.org/10.18637/jss.v067.i01

[4] Daniel J. Beal and Howard M. Weiss. 2003. Methods of Ecological Momentary Assessment in Organizational Research. *Organizational Research Methods* 6, 4 (2003), 440–464. https://doi.org/10.1177/1094428103257361

[5] S. L. Beilock and M. S. Decaro. 2007. From poor performance to success under stress: working memory, strategy selection, and mathematical problem solving under pressure. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 33, 6 (2007), 983–998.

[6] Niels van Berkel, Matthias Budde, Senuri Wijenayake, and Jorge Goncalves. 2018. Improving Accuracy in Mobile Human Contributions: An Overview. In *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers (UbiComp '18)*. ACM, New York, NY, USA, 594–599. https://doi.org/10.1145/3267305.3267541

[7] Niels van Berkel, Denzil Ferreira, and Vassilis Kostakos. 2017. The Experience Sampling Method on Mobile Devices. *Comput. Surveys* 50, 6, Article 93 (2017), 40 pages. https://doi.org/10.1145/3123988

[8] Niels van Berkel, Jorge Goncalves, Simo Hosio, and Vassilis Kostakos. 2017. Gamification of Mobile Experience Sampling Improves Data Quality and Quantity. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 3, Article 107 (2017), 21 pages. https://doi.org/10.1145/3130972

[9] Niels van Berkel, Jorge Goncalves, Lauri Lovén, Denzil Ferreira, Simo Hosio, and Vassilis Kostakos. 2019. Effect of Experience Sampling Schedules on Response Rate and Recall Accuracy of Objective Self-Reports. *International Journal of Human-Computer Studies* (2019). https://doi.org/10.1016/j.ijhcs.2018.12.002

[10] Benjamin M. Bolker, Mollie E. Brooks, Connie J. Clark, Shane W. Geange, John R. Poulsen, M. Henry H. Stevens, and Jada-Simone S. White. 2009. Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in Ecology & Evolution* 24, 3 (2009), 127–135. https://doi.org/10.1016/j.tree.2008.10.008

[11] Catherine E. Connelly, David Zweig, Jane Webster, and John P. Trougakos. 2011. Knowledge hiding in organizations. *Journal of Organizational Behavior* 33, 1 (2011), 64–88. https://doi.org/10.1002/job.737

[12] S. Consolvo and M. Walker. 2003. Using the experience sampling method to evaluate ubicomp applications. *IEEE Pervasive Computing* 2, 2 (2003), 24–31. https://doi.org/10.1109/MPRV.2003.1203750

[13] N. Cowan. 2005. *Working Memory Capacity.* Psychology Press.

[14] Nelson Cowan. 2010. The Magical Mystery Four: How Is Working Memory Capacity Limited, and Why? *Current Directions in Psychological Science* 19, 1 (2010), 51–57. https://doi.org/10.1177/0963721409359277 PMID: 20445769.

[15] M. Csikszentmihalyi, R. Larson, and S. Prescott. 1977. The ecology of adolescent activity and experience. *Journal of Youth and Adolescence* 6, 3 (1977), 281–294.

[16] Mihaly Csikszentmihalyi and Reed Larson. 2014. *Validity and Reliability of the Experience-Sampling Method.* Springer Netherlands, Dordrecht, 35–54. https://doi.org/10.1007/978-94-017-9088-8_3

[17] Fred J. Damerau. 1964. A Technique for Computer Detection and Correction of Spelling Errors. *Commun. ACM* 7, 3 (1964), 171–176. https://doi.org/10.1145/363958.363994

[18] Tilman Dingler, Albrecht Schmidt, and Tonja Machulla. 2017. Building Cognition-Aware Systems: A Mobile Toolkit for Extracting Time-of-Day Fluctuations of Cognitive Performance. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 3, Article 47 (2017), 15 pages. https://doi.org/10.1145/3132025

[19] Carsten Eickhoff and Arjen P. de Vries. 2013. Increasing cheat robustness of crowdsourcing tasks. *Information Retrieval* 16, 2 (2013), 121–137. https://doi.org/10.1007/s10791-011-9181-9

[20] R. W. Engle, S. W. Tuholski, J. E. Laughlin, and A. R. Conway. 1999. Working memory, short-term memory, and general fluid intelligence: a latent-variable approach. *Journal of Experimental Psychology: General* 128, 3 (1999), 309–331.

[21] Hossein Falaki, Ratul Mahajan, Srikanth Kandula, Dimitrios Lymberopoulos, Ramesh Govindan, and Deborah Estrin. 2010. Diversity in Smartphone Usage. In *Proceedings of the 8th International Conference on Mobile Systems, Applications, and Services (MobiSys '10)*. ACM, New York, NY, USA, 179–194. https://doi.org/10.1145/1814433.1814453

[22] R. Frank Falk and Nancy B Miller. 1992. *A primer for soft modeling*. University of Akron Press.

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

[24] Gary Hsieh, Ian Li, Anind Dey, Jodi Forlizzi, and Scott E. Hudson. 2008. Using Visualizations to Increase Compliance in Experience Sampling. In *Proceedings of the 10th International Conference on Ubiquitous Computing (UbiComp '08)*. ACM, New York, NY, USA, 164–167. https://doi.org/10.1145/1409635.1409657

[25] Ira E. Hyman, S. Matthew Boss, Breanne M. Wise, Kira E. McKenzie, and Jenna M. Caggiano. 2009. Did you see the unicycling clown? Inattentional blindness while walking and talking on a cell phone. *Applied Cognitive Psychology* 24, 5 (2009), 597–607. https://doi.org/10.1002/acp.1638

[26] M Iida, P. E. Shrout, J.-P Laurenceau, and Niall Bolger. 2012. Using diary methods in psychological research. (2012), 277–305.

[27] Shamsi T. Iqbal and Brian P. Bailey. 2005. Investigating the Effectiveness of Mental Workload As a Predictor of Opportune Moments for Interruption. In *CHI '05 Extended Abstracts on Human Factors in Computing Systems (CHI EA '05)*. ACM, New York, NY, USA, 1489–1492. https://doi.org/10.1145/1056808.1056948

[28] Chang-Jae Kim, Sang-hyun Hong, Byung-Sam Kim, Joon-Pyo Cheon, Yoonki Lee, Hyun-Jung Koh, and Jaemin Lee. 2008. Comparison of various tests designed to assess the recovery of cognitive and psychomotor function after ambulatory anesthesia. *Korean Journal of Anesthesiology* 55, 3 (2008), 291–297.

[29] Aniket Kittur, Ed H. Chi, and Bongwon Suh. 2008. Crowdsourcing User Studies with Mechanical Turk. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '08)*. ACM, New York, NY, USA, 453–456. https://doi.org/10.1145/1357054.1357127

[30] Predrag Klasnja, Beverly L. Harrison, Louis LeGrand, Anthony LaMarca, Jon Froehlich, and Scott E. Hudson. 2008. Using Wearable Sensors and Real Time Inference to Understand Human Recall of Routine Activities. In *Proceedings of the 10th International Conference on Ubiquitous Computing (UbiComp '08)*. ACM, New York, NY, USA, 154–163. https://doi.org/10.1145/1409635.1409656

[31] Kostadin Kushlev, Jason Proulx, and Elizabeth W. Dunn. 2016. "Silence Your Phones": Smartphone Notifications Increase Inattention and Hyperactivity Symptoms. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM, New York, NY, USA, 1011–1020. https://doi.org/10.1145/2858036.2858359

[32] Donald A. Laird. 1925. Relative Performance of College Students as Conditioned by Time of Day and Day of Week. *Journal of Experimental*

Psychology 8, 1 (1925), 50.

[33] Reed Larson and Mihaly Csikszentmihalyi. 2014. *The Experience Sampling Method*. Springer Netherlands, Dordrecht, 21–34. https://doi.org/10.1007/978-94-017-9088-8_2

[34] Neal Lathia, Kiran K. Rachuri, Cecilia Mascolo, and Peter J. Rentfrow. 2013. Contextual Dissonance: Design Bias in Sensor-based Experience Sampling Methods. In *Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp ‘13)*. ACM, New York, NY, USA, 183–192. https://doi.org/10.1145/2493432.2493452

[35] V. R. LeBlanc, M. M. McConnell, and S. D. Monteiro. 2015. Predictable chaos: a review of the effects of emotions on attention, memory and decision making. *Advances in Health Sciences Education. Theory and Practice* 20, 1 (2015), 265–282.

[36] K. O. McCabe, L. Mack, and W. Fleeson. 2012. *A guide for data cleaning in experience sampling studies*. Guilford Press, New York, NY, US, 321–338.

[37] Abhinav Mehrotra, Jo Vermeulen, Veljko Pejovic, and Mirco Musolesi. 2015. Ask, but Don't Interrupt: The Case for Interruptibility-aware Mobile Experience Sampling. In *Adjunct Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2015 ACM International Symposium on Wearable Computers (UbiComp/ISWC'15 Adjunct)*. ACM, New York, NY, USA, 723–732. https://doi.org/10.1145/2800835.2804397

[38] M. R. U. Meyer, C. Wu, and S. M. Walsh. 2016. Theoretical Antecedents of Standing at Work: An Experience Sampling Approach Using the Theory of Planned Behavior. *AIMS Public Health* 3, 4 (2016), 682–701.

[39] George A Miller. 1956. The Magical Number Seven, Plus or Minus Two: Some limits on our capacity for processing information. *Psychological review* 63, 2 (1956), 81.

[40] Minitab. 2014. How to Interpret a Regression Model with Low R-squared and Low P values. https://bit.ly/2otiSw5

[41] Martin Pielot, Tilman Dingler, Jose San Pedro, and Nuria Oliver. 2015. When Attention is Not Scarce - Detecting Boredom from Mobile Phone Usage. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp ‘15)*. ACM, New York, NY, USA, 825–836. https://doi.org/10.1145/2750858.2804252

[42] Suzanne Prescott and Mihaly Csikszentmihalyi. 1981. Environmental effects on cognitive and affective states: The experiential time sampling approach. *Social Behavior and Personality: an international journal* 9, 1 (1981), 23–32. https://doi.org/10.2224/sbp.1981.9.1.23

[43] Robert W. Reeder, Adrienne Porter Felt, Sunny Consolvo, Nathan Malkin, Christopher Thompson, and Serge Egelman. 2018. An Experience Sampling Study of User Reactions to Browser Warnings in the Field. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI ‘18)*. ACM, New York, NY, USA, Article 512, 13 pages. https://doi.org/10.1145/3173574.3174086

[44] Harry T. Reis and Shelly L. Gable. 2000. Event-sampling and other methods for studying everyday experience. *Handbook of Research Methods in Social and Personality Psychology* (2000), 190–222.

[45] Stephanie Rosenthal, Anind K. Dey, and Manuela Veloso. 2011. Using Decision-Theoretic Experience Sampling to Build Personalized Mobile Phone Interruption Models. In *Pervasive Computing*, Kent Lyons, Jeffrey Hightower, and Elaine M. Huang (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 170–187.

[46] James A. Russell. 1980. A circumplex model of affect. *Journal of Personality and Social Psychology* 39, 6 (1980), 1161.

[47] Ulrich Schimmack. 2003. Affect Measurement in Experience Sampling Research. *Journal of Happiness Studies* 4, 1 (2003), 79–106. https://doi.org/10.1023/A:1023661322862

[48] Christina Schmidt, Fabienne Collette, Christian Cajochen, and Philippe Peigneux. 2007. A time to think: Circadian rhythms in human cognition. *Cognitive Neuropsychology* 24, 7 (2007), 755–789. https://doi.org/10.1080/02643290701754158 PMID: 18066734.

[49] Christie Napa Scollon, Chu-Kim Prieto, and Ed Diener. 2009. *Experience Sampling: Promises and Pitfalls, Strength and Weaknesses*. Springer Netherlands, Dordrecht, 157–180. https://doi.org/10.1007/978-90-481-2354-4_8

[50] S. Shiffman, A. A. Stone, and M. R. Hufford. 2008. Ecological Momentary Assessment. *Annual Review of Clinical Psychology* 4 (2008), 1–32.

[51] A. A. Stone, R. C. Kessler, and J. A. Haythornthwaite. 1991. Measuring daily events and experiences: decisions for the researcher. *Journal of Personality* 59, 3 (1991), 575–607.

[52] Khai N. Truong, Thariq Shihipar, and Daniel J. Wigdor. 2014. Slide to X: Unlocking the Potential of Smartphone Unlocking. In *Proceedings of the 32Nd Annual ACM Conference on Human Factors in Computing Systems (CHI ‘14)*. ACM, New York, NY, USA, 3635–3644. https://doi.org/10.1145/2556288.2557044

[53] Nash Unsworth, Richard P. Heitz, Josef C. Schrock, and Randall W. Engle. 2005. An automated version of the operation span task. *Behavior Research Methods* 37, 3 (2005), 498–505. https://doi.org/10.3758/BF03192720

[54] Aku Visuri, Niels van Berkel, Chu Luo, Jorge Goncalves, Denzil Ferreira, and Vassilis Kostakos. 2017. Challenges of quantified-self: encouraging self-reported data logging during recurrent smartphone usage. In *Proceedings of the 31st British Computer Society Human Computer Interaction Conference.*

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

[56] R. West, K. J. Murphy, M. L. Armilio, F. I. Craik, and D. T. Stuss. 2002. Effects of time of day on age differences in working memory. *Journal of Gerontology* 57, 1 (2002), 3–10.

[57] Ladd Wheeler and Harry T. Reis. 1991. Self-Recording of Everyday Life Events: Origins, Types, and Uses. *Journal of Personality* 59, 3 (1991), 339–354. https://doi.org/10.1111/j.1467-6494.1991.tb00252.x

[58] David L. Woods, Mark M. Kishiyama, E. William Yund, Timothy J. Herron, Ben Edwards, Oren Poliva, Robert F. Hink, and Bruce Reed. 2011. Improving digit span assessment of short-term verbal memory. *Journal of Clinical and Experimental Neuropsychology* 33, 1 (2011), 101–111. https://doi.org/10.1080/13803395.2010.493149

[59] J. C. Cassandra Wright, M. Paul Dietze, A. Paul Agius, Emmanuel Kuntsche, Robin Room, Michael Livingston, Margaret Hellard, and S. C. Megan Lim. 2017. An Ecological Momentary Intervention to Reduce Alcohol Consumption in Young Adults Delivered During Drinking Events: Protocol for a Pilot Randomized Controlled Trial. *JMIR Research Protocols* 6, 5 (2017), e95. https://doi.org/10.2196/resprot.6760

[60] Yulong Yang, Gradeigh D. Clark, Janne Lindqvist, and Antti Oulasvirta. 2016. Free-Form Gesture Authentication in the Wild. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI ‘16)*. ACM, New York, NY, USA, 3722–3735. https://doi.org/10.1145/2858036.2858270

[61] Zhen Yue, Eden Litt, Carrie J. Cai, Jeff Stern, Kathy K. Baxter, Zhiwei Guan, Nikhil Sharma, and Guangqiang (George) Zhang. 2014. Photographing Information Needs: The Role of Photos in Experience Sampling Method-style Research. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI ‘14)*. ACM, New York, NY, USA, 1545–1554. https://doi.org/10.1145/2556288.2557192