International Journal of Human-Computer Studies 125 (2019) 118–128

Contents lists available at ScienceDirect

International Journal of Human-Computer Studies

journal homepage: www.elsevier.com/locate/ijhcs

Effect of experience sampling schedules on response rate and recall accuracy of objective self-reports

Niels van Berkel[a,*], Jorge Goncalves[a], Lauri Lovén[b], Denzil Ferreira[b], Simo Hosio[b], Vassilis Kostakos[a]

[a] Interaction Design Lab, The University of Melbourne, Australia
[b] Center for Ubiquitous Computing, University of Oulu, Finland

ARTICLE INFO

Keywords:
Experience sampling method
ESM
Ecological momentary assessment
EMA
Self-report
Smartphone
Contingency
Response rate
Accuracy
Mobile questionnaires
Data quality
Validation

ABSTRACT

The Experience Sampling Method is widely used to collect human labelled data in the wild. Using this methodology, study participants repeatedly answer a set of questions, constructing a rich overview of the studied phenomena. One of the methodological decisions faced by researchers is deciding on the question scheduling. The literature defines three distinct schedule types: randomised, interval-based, or event-based (in our case, smartphone unlock). However, little evidence exists regarding the side-effects of these schedules on response rate and recall accuracy, and how they may bias study findings. We evaluate the effect of these three contingency configurations in a 3-week within-subjects study (N = 20). Participants answered various objective questions regarding their phone usage, while we simultaneously establish a ground-truth through smartphone instrumentation. We find that scheduling questions on phone unlock yields a higher response rate and accuracy. Our study provides empirical evidence for the effects of notification scheduling on participant responses, and informs researchers who conduct experience sampling studies on smartphones.

1. Introduction

To collect data on human behaviour 'in the wild', researchers have adopted the use of *in situ* data collection methods. One popular method is the Experience Sampling Method (ESM) (Larson and Csikszentmihalyi, 1983), also known as Ecological Momentary Assessment (EMA). In an ESM study, participants are repeatedly asked a set of questions during their daily lives. Typically, data collection occurs multiple times per day over a period of several days or weeks, allowing for a detailed analysis of the phenomena under investigation (Larson and Csikszentmihalyi, 1983). An important mechanism in experience sampling is the adoption of notifications, used to inform participants when a questionnaire should be answered. This stands in contrast with the diary study, in which participants are traditionally not actively informed to provide data. Participants in an ESM study can either comply with these requests for information, dismiss, or ignore it (deliberately or unwittingly). Study notifications – *i.e.*, prompts to answer a questionnaire – can be triggered by a variety of schedule configuration alternatives, categorised into three contingency categories (Barrett and Barrett, 2001; Wheeler and Reis, 1991): randomly distributed (signal contingent), following a time interval (interval contingent), or triggered by a specific event (event contingent).

In ESM studies, it is crucial that the response rate is high, otherwise the data may be considered unreliable. Similarly, researchers strive to ensure participant input is as accurate as possible to ensure reliability of data analysis and results. Yet, it remains challenging to compare the effect of the three scheduling types on participants' responses due to the effect of other methodological parameters. For example, participants' response rate has been said to depend on participants' attachment to the study results (Larson and Csikszentmihalyi, 1983), number of notifications (Consolvo and Walker, 2003), communicating response rates to participants (Barrett and Barrett, 2001), effort required to complete a questionnaire (Consolvo and Walker, 2003), and notification expiry time. Surprisingly, few of these have been evaluated systematically, and researchers are limited to using their intuition about their effects when designing studies. Furthermore, the specific event responsible for triggering notifications in an event contingent-based schedule can be virtually anything – increasing the complexity of comparing these different contingencies. In this work, we therefore opt to choose an event which is actively used in HCI studies: unlocking of the participant's smartphone (see *e.g.*, (van Berkel et al., 2016; Harbach et al., 2014; Niforatos and Karapanos, 2014)).

* Corresponding author.
E-mail address: n.vanberkel@student.unimelb.edu.au (N. van Berkel).

https://doi.org/10.1016/j.ijhcs.2018.12.002
Received 23 November 2017; Received in revised form 10 September 2018; Accepted 2 December 2018
Available online 03 December 2018

In this study, we compare three ESM contingency types to assess whether they affect participants' response rate and accuracy. Measurement of ESM response accuracy is notoriously challenging because self-report data often lacks absolute ground truth (*e.g.*, emotional state). Because our study investigates quantifiable and objective aspects of smartphone usage (*e.g.*, usage duration), we are able to compare participants' reports of their personal smartphone usage against ground truth data of their actual usage. Our findings help researchers overcome the intrinsic challenges of collecting participant data using the ESM on smartphones. While the choice for a certain notification contingency configuration sometimes follows from the study design and not *vice versa*, our results quantify the effects of these methodological decisions on participants' response rate and recall accuracy.

2. Related work

The methods to study human behaviour takes many forms, ranging from lab interviews to ethnographic fieldwork. A common way of collecting intrinsic data such as human emotions or thoughts, is to intermittently ask study participants to answer a set of questions during their daily activities. This repetition leads to a more complete overview of the participants' daily life if compared to a single one-time questionnaire, as "*little experiences of everyday life fill most of our waking time and occupy the vast majority of our conscious attention*" (Wheeler and Reis, 1991). For such repeated inquiries, the Experience Sampling Method (Larson and Csikszentmihalyi, 1983) is a popular choice. Three key characteristics of the ESM are highlighted in Larson's seminal publication; "*it obtains information about the private as well as the public parts of people's lives, it secures data about both behavioural and intrapsychic aspects of daily activity, and it obtains reports about people's experience as it occurs, thereby minimizing the effects of reliance on memory and reconstruction*" (Larson and Csikszentmihalyi, 1983). Since its introduction, the method has been used to study a wide range of topics (*e.g.*, substance usage (Shiffman, 2009), educational practices (Muukkonen et al., 2008), evaluation of technology (Ickin et al., 2012)) and has seen increased uptake in the HCI community (van Berkel et al., 2017), indicating the method's wide applicability.

2.1. ESM study configuration

A core element of the ESM is to notify participants *when* to complete a certain questionnaire. Albeit the ESM in its original form only describes the use of randomly triggered notifications, the current consensus is to distinguish three main ESM notification trigger types (Barrett and Barrett, 2001; van Berkel et al., 2017; Wheeler and Reis, 1991):

- Signal contingent, in which the timing of notifications is randomised over a predefined duration;
- Interval contingent, in which timing of notifications follows a predefined time schedule;
- Event contingent, in which specific (measurable) events result in the presentation of a notification. These events can be related to the usage of the device (*e.g.*, unlocking the device (van Berkel et al., 2016)), or following a change in readings from a device sensor (*e.g.*, the microphone (Lathia et al., 2013)).

Literature highlights various concerns for each of these respective notification trigger types. Signal contingent notifications are unlikely to capture relatively rare events (*e.g.*, interactions with friends or partners (Wheeler and Reis, 1991)), therefore requiring a longer study duration to collect a usable quantity of data. Interval contingent notifications are likely to be distant in time from the event being recorded, introducing retrospection bias (Wheeler and Reis, 1991). In addition, the consistent repetition might result in participants being able to predict the occurrence of the next notification, possibly introducing cognitive bias (Consolvo and Walker, 2003). Event contingent notifications rely solely on the event that is being observed, and in the case of an uncommon event only a few notifications will be sent to study participants. Event contingent notifications can also be prone to design bias as the result of contextual dissonance (Lathia et al., 2013) — the choice of event triggering the notification directly influences the number of notifications. In addition, the ability to anticipate incoming alerts may result in behavioural reactivity (Hormuth, 1986).

In addition to these trigger-specific concerns, the literature also describes two concerns related to the general experience sampling methodology (van Ballegooijen et al., 2016; Santangelo et al., 2013). By continuously asking participants to focus on one aspect of their lives, the method may not only measure but also influence the participant's behaviour. This may have a positive outcome for the study participant (*e.g.*, by acting as an intervention technique for mental health patients (van Ballegooijen et al., 2016)), however it can also negatively affect the reliability of study results – this phenomenon is called *measurement reactivity*. Furthermore, *response fatigue* describes the notion that both the number and quality of participant responses decreases over time due to a decreasing interest in contributing to the study. We discuss both these concerns in more detail below.

2.2. Response rate

A study's response rate (*i.e.*, compliance rate) is the ratio of successfully answered questionnaire notifications over the received notifications. Many factors can potentially influence a participant's response rate. Typically, a higher response rate reflects a more accurate observation of the participant's experience, as a larger portion of the studied phenomenon is captured. Factors that have been identified to influence participants' response rate include: participant attachment (Larson and Csikszentmihalyi, 1983), study fatigue (Stone et al., 1991), technical difficulties (*e.g.*, failure in transmission, crashes) (Consolvo and Walker, 2003), motivational elements (*e.g.*, gamification) (van Berkel et al., 2017), as well as study design choices such as notification expiry time (allowing more time to answer a questionnaire before it expires).

Various studies have investigated ways of increasing response rate in experience sampling studies. Litt et al. (Litt et al., 1998) describe the use of "booster" telephone calls, in which participants received phone calls to encourage compliance. Naughton et al. (Naughton et al., 2015) increase response rate by an average of 16% by sending a follow-up reminder after 24 h, with little benefit reported on extending reminders to 48 h. Kapoor and Horvitz (Kapoor and Horvitz, 2008) test four different sampling strategies, ranging in sophistication from randomly timed notifications to the use of a predictive interruptibility model. Employed on a desktop computer, the system identifies usage behaviour (*e.g.*, typing) as well as external signals (*e.g.*, calendar events). Interestingly, the different probing mechanisms led to a high difference in average number of notifications (4.65 notifications for '*decision-theoretic dynamic probing*', 21.35 notifications for random probing). In 73.08% of random notifications, participants indicated they were too busy to respond to the probe, whereas participants in the decision-theoretic dynamic probing condition indicated to be too busy in 46.63% of cases. Although participants indicated to be less busy in this model, the high difference in notifications highlights a potential pitfall of a dynamic notification scheme. Lathia et al. (Lathia et al., 2013) compare the effect of four different sensor-based notification triggers (following a 'social event' (text message or phone call), change in location, microphone detecting silence, and microphone detecting noise). Their results show that these event-based sampling techniques do not collect responses distributed evenly over the course of day, introducing bias into the data collection. In addition, there are significant differences in the level of self-reported affect between these sensor-based schedules. While not further dissecting these differences, the authors note that "*our design parameters influence the outcome, and any inferences made on such*

data should take into account that design will influence the view that researchers will build of their participants" (Lathia et al., 2013). Our study builds upon this work by collecting self-reports for which ground-truth data is available, allowing for the measurement of participant accuracy.

2.2.1. Response fatigue

The gradual decline of response rate over the duration of the study is a well-documented phenomenon (*e.g.*, (Naughton et al. (2015); Stone et al. (1991))). Hsieh et al. (Hsieh et al., 2008) were one of the first to study this behaviour, and improved participant compliance over time by providing visual feedback to the participant through a comparative study (visual feedback *vs* no feedback).

Fuller-Tyszkiewicz et al. (Fuller-Tyszkiewicz et al., 2013) investigate response fatigue and find that "*demands of ESM protocol undermine quantity more so than quality of obtained data*" (Fuller-Tyszkiewicz et al., 2013). Although the sample size of the study was considerable (N = 105), the study duration was limited to 7 days – with many ESM studies being of longer duration. Reynolds et al. (Reynolds et al., 2016) report on a study on parent-child relationships, using the methodologically related diary method over a period of 56 days. Participants were requested to complete one diary entry per day (as opposed to multiple daily entries in an ESM study). As the study progressed, participants completed less diary entries on weekdays, with no change on weekends. As child and parent separately reported on the same events, the authors analysed the change in agreement over time and found that the agreement between parent and child slightly declined as the study progressed. As such, response fatigue did not only affect response rate, but also the quality of the participant response (Reynolds et al., 2016).

2.3. Participant response accuracy

The ESM is specifically designed to increase participant response accuracy by reducing reliance on participants' long term memory to reconstruct past events (Larson and Csikszentmihalyi, 1983). Participant reflection on past events through long term memory has been shown to be unreliable, reducing the validity of collected research data (Iida et al., 2012). However, the accuracy of ESM responses itself remains understudied (van Berkel et al., 2018), despite being referred to as the gold standard in measurement of affective experience of daily life and is used as a benchmark against other methods (Krueger and Schkade, 2008). As affective experience is inherently personal, establishing a baseline to measure accuracy is challenging. In their analysis of reliability and validity of experience sampling for children with autism spectrum disorders, Chen et al. (Chen et al., 2015) ran a 7-day long study in which participants answered questions concerning their activity, experience, and emotion. Through a split-week analysis (comparing data from the start to data from the end of the week), the authors find no significant difference between quality of experiences and associated emotions, suggesting a high internal reliability.

3. Experimental design

We designed an experiment where participants use their smartphones to answer a set of questions regarding their perceived smartphone usage. In addition, we collected ground truth data by logging actual smartphone usage in the background. ESM notifications were presented according to one of the following experimental conditions:

- **Condition A - Signal contingent (random)**. The presentation of notifications is randomised (using a uniform randomisation scheme) over the duration of the day. We enforce a set limit of six notifications per day, without regards for the fact whether the notification is unanswered. The time-window within which the timing of the notifications is randomised ranges between 10:00 to 20:00 to avoid night-time alerts.
- **Condition B - Interval contingent (bi-hourly)**. Notifications are scheduled to occur every two hours between 10:00 and 20:00 (*i.e.*, at 10:00, 12:00, […], 20:00). This results in a total of six notifications presented to our participants over the course of the day.
- **Condition C – Smartphone unlock (*i.e.*, following an unlock event)**. Notifications are presented when the participant unlocks the mobile phone, considering pre-configured time sets. Identical to Conditions A and B, we send a maximum of six notifications per day. A time constraint is added to ensure notifications are spread over the duration of the day, consisting of 6 time-bins of equal duration (*e.g.*, 10:00 to 11:40, 11:40 to 13:20, […], 18:20 to 20:00). In case the phone has already been unlocked in the current time-bin, no new notification is shown — regardless of whether the notification was answered. If the phone of the participant is not unlocked in a given time-bin, no notification is shown. We chose the unlock event in our notification configuration because this event is widely used in the literature (*e.g.*, (van Berkel et al., 2016; Fischer et al., 2011; Harbach et al., 2014; Zhang et al., 2016)) and acts as a precedent to many other smartphone interactions. As such, the design of our study does not aim to encompass all notification events or generalise to all possible event contingent configurations.

All conditions have a maximum of six notifications per day. We applied a notification expiry time of 15 min after which the notifications were dismissed and no longer shown to the participant. In addition, we logged whether notifications expired or actively dismissed by the participant. Finally, participants were not able to initiate data submissions themselves.

3.1. Study parameters

We use a balanced Latin Square to distribute our participants over the three conditions, reducing the risk of order effects (*e.g.*, fatigue, loss of interest, learning) on our data. As we have an odd number of conditions in our study, we apply a double Latin Squares design (*i.e.*, second square is a mirror of the first) in the distribution of our participants. For example, some participants begun our study in Condition C for a week, then switched to B for a week, and finally to A for another week. We balanced participants equally over each 'square' in the double Latin Square design. Our 21-day long deployment is in line with recommendations from Stone et al. (Stone et al., 1991) — data quality deteriorates after a period of 2–4 weeks. Each condition is applied for seven full days before a new condition is introduced, resembling an authentic study configuration as opposed to randomly distributing the conditions over the entire duration of this study.

3.2. ESM questionnaire

All the ESM questions are identical in all conditions. The questionnaire consists of three questions regarding recent smartphone usage, one question on the level of interruption of the notification, and one self-assessment of the response accuracy. The questions on smartphone usage include the number of unique applications used (shown to be context dependent (Do et al., 2011), thus changing throughout the day), duration of smartphone usage (both diverse across users (Falaki et al., 2010) and challenging for users to estimate accurately (Andrews et al., 2015)), and number of times the phone was turned on (challenging for users to accurately estimate (Andrews et al., 2015)). We explicitly instructed participants to include the number of locked screen-glances (*e.g.*, checking notifications, time) in their answers. We base our questions regarding the level of interruption of notification timing on (Fischer et al., 2011).

We phrase our question to specifically mention the amount of time since the last notification was presented, regardless of the fact whether this notification was answered by the participant. This also makes it clear for the participant as to what exactly is being asked. For example: "*How many unique applications have you used since 12:00?*" . For the first notification of each day, we ask participants to answer the question for the timespan following 08:00 up to the current time. The question set consists of five questions in total. As aforementioned, three of those questions target smartphone usage (number of unique applications, duration of phone usage in minutes, and number of times the screen of the phone was turned on), presenting the participant with a numeric input field. The final two questions focus on the questionnaire itself ("*How confident are you in your answers?*" and "*Was this an appropriate time to interrupt you with these questions?*"). These last two questions present the participant with a 5-point Likert-scale ranging from [not at all confident / very inappropriate] to [very confident / very appropriate].

Finally, participants completed an exit questionnaire online after the study ended to collect information on general smartphone usage, notification-interaction habits, and the participants' perception in answering the ESM questions.

3.3. Data collection

Data was collected using a custom-made application that was installed through the Google Play Store (private beta release). The application was running continuously in the background of the participants' personal device. Using this application, based on the AWARE mobile instrumentation framework (Ferreira et al., 2015), we collect not only the participants' responses but also log, *inter alia*, the following data: phone usage, application usage, ESM answer, and ESM question status (*i.e.*, expired, answered, dismissed). The application is also responsible for sending notifications to participants. Notifications appeared as regular notifications on the participants phone, with the questionnaire opening when the notification is touched.

4. Method

4.1. Recruitment

A total of 24 participants participated in this study. We omit contributions from 4 participants who encountered miscellaneous device incompatibility issues leading to partial data loss, leaving 20 participants in total (average age 26.4 ($\pm$ 4.6) years old, 5 women, 15 men). As a result, our Latin Square design (Section 3.1) is no longer complete. We report the final distribution of participants over conditions in Table 1 below. Participants were recruited from our university campus using mailing lists and had a diverse educational background. To minimise the novelty effect and ensure proper device operation, participants used their personal device.

4.2. Procedure

We invited all participants for an individual training session in our lab. During this session, we explained our interest in the way they use their smartphone in daily life. We did not inform them of the different contingency types that would be evaluated. To set expectations on the level of participant strain, we explained that we would send a maximum of six notifications per day between 10:00 and 20:00. In addition, we explained each of the five ESM questions one-by-one to ensure participants understood the questions we asked them to answer. Following the explanation of the ESM questions, we informed participants about the data logging capabilities of our application and installed the necessary software on the participant's phone. We explained to our participants how we interpret 'smartphone usage' for the purpose of this study: any moment at which the smartphone screen is turned on. Participants were given practical examples to ensure their reports are in line with our expectations (*e.g.*, listening to music with the screen turned off is not considered as usage). Due to technical limitations of Android OS, we are unable to distinguish between 'real' smartphone usage and the user putting away their phone after usage without turning off the screen (*i.e.*, resulting in a timeout). We explicitly instructed participants to include the number of locked screen-glances (*e.g.*, checking notifications, time) in their answers. All deployments lasted for 21 days, and were followed by an exit questionnaire. Participants received two movie vouchers as compensation.

Table 1
Final distribution of participants across conditions over study period.

	A - Signal	B - Interval	C - Unlock
Week 1	6	7	7
Week 2	7	8	5
Week 3	7	5	8

4.3. Data analysis

We aim to provide Bayesian statistics as currently discussed in the broader HCI community. This brings forward a focus on effect size and uncertainty in an attempt to "*convey uncertainty more faithfully*" (Matthew Kay et al., 2016) and provide a more useful presentation of our results for both academics and practitioners. To this end, we use Bayes factors over *p*-values for hypothesis testing and credible intervals (using highest posterior density) over confidence intervals. The use of *p*-values has faced increased scrutiny (*e.g.*, (Kaptein and Robertson, 2012; Wagenmakers, 2007)), mainly due to the incorrect interpretation and application by researchers.

Bayes factors provide an intuitive way of comparing hypotheses by quantifying the likelihood of competing hypotheses as a ratio. This enables a direct comparison between models. For example, a Bayes factor of 15 indicates that the collected data is 15 times more probable if H1 were true than if the null hypothesis H0 were true. In formula form this becomes: $BF_{10} = \frac{p(D \mid H1)}{p(D \mid H0)}$, with D denoting data observed in the study. Throughout our analysis we use the *BayesFactor* package (Morey et al., 2014) for R. Given our within-subjects design, our analysis consists mainly of 'Bayesian ANOVA', as described in (Rouder et al., 2012). The BayesFactor provides default priors for common research designs. In our models, we consistently treat participants as a random factor as per the study's design. Using posterior estimation, we compare the study conditions through credible intervals. Credible intervals are calculated using the 95% Highest Posterior Density interval (HPD) through the *bayesboot* package (Bååth, 2016). We adopt the interpretation of Bayes factors' significance as put forward by Jeffreys (Jeffreys, 1998) and refined by Lee and Wagenmakers (Lee and Wagenmakers, 2014).

5. Results

A theoretical maximum of 2520 notifications (840 per condition) could be sent during the study (6 notifications per day, 20 participants, 21 days). A total number of 2313 notifications were issued during the study. This difference is due to the phone running out of power, the phone being turned off, or the participant not using the device during a given timeslot (Condition C) – occurrences which are beyond our control. Our results highlight a considerable effect of condition on response rate, two out of three accuracy measures, and self-reported confidence levels. Study condition did not affect *perceived* level of interruption. We discuss the detailed results of each dependent variable below.

Table 2
Notification behaviour and response rate.

	A - Signal	B - Interval	C - Unlock
Notifications sent	814	806	693
Response rate	51.6%	51.6%	76.8%
Notifications answered	420	416	532
95% HPD	[47.14, 56.89]	[47.18, 56.73]	[72.70, 81.83]
Expired	46.5%	47.8%	22.1%
Dismissed	1.9%	0.6%	1.1%

5.1. Response rate

A total of 2313 ESM notifications were triggered over the duration of the study, with each notification consisting of 5 questions. The average response rate over all conditions was 59.7%. Many unanswered ESM questionnaires were the result of notification expiration (39.3% of total), with only 1.0% being actively dismissed by a participant. We sent a total of 814 randomly timed notifications in Condition A, and 806 scheduled notifications were sent for Condition B. For Condition C, a total of 693 notifications were sent upon phone unlock (out of a total of 4190 unlock events between 10:00 and 20:00 for days on which Condition C was active, 16.5%). Table 2 presents an overview of response rate results as split per condition.

We compute the effect of condition on response rate. A JZS Bayes factor within-subjects ANOVA with default prior (non-informative Jeffreys prior) revealed a Bayes factor of 1.37×10^{20}, providing extreme evidence for the existence of an effect of condition on response rate. We calculate a 95% HPD per condition and report these in Fig. 1 and Table 2.

Fig. 1 shows the overall effect of the three conditions on the participants' response rate. As can been seen from Fig. 2, Condition C results in the highest response rate for 17 out of 20 participants. This shows that the effect is consistent across participants, even for those that have an *overall* lower tendency to respond to ESMs.

5.1.1. Effect of day and time on response rate

Our study lasted 21 days, with the ESM response rate remaining fairly constant throughout the study. In Fig. 3, a 'wave' effect is visible, indicating minor variance among participant clusters (study condition changed every 7 days). We compute the effect of relative study day on response rate, regardless of study condition. A JZS Bayes factor within-subjects ANOVA with default prior revealed a Bayes factor of 0.08, providing moderate evidence for the lack of effect of study day on response rate. The response rate changes over time of day, as visible in Fig. 3. Condition B display a sharp drop in response rate in the morning hours when compared to the other conditions.

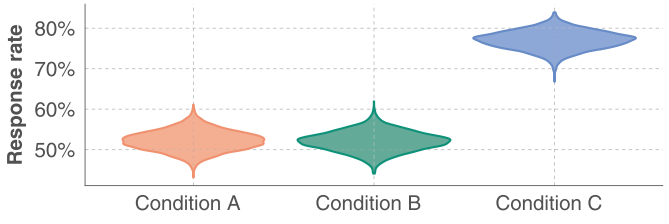

Fig. 1. Posterior distribution of response rate.

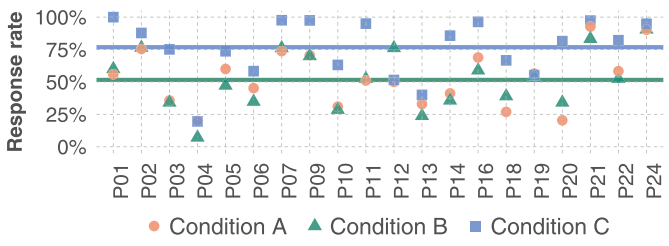

Fig. 2. Average response rate per participant and condition means (Condition A and B overlap).

5.1.2. Recall accuracy

We measure the recall accuracy of participants by comparing participants' answers against the actual data as logged passively on participants' phones. For the three questions focused on recall accuracy, we calculate both the mean average error (MAE) and the root-mean-square error (RMSE). MAE and RMSE are traditionally used to measure accuracy of a model's prediction. Lower scores indicate higher accuracy.

5.1.3. Usage duration

We asked participants to report the duration of their smartphone usages in minutes. We calculate the sum duration of phone usage activities for the time period as asked by each specific question. This is measured by calculating the time difference between the screen of the device being unlocked and the screen of the device being turned off. See Table 3 for an overview of these results.

For each answer provided by the participant we calculate the accompanying error ratio and apply a logarithmic function to normalise the data from the scale [0, Inf] to the scale [-Inf, Inf]. We then compute the effect of condition on this calculated error ratio. A JZS Bayes factor for a within-subjects ANOVA model with default prior revealed a Bayes factor of 0.02, providing very strong evidence for no effect of condition on the participants' error. The posterior distribution is shown in Fig. 4, and highlights the overlap between conditions. We calculate the 95% HPD per condition and report these numbers in Table 3. Running the same test for effect of relative study day on the error value shows a Bayes factor of 0.01, providing very strong evidence that the participants' answers were not affected by the longevity of the study.

Fig. 5 shows a scatterplot of self-reports on screen usage against the recorded observations (in minutes). To visualise the distribution of these two variables, we add marginal histograms to both axes. The value of self-reported observations is strongly clustered near 'round' numbers (*e.g.*, 5 min., 10 min., etc.), with a high number of usage durations reported in between the 0 and 5 min mark. We find evidence that participants *underestimate* their smartphone usage duration.

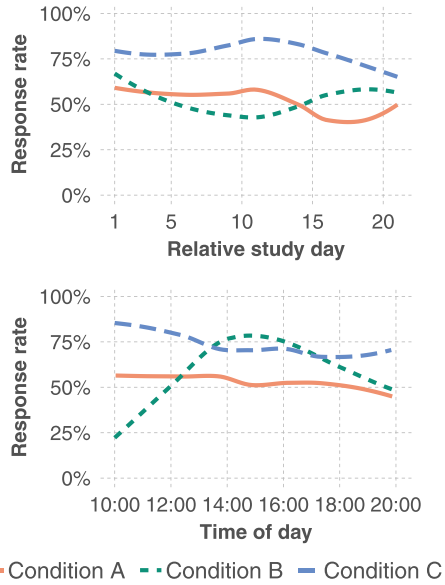

Fig. 3. Response rate over study day and time of day.

5.1.4. Number of unique applications

Similar to the calculation of smartphone usage duration, we calculate the number of unique applications used by our participants in the specified time window. We first eliminate system applications (*e.g.*, launchers, lockscreens) from out data as recommended in literature (Jones et al., 2015). Then we compare the observed number of unique application with the participant reports and calculate the corresponding MAE and RMSE per condition (Table 4).

The mean MAE was lowest in Condition C. First, we calculate the error ratio and again apply a logarithmic function to normalise the data to the scale [-Inf, Inf]. We compute the effect of condition on the error ratio in the reports submitted by our participants. A JZS Bayes factor for a within-subjects ANOVA model with default prior revealed a Bayes factor of 25.9, providing strong evidence for the effect of condition. We calculate the 95% HPD per condition and report these in Fig. 6 and Table 4. We also calculate the effect of relative study day on error ratio and find a Bayes factor of 0.001, providing extreme evidence for a lack of effect of study day on error ratio.

The scatterplot in Fig. 7 shows the self-reported observations against recorded observations with regards to unique applications used. For the majority of measurement periods, participants used a low number of unique applications (less than 5). The scatterplot indicates that participants are underestimating their application usage when they use a larger number of applications.

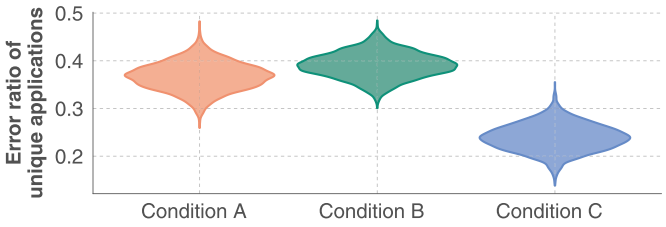

Fig. 6. Posterior distribution of error ratio for self-reported number of unique apps.

5.1.5. Screen-On frequency

Once again, we compare the answers as provided by our study participants with the collected ground truth data. We sum the number of times the screen was turned on in the time windows as specified in the questions. Then, we calculate the corresponding MAE and RMSE per condition. These results are presented in Table 5.

Table 3
Observed and reported screen usage duration (in minutes), including MAE and RMSE.

	A - Signal	B - Interval	C - Unlock
Observed mean (SD)	24.55 (± 28.47)	27.36 (± 29.05)	17.10 (± 24.97)
Self-reported mean (SD)	15.32 (± 16.38)	17.34 (± 16.93)	10.17 (± 18.45)
Mean error ratio	0.32 (± 1.04)	0.33 (± 0.96)	0.37 (± 1.15)
95% HPD	[0.21, 0.42]	[0.24, 0.43]	[0.27, 0.47]
MAE	14.99	16.66	12.81
RMSE	24.98	26.56	25.31

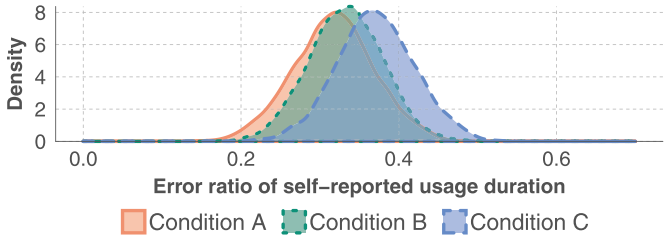

Fig. 4. Posterior distribution of error ratio for self-reported usage duration.

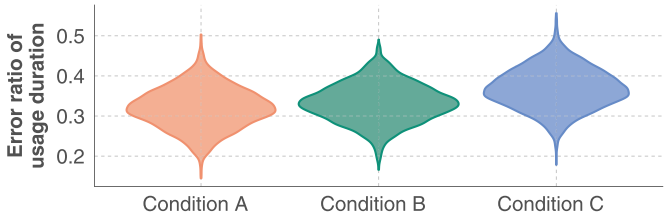

Fig. 5. Recorded *vs* self-reported screen usage duration (in minutes).

Table 4
Observed and reported unique applications used, including MAE and RMSE.

	A - Signal	B - Interval	C - Unlock
Observed mean (SD)	5.17 (± 3.30)	5.43 (± 2.88)	3.51 (± 2.58)
Self-reported mean (SD)	3.29 (± 2.95)	3.40 (± 2.57)	2.45 (± 1.72)
Mean error ratio	0.37 (± 0.55)	0.39 (± 0.48)	0.24 (± 0.55)
95% HPD	[0.31, 0.43]	[0.34, 0.44]	[0.18, 0.30]
MAE	2.67	2.58	1.80
RMSE	4.06	3.63	2.65

The mean MAE was lowest in Condition C. We calculate the error ratio and apply a logarithmic function to normalise the data, and the compute the effect of condition on the error ratio. A JZS Bayes factor within-subjects ANOVA with default prior revealed a Bayes factor of 6.34×10^9, providing extreme evidence for an effect of condition on participant error. We calculate the 95% HPD per condition and report these numbers in Fig. 8 and Table 5. Additionally, a within-subjects ANOVA indicated extreme evidence for the lack of effect of study day on error rate (Bayes factor = 0.004).

The scatterplot in Fig. 9 shows the frequency of screen turn on events, as collected from participant self-reports and collected ground truth. We see that especially for higher frequencies in screen on events, participants underestimate these events. The histogram of self-reported observations shows a less visible version of the trend shown in Fig. 5, where self-reports are clustered around 'round' numbers.

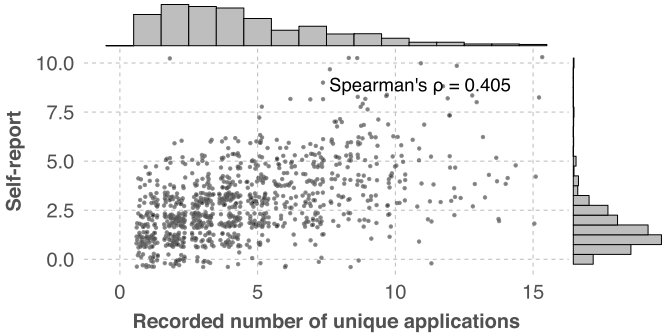

Fig. 7. Recorded *vs* self-reported number of apps.

Table 5
Observed and reported number of times screen turned on, including MAE and RMSE.

	A - Signal	B - Interval	C - Unlock
Observed mean (SD)	10.67 (± 11.49)	10.44 (± 7.49)	4.99 (± 6.58)
Self-reported mean (SD)	6.08 (± 6.15)	6.63 (± 4.89)	4.10 (± 4.32)
Mean error ratio	0.43 (± 0.67)	0.38 (± 0.67)	0.12 (± 0.74)
95% HPD	[0.36, 0.50]	[0.31, 0.44]	[0.06, 0.19]
MAE	5.93	5.70	3.65
RMSE	10.36	8.25	6.99

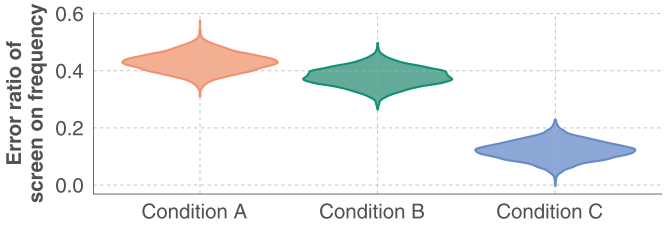

Fig. 8. Posterior distribution of error ratio for self-reported times screen on.

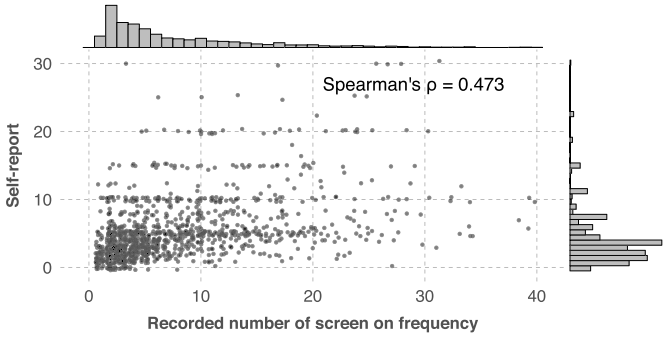

Fig. 9. Recorded *vs* self-reported screen on frequency.

5.1.6. Confidence

Each ESM questionnaire asked participants to rate their confidence in the accuracy of their answers. Mean confidence scores are 3.28 (± 0.97), 3.20 (± 1.01), and 3.54 (± 0.99) for Conditions A, B, and C respectively. We compute the effect of condition on the reported confidence. A JZS Bayes factor within-subjects ANOVA with default prior revealed a Bayes factor of 5.10×10^5, providing extreme evidence for an effect of condition on reported confidence. We calculate the 95% HPD per condition: [3.19, 3.38], [3.10, 3.29], and [3.46, 3.62] for Conditions A, B, and C respectively. Fig. 10 shows the distribution of the posterior. In addition, we compute the effect of relative study day on participant's confidence level. This results in a Bayes factor of < 0.001, providing extreme evidence for the lack of an effect of study day on reported confidence.

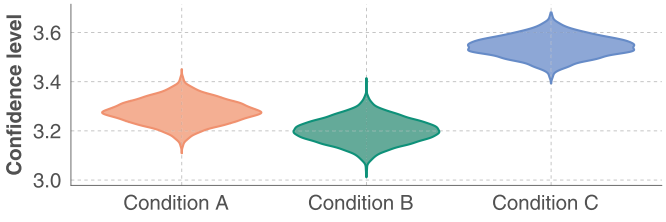

Fig. 10. Posterior distribution for self-reported confidence levels.

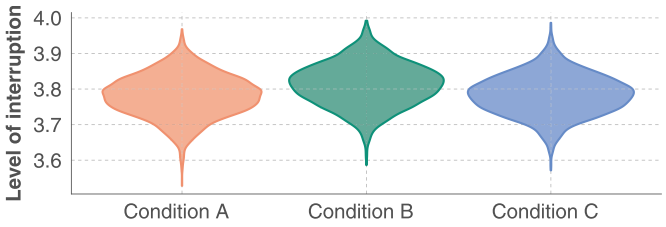

Fig. 11. Posterior distribution for self-reported level of interruption.

5.1.7. Level of interruption

Final question of each questionnaire asked participants to rate the level of interruption. Mean scores for each condition are 3.78 (± 1.20), 3.81 (± 1.24), and 3.78 (± 1.27) for Condition A, B, and C respectively. We compute the effect of condition on the perceived level of interruption for each questionnaire. A JZS Bayes factor within-subjects ANOVA with default prior revealed a Bayes factor of 0.11, providing moderate evidence for a lack of effect of condition on perceived interruption. We calculate the 95% HPD per condition: [3.67, 3.90], [3.69, 3.92], and [3.67, 3.89] for Conditions A, B, and C respectively (see Fig. 11). We also compute the effect of relative study day on the perceived level of interruption, and find a Bayes factor of 0.005, providing extreme evidence for the lack of an effect of study day on perceived interruption.

5.1.8. Notification trigger

The time of notification trigger has a direct effect on the elapsed time since 'asked upon time' (time as displayed in the questions; *e.g.*, '*[…] used since [xx.xx]?*'). Elapsed time thus indicates the timespan for which participants must answer a question. The notification trigger is randomised in Condition A, thus resulting in a wide range of elapsed time durations. Condition B presents a notification every 2 h, providing a constant elapsed time of 120 min. In Condition C, notifications are triggered depending on when participants use their phone. Condition C is therefore the only condition in which notification presentation is not controlled by a system designed schedule.

We visualise the effect of condition on this elapsed time using a density plot (Fig. 12). Condition B is omitted from the plot as it is centred at the 120 min mark, and including it reduces the readability of the graph. Condition A has an average elapsed time of 113 min, Condition B an average of 122 min, and Condition C results in the shortest elapsed time of 87 min. The density of Condition C shows a strong tendency for a shorter elapsed time when compared with Condition A. This difference can be accounted for by the frequent (but brief duration) of smartphone usage, as has been reported in the literature (Böhmer et al., 2011; Ferreira et al., 2014). By multiplying the average elapsed time with the number of daily notifications and the condition's respective response rate, we find the average daily time covered by the answered ESM questions; 349.85 min for Condition A, 377.71 min for Condition B, and 400.90 minutes for Condition C.

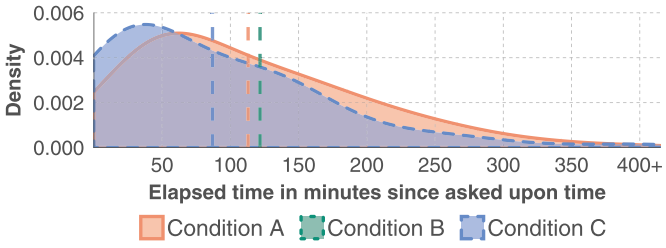

Fig. 12. Condition differences in asked upon time and notification presentation. Vertical lines indicate condition means. Condition B omitted for visibility.

6. Discussion

Previous work has identified various factors that influence participant responses in experience sampling (*e.g.*, motivation (Larson and Csikszentmihalyi, 1983), feedback visualisations (Hsieh et al., 2008), gamification (van Berkel et al., 2017)). Here, we systematically investigate the effect of a notification schedule (*i.e.*, contingency) on response rate, recall accuracy, and participants' perception (interruption and confidence). Based on the literature indicating that various personal factors influence response rate (*e.g.*, participant attachment (Larson and Csikszentmihalyi, 1983), study fatigue (Stone et al., 1991)), we use a within-subjects design to minimise the effect of individual motivation. Our results show a consistently higher response rate when using the smartphone unlock event strategy *versus* a signal or interval configuration. Furthermore, our results indicate that recall accuracy is higher in the unlock-based condition for two out of three recall metrics. To assist researchers in both their study design and data analysis, we highlight insights on the effect of different contingency configurations on response rate and recall accuracy.

We use Bayesian analysis instead of a frequentist approach. While the latter is more common within the HCI-community, the former has certain key advantages. Using frequentist statistics (and accompanying *p*-values) often leads to misinterpretation by researchers (Kaptein and Robertson, 2012) and to a focus on the *existence of an effect* rather than the *size of the effect*. In addition, Bayesian statistics have been hauled as better-suited for the HCI community – in which relatively small sample sizes are typical (Caine, 2016). Using a Bayesian approach, researchers can more precisely compare novel conditions (*e.g.*, a more complex event contingent notification schedule) against known conditions (our unlock event based configuration) (Matthew Kay et al., 2016).

6.1. Response rate

Our results show a considerable difference in the response rate between conditions, with a response rate of 51.6% for both Condition A and B, and a response rate of 76.8% for Condition C. As a result of the contingency configuration as driven by the participants' smartphone usage, Condition C also resulted in the fewest notifications send out: 693 notifications *versus* 814 and 806 in Conditions A and B respectively. Despite the overall lower number of notifications, the higher response rate in Condition C eventually results in 27% more answered questionnaires when compared to Conditions A and B: 532 questionnaires for Condition C, 420 for A, and 416 for B.

The increase in response rate in Condition C highlights the relevancy of work on user interruptibility for experience sampling studies. There is a plethora of ways in which an event contingent notification scheme can be configured, including phone usage (*e.g.*, screen unlock (van Berkel et al., 2016)), sensor readings (*e.g.*, location (Froehlich et al., 2006)), and events external to the device (*e.g.*, experiencing a headache (Kikuchi et al., 2007)). While the smartphone unlock is only one possible event contingent study protocol, its premise is straightforward; we only notify a participant when we know the smartphone is actively used. As a result, we rely on participants to actively use their smartphones throughout the day. Our results show the advantage of this strategy for both researchers (higher response rate and more completed responses) and study participants (decrease in total number of notifications).

Zhang et al. (Zhang et al., 2016) examined *unlock journaling* as a diary entry technique and compared it to notification reminders and presenting no reminders at all (participants created diary entries on their own). While the diary method is distinct from ESM, the results are in line with our findings. Unlock journaling allows participants to input data while unlocking their phone, as opposed to opening a notification or separate application. This input 'field' was displayed every time the participants started using their phone, and resulted in a higher response rate than the other (aforementioned) conditions, as well as a reducing experienced intrusiveness. Our results also show an increased response rate when presenting an alert as soon as the participant starts using their phone. While Zhang et al. (Zhang et al., 2016) offered the possibility to answer a question on each phone unlock, we limited our study to a maximum of six notifications per day. This design choice helps to keep the study unobtrusive and ensures that the topic at hands is not continuously brought to the attention of the participant – a potential source of measurement reactivity (van Ballegooijen et al., 2016; Reynolds et al., 2016).

Participant response rate, although not always reported by researchers (van Berkel et al., 2017), can differ considerably between participants. This is problematic, as the general interpretation of collected ESM answers will be skewed towards those participants that *did* actively contribute. As a result, researchers are sometimes forced to remove from the analysis participants with a low number of contributions (*e.g.*, (Epp et al., 2011)). Fig. 2 shows the response rate *per participant* in our study, including the mean response rate over all participants per condition. Just like in related work, the response rates in our study differ considerably between participants. However, Fig. 2 shows that for 17/20 participants, Condition C outperforms both Conditions A and B. This shows that Condition C's notification configuration works across a wide range of participants, including both active contributors as well as those with an overall lower response rate.

Context is an important aspect to consider when conducting ESM studies. As the context of a participant regularly changes throughout the day, it is common practice for ESM questionnaires to disappear after a given amount of time – better known as notification expiry time (van Berkel et al., 2017). In this study, a 15 min expiry time was used. Given that notifications were therefore disappearing in 15 min, it is likely that participants in Condition A and B have simply missed some of these notifications altogether (rather than actively ignoring them) – contributing to the lower response rates in these conditions. However, keeping these notifications alive for longer durations of time goes against the *in-situ* nature of Experience Sampling. This approach therefore aligns with current practice of researchers utilising the ESM.

Summarised, a study's contingency configuration can significantly affect response rates. In our study, delivering notification when participants unlocked their smartphones achieved the highest response rates. This contingency resulted in fewer sent notifications (compared to the other conditions), and a higher number of answered notifications. Remarks from study participants highlight how a participant's typical smartphone behaviour may affect response rates, and why many ESM notifications typically go unanswered; "*I usually don't notice notifications when I am away from home/work, because I used to keep my phone in the bag and check it only when I need it.*" (P12), and "*I keep my phone in silence all the time, to not get distracted by the notifications too often.*" (P19). A participant with Condition C scheduled for the last week remarks; "*By the end, the notifications were appearing right when I was free and could easily answer the question.*" (P02). Based on these participant comments, presentation of notifications following phone unlock has a positive effect on both availability and visibility (*i.e.*, participants are aware that a notification was received and are available and willing to answer it). We assume that in Condition C, participants noticed all notifications (visibility of 100%), and that those questionnaires that went unanswered are primarily the result of being interrupted in an ongoing activity (76.8% availability). For Conditions A and B, which feature almost identical response rates, almost half of the notifications either went by unnoticed or arrived at a time at which the participant was unable or unwilling to answer.

6.1.1. Day and time

Studies employing ESM typically run for a relatively short amount of time, following the observation that data quality decreases after a period of two to four weeks (Stone et al., 1991). As shown in Fig. 3, a small reduction in overall response rate (across study conditions) is visible across the total duration of the study. Dividing the study in 1-

week bins shows a response rate of 62.3%, 58.3%, and 56.6% for study week 1, 2, and 3 respectively. Fig. 3 shows participants' response rate over the duration of the day. Condition B, in which questionnaires are presented every other hour (10:00, 12:00, […]), shows a very low response rate in the morning, which increases in the afternoon (even surpassing Condition C), before dropping slightly in the evening. Poppinga et al. (Poppinga et al., 2014) investigate receptivity of smartphone notifications, and their results also indicate a low response rate in morning hours, surpassed only by a lower response rate during the night. However, this does not explain the stark difference between conditions for these hours. Condition B is the only condition in which participants could precisely anticipate notifications, potentially causing a lower response rate as participants did not want to be interrupted at the start of their day by an (expected) ESM questionnaire.

The exit questionnaire answers also highlight a difference between participants. Several participants commented that their willingness to answer diminished during the study – "*First week was ok, later my willingness to answer was decreasing fast. I was happy when it ended.*" (P12) and "*Last couple of days I had to motivate myself by remembering the goal and that it's about a real study.*" (P10). Others mentioned that their willingness to respond remained the same; "*I think I had good motivation throughout the study.*" (P13) and "*My willingness to answer notifications didn't change during the study*." (P03). Unsurprisingly, none of the participants commented that their response rate increased over time.

6.2. Recall accuracy

Recall accuracy reflects participants' ability to accurately reflect on their mobile phone usage. Previous work shows the inherent difficulty of reflecting on one's own mobile phone usage (Andrews et al., 2015). As shown in Figure, Condition C resulted in the shortest time elapsed between notification onset and the time as presented in the question. Revisitation of phone usage often occurs in short time intervals (van Berkel et al., 2016; Jones et al., 2015), and frequently throughout the day (Böhmer et al., 2011; Ferreira et al., 2014). As a direct result of this participant behaviour, Condition C resulted in participants reflecting on a shorter period of time. This naturally affects their recall accuracy, as a shorter time period allows for easier reconstruction. Researchers requiring a certain minimum time period to be covered in ESM questionnaires can impose a minimum time difference between notifications.

Comparing between the study conditions, the results indicate that the Condition C resulted in the most accurate data on two of the three metrics ('unique applications used', 'number of screen on events', but not for 'usage duration'). For the 'usage duration' metric, no differences exist between conditions. Furthermore, although participants' self-reported confidence scores are relatively similar between conditions (3.28, 3.20, and 3.54 for Conditions A, B, and C respectively), our results indicate that these self-reported confidence scores do provide an indication of the accuracy of the participants' answers.

Our results show that participants' recall accuracy was significantly affected by the notification schedule. Because the event contingent schedule is driven by the behaviour of our study participants, the time window over which they were asked to reflect on their notifications differed between conditions (Fig. 12). A potential drawback of this result is that each individual data point provides a shorter area of duration when compared to the other conditions. However, given the considerable difference in response rate between these conditions, Condition C still resulted in the most complete average daily coverage. Participants provided a variety of comments when asked about their accuracy. Several participants suspect that their accuracy decreased over time; "*I believe that the accuracy might have decreased with time.*" (P10), "*I guess first week I was more motivated, it was something new and I wanted to provide high quality results*" (P12), whereas others indicated belief that their accuracy increased over time due to the learning effect; "*It became easier to answer questions accurately after one week*" (P18).

Recall accuracy, or more generally 'answer quality', has received less (methodological) attention in the ESM literature compared to response rate – most likely given the challenge in assessing the quality of answers. Literature describes that, participants that feel valued, believe the study is of importance, are not highly burdened by the study, and feel responsible to the researcher produce the highest data quality (Conner and Lehman, 2012; Larson and Csikszentmihalyi, 1983). Conner & Lehman (Conner and Lehman, 2012) state that fixed schedules (*i.e.*, interval contingent) result in the least burden to participants, while variable sampling (*i.e.*, signal contingent) introduces a high risk for participant burden. Our results quantify the experienced level of interruption through self-reports, and do not show a difference in experienced level of interruption between conditions. These self-reports are, however, limited to moments at which the participant completed a questionnaire. The event contingent configuration notified participants when they unlocked their phone, with a set maximum of notifications equally distributed over the day – as such, this schedule was both fixed and variable. Even though questionnaire notifications arrived at moments in which participants were about to use their phone, this did not lead to a decrease in self-reported level of interruption.

The analysis of recall accuracy in this study is limited to objective data (smartphone usage). It is not clear whether or how different notification schedules would influence more subjective data such as emotions or thoughts. Usage of smartphones has been shown to be far from an emotionally neutral activity (*e.g.*, smartphone overuse (Lee et al., 2014)), especially when it concerns the use of social media networks (Andreassen, 2015). Using an unlock-based trigger to show an ESM questionnaire could therefore result in undesired side effects. As participants were able to provide an indication on the accuracy of their answers, researchers could consider asking participants to report their confidence after completing a questionnaire. This self-reported accuracy label can be used as a parameter in assessing the accuracy of participant answers when missing ground truth data.

6.2.1. Numerical bias

As seen in Fig. 5, participants favoured numeric responses ending in 0 and 5. This preference for 'round' numbers has been previously identified in literature (*e.g.*, preference for round numbers in a guessing task (Ross and Engen, 1959)). The effect is less pronounced in Fig. 7, and not visible in Fig. 9– intuitively this is the result of the larger scale among which participants make their estimation. This is in line with results from Baird et al. (Baird et al., 1970). Upon reflection in the exit questionnaire, some of the participants were aware of their 'numerical bias', leaving comments as "*I tried to be accurate but especially with bigger numbers I had to estimate nearest 5 mins.*" (P07) The literature discusses various sources of potential bias in different configurations of experience sampling (*e.g., contextual dissonance* for sensor triggered notifications Lathia et al. (Lathia et al., 2013)). To the best of our knowledge, the effect of numerical bias has not been previously studied in the context of ESM. Researchers should be aware of their participants' bias towards round numbers, especially with a large answer range.

6.3. Interval-Informed event contingency

Although resulting in the highest response rate in our study, a smartphone-unlock trigger may not translate well to other studies. First, the phenomenon under investigation may simply require a different type of sensing, for instance by relying on the microphone or accelerometer. Second, participants may anticipate incoming alerts, potentially resulting in behavioural reactivity (Hormuth, 1986). However, this depends highly on the chosen sensor. Previous work shows both a high daily frequency (Böhmer et al., 2011; Ferreira et al., 2014) and revisitation habit (van Berkel et al., 2016) of smartphone usage. In our case, while study participants may realise that unlocking a device could lead to an incoming notification, the number of notifications is typically only a small subset of the total number of smartphone unlocks during

the day (16.5% in our study, considering only unlocks inside the time bin). Third, as discussed by Lathia et al. (Lathia et al., 2013), event contingent triggers may lead to design bias as the result of contextual dissonance. We partially offset this problem by not requiring participant input on every unlock but instead offer a schedule which limits and distributes the event contingent triggers over the duration of the day. Future work should explore how other methodological parameters such as questionnaire length, complexity of the task, or number of daily notifications affect the widely-used unlock event. For example, as smartphone usage sessions are usually short in duration (van Berkel et al., 2016), participants might be willing to answer a short questionnaire upon unlocking their phone. However, as the length of the questionnaire increases – the increased response rate as established in this study may disappear. As smartphone usage is diverse in nature, changing usage styles may also play an important role in answering ESM requests. For example, usage sessions driven by proactive usage (*i.e.*, the participant decided to use the phone without any trigger) rather than notification-driven usage may indicate that the participant has time available to answer a questionnaire.

As shown in Fig. 3, our results show that the smartphone-unlock trigger was able to obtain a high response rate across the duration of the day and thus cover the entire selected time period. These results are limited to only one event (phone unlock) and do not necessarily generalise to all possible configurations of an event contingent schedule. We highlight the opportunity for researchers to combine different contingency types in the design of their study. As in our study, a time schedule can be orthogonal to the event contingency — in our case to both prevent ESM notifications each time the phone was unlocked and to balance ESMs over the duration of the day. We term this specific event contingent configuration *interval-informed event contingency*. This type of configuration has proven useful before (*e.g.*, (Khan et al., 2009)), as it allows for sampling at the occurrence of a given event with reassurance that participants are not overburdened and questions are spread over duration of the day. Logically, researchers can randomise the presentation of notifications following an event. While this does not guarantee the presentation of notifications over the duration of the day, it does eliminate the possibility for participants to infer the next ESM notification. We term this notification strategy *signal-informed event contingency*. Both of these two contingency combinations ensure that predictability of an incoming questionnaire is considerably reduced or even removed, while still ensuring that participants receive the notification following the specified event. In the case of a smartphone unlock event, as was tested in this study, this results in an increased visibility to the participant.

6.4. Limitations

We recognise several limiting issues in the work. Our participant sample consisted solely of university students, while commonly used among researchers, this population's smartphone usage behaviour may be atypical. Second, the experience sampling methodology is also applied to ask participants to reflect on their emotional state. In this study, we measured an objective metric, smartphone usage, to answer our research questions on recall accuracy. Validation is required to support our findings in the context of quantifying emotions. Due to technical limitations, we were unable to distinguish between actual smartphone usage and the state of the smartphone screen. As a result, participants who did not turn off the screen of their phone but instead waited for the screen to turn itself off (*i.e.*, timeout time) may have underestimated their recorded smartphone usage. We obviated this problem by providing clear instructions to our participants during individual intake sessions. Furthermore, due to the within-subjects design any atypical smartphone behaviour of individual participants did not affect the comparison between conditions. Third, due to the design and duration of our study, we are unable to make any conclusions on the long-term effects of the tested notification schedules. Finally, the results of our event contingent condition only apply to the chosen event (*i.e.*, smartphone unlock) and cannot be generalised to other events. The event of smartphone unlock was chosen as it represents a commonly used event, rather than a representation of all possible event triggers. We expect that the use of an event that is less indicative of the participant's availability to using their phone (*e.g.*, a change in location) would result in a different outcome.

7. Conclusion

In this paper, we systematically compared three established ESM contingency strategies. We conducted a 21-day field study with 20 participants, combining mobile instrumentation and experience sampling to measure the response rate and recall accuracy of participants in answering questions about their smartphone usage. Our results show that *smartphone-unlock* triggering leads to an increased response rate. Albeit sending less notifications, and thus reducing participant strain, a higher absolute number of questionnaires was completed than in the other two conditions. The recall accuracy of our participants was considerably higher in the *smartphone-unlock* condition as well. We observe a reduced recall time for this condition as the result of frequent mobile usage sessions throughout the day. These results align with the premise of experience sampling, in which participants' accuracy suffers less from recall bias when compared to other methods due to a shorter 'reflection period' .

Our findings on response rate and recall accuracy quantify the effects of this methodological configuration for experience sampling studies. We show that the different scheduling configurations do effect participants' response rate and recall accuracy. This is an important observation for researchers utilising this method. The questions presented to our participants focused on objective and measurable information on their smartphone usage rather than for example the participant's emotional states, allowing us to obtain highly reliable ground truth data. Using a notification scheme based on the participants' smartphone usage, our results indicate that researchers can considerably increase both response rate and recall accuracy of their ESM studies while lowering participant burden. In future work, we aim to predict the accuracy and response rate of participants. This would allow researchers to either prime participants at the most opportune moment or take expected accuracy into account during data analysis.

Acknowledgement(s)

This work is partially funded by the Academy of Finland (Grants 276786-AWARE, 286386-CPDSS, 285459-iSCIENCE, 304925-CARE), the European Commission (Grant 6AIKA-A71143-AKAI), and Marie Skłodowska-Curie Actions (645706-GRAGE).

References

Andreassen, C.S., 2015. Online social network site addiction: a comprehensive review. *Curr. Addict. Rep.* 2 (2), 175–184. https://doi.org/10.1007/s40429-015-0056-9.

Andrews, S., Ellis, D.A., Shaw, H., Piwek, L., 2015. Beyond self-report: tools to compare estimated and real-world smartphone use. PLoS One 10 (10). https://doi.org/10.1371/journal.pone.0139004.

R. Bååth. 2016. bayesboot: An Implementation of Rubin's (1981) Bayesian Bootstrap. *R package version 0.2.1* https://cran.r-project.org/web/packages/bayesboot/index.html.

Baird, J.C., Lewis, C., Romer, D., 1970. Relative frequencies of numerical responses in ratio estimation. *Percept. Psychophys.* 8 (5), 358–362. https://doi.org/10.3758/bf03212608.

van Ballegooijen, W., Ruwaard, J., Karyotaki, E., Ebert, D.D., Smit, J.H., Riper, H., 2016. Reactivity to smartphone-based ecological momentary assessment of depressive symptoms (MoodMonitor): protocol of a randomised controlled trial. BMC Psychiatry 16 (1), 359. https://doi.org/10.1186/s12888-016-1065-5.

Barrett, L.F., Barrett, D.J., 2001. An introduction to computerized experience sampling in psychology. *Soc. Sci. Comput. Rev.* 19 (2), 175–185. https://doi.org/10.1177/089443930101900204.

van Berkel, N., Budde, M., Wijenayake, S., Goncalves, J., 2018. Improving Accuracy in Mobile Human Contributions: An Overview. In: Adjunct Proceedings of the 2018

ACM International Joint Conference on Pervasive and Ubiquitous Computing, pp. 594–599. UbiComp'18 Adj.. https://doi.org/10.1145/3267305.3267541.
van Berkel, N., Ferreira, D., Kostakos, V., 2017a. The experience sampling methods on mobile devices. *ACM Comput. Surv.* 50 (6), 93. 91-93:40. https://doi.org/10.1145/3123988.
van Berkel, N., Goncalves, J., Hosio, S., Kostakos, V., 2017b. Gamification of mobile experience sampling improves data quality and quantity. *Proc. ACM on Interact. Mob. Wearable Ubiquitous Technol.* 1 (3), 107. 101-107:121. https://doi.org/10.1145/3130972.
van Berkel, N., Luo, C., Anagnostopoulos, T., Ferreira, D., Goncalves, J., Hosio, S., Kostakos, V., 2016. A systematic assessment of smartphone usage gaps. In: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, pp. 4711–4721. https://doi.org/10.1145/2858036.2858348.
Böhmer, M., Hecht, B., Schöning, J., Krüger, A., Bauer, G., 2011. Falling asleep with Angry Birds, Facebook and Kindle: a large scale study on mobile application usage. In: Proceedings of the 13th International Conference on Human Computer Interaction with Mobile Devices and Services. ACM. pp. 47–56. https://doi.org/10.1145/2037373.2037383.
Caine, K., 2016. Local Standards for Sample Size at CHI. In: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems. ACM. pp. 981–992. https://doi.org/10.1145/2858036.2858498.
Chen, Y.-W., Cordier, R., Brown, N., 2015. A preliminary study on the reliability and validity of using experience sampling method in children with autism spectrum disorders. *Develop. Neurorehabil.* 18 (6), 383–389. https://doi.org/10.3109/17518423.2013.855274.
Conner, T.S., Lehman, B.J., 2012. Getting started: launching a study in daily life. In: Mehl, M.R., Conner, T.S. (Eds.), Handbook of Research Methods For Studying Daily Life. Guilford Press, New York, pp. 89–107.
Consolvo, S., Walker, M., 2003. Using the experience sampling method to evaluate Ubicomp applications. IEEE Pervasive Comput. 2 (2), 24–31. https://doi.org/10.1109/MPRV.2003.1203750.
Do, T.M.T., Blom, J., Gatica-Perez, D., 2011. Smartphone usage in the wild: a large-scale analysis of applications and context. In: Proceedings of the 13th International Conference on Multimodal Interfaces. ACM. pp. 353–360. https://doi.org/10.1145/2070481.2070550.
Epp, C., Lippold, M., Mandryk, R.L., 2011. Identifying emotional states using keystroke dynamics. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM. pp. 715–724. https://doi.org/10.1145/1978942.1979046.
Falaki, H., Mahajan, R., Kandula, S., Lymberopoulos, D., Govindan, R., Estrin, D., 2010. Diversity in smartphone usage. In: Proceedings of the 13th International Conference on Human Computer Interaction with Mobile Devices and Services. ACM. pp. 179–194. https://doi.org/10.1145/1814433.1814453.
Ferreira, D., Goncalves, J., Kostakos, V., Barkhuus, L., Dey, A.K., 2014. Contextual experience sampling of mobile application micro-usage. In: Proceedings of the 16th International Conference on Human-Computer Interaction with Mobile Devices and Services. ACM. pp. 91–100. https://doi.org/10.1145/2628363.2628367.
Ferreira, D., Kostakos, V., Dey, A.K., 2015. AWARE: mobile context instrumentation framework. *Front. in ICT* 2 (6), 1–9. https://doi.org/10.3389/fict.2015.00006.
Fischer, J.E., Greenhalgh, C., Benford, S., 2011. Investigating episodes of mobile phone activity as indicators of opportune moments to deliver notifications. In: Proceedings of the 13th International Conference on Human Computer Interaction with Mobile Devices and Services. ACM. pp. 181–190. https://doi.org/10.1145/2037373.2037402.
Froehlich, J., Chen, M.Y., Smith, I.E., Potter, F., 2006. Voting with your feet: an investigative study of the relationship between place visit behavior and preference. In: Dourish, P., Friday, A. (Eds.), UbiComp 2006: Ubiquitous Computing. Springer, Berlin Heidelberg, pp. 333–350.
Fuller-Tyszkiewicz, M., Skouteris, H., Richardson, B., Blore, J., Holmes, M., Mills, J., 2013. Does the burden of the experience sampling method undermine data quality in state body image research. Body Image 10 (4), 607–613. https://doi.org/10.1016/j.bodyim.2013.06.003.
Harbach, M., von Zezschwitz, E., Fichtner, A., De Luca, A., Smith, M., 2014. It's a hard lock life: a field study of smartphone (Un)Locking behavior and risk perception. *Symp. Usable Priv. Secur.* 213–230.
Hormuth, S.E., 1986. The sampling of experiences *in situ*. J. Pers. 54 (1), 262–293. https://doi.org/10.1111/j.1467-6494.1986.tb00395.x.
Hsieh, G., Li, I., Dey, A., Forlizzi, J., Hudson, S.E., 2008. Using Visualizations to Increase Compliance in Experience Sampling. In: Proceedings of the 10th International Conference on Ubiquitous Computing. ACM. pp. 164–167. https://doi.org/10.1145/1409635.1409657.
Ickin, S., Wac, K., Fiedler, M., Janowski, L., Hong, J.-H.H., Dey, A.K., 2012. Factors influencing quality of experience of commonly used mobile applications. *Commun. Mag., IEEE* 50 (4), 48–56. https://doi.org/10.1109/MCOM.2012.6178833.
Iida, M., Shrout, P.E., Laurenceau, J.-P.P., Bolger, N., 2012. Using diary methods in psychological research. In: Cooper, H. (Ed.), APA Handbook of Research Methods in Psychology. American Psychological Association, pp. 277–305.
Jeffreys, H., 1998. The Theory of Probability. OUP, Oxford.
Jones, S.L., Ferreira, D., Hosio, S., Goncalves, J., Kostakos, V., 2015. Revisitation analysis of smartphone app use. In: Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing, pp. 1197–1208. https://doi.org/10.1145/2750858.2807542.
Kapoor, A., Horvitz, E., 2008. Experience sampling for building predictive user models: a comparative study. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM. pp. 657–666. https://doi.org/10.1145/1357054.1357159.
Kaptein, M., Robertson, J., 2012. Rethinking statistical analysis methods for CHI. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI). ACM. pp. 1105–1114. https://doi.org/10.1145/2207676.2208557.
Kay, M., Haroz, S., Guha, S., Dragicevic, P., 2016a. Special interest group on transparent statistics in HCI. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI). ACM. pp. 1081–1084. https://doi.org/10.1145/2851581.2886442.
Kay, M., Nelson, G.L., Hekler, E.B., 2016b. Researcher-centered design of statistics: why bayesian statistics better fit the culture and incentives of HCI. In: Conference on Human Factors in Computing Systems. ACM. pp. 4521–4532. https://doi.org/10.1145/2858036.2858465.
Khan, V.J., Markopoulos, P., Eggen, B., 2009. An experience sampling study into awareness needs of busy families. In: Conference on Human System Interactions, pp. 338–343. https://doi.org/10.1109/HSI.2009.5091002.
Kikuchi, H., Yoshiuchi, K., Ohashi, K., Yamamoto, Y., Akabayashi, A., 2007. Tension-type headache and physical activity: an actigraphic study. Cephalalgia 27 (11), 1236–1243. https://doi.org/10.1111/j.1468-2982.2007.01436.x.
Krueger, A.B., Schkade, D.A., 2008. The reliability of subjective well-being measures. *J. Public Econ.* 92 (8-9), 1833–1845. https://doi.org/10.1016/j.jpubeco.2007.12.015.
Larson, R., Csikszentmihalyi, M., 1983. The experience sampling method. In: Csikszentmihalyi, M. (Ed.), Flow and the Foundations of Positive Psychology. Wiley Jossey-Bass, Springer Science + Business Media Dordrecht 2014, pp. 41–56.
Lathia, N., Rachuri, K.K., Mascolo, C., Rentfrow, P.J., 2013. Contextual dissonance: design bias in sensor-based experience sampling methods. In: International Joint Conference on Pervasive and Ubiquitous Computing. ACM. pp. 183–192. https://doi.org/10.1145/2493432.2493452.
Lee, M.D., Wagenmakers, E.-J., 2014. Bayesian Cognitive Modeling: A Practical Course. Cambridge University Press, Cambridge.
Lee, U., Lee, J., Ko, M., Lee, C., Kim, Y., Yang, S., Yatani, K., Gweon, G., Chung, K.-M.M., Song, J., 2014. Hooked on smartphones: an exploratory study on smartphone overuse among college students. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM. pp. 2327–2336. https://doi.org/10.1145/2556288.2557366.
Litt, M.D., Cooney, N.L., Morse, P., 1998. Ecological momentary assessment (EMA) with treated alcoholics: Methodological problems and potential solutions. Health Psychol. 17 (1), 48–52.
R.D. Morey, J.N. Rouder and T. Jamil. 2014. BayesFactor: Computation of Bayes factors for common designs. *R package version 0.9.12-2* https://cran.r-project.org/web/packages/BayesFactor/index.html.
Muukkonen, H., Hakkarainen, K., Inkinen, M., Lonka, K., Salmela-Aro, K., 2008. CASS-methods and Tools for Investigating Higher Education Knowledge Practices. In: International Conference on International Conference for the Learning Sciences. International Society of the Learning Sciences. pp. 107–114.
Naughton, F., Riaz, M., Sutton, S., 2015. Response parameters for SMS text message assessments among pregnant and general smokers participating in SMS cessation trials. *Nicot. Tob. Res.* 18 (5), 1210–1214. https://doi.org/10.1093/ntr/ntv266.
Niforatos, E., Karapanos, E., 2014. EmoSnaps: a mobile application for emotion recall from facial expressions. *Pers. Ubiquitous Comput.* 19 (2), 425–444. https://doi.org/10.1007/s00779-014-0777-0.
Poppinga, B., Heuten, W., Boll, S., 2014. Sensor-based identification of opportune moments for triggering notifications. *Pervasive Comput.* 13 (1), 22–29. https://doi.org/10.1109/MPRV.2014.15.
Reynolds, B.M., Robles, T.F., Repetti, R.L., 2016. Measurement reactivity and fatigue effects in daily diary research with families. Dev. Psychol. 52 (3), 442–456. https://doi.org/10.1037/dev0000081.
Ross, B.M., Engen, T., 1959. Effects of round number preferences in a guessing task. J. Exp. Psychol. 58 (6), 462–468.
Rouder, J.N., Morey, R.D., Speckman, P.L., Province, J.M., 2012. Default Bayes factors for ANOVA designs. J. Math. Psych. 56 (5), 356–374. https://doi.org/10.1016/j.jmp.2012.08.001.
Santangelo, P.S., Ebner-Priemer, U.W., Trull, T.J., 2013. Experience sampling methods in clinical psychology. In: Comer, J.S. (Ed.), The Oxford Handbook of Research Strategies for Clinical Psychology, pp. 188–210.
Shiffman, S., 2009. Ecological momentary assessment (EMA) in studies of substance use. Psychol. Assess. 21 (4), 486–497. https://doi.org/10.1037/a0017074.
Stone, A., Kessler, R., Haythomthwatte, J., 1991. Measuring daily events and experiences: decisions for the researcher. J. Pers. 59 (3), 575–607. https://doi.org/10.1111/j.1467-6494.1991.tb00260.x.
Wagenmakers, E.-J., 2007. A practical solution to the pervasive problems of *p* values. *Psychon. Bull. Rev.* 14 (5), 779–804. https://doi.org/10.3758/bf03194105.
Wheeler, L., Reis, H.T., 1991. Self-recording of everyday life events: origins, types, and uses. J. Pers. 59 (3), 339–354. https://doi.org/10.1111/j.1467-6494.1991.tb00252.x.
Zhang, X., Pina, L.R., Fogarty, J., 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. pp. 5658–5664. https://doi.org/10.1145/2858036.2858360.