Investigating Episodes of Mobile Phone Activity as Indicators of Opportune Moments to Deliver Notifications

Joel E. Fischer, Chris Greenhalgh, Steve Benford
The Mixed Reality Laboratory
University of Nottingham
Nottingham, NG8 1BB, UK
{jef, cmg, sdb@cs.nott.ac.uk}

ABSTRACT
We investigate whether opportune moments to deliver notifications surface at the endings of episodes of mobile interaction (making voice calls or receiving SMS) based on the assumption that the endings collocate with naturally occurring breakpoint in the user's primary task. Testing this with a naturalistic experiment we find that interruptions (notifications) are attended to and dealt with significantly more quickly after a user has finished an episode of mobile interaction compared to a random baseline condition, supporting the potential utility of this notification strategy. We also find that the workload and situational appropriateness of the secondary interruption task significantly affect subsequent delay and completion rate of the tasks. *In situ* self-reports and interviews reveal complexities in the subjective experience of the interruption, which suggest that a more nuanced classification of the particular call or SMS and its relationship to the primary task(s) would be desirable.

Author Keywords
Interruptions, interruptibility, receptivity, context-awareness, experience-sampling, ESM, studies in the wild

ACM Classification Keywords
H5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

General Terms
Human Factors.

INTRODUCTION
Interruptions have a profound impact on our attentional orientation in everyday life. Recent advances in mobile information technology increase the number of potentially disruptive notifications on mobile devices by an increasing availability of services. For example, as well as the more familiar notifications of direct communication attempts such as an incoming voice call or SMS, the user may also be notified of a friend nearby using a location-based service, or a status update on a social network service. A side effect is that the mobile device's disruptive potential is increased. This can be a particular problem for mobile users as their context is apt to change radically over time, which increases the possibility of an interruption being inappropriate.

This paper seeks to inform the design of systems that manage interruptions by detecting or predicting opportune moments for interruption delivery [17,18,1,13] so as to minimise the detrimental effects of interruptions. The identification of breakpoints in the primary task has been shown to approximate such moments in the laboratory [17,18,1]. However, it is in the relative chaos of everyday activity where we must routinely identify them if we are to apply this concept in practical systems.

Previous work has shown that the episodic nature of the human everyday experience [27] provides opportune moments for interruptions [1], and that transitions between physical activities are indicative of such breakpoints in mobile experience [13]. In this paper, we explore the hypothesis that episodes of mobile phone use indicate opportune moments to deliver notifications as the attention shifts to the mobile interaction episode at the beginning and away from it at the end. An opportunistic notification delivery mechanism similar to the *defer-to-breakpoint* interruption management strategy [21,20,18] would then defer interruptions until the end of an episode of mobile interaction, which might provide an opportune moment before the user attention shifts away from the device.

After expanding on the background and motivation for this work we present a naturalistic study to test this hypothesis, followed by a discussion of emergent issues from the qualitative follow-up. The experiment also allows us to study the effect of the interrupting task's workload and situational appropriateness on participants' responses.

BACKGROUND AND MOTIVATION
Interruption has been defined as "an externally generated randomly occurring, discrete event that breaks continuity of cognitive focus on a primary task" [3]. Even though interruptions may also be caused internally [21], research in interruption management usually focuses on effects of external interruptions and strategies to deal with them. We acknowledge that interruptions may be essential to the ways

MobileHCI 2011, Aug 30–Sept 2, 2011, Stockholm, Sweden.

we communicate in the workplace [23,16], and that we have refined strategies to account for interruptions in private life [25]. Yet, the potential detrimental effects of interruptions on task performance [4,5,7], task resumption rate [23] and emotional state [1] justify the need for systems that mediate interruptions in order to minimise their cost.

Opportune moments for interruptions

An influential body of work has associated opportune moments for interruptions to naturally occurring breakpoints in the primary cognitive task. Breakpoints reflect transient reduction in cognitive task processing.

Miyata and Norman [21] relate the user's memory load at different stages of the primary user task to the disruptiveness of interruptions. They posit that interruptions would be least disruptive if they occurred after evaluation and before forming a new goal. "If the change occurs at the conclusion of the current task or at a natural breaking point, then there is probably no difficulty." [21: 275].

Since then, a host of laboratory empirical work has largely validated their assumption. For instance, one study has found that the time to attend to an interruption was significantly longer when participants were interrupted between activities within a task, than when interrupted between tasks or before starting or after ending the task [19]. Another study has shown that the time to resume the primary task of programming a VCR after an interruption task was lowest when the interruption occurred right before a new task stage [22]. In a study that looked at the effects of interruptions by instant messaging, participants attended significantly more slowly to messages delivered during a cognitively more demanding task. The study concludes that the optimal design solution for a system that manages interruptions is to queue interruptions and deliver them at naturally occurring task completions [4].

More recently, interruption management has advanced by drawing on models of event perception from neuropsychology that posit that the brain structures our everyday experience into temporally bounded episodes [27]. The authors show that patterns of brain activity while watching a video match the pattern in which participants recalled events from the video on both a coarse and a fine level of event hierarchy [27]. An experiment showed that coarse and fine breakpoints occur between tasks and sub-tasks and that the more opportune moments for interruptions lay at coarse breakpoints [1].

More precise alignment of the workload of the primary task with opportune moments for interruptions has been achieved by using pupillary response as an indicator of workload in interactive tasks [17]. Then, opportune moments for interruptions could be predicted from interaction in real-time [18]. However, the need to monitor the user's primary task in order to predict breakpoints requires heavy instrumentation with both software [e.g. 17,18,14,15] and hardware sensors [16, 2] in laboratory environments. Few studies have taken on a more naturalistic approach or looked at the effects of interruption timing in mobile settings.

The progression of mobile technology facilitates ever more capable computing and sensing platforms, which provide opportunities to transfer approaches in context-aware interruption management from the stationary desktop environment [14,15,17,18] to mobile settings. For example, Ho and Intille report on a study [13] that tested the receptivity to interruption at transitions in physical activity. Body-worn accelerometers sensed transitions such as from sitting to standing/walking. Participant's self-reported receptivity at these breakpoints in physical activity was significantly higher than at random other times [13].

However, the temporal and spatial mobility of mobile device users, the large range of possible egocentric mobile device positions (e.g. in hand, in pocket, in bag, on desk etc.), and a desire to avoid invasive (e.g. body-worn) sensors make it extremely difficult to control or observe their primary task. This leaves an unanswered need to identify opportune moments routinely.

Inspired by the presented research on breakpoints, we present a naturalistic experiment that studies the effects of timing interruptions in relation to mobile phone activity, in particular making voice calls and received SMSs. We deliberately sacrifice control of the primary task for ecological validity. In addition to measuring reaction, we explore contextual richness through interviews and look at the impact of the interruption task, as follows.

Secondary task type influence

Whereas the nature of the primary task has had much attention in the literature, the effects of the nature of the secondary task, or the cognitive task or otherwise activity initiated by the interruption, on disruptiveness have been neglected. Latorella [20] develops a view of *interruption as a process*. The advantage of this model is that the interruption task itself is considered. The complexity of the interruption task in terms of information processing and memory demands has been reported to affect the disruptiveness [12], and the observation of two mobile professionals showed that in over 40% of their interruptions they engaged in a new activity as a result [23].

Clearly, an interruption may not only affect the original primary task, but it may become the starting point for a new primary task, effectively becoming a task switch. In addition, the attention demanding nature of mobility, where mobile HCI tasks may often compete with tasks such as orienting and navigating, may lead to fragmentation of mobile HCI into second-long bursts [24], indicating that length and attention resource demands of the interrupting task may play a significant role for mobile settings.

In this study, in addition to interruption timing, we look at the effect the type of interruption task has on the perceived workload of the interruption task and the resulting perceived burden to complete the task.

Operationalisation	Prior work
Time to attend to an interruption	[19, 5]
Time to resume the primary task	[22, 1]
Time on the primary task (completion time)	[1, 4]
Time on the interruption	[1]
Pupillary response	[17]
Forgetting the primary task goal	[5]
Self-reported receptivity rating	[13]
Self-reported emotional state	[1]

Table 1: Dependent measures to assess interruption timing in related experimental work.

RESEARCH QUESTION AND HYPOTHESES

To test the effects of interrupting after representative episodes of mobile interaction, the primary tasks of calling and reading SMS were chosen to test our hypotheses, because they are arguably among the most common examples of episodes of mobile interaction. This approach provides an alternative to the constraint of using bodily worn sensors in experimentation [13]. Our principal research question is as follows:

RQ: Does the end of an episode of mobile interaction represent an opportune moment for an interruption?

To answer the research question, a naturalistic experiment was designed that relies on an application on a mobile phone to infer opportunities for interruptions from phone activity. We formulate testable hypotheses for a mix of behavioural (H1,2,4) and self-reported (H3,5) dependent measures inspired by related work (see table 1):

H1: People will be quicker to accept the notification of an interruption at the end of an episode of mobile interaction than at random other times.

H2: People are significantly more responsive to interruptions at the end of an episode of mobile interaction than at random other times.

H3: People will perceive completing the task at random times as a higher burden than after episodes of mobile interaction, and people will rate the appropriateness of the timing of a notification after an episode of mobile interaction higher than at random other times.

Whereas H1-3 are aimed at testing the impact of the *timing* strategy (independent variable (IV) 1), H4-5 are aimed at testing the influence of the *task type* (IV2) of the interruption. Due to the dynamic nature of context whilst being mobile [24], we assume that attentional and cognitive demand of the interruption task (as indicated by perceived workload through NASA TLX assessment), and its social and situational appropriateness influence the perceived disruptiveness of an interruption and the completion rate of the task.

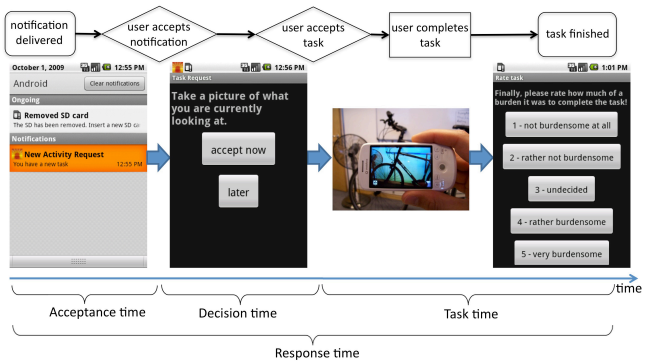

Figure 1. Temporal metrics (bottom) to analyse user behaviour (top decision flow) in phases of the interruption.

H4: Interruption tasks with a higher perceived workload, and/or situational inappropriateness are delayed longer before being started and have a lower completion rate.

H5: Interruption tasks with a higher perceived workload, and/or situational inappropriateness are perceived as more burdensome to complete and less appropriate when mobile.

These hypotheses may support the assumption, which inspired this experiment: that cognitive breakpoints are located at the endings of episodes of interaction.

EXPERIMENT DESIGN

In a 3x3 within-subjects design, we manipulated *task type* (multiple-choice, free-text and photo) and *timing* (random, opportune: after SMS, after call). We employed the experience-sampling method over a period of two weeks and post-hoc interviews and the NASA-TLX questionnaire to assess perceived workload of the tasks.

Methods

The experience-sampling method (ESM) has been designed to collect subjective assessments of experience *in situ,* over a variable period of time and where participants are locally dispersed [6]. In addition to self-reported ratings of the appropriateness of the timing and the burden to complete the task, we also collected behavioural data describing device usage as ground truth for parametric data analysis. To reflect and study different phases of the interruption process in more detail, we computed several temporal metrics from timestamps (see figure 1), again similar to dependent measures in related work (see table 1).

- First, *acceptance time* is the time between notification delivery and the participant's acceptance of the notification. So as not to convolute acceptance time by *task type*, a generic notification "new activity request" had to be clicked to accept the notification after pulling down a task bar equivalent to checking the SMS inbox.
- Then, *decision time* is the time between the task type being displayed to the participant and accepting the task.
- Then, the *task time* is the time the user spent on the task and the rating of the burden of the task, which concluded every task.

- Finally, the *response time* is the sum of the three times above, the interruption process from notification delivery to completed response.

After the study period, we conducted an assessment of the perceived workload of the task types by means of the NASA-TLX questionnaire. What is the perceived contribution of each workload factor to the overall workload by task and for which factors do the tasks differ? This also serves the purpose of a manipulation control for the intended task design; did the manipulation of the IV *task type* succeed?

To contextualise the quantitative findings, we concluded the study with semi-structured interviews around themes such as appropriateness and disruptiveness of task and timing, anecdotal experience of interruptions in context, and social implications of the interruptions.

Procedure, App(aratus), and Manipulation Control

After instructing participants about the procedure and gaining their informed consent we gave each participant a mobile phone running the experiment application and asked them to use it for two weeks with their own SIM card as their everyday phone. We told participants to attend to the experiment notifications how they would normally attend to their personal messages and discouraged attending to the notifications when deemed unsafe, e.g. when driving a car.

MActivityMonitor

The app *M(obile)ActivityMonitor* was designed for Android 1.5 and sent random and user activity triggered notifications by monitoring broadcast events such as when the user made a phone call or received an SMS. To monitor experiment progress remotely and to minimise the risk of data loss, collected data was only transmitted to a server when the phone connected to Wi-Fi, to minimise participants' costs.

The app would send around six SMS-style notifications to the participant's phone between 9am and 9pm. Three messages were sent at a (pseudo-) random time with at least one hour in between. Additionally, notifications were sent after the user had completed or attempted to make a phone call, and after they had opened a new text message from their inbox. An algorithm attempts to balance the distribution of notifications over the day so that the participants could not predict notification delivery. It determines if the historic pattern of opportune moments of the participant's previous days shows enough opportunities over the course of the day to defer to a later moment. In case the participants did not respond to the notification the notification timed out (disappeared) after 30 minutes.

Task design

When participants clicked on the generic notification "new activity request" (see figure 1), they were prompted to complete one of three tasks:

1. A multiple-choice task: "How good was the timing of the interruption of this task when you first noticed it?"
2. A free text task: "What are you doing at the moment?"
3. A photo task: "Take a picture of what you are looking at." (see figure 1).

The tasks were designed to impose varying attentional and cognitive demand. In keeping with the requirements for relatively short episodes of interaction on mobile devices [24] and repeated prompting in an ESM study [6], none of them should take longer than one minute to complete. Task order and balance was counterbalanced in order to avoid learning effects and predictability of task type.

In addition, tasks each had varying characteristics. The multiple-choice task (MC) was designed to be the quickest to complete so it would absorb attentional resources for the shortest time, but it did require some cognitive resources to reason about the appropriateness of the timing of the interruption. The free text task (FT) was designed to absorb the most attentional resources, as it required typing on the phone's virtual keyboard. It was probably most demanding cognitively too, as it required the participant to reflect on what they were doing at the moment and to compose it into a short statement. The photo task (PH) added an extra quality. Instead of interacting solely with the device screen, participants interacted with the environment *through* the device by being forced to select a motif/subject and take a photo. Hence, we expected this task to be most confounded by the social context of the participant's current setting.

Task design manipulation control: NASA-TLX

We assessed our task design by NASA-TLX. Participants rated the procedure's six factors of workload (mental demand, physical demand, temporal demand, performance, effort and frustration) after the study for each of the three task types. A repeated measures ANOVA showed that the mean aggregated workload differed significantly by task, with $F(2, 40) = 13.19$; $p < .01$. Pairwise comparisons by the Bonferroni procedure showed that the mean workload of the FT task (57.8) was significantly higher than the mean workload of the MC task (31.0; $p < .01$) and significantly higher than the mean workload of the PH task (40.3; $p < .01$). MC task and PH tasks did not differ significantly. This supports the intended manipulation of the IV *task type*.

In order to compare the amount of each of the six factor's contribution (e.g. temporal demand etc.) to the perceived workload of the tasks, a further analysis of the contribution of the individual workload factors to overall workload (see figure 2) was conducted. It showed a significant effect of the individual factor, with $F(5, 95) = 6.38$; $p < .01$ *Task type* also contributed significantly, with $F(2, 38) = 10.57$; $p < .01$. The interaction of the factors was not significant.

Pairwise comparison showed that the mean rating of *temporal demand* for the MC task (79.0) was significantly lower ($p < .01$) than for the FT task (205.8) and the photo task (149.3; $p < .01$). Also, *effort* of the FT task (185.3) was rated significantly higher ($p < .01$) than *effort* for the MC task (88.3) and significantly higher ($p < .05$) than the mean rating of effort for the PH task (125.0).

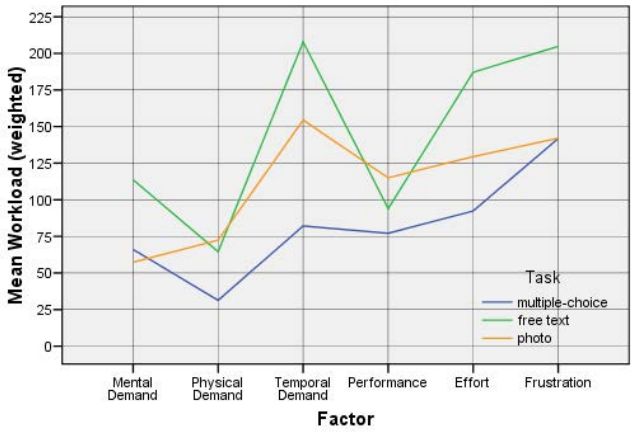

Figure 2. Workload contribution of factors by task type.

Participants

20 participants, (10 male, 10 female) were recruited through email lists and subsequent snowballing. The participants were between 21 and 48 years old (M=30, median = 27.5). 10 participants were postgraduate students, five were employed at the university, and three were employed in sales, one in health and one in the environmental sector. Participation was reimbursed with £20.

ANALYSIS AND RESULTS

Each of our 20 participants took part in our experiment for two weeks. In total, they received 2002 notifications and completed 1380 of the tasks (i.e. a response rate of 68.9%). Table 2 shows the distribution of the IVs across the notifications and responses.

To check if the distribution of messages (sent vs. responded) was biased by the timing strategy (random vs. opportune) by which they were sent, we conducted a chi-square analysis on the resulting contingency table. The analysis showed that the association between the distribution of messages and the timing strategy was significant, with χ^2 (1) = 11.7, *exact p* = .001. A ϕ value of .076 indicates a weak association.

The presence of an association thus established, we tested whether the messages sent at a random time were more likely not to get responded upon than the ones sent at hypothesised opportune times by analysing the distribution of the IV timing among the non-responses. Non-responses to notifications sent at a random time (377) outweighed non-responses to notifications sent at an opportune time (245). A chi-square goodness-of-fit test showed that this distribution of frequencies was significant, with χ^2 (1) = 28.01, *exact p* <.001. However, the opposite was not true.

The distribution of notifications that were responded upon sent at random times (723) and at opportune times (657) were not significantly biased towards random or opportune timing, with χ^2 (1) = 3.16, *exact p* = 0.08.

To summarise, participants were significantly more likely *not* to respond to a notification if it was sent at a random time than at an opportune time. However, there was no significant difference by timing of the notifications that did receive a response.

IVs			Task Type							
	Levels		*Multiple-choice*		*Free-text*		*Photo*		Total	
			s	r	s	r	s	r	s	r
Timing	*Random*		359	262 73%	370	260 70.3%	271	201 54.2%	1100	723 65.7%
	Op-por-tune	*Sms*	154	132 85.7%	134	101 74.5%	140	85 60.7%	428	318 74.3%
		Call	154	126 81.8%	168	124 73.8%	152	89 58.6%	474	339 71.5%
	Total		667	520 78%	672	485 72.2%	563	375 66.7%	2002	1380 68.9%

Table 2. Distribution of sent (s) and responded upon (r) notifications and response rates across levels of the IVs.

Furthermore, in order to test if participants were more likely to complete tasks of a certain type, we conducted a further chi-square goodness-of-fit test on the distribution of frequencies of *task type* among the responded upon notifications. The null hypothesis that the three tasks are equally likely to be completed proved to be significant and thus has to be rejected, with χ^2 (2) = 24.89; exact $p < .001$. Participants were 5.8% more likely to complete an MC task over a FT task, and 11.3% more likely to complete an MC task over a PH task, and still 5.5% more likely to complete a FT task over the PH task (see table 2). This supports the part of our hypothesis H4 that tasks with a higher workload and/or situational inappropriateness receive a lower completion rate.

Behavioural data

The four primary behavioural dependent variables *acceptance time*, *response time*, *decision time* and *task time* were computed from timestamps recorded each time when a participant went through the process of responding to a notification (see figure 1).

Whereas repeated-measures ANOVA would be the familiar choice of data analysis technique, it has a major drawback: it requires participants to have equal numbers of repeated measurements [11]. In a study where measurements are collected on the individual level the analyst would have to shrink all datasets to the size of the one with the fewest repeated measures, or to exclude sparse datasets entirely. In any case, this would affect a loss of richness of the data and may even lead to false conclusions. We adopt linear mixed models (LMM) as an alternative approach, which has been applied to HCI research before [26]. LMM is a disaggregate procedure which does not require equal amounts of measurements per subject and condition, and the variances do not need to be uniformly distributed, as it computes its estimates from maximum likelihood and not from ANOVA [11]. LMM have the advantages that variance in the data is not lost by averaging as in an aggregate procedure such as repeated-measures ANOVA, and that they account for the individual participant as a random effect, i.e. *participant* can be included as part of the model to reveal if individual differences have any significant effects on the result.

IV	DV	Test	Value	df	p-value
Timing (fixed effect)	Acceptance t.	F	104.59	2, 1526.9	< 0.001
	Response t.	F	73.71	2, 1374.9	< 0.001
	Decision t.	F	0.84	2, 1464.6	0.431
	Task time (t.)	F	1.33	2, 1362.0	0.264
Task type (fixed effect)	Acceptance t.	- Task type unknown at time of acceptance -			
	Response t.	F	13.03	2, 1360.8	< 0.001
	Decision t.	F	16.38	2, 1450.5	< 0.001
	Task time	F	875.43	2, 1353.7	< 0.001
Partici-pant (random effect)	Acceptance t.	Wald Z	2.51	-	< 0.05
	Response t.	Wald Z	2.44	-	< 0.05
	Decision t.	Wald Z	2.63	-	< 0.01
	Task time	Wald Z	2.76	-	< 0.01

Table 3: LMM results of behavioural effects of IVs.

Dependent measures were log-normalised to meet the assumption of normality. For *acceptance time*, the only fixed effect was *timing,* as *task type* was unknown to the participant at the time of accepting the notification. For the other three cases, the IVs *task type* and *timing* were modelled as fixed effects*; participant* was always included as a random effect. Note that in mixed models, Satterthwaite's approximation of degrees of freedom may yield non-integer denominator degrees of freedom [26].

In addition to results from LMM (table 3) we report pairwise comparisons from the Bonferroni procedure for significant effects. We use log-normalised values to compute significance levels but provide median values in seconds for the sake of readability and sense-making.

Acceptance time

Timing had a significant main effect on *acceptance time* (see table 3). In addition *participant* was a significant random effect. Further computation according to [11] showed that the percentage of variance in acceptance time explained by between-subjects effects was 5.9% in the employed default variance component model. Pairwise comparison showed that acceptance time was significantly higher (at the .01 level) when the notification was delivered at a random time (median (med.) = 36s) than when the notification was delivered after the participant had read an SMS (med. = 19s) or had made a phone call (med. = 10s). *Acceptance time* for the opportune conditions (SMS vs. call) also differed at the .01 level.

The result that acceptance time is significantly higher for random than opportune times support our hypothesis H1 that people attend to notifications on their mobile phones significantly quicker when they have just completed an episode of interaction.

Response time

Both the manipulation of *timing* and *task type* had a significant main effect on *response time* (see table 3). Again, *participant* was a significant random effect. The variance in *response time* attributable to *participant* was 6.2%. Pairwise comparison showed that *response time* for random timing of notifications (med. = 66s) was significantly higher (at the .01 level) than after reading an SMS (med. = 38s) or after making a call (med. = 29s). Response time did not differ significantly for notifications after reading an SMS or making a call.

Furthermore, *response time* was significantly lower (at the .01 level) for the multiple-choice (MC) task type (med. = 29s) than for the free-text (FT) task (med. = 53s) or the photo (PH) task (med. = 48s). As response time is a composite temporal metric (see figure 1), this may be explained by the significantly shorter *decision* and *task time* for the MC task, as discussed below.

Results regarding response time support our hypothesis H2. People's response time to notifications send after completing an episode of mobile interaction is significantly lower than to notifications send at random times.

Decision time

Task type had a significant effect on *decision time* (see table 3). *Timing* did not have a significant effect on *decision time*. *Participant* was a significant random effect. Individual differences explained 15.3% of the variance in *decision time*. Pairwise comparison showed that the mean decision time for the MC task (4s) was significantly lower (at the .01 level) than for the PH task (3s). The difference to the mean for the FT task (17s) was not significant for either task.

The fact that decision time was significantly lower for the MC task than for the other tasks completes the support of our hypothesis H4 that tasks with a higher workload and/or social inappropriateness are delayed longer before being started and have a significantly lower completion rate.

Task time

The results regarding task time further stress achievement of the goal of task design: tasks with distinct characteristics. With respect to task time, the tasks differed significantly.

Task type had a significant effect on *task time* (see table 3). *Timing* did not have a significant effect on *task time*. Again, individual differences contributed by *participant* showed to be a significant random effect. The variance in *task time* attributable to *participant* was quite high (20%). The *task time* spent on the MC task (med. = 3.4s) was significantly lower (at the .01 level) than the time spent on the FT task (med. = 17.6s) or the PH task (med. = 12.5s). Likewise, the difference between the time spent on the FT and the PH task was significant at the .01 level.

Self-reported data

We collected ratings of the participants' perception of the appropriateness of the timing of the notification by means of the MC task and ratings of the perceived burden of completing the task (see figure 1) at the end of every task. Both dependent measures were Likert scales with 5 ranks (burden: from 'not burdensome at all' to 'very burdensome'; timing: from 'very good' to 'not good at all'). To analyse the data, we obtained the median rating per participant per category, and conducted nonparametric Friedman tests for ordinal repeated measures.

Appropriateness of timing

Participants' self-reported *appropriateness of timing* did not differ significantly by *timing* (χ^2 (2) = 5.65, exact p = 0.068). The median rating for the random notifications was 'undecided', whereas for the conditions SMS and call it was between 'rather not good' and 'undecided'.

Burden of response

Participants' self-reported *burden of completing the task* did not differ by *task type* (χ^2 (2) = 4.51, exact p = 0.1) or by *timing* (χ^2 (2) = 0.46, exact p = 0.8). The median reported burden to complete the MC and the PH task was 'rather not burdensome', whereas for the FT task it was 'undecided'. The median burden for randomly timed notifications was 'undecided', whereas for after an SMS it was 'rather not burdensome' and for call it was between 'rather not burdensome' and 'undecided'.

Summary

The results of the nonparametric tests on the self-reported perception of the burden of completing the task and the perception of the appropriateness of timing do not support our hypotheses H3 and H5. The *timing* of the notifications did not make a difference in how much of a burden participants saw in completing the task, or in how appropriate they rated the timing of the notification (H3). Also, the *task type* did not influence the perception of the burden of completing the tasks in a significant way (H5).

Interview data

After the experiment 18 participants were interviewed in a semi-structured fashion. Interview responses were also coded for statistical analysis that we report here.

None of the participants felt that they could predict the timing of a notification in advance, but eight participants noticed the notifications were triggered by their phone activity and five of them correctly identified phone calls and SMS as triggering the notifications. In keeping with the results from the ESM, a Friedman test of the rankings of the appropriateness (best, medium, worst) and the disruptiveness (most, medium, least) of the three types of timing (random, SMS, call) during the interview failed to produce significant results.

Despite the statistical insignificance, the random condition was still ranked as the most disruptive condition 9 out of 16 times (6 times as least disruptive), and the least appropriate 7 out of 17 times (6 times as most appropriate). The *SMS* condition trumped the other ones in terms of appropriateness (most: 8, least: 5 out of 16 times) and least disruptiveness (least: 8, most: 2 out of 15 times).

In contrast to the *in situ* ratings of the burden to complete a task, but in accordance with intended task design and the findings on task workload and task time, participants reported in the interviews that they perceived the tasks as quite distinct from one another. When asked to rank the burden of the tasks in the interview again, the free text task was ranked as the most burdensome in 14 out of 17 cases (82%), the photo task was ranked in the middle with 10 mentions of medium burden (59%) and the multiple choice task was in 13 cases considered the least burdensome (76%). A Friedman test showed that burden was ranked significantly different for the tasks, with χ^2 (2) = 24.8, exact p < .001, Kendall's W = 0.73. Pairwise comparisons by Wilcoxon's test also showed the three tasks were all ranked significantly different from each other on a .01 level.

Summary and hypotheses

Effects of timing

Hypothesis H1 (quicker acceptance) is supported by the finding that *acceptance time* is significantly higher for random than opportune times. Hypothesis H2 (quicker completion) is supported by the finding that t*iming* had a significant effect on *response time*. In relation to hypothesis H3 (perceived appropriateness of timing) no significant effect was found in the self-reported or interview data.

Effects of task type

Hypothesis H4 (task delay and non-completion) is supported by the finding that the *task type* significantly affected the likelihood of completing the task, and the finding that the *decision time* was significantly lower for the MC task than for the other tasks. In relation to hypothesis H5 (perceived burden) no significant effect was found in the self-reported or interview data.

To summarise the results, our hypotheses related to the participants' behaviour were supported by the analyses, whereas the hypotheses related to the participants' self-reports were not supported by the analyses. In the following we unpack this disparity by discussing the findings from the interviews at the end of the study.

DISCUSSION

Now, we discuss findings with qualitative descriptions from interviews, illustrate with participants' quotes and relate issues back to wider concerns on interruption management.

Contextual sensitivity to the timeliness of interruptions

In the interviews there was substantial disagreement between the participants regarding the timing strategy of the notifications. This is reflected in the lack of significant support for hypothesis H3. Here we lay out some of the factors that participants reported as relevant.

Present-at-hand

The activities of making a call and reading an SMS were both characterized by holding the phone in hand. Participants mentioned this as being beneficial for dealing with the interruption task.

If I've already got it in my hand, in that position there's more chance of me responding to it. If it gave me time to put the phone down, then chance is less of me responding immediately, because I went to a different task.

In reference to the discussion of *present-at-hand* vs. *ready-to-hand* [9] it seems that to time the mobile interruption appropriately means to time it so that the device is still present-at-hand (i.e. in hand), but no longer ready-to-hand (i.e. in use). To exemplify, if the device still mediates another task such as calling or texting the notification is

likely to be disruptive. One might conjecture that the mobile phone no longer being ready-to-hand while still being present-at-hand may indicate an opportune moment to deal with new incoming messages.

Locus of control

However, as suggested by the insignificant differences in participants' *in situ* self-reports, phone activity related timing was not always preferred to random timing. One person explained their preference of the random condition because it more often correlated with them having "actively chosen to look" at their phone, thus raising issues with balancing control of awareness and interruption [15].

Task context

In addition, it may not have been the random timing *per se* that people preferred, but they may have found it less bad compared to situations where the phone activity related notifications actually interrupted their phone activity.

While I'm reading a text it is quite annoying, it's like a little child poking you while you're doing something else, but pretty good after I sent one. Random ones…changed over time. (…) After a phone call was quite interruptive because sometimes you want to call someone else, or you didn't reach the person and need to call back. Then came the message. Was quite interrupting.

Apparently, the notifications triggered by phone activity were more prone to interrupt phone activity that consisted of multiple sub-tasks; such as making several phone calls or exchanging several SMS in succession.

When I was sending multiple text messages…Just the tasks in general were more annoying then. When I was having almost like an IM style text conversation with someone who expects a reply.

This provides further support of the disruptiveness of a notification delivered when the device is ready-to-hand.

Task coherence

Further anecdotes from the interviews suggest that similarity in the activities of texting and replying to a task request may have made the SMS-triggered task notification more appropriate than a call-triggered notification.

Best is after a text because chances are you still have your keyboard out. With the phone you're just holding it to your ear, then you put it away.

The mode of interaction when composing an SMS and when responding to the notifications has similar physical requirements – having eyes and hands free. On the contrary, people often reported to use their mobile phones for calling when this requirement was violated, for example whilst driving or walking, or generally physically moving between activities, e.g. on their way to the car to confirm a meeting.

Behavioural vs. self-reported evaluation of timeliness

Whereas the quantitative analyses of the behavioural data support our assumption that opportune moments for interruptions are located at the endings of episodes of mobile interaction, the analysis of the self-reports fails to give further support. It appears that the benefit of the presented interruption delivery method may be on the side of the *sender*: The interruption gets tended to and responded upon significantly quicker, which suggest that endings of episodes of mobile interaction are indeed opportune moments to *deliver,* rather than to *receive* an interruption. The experience of receiving an interruption is subjective and situated amidst a host of environmental factors (e.g. arrangement of space [2], cultural [25] and organisational norms and practices [16]), psychological factors (e.g. mental workload [1], attentional focus [14]), and factors pertaining to the interruption (e.g. content [10] and sender [8]). This participant's reasoning about the experience of the interruption delivery method illustrates the difficulty in predicting opportune moments for interruptions purely based on sensing phone activity:

It was quite good when I got a text message that wouldn't require a response. It was a big difference there between if you wanted to carry on with another text message or wanted to make another call or if it was the end of a conversation. If it was at the end it was a quite good time and no problem at all, but if it was in the middle of a conversation or the middle of texting, if you're doing 2 to 3 texts, that didn't seem good. I suppose the end of a conversation, yeah, that's possibly good.

As the participant illustrates, the ongoing information exchange used by our application to piggyback onto, may have already informed the intention of carrying out a new activity, which is a good example of how an interrupting task can become the onset of a new primary task [23]. Hence, even though endings of interactional episodes were assumed to collocate with cognitive breakpoints, the interviews show that a) the breakpoint may have been missed and the user is already in a state of processing a subsequent task, or b) the breakpoint is at a sub-task and may still be considered more disruptive than a randomly timed notification.

Experience of the interruption tasks

We now relate the behavioural findings to the participant's perceptions of the interrupting task and its burden, and briefly survey the range of reported factors.

We found that interruption tasks with a higher workload are delayed longer. The NASA-TLX analysis (see figure 2) showed significant differences in *temporal demand* and *effort* for the three task types, which accords with some of the interview comments. The FT task was reported as taking the most time, and requiring more cognitive resources than the other tasks (paralleling the tasks' assessment of workload and actual *task time*). 12 people said they deferred the FT task when asked if the task type influenced if they accepted the task right away or if they deferred it. The MC task was generally preferred to the other two tasks for taking the "least effort", and being "quick". However it also required reflecting on the moment of interruption, implying a degree of mental demand:

It wasn't just the time the task took. It was a little more effort to sit and think about the MC task, whereas a photo task you didn't have to think, you could just take a photo of anything. Whereas for the MC, I had to put more thought into it.

However, the other workload factors show less differentiation and in some case contrasting ordering (e.g. *performance*, for which the FT mean workload is less than that for PH), or may be confounded by social, affective and environmental factors, which were also reported to affect the appropriateness of completing a task. For example, the free-text task was reported to be inappropriate while driving or walking, or difficult to do in sunlight because of diminished visibility of the screen.

Highlighting an affective aspect to the tasks, the photo task was described as being "fun", "interesting", "enjoyable", "engaging with the environment" and an "easy" task, which may well mitigate the perceived burden of competing the task. In contrast, the MC task did not allow creative completion, which may have made it less enjoyable.

As anticipated, we also found that the photo task introduces an element of social accountability, which affects the timing and completion rate of that task. The PH task differed from the other two tasks in that it did not only involve interaction with the device but with the environment *through* the camera; a fact that may have rendered the task socially inappropriate in some situations, as this participant points out:

Probably the photo task I would defer to a later time. Depending on who I was with. So if I sat in a meeting and it goes off… to take a photograph of the person that I'm looking at, it's not very socially acceptable, is it?

In addition, as opposed to real world interruptions, the study relied on fabricated content of the interruptions (the tasks). However, we know from other studies that factors such as the content [10] or the sender of the interruption [8] play a significant role in how receptive a person will be to the interruption.

In contrast to the analysis of behavioural data, self reports did not support that the burden of response also differed by task type. The interviews show that lack of significance may be due in part to the multi-dimensional character of *task burden*, including multiple workload factors, environmental factors, social accountability, and affect, which precludes the effective use of a single measure.

Practical considerations

Finally, we highlight some pragmatic issues and observations of the presented interruption strategy.

A challenge for systems that defer potential interruptions to an anticipated opportune moment is posed by the fact that the content of the interruption may be urgent or time critical to the recipient. Therefore, we assume that most people do not want a mediating service that interferes with their first-order communication, such as phone calls, SMS and email. Consequently, either system design must incorporate the difficult problem of robust semantic content analysis, or its application must be limited to non-time critical messages.

Alternatively, the presented interruption strategy could be applied to mobile applications that aggregate and deliver information from the user's second-order communication networks, such as social networks' *activity streams,* or other information sources the user has subscribed to, such as RSS feeds. The mechanism could also be used by services that deliver a dedicated user-experience or prompts for interaction, such as location-based services or games. In future work, we may investigate a prototypical application that mediates and manages *genuine* interruptions from the user's second-order communication network.

In summary, the presented strategy mediates interruptions by *deferring* them until an episode of interaction provides an opportune moment and messages are made available in an inbox-and-notification style, which has been called a *negotiated* strategy of coordinating interruptions [20], where the message is made available and the user tends to the message content at their own pace. In McFarlane's typology [20], our employed strategy represents a mix of *mediated* and *negotiated* interruption coordination.

CONCLUSIONS

Using a naturalistic study to test novel but simple interruption coordination based on sensing mobile phone activity we find that mobile users tend to accept and reply to notifications significantly more quickly after they finish an episode of mobile interaction than at random other times. This suggests that the presented strategy may be effective for applications that aggregate and deliver content proactively, or for systems that manage interruptions from the user's second-order communication network.

However, *in situ* self-reports did not show the subjective experience of activity-triggered timing to be superior to the random condition. The qualitative analysis exposes some of the situated complexities of interruption handling that can influence whether the phone activity-triggered notification is considered timely. In particular, three major task/activity contexts are revealed that influence perceived timeliness, i.e. whether at the moment of interruption the user

- a) just finished a task – physically but esp. cognitively – and is therefore available to an interruption (best case);
- b) has only finished a sub-task within a larger activity (intermediate case); or
- c) has already instituted, or started planning [21] a new task, which is therefore being interrupted (worst case).

On the one hand, findings a), b) support the assumption that cognitive breakpoints may be located at the endings of episodes of mobile interaction, due to parallel findings that breakpoints higher in the task hierarchy may be more opportune than between sub-tasks [1]. On the other hand, finding c) qualifies the assumption by uncovering that breakpoints and endings of mobile episodes do not always collocate, which means that opportune moments may have been missed or not reached yet. Distinguishing these cases is a question for future work, which may also be inspired by the consideration whether the device was still ready-to-hand, present-at-hand, or neither. Whereas if the device was

still ready-to-hand [9], e.g. to mediate a phone call, a notification would likely be perceived as disruptive, having the device still present-at-hand may provide a more opportune moment than when it has been put away.

With regards to the interrupting task, we find that its character has a significant effect on the time to decide whether to accept the task and the overall completion rate. We observe that the appropriateness of completing an interruption task depends not only on the factors that comprise workload (esp. *temporal demand, effort, frustration*), but also its situated social accountability (e.g. taking photos in a meeting), and cognitive and attentional demands (e.g. typing while walking) contribute to the burden of dealing with an interruption task, while affective factors may mitigate the sense of burden (e.g. sense of fun).

ACKNOWLEDGMENTS

We thank all our participants for taking part. Joel Fischer is supported by the EPSRC (EP/I011587/1).

REFERENCES

1. Adamczyk, P. D. and B. P. Bailey. If not now, when?: the effects of interruption at different moments within task execution. *Proc. CHI 2004*. ACM Press (2004).
2. Avrahmi, D., Forgarty, J., Hudson, S.E. Biases in human estimation of interruptibility: effects and implications for practice. *Proc. CHI 2007*, ACM Press (2007).
3. Coragio, L. Deleterious effects of intermittent interruptions on the task performance of knowledge workers: a laboratory investigation. *Unpublished doctoral dissertation*, University of Arizona (1990).
4. Cutrell, E. B., Czerwinski, M., Horvitz E. Effects of instant messaging interruptions on computing tasks. *Ext. abstracts CHI 2000*, ACM Press (2000).
5. Cutrell, E. B., Czerwinski, M., Horvitz E. Notification, Disruption, and Memory: Effects of Messaging Interruptions on Memory and Performance. *Proc. INTERACT 2001*, IOS Press (2001).
6. Csikszentmihalyi, M. , Larson, R., Prescott, S. The ecology of adolescent activity and experience. *Journal of Youth and Adolescence*, 6, 3 (1977), 281-294.
7. Czerwinski, M., Cutrell, E., Horvitz, E. Instant messaging: Effects of relevance and timing. People and computers XIV: *Proc. HCI 2000*, British Computer Society (2000).
8. Dabbish, L. A. Coordinating Initiation and Response in Computer-Mediated Communication. PhD thesis. Pittsburgh, PA, Carnegie Mellon University, 2006.
9. Dourish, P. *Where the action is.* MIT Press, Cambridge, MA, USA, 2001.
10. Fischer, J.E., Yee, N., Bellotti, V., Good, N., Benford, S., Greenhalgh, C. Effects of content and time of delivery on receptivity to mobile interruptions. *Proc. MobileHCI 2010*, ACM Press (2010).
11. Garson, G. D. Linear Mixed Models. Available at: http://faculty.chass.ncsu.edu/garson/PA765/multilevel.htm. Last accessed on 18.09.2010.
12. Gillie, T., Broadbent, D. What makes interruptions disruptive? A study of length, similarity, and complexity. *Psych. Research* 50 (1989), 243-250.
13. Ho, J., Intille, S.S. Using context-aware computing to reduce the burden of interruptions from mobile devices. *Proc. CHI 2005*, ACM Press (2005).
14. Horvitz, E.J., Apacible J. Learning and reasoning about interruption. *Proc. ICMI 2003*, ACM Press (2003).
15. Horvitz, E.J., Apacible, J., Subramani, M. Balancing awareness and interruption: investigation of notification deferral policies. *Proc. UM 2005*, Springer (2005).
16. Hudson, J. M., Christensen, J., Kellogg, W. A., Erickson, T. "I'd be overwhelmed, but it's just one more thing to do": availability and interruption in research management. *Proc. CHI 2002*, ACM Press (2002).
17. Iqbal, S.T., Adamczyk, P.D., Zheng, X.S., Bailey, B.P. Towards an Index of Opportunity: Understanding changes in mental workload during task execution. *Proc. CHI 2005*, ACM Press (2005).
18. Iqbal, S.T., Bailey, B.P. Understanding and developing models for detecting and differentiating breakpoints during interactive tasks. *Proc. CHI 2007*, ACM Press (2007).
19. Latorella, K.M. Effects of modality on interrupted flight deck performance: Implications for data link. *NASA Langley Technical Report Server* (1998).
20. McFarlane, D.C., Latorella, K.A. The Scope and Importance of Human Interruption in Human-Computer Interaction Design. *HCI* 17, 1 (2002), 1-62.
21. Miyata, Y., Norman, D. A. Psychological issues in support of multiple activites. In Norman. D. A. (ed.) User-Centered System Design. Hillsdale, NJ, Lawrence Erlbaum Associates (1986), 265-284.
22. Monk, C.A., Boehm-Davis, D.A., Tafton, J.G. The attentional cost of interrupting task performance at various stages. *Proc. HFES 2002*, Human Factors and Ergonomics Society (2002).
23. O'Conaill, B., Frohlich, D. Timespace in the workplace: dealing with interruptions. *Proc. CHI 1995*, ACM Press (1995).
24. Oulasvirta, A., Tamminen, S., Roto, V., Kuorelahti, J. Interaction in 4-second bursts: the fragmented nature of attentional resources in mobile HCI. *Proc. CHI 2005*, ACM (2005).
25. Tolmie, P., Crabtree, A., Rodden, T., Benford, S. "Are you watching this film or what?" Interruptions and the juggling of cohorts. Proc. CSCW 2008, ACM Press (2008).
26. Wang, H.-C., Fussell, S.R., Setlock, L.D. Cultural Difference and Adaptation of Communication Styles in Computer-Mediated Group Brainstorming. *Proc. CHI 2009*, ACM Press (2009).
27. Zacks, J.M., Braver, T.S., Sheridan, M.A., Donaldson, D.I., Snyder, A.Z., Ollinger, J.M., Buckner, R.L., Raichie, M.E. Human brain activity time-locked to perceptual event boundaries. *Nature Neuroscience* 4, 6 (2001), 651-655.