Final author version. Accepted on April 19, 2007, for publication in Personal and Ubiquitous Computing.

Considerate Home Notification Systems: A Field Study of Acceptability of Notifications in the Home

Martijn H. Vastenburg, David V. Keyson, Huib de Ridder

Faculty of Industrial Design Engineering, Delft University of Technology, Landbergstraat 15, 2628 CE, Delft, The Netherlands

{m.h.vastenburg, d.keyson, h.deridder}@tudelft.nl

Communicating author:
Martijn Vastenburg, m.h.vastenburg@tudelft.nl, tel. +31-15-2784960, fax +31-15-2787316

Abstract A field study in ten homes was conducted to understand what influences users' acceptability of notifications in the home environment. The key finding is that perceived message urgency is the primary indicator of acceptability of notifications in the home – if people think a message is urgent, they want the message to be shown immediately, regardless of what they are doing at the time of notification. The study also shows that the acceptability of low-urgent and medium-urgent messages could be improved by taking into account mental activity load at the time of notification. No effect of physical activity was found on acceptability. The results suggest that to improve the scheduling of notifications in the home, notification systems need a mechanism assessing both the message urgency and the mental activity load, whereas physical activity can be ignored. From a methodological point of view, it is difficult to measure acceptability of notifications in a realistic setting, given the need to balance experimental control with realistic context. The present paper suggests a way to introduce controlled notifications and subjective measurements of acceptability in homes.

Keywords: Interruptions, notification messages, considerate home environments, ubiquitous computing, user engagement

1 Introduction

Information and communication technology can help people stay up to date with events in the world. Traffic updates are sent to commuters using voice messages on mobile phones, new mail is announced using auditory signals on computers, and washing machines use irritating beeps to indicate the laundry is ready. Increasingly, people at home are connected to networked information services [1]. Medicine reminders, burglar alarms, and weather and news update services all notify users in their homes of possibly interesting events. These notifications can be helpful and appreciated, but they can also be inconvenient and distract the user. Since the number of information services present in everyday life appears to be growing, people might soon be overwhelmed with notifications.

To avoid overwhelming users with notifications at unwanted moments, notification systems need to be made aware of their environment. Ideally, notification systems should sense the state of its users and their environment, reason about the value of the notification message content, and decide the best time and form for presenting messages. An understanding of how the acceptability of notifications is influenced by contextual factors is needed to design considerate notification systems.

In the domain of task-oriented work environments, many results are available on how interruptions affect people and how systems can optimally choose timing and modality of notifications, as summarized in section 2. Both the objective impact of interruptions on task performance and the subjective acceptability of interruptions have been studied. However, it is not known if the results of these studies also apply to the home environment; the acceptability of notifications in the home could differ significantly from the work environment. To create a considerate mechanism for scheduling and presenting notifications in the home, we need to know what factors influence the acceptability of notifications by the user in the home. As a first step, the present study concentrates on two factors: engagement in activities and message urgency.

This paper is organized as follows. Section 2 describes related work. In section 3, the initial model of acceptability of notifications is described, including the expected results. The present study incorporates research methodologies for on-line registration of user experiences under natural circumstances. The resulting field study design, in which a laptop with notification and questionnaire software was placed in the houses of ten participants, is described in section 4. Section 5 describes the key findings from the study, while the remainder of the article is used to discuss the results and future steps.

2 Related work

2.1 Considerate computing

Notification systems provide access to, and draw user attention to, information secondary to the current user activities [2]. Because the primary activity is interrupted by the notification, task performance may decrease. Prior studies in the area of human interruptibility and notification systems have in common the goal of increasing task performance, e.g. [3-5]. Typical application domains are air traffic control and office work [6, 7].

Since human attention is a scarce resource, each notification message can be considered a potential threat to task performance. Attention can be viewed as a constrained resource that

can be traded for some utility [2, 3, 8]. The attentive user interface paradigm [9, 10] and the considerate computing paradigm [11] aim at avoiding overloading the user by adapting system behavior based on the sensed user attention focus. Attentive or considerate user interfaces generally calculate the cost in terms of user attention and the benefit in terms of subjective or objective performance factors, in order to predict acceptability and select the optimal timing of the interruptions.

The cost of notifications in terms of user attention can be reduced by adapting the presentation of messages to the user state. Presentation in the users' periphery minimizes the impact of interruptions on ongoing activities [2, 12]. In the case of aware notification systems, non urgent messages could be presented in the periphery of the user, while urgent messages could be presented in the foreground.

The cost of interruptions can also be reduced by adapting the timing of notifications. The cost in between tasks is lower, because supposedly people may be between evaluation of the last activity and formation of a new goal [13]. In a study on notification systems for mobile devices, scheduling of messages was linked to transitions in physical activity, under the assumption that changes in physical activity can be used as an indicator of user activity switches [14]. Notifications that were delivered at activity transitions were generally more easily accepted by the participants.

2.2 Measurement of impact

To assess and model the acceptability of notifications, a mechanism is needed to measure acceptability. Traditionally, studies of interruptions are based on objective measurements of effects of interruptions on task performance. More recently, subjective measurements have been used to measure acceptability of interruptions: video tagging [15, 16], rating scales [14], and self-reports by sticking up fingers [17].

Video tagging was used to study the interruptibility of office workers [15]. Participants performed 5 one-hour sessions in their offices. Sessions were taped on video, and system events were captured. After each session, subjects were asked to tag and assess the video. As a major advantage of post hoc video tagging, the setup does not interfere with user activities, resulting in more natural and realistic user behavior. It might however be difficult for participants to rate their interruptibility after the session, since users would have to recall situations based on the video.

Kern et al. used video tagging to study mobile interruptibility [16]. A series of 94 realistic everyday-life interruptions were captured on short videos using an actor. A group of 24

subjects were asked to annotate the video clips. The experiment focused on individual differences in interruptibility, therefore the researchers wanted all participants to rate the same situations. Results indicated for example that in judging interruptibility, women are more likely to consider their social context than men. Although video tagging made it possible to collect multiple user ratings for a single situation, it is not clear whether the participants were able to relate to the videos and judge the situations accordingly.

In work by Hudson et al. [17], a different approach was taken in examining the interruptibility of office workers; instead of post hoc rating the interruptibility, participants had to rate their interruptibility immediately. Four staff members were monitored for 14-22 working days. Audio and video recordings served as a source for 'simulated sensors', which registered, for example, the number of people in the room. Subjects were asked to rate their interruptibility approximately two times per hour. Subjects had to hold up fingers to indicate the rating. This way the disturbance caused by the alerts and responses was minimized. The subjects were asked to give an in-situ self-report after each alert ("beeper study").

In these studies, subjective data on the impact of interruptions were collected and used to construct computer models that help improve the coordination and presentation of interruptions. These computer models consider not only task performance, but a whole range of factors that the users themselves found relevant. Subjective measures seem appropriate for an exploratory study in the home environment; a range of relevant factors can easily be measured.

2.3 Notifications in the home

Notifications are not restricted to work environments, they also occur in the home. Interruptions have been studied before in the home environment [18], with a focus on task performance (preparing *punch* in the kitchen). It is however hard, if not impossible, to express the effect of interruptions and notifications solely in terms of task performance. As an example, a mobile phone will play a low-battery warning regardless of the current context. If the battery is empty in the middle of the night, the warning could result in a disturbed night's sleep [19]. Should the effect of this interruption be modeled only in terms of a decreased 'performance' in sleeping? In the home, apparently, other factors than the task performance factor come into play [20, 21].

As a first step towards developing aware home notification systems, an understanding of how people experience notifications at home is needed. In a previous experiment by Vastenburg, Keyson and De Ridder [22], which served as a pilot for the present study,

subjective data were collected and analyzed in an exploratory field study. For each inflicted interruption, participants were asked to describe their activities, rate their state and context, and judge the value and urgency of the notification messages as well as the acceptability of the interruptions. The results indicated a strong positive relation between the user-rated urgency of messages and the acceptability of interruptions. Unexpectedly, no significant relation was found between user engagement in activities and interruptibility. The degree of user engagement in activities is however expected to influence interruptibility; for example, when a user is working on a highly-urgent task, the acceptability of interruptions would be lower. The methodology used in the experiment might have blurred the concepts of user engagement in activities and message urgency. An adapted version of the procedure will be used in the present experiment, in which 1) the length of the experiment is extended from 1 to 3 evenings per participant, in order to investigate the 'novelty effect' of the prototype; 2) participants are instructed to rate their context *before* the interruption message is shown, in order to prevent an effect of the message on the user ratings of context variables; and 3) participants are reminded to consider both context and the notification message when assessing acceptability.

3 Initial model

The goal of the user study described here is to gain insight into attentional, social, and urgency aspects relevant to the acceptability and preferred timing of notifications in a living room setting. The knowledge gained will be used to reflect on an initial model for predicting the best time to present messages in a considerate home notification system. The initial model itself is based on the results of the study described in [22].

The initial model represents a cost-benefit tradeoff for notifications (figure 1); cost in terms of interrupted user activities, benefit in terms of the value of the notification messages. To get a better understanding of the mechanism, we asked users to rate their engagement in current activities, the urgency of the message, and the acceptability and preferred timing of the notification.

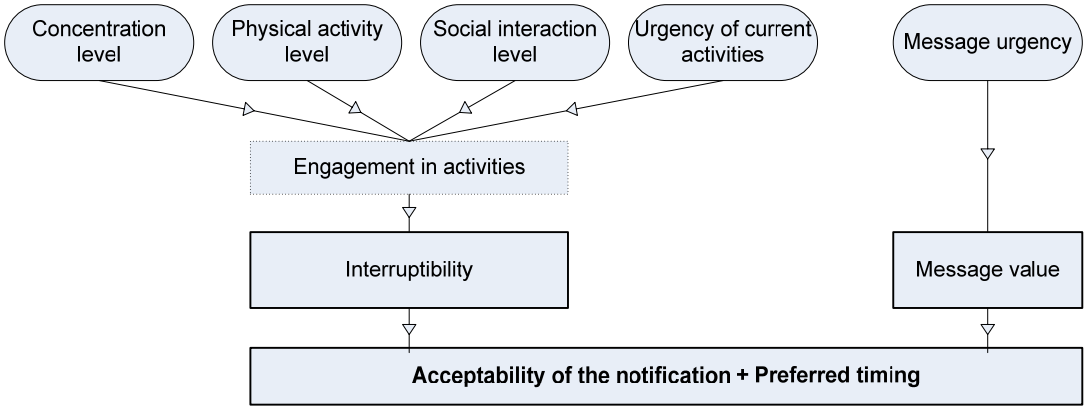

Fig. 1 Initial model of acceptability of notifications. The subjective general acceptability and preferred timing are linked to both user activity related factors and message urgency

In the model as shown in figure 1, *engagement in activities* indicates the involvement of the user in his/her current activities. Engagement is measured using subjective ratings of *concentration level*, *physical activity level*, *social interaction level*, and *urgency of current activities*. *Message urgency* is a subjective rating of the urgency of the notification message, which needs to be judged independent of the current user activities. *Acceptability of the notification* and *preferred timing* are subjective ratings of the general acceptability and preferred timing considering the message and the user activities at the time of interruption.

In the perspective of the taxonomy of McFarlane [4], there are eight factors underlying human interruptibility. The variation in interruptions used in our present study was restricted. The source of interruption is fixed to a computer; the method of coordination is set to scheduled interruptions, the method of expression is fixed to a plain depiction of the message on a computer screen, and the channel of conveyance is set to a computer screen. Accordingly, variation in acceptability ratings in our study is a consequence of four factors: 1) the individual characteristic of the person receiving the interruption, 2) the meaning of the interruption (i.e., the type of interruption, for example an alarm), 3) the human activity changed by the interruption, and 4) the effect of the interruption (the impact of the interruption, e.g., start a new activity).

3.1 Expected results

The acceptability of notifications is expected to be positively related to the user rated urgency of the notification message, but negatively related to the engagement of the user in his/her activities (figure 2). A more urgent message is expected to lead to a higher perceived benefit, and consequently to a higher acceptability of the notification. A higher level of engagement of

the user in his/her activities is expected to lead to a higher perceived cost, and consequently to a lower acceptability.

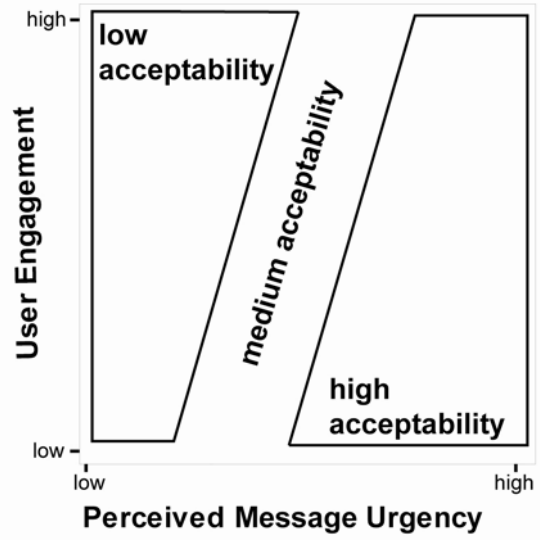

Fig. 2 Expected acceptability of notifications. Acceptability is expected to be positively related to the user rated urgency of the messages, but negatively related to the engagement in user activities

4 User study

Data was collected over 30 sessions (10 participants x 3 sessions). In each session, 12 notification messages were scheduled, for a total of 360 scheduled notifications.

4.1 Participants

Ten subjects (6 women, 4 men) participated in the study, age ranged from 25 to 56 (mean age 33 years). Participants were selected based on their home situation, being not living alone and no children at home. All participants were employed; nine out of ten had finished academic education.

4.2 Procedure

Test subjects participated at home. A laptop and a webcam were installed in the living room of the participant. The laptop was used to activate notification messages based on scenarios, as described in section 4.3 and 4.4, and to present the questionnaires. The webcam was used to log motion activity, and to capture the people present in the room at the time of interruption. Participants could delete the webcam pictures before returning the laptop at the end of the experiment. A microphone was used to log audio activity. The experimenter left the scene after placing the equipment and instructing the participants.

Participants selected three evenings within one week for the experiment. Participants were asked to do whatever they would do regularly, so user activity was not a controlled condition.

Since the study took about 18 hours per participant in total, a natural dispersion in user engagement was expected. Notifications were given approximately two times per hour. When a notification was activated, a bell sound was played, and the first part of the questionnaire had to be filled in at the laptop. Then, the notification message was immediately shown, and the second part of the questionnaire was presented. Participants were instructed to fill in the questionnaires themselves; partners were not allowed to do so. The bell sound and volume were not varied during the study. The study started when the subject arrived home from work or around 16:00 at a non-working day, and ended at bed-time.

4.3 Notification messages

A fixed set of 36 informational and alerting messages was created and classified beforehand by a panel consisting of three product designers. The panel was instructed to create a diverse set of notification messages, consisting of 12 *low-urgent*, 12 *medium-urgent* and 12 *high-urgent* messages. The panel defined and selected messages on the basis of plausibility, such that it was reasonable to expect that participants could relate to the messages in terms of their living situation. Table 1 shows a sample of the messages that were used in the experiment; the final selection and classification of messages was approved by all panel members.

Each single message might result in entirely distinct user ratings when presented to different participants or in different situations, based on the message structure, style, phraseology and by relationships between messages. In recognizing individual differences and subjective factors, emphasis was placed on creating a set of messages that would produce a wide spectrum of levels of perceived urgencies, from low-urgency to high-urgency across subjects, rather than attempting to create a set of messages that would be equally perceived by all subjects. In analyzing the results, message urgency was therefore based on perceived urgency rather than induced urgency.

Table 1 Sample of notification messages that were used in the experiment. The messages, originally in Dutch, were defined and classified by a panel of three product designers

Classification	Notification message
low-urgent	• To save energy, the thermostat should be set lower. • Don't forget to water the plants
medium-urgent	• The video tape needs to be returned to the video shop today. • The battery of your mobile phone is almost empty.
high-urgent	• Do not forget to take your medications now. • The smoke detector in the shed has detected smoke.

4.4 Notification scenarios and activation

Notification scenarios were created, one scenario per participant per day. The scenarios stated the order of the notification messages and the scheduling of the interruptions. The order of 36 pre-defined messages was randomized for each participant, using a combination of 4 low-urgent, 4 medium-urgent and 4 high-urgent messages for each day.

Notifications were only activated after motion was detected by the webcam, thereby reducing the chance of presenting notifications when no participant was present. Half of the notifications were activated immediately after motion was detected, and the rest of the notifications were activated five minutes after motion was detected. The five minute delay, which was scheduled randomly, was used to distribute the moment of interruption in relation to motion activity. An extra delay of 10 up to 30 minutes was scheduled between notifications.

At the time of activation, a bell sound was played to indicate a new message. Participants were instructed to fill in the questionnaire immediately, even if the timing was inconvenient. If the participant did not respond within 5 minutes, which only occurred when no participant was present at the time of notification, the questionnaire was removed, and a new notification was scheduled. Participants were asked to end the session only when they went to bed; when the session was closed, all remaining non-activated messages were skipped.

4.5 Questionnaire

A questionnaire was used to collect subjective data. The questionnaire consists of 3 parts. After hearing the notification bell, participants were asked to rate their engagement in activities (figure 3). Then, the notification message was shown (figure 4). In part 2, participants were asked to rate the message, without considering the current activities (figure

5). In part 3, participants were asked to rate the acceptability and the preferred timing of the notification, considering the message and the activities at the time of the bell. All questionnaire items are directly related to the model depicted in figure 1.

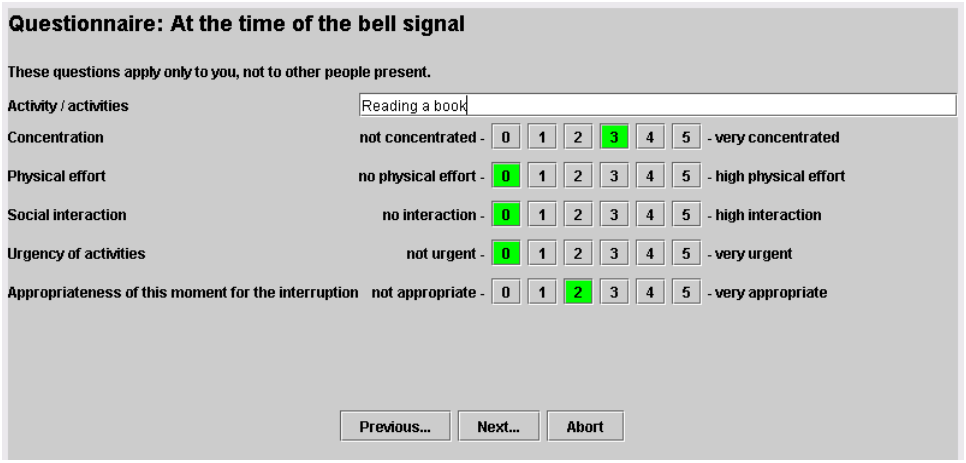

Fig. 3 Questionnaire part 1 (originally in Dutch). In the first part of the questionnaire, participants had to rate their current activities, before the notification message was shown

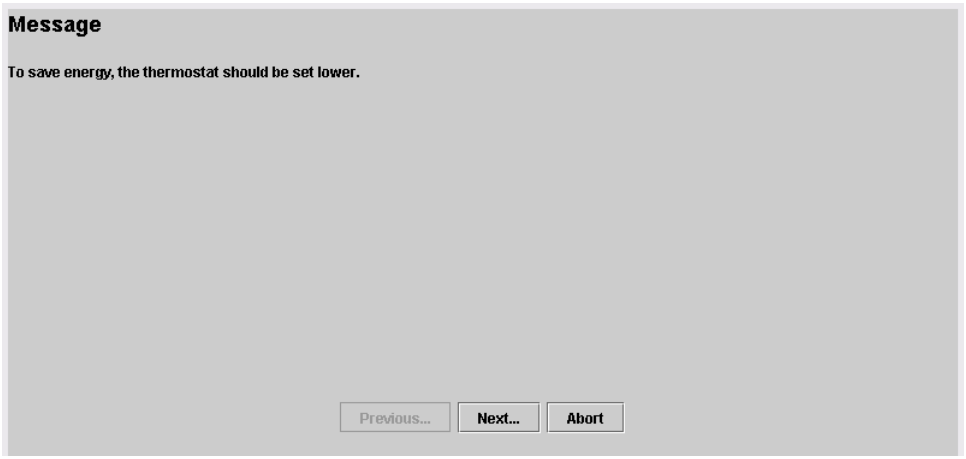

Fig. 4 Presentation of the notification message (originally in Dutch). The message was shown after participants had rated their current activities

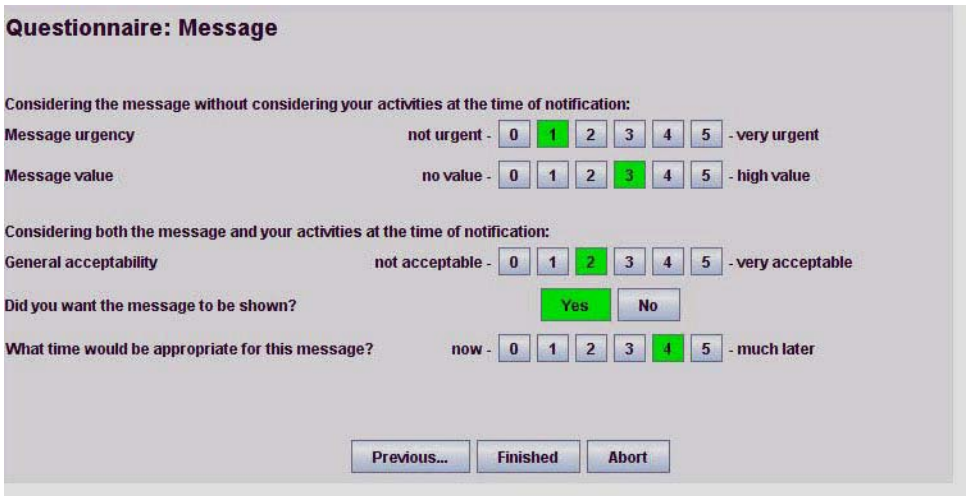

Fig. 5 Questionnaire part 2 and 3 (originally in Dutch). Participants were asked to rate the urgency and value of the message in part 2. In part 3, participants were asked to rate the acceptability and preferred timing of the notification

Preferred timing was presented using two sub-questions: participants had to indicate if they wanted the message to be shown at all, and, if the answer was positive, users were asked to indicate the preferred timing on a scale from *now* to *much later*.

In the previous study [22], the notification message was shown before the questionnaire. The user ratings of engagement in activities might have been influenced by the relevance of the notification message; when a highly urgent message ("Smoke has been detected in the shed.") was shown, participants might have rated their current activities as less urgent. For the present experiment it was decided to adapt the questionnaire display and first ask users to rate their degree of engagement in the current activities, before showing a notification message.

5 Results

5.1 Factors of acceptability

To understand the factors underlying acceptability and preferred timing, and to consider the interrelationships between the items of the initial model as depicted in figure 1, a factor analysis was conducted on the results of the questionnaire. A total number of 231 completed questionnaires were collected in the field study. Table 2 shows the results of the factor analysis with principal components using SPSS [23] with Varimax rotation and Kaiser Normalization, all component loadings <.20 are suppressed.

Table 2 Rotated component matrix using Varimax rotation, component loadings <.20 suppressed. The four components that emerged from the factor analysis were labeled as C1: message urgency, C2: user engagement in activities, C3: social interaction and C4: physical activity

			Component			
			C1	C2	C3	C4
User engagement in activities	Q1	Concentration		.80		
	Q2	Physical activity				.92
	Q3	Social interaction	-.23	.35	.49	-.30
	Q4	Urgency of activities		.74		.40
	Q5	Interruptibility		-.85		
Message urgency	Q6	Message urgency	.94			
	Q7	Message value	.94			
Acceptability	Q8	General acceptability	.86	-.33		
Preferred timing	Q9a	Timing A	.78		.53	
	Q9b	Timing B	.21		.88	
% of Variance			32%	22%	14%	11%

The four components that emerged from the factor analysis were labeled as *message urgency (C1)*, *user engagement in activities (C2)*, *social interaction (C3)* and *physical activity (C4),* based on the factor loadings as depicted in table 2. These four components explained a cumulative percentage of variance of 79%.

Variation in *general acceptability (Q8)* could be explained using only the components *message urgency (C1)* and *user engagement in activities (C2)*. The factor analysis shows high factor loadings of *message urgency (Q6)*, *message value (Q7)* and *general acceptability (Q8)* on component 1, which is consistent with our expectations; a positive correlation between message urgency, message value and general acceptability is expected. Also, as expected, general acceptability is negatively influenced by *user engagement in activities (C2)*.

5.2 General acceptability

Figure 6 shows the acceptability of notifications plotted against message urgency (C1) and user engagement in activities (C2) that were shown relevant to the acceptability in the factor analysis. Each of the 231 items in the graph represents a single interruption of a single subject. The notifications are labeled by acceptability level as rated in Q8, reduced to 3 levels (low=0/1, medium=2/3, high=4/5). Messages that are rated high-urgent, as shown by the horizontal axis, tend to be highly acceptable. Medium-urgent messages tend to be moderately acceptable, and low-urgent messages tend to be unacceptable. The factor analysis also showed a negative factor loading of acceptability (Q8) on *user engagement in activities* (C2). This relation can also be seen in the figure; acceptability is rated higher for low levels of engagement in activity than for high levels.

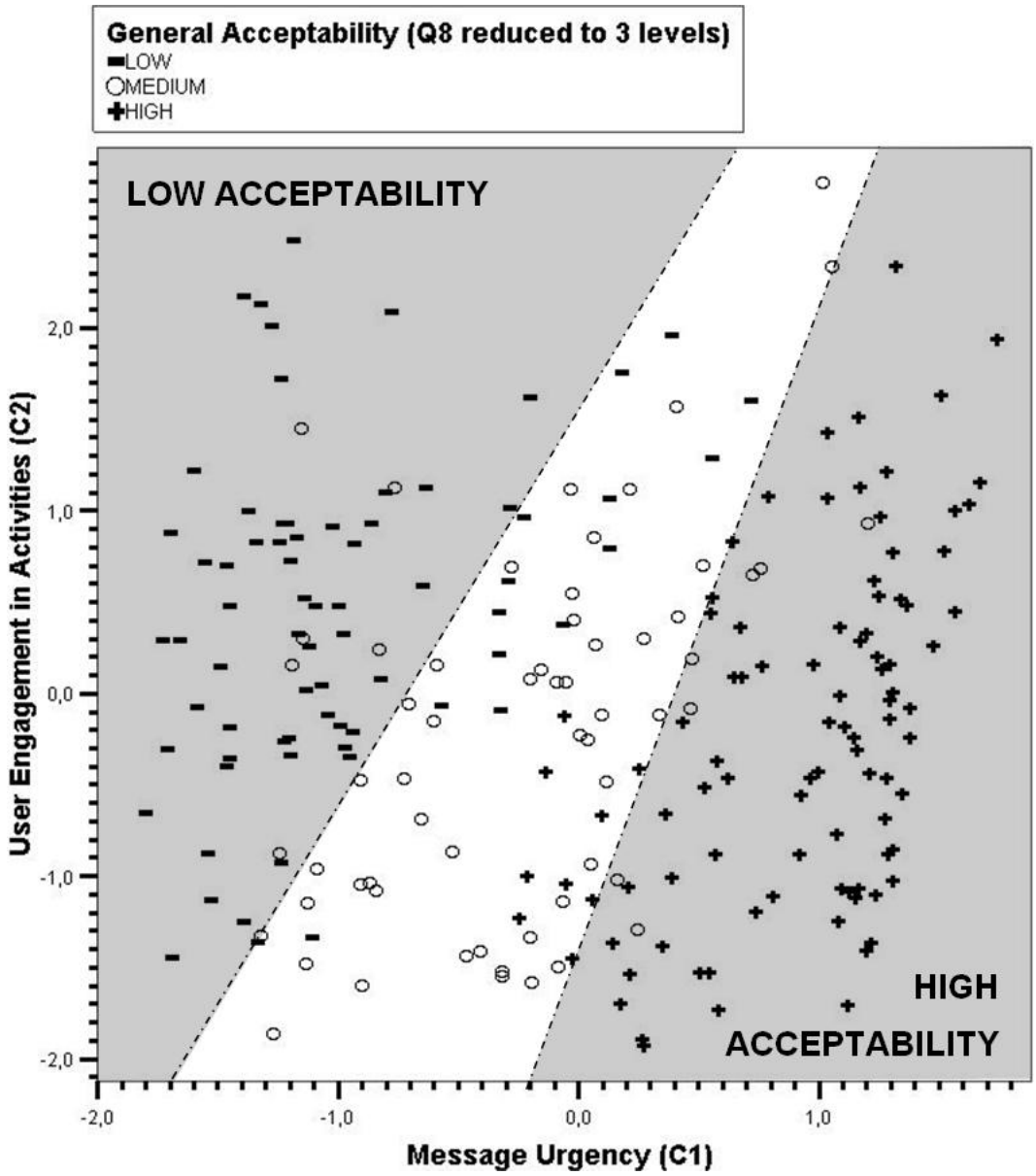

Fig. 6 Subjective acceptability ratings plotted against the acceptability-related components from the factor analysis. Messages with high urgency tend to be highly acceptable ("+"), the acceptability of medium and low-urgent messages was more difficult to predict. The three acceptability classes separated in the figure by the dashed lines result from a linear discriminant analysis: *C2=1.99C1 + 1.51* (low/medium) and *C2=3.28C1 - 1.78* (medium/high). The resulting classification resembles the expected outcome as depicted in figure 2

The experimental outcome resembles the expected outcome as depicted in figure 2. Using linear discriminant analysis in SPSS [23], the data set was classified into three clusters based on the ratings on Q8. Based on these clusters, 84.0% of the cases could be correctly classified. The clustering was accurate for the high-acceptability cases (90,7% correct), while 76.7% of the medium-acceptable cases were correctly classified, and 81,0% of the low-acceptable cases.

The dashed discriminant lines depicted in figure 6 are roughly parallel, which suggests that general acceptability can be described by means of a simple linear model. Therefore, multiple linear regression was applied to investigate a possible linear relationship between Q8 (general acceptability) and components C1 and C2. The multiple linear regression showed that a significant proportion (84%) of the variance in general acceptability could be accounted for by a linear combination of C1 (message urgency) and C2 (user engagement in activities): *general acceptability=1.58C1 - 0.60C2 + 2.74* (R^2=.84, $F_{(2,228)}$=604.3, p<.001). Figure 7 shows the subjective acceptability ratings plotted against the acceptability related components from the factor analysis, combined with the linear model.

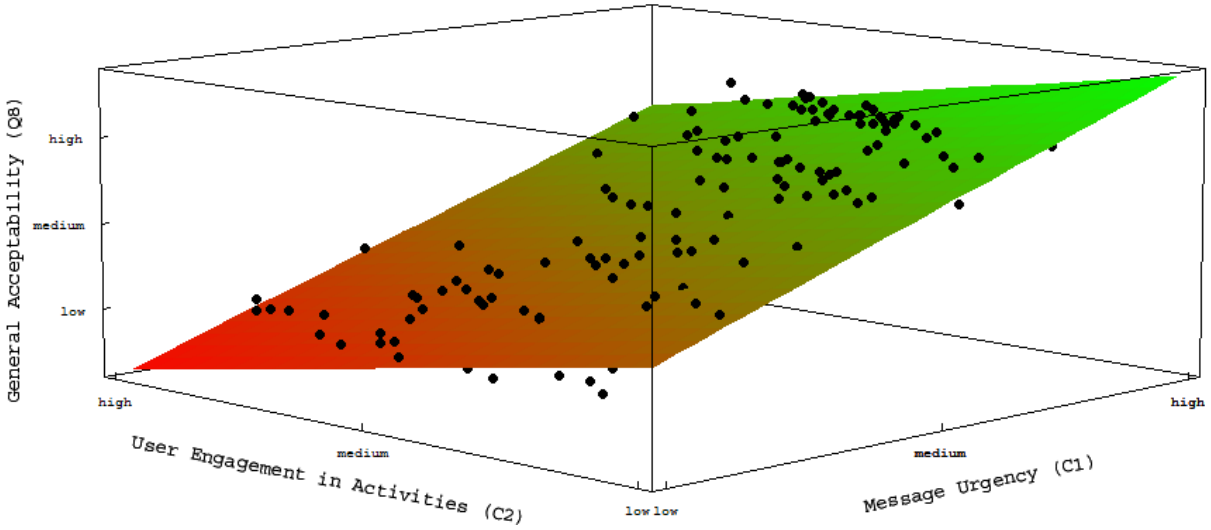

Fig. 7 Multiple linear regression of subjective acceptability ratings against the acceptability-related components from the factor analysis. General acceptability can be described using a simple linear equation: Q8=1.58C1 - 0.60 C2 + 2.74 (R^2=.84, $F_{(2,228)}$=604.3, p<.001)

5.3 Preferred timing

The questionnaire results show participants wanted to see all high-acceptable messages immediately, medium-acceptable messages should be postponed, and low-acceptable messages should not be presented at all. Considering the scheduling issue, we asked participants two questions: 1) did you want to see the message (Q9a), and if so, what would be the best time to present the message (Q9b). The results of Q9a and Q9b are combined in table 3; negative answers to Q9a are represented by a 'never' score in the table. In case of a positive answer to Q9a, the subjective preferred timing is listed in the table, ranging from *now* to *much later*. A significant relationship was found between acceptability (Q8) and preferred timing (Q9) (χ^2=225.64, df=30, p<0.0005). This relation suggests that a lower acceptability results in a higher desire to postpone, or even to skip, notification messages, and vice versa.

Table 3 Cross-tabulation count of general acceptability (Q8) and preferred timing (Q9). The table shows the relation between preferred timing and acceptability

		Preferred timing (Q9)							Total
		now (0)	(1)	(2)	(3)	(4)	much later (5)	never	
General Acceptability (Q8) of the notification message	0 (not acceptable)	0	0	1	1	0	2	22	26
	1	0	2	2	3	0	3	13	23
	2	1	2	3	6	1	1	5	19
	3	8	2	10	3	1	1	2	27
	4	22	12	0	2	1	0	0	37
	5 (very acceptable)	52	4	0	1	1	0	0	58
Total		83	22	16	16	4	7	42	190

5.4 Subjective ratings of message urgency

The subjective scores on message urgency range from low-urgent to high-urgent (figure 8), which reflects the effort of the panel of product designers in creating a diverse and plausible set of notification messages. The horizontal axis shows the induced message urgency, i.e., the message urgency which was pre-classified by the panel. The vertical axis shows the message urgency component (C1), i.e., the subjective message urgency scores of the participants. Participants were consistent in their urgency ratings for highly-urgent messages. A higher degree of variation was observed for low-urgent and medium-urgent messages. Alarm messages were rated highest on urgency, including "*Somebody is touching your car.*" and "*The bath is running over.*".

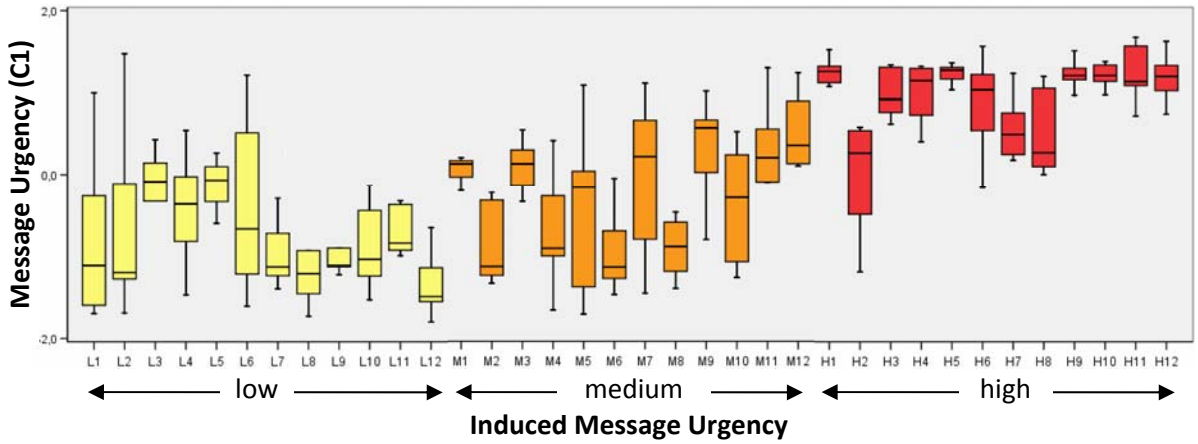

Fig. 8 Subjective message urgency of the individual notification messages. The horizontal axis shows the 36 messages used in the experiment, i.e., 12 pre-classified as low-urgent ("low"), 12 medium-urgent ("medium") and 12 high-urgent ("high") messages. The vertical axis shows the message urgency ratings by the participants, outliers are hidden. Inter-subject variability tended to be lower for messages judged as high-urgent. A higher degree of variation was observed for messages not judged as high-urgent

A one-way ANOVA indicated a significant effect of induced message urgency on subjective message urgency ($F_{(2,228)}$=123.3, p<.0005), suggesting that the classification of the panel and the ratings participants are similar. The subjective message urgency for the high-urgent messages (M=.94; SD=.53) was significantly higher (t=-11.3; p<.005) than for the medium-urgent messages (M=-.30; SD=.80). Likewise, the subjective urgency for the medium-urgent messages was significantly higher (t=-3.37; p<.005) than for the low-urgent messages (M=-.73; SD=.75).

5.5 Engagement and activities

No relation was found between user activities and interruptibility. Participants were instructed to enter all activities at the time of interruption. Examples of typical activities people were engaged in can be seen in table 4. Only in 4 out of 231 cases, multiple activities were mentioned. Similar activities (watching TV, on the phone) appear in varying degrees of interruptibility. Consequently, no one-to-one relation between user activities and interruptibility can be defined.

Table 4 Typical user activities clustered by interruptibility

Interruptibility (Q5)	Activities
HIGH	watching TV, just entered the room, watching commercial break, finishing phone conversation, closing the window
MEDIUM	cooking, using computer, watching TV, listening to music, brushing teeth, cleaning the house
LOW	working, going to the bathroom, watching soccer, on the phone, cooking, welcoming guest, eating

In the experiment, the user activities were not controlled. A variation in activities, and consequently in user engagement in activities (C2) levels, is expected. A normality test confirmed that the distribution of C2 levels resembles a normal distribution (Kolmogorov-Smirnov, sign. 0.25); the variation in user engagement in activities in the experiment resembled a normal distribution.

6 Discussion

6.1 General acceptability

High-urgent messages were found to be acceptable, no matter what. Based on the results of the pilot study [22], this dominating effect of message urgency was expected. Whereas the effect of user engagement in activities was not clear in the pilot study, the present study does show that acceptability of low-urgent and medium-urgent messages may be improved by creating a system that is aware of user activities, and that adapts the presentation and timing to the activity context. Based on the results of the present study, one might conclude an effective way to predict acceptability of notifications is to consider only message urgency and user engagement in activities.

Figure 9 presents an updated model of acceptability being a simplified version of the initial model (figure 1). Physical activity level, social interaction level and urgency of user activities did not correlate to acceptability and preferred timing; therefore these factors have been removed from the model. Concentration level (Q1) has been generalized to attention level. Message urgency (Q6) and message value (Q7) were highly correlated; these factors have been combined in the updated model.

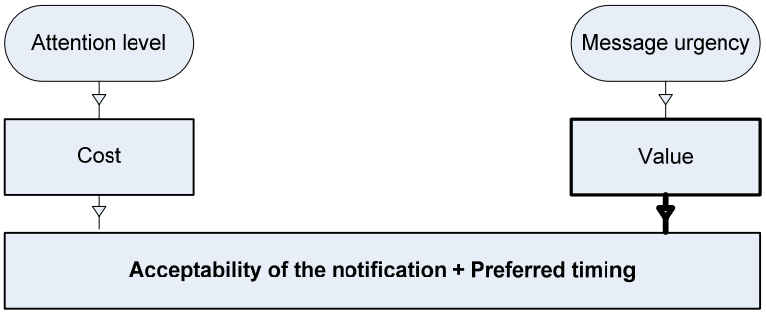

Fig. 9 Updated model of acceptability of notifications. The subjective acceptability and preferred timing are linked to the attention level and the perceived message urgency. The bold arrow indicates message urgency to be the primary indicator of acceptability and preferred timing

6.2 Preferred timing

The preferred timing for presenting messages appeared to be directly related to acceptability (table 3). High-acceptable messages were requested to be shown immediately, while non acceptable messages were to be postponed or not shown at all. Apparently the preferred timing depends on acceptability: immediate interruptions are accepted for highly acceptable messages only; low and medium acceptable messages should be presented at a later point in time.

For 42 out of 190 notifications, participants indicated they did not want the message to be shown at all. Based on the questionnaire data and the exit interviews, two possible explanations come to mind. First of all, the notification system in the experiment did not adapt the presentation style to the content of the message. A realistic notification system might adapt the presentation style to the messages; non-urgent messages could be presented in a non-obtrusive style. Since adaptation was not possible in the experiment, participants might have selected messages not to be presented at all. A second explanation might stem from the experimental setup. Some participants indicated they found it hard to empathize with the messages. For example, after seeing the message "Reminder: waste paper will be collected tonight", one participant said waste paper was not collected in his neighborhood at all, so he rejected the message.

6.3 Reflection on experiment design

In the previous study [22], no significant relation between user engagement and acceptability of notifications was found. In the present experiment, the length of the experiment was extended from 1 to 3 evenings per participant, the questionnaire was redesigned by asking users to first rate their degree of engagement before presenting a notification message, and participants were reminded to consider both context and the notification message when assessing acceptability.

The effect of extending the experiment from 1 to 3 evenings per participant can be seen in table 5 showing a linear regression between Q8 (acceptability) and components C1 (message urgency) and C2 (engagement) per evening. The table shows that the effect of C2 on Q8 decreases in time, from -.71 on day 1 to -.51 on day 3, whereas the effect of C2 remains significant. A possible explanation for this reduction in time could be that user engagement was boosted on the first night because of the novelty of the system. On day 2 and 3, when participants get acquainted with the system, the novelty effect diminishes. The changes in user ratings in time underline the need for longitudinal studies, and confirm our choice to extend the experiment.

Table 5 Linear prediction equations for general acceptability based on multiple linear regression, per day and overall. Q8 (general acceptability) could be accounted for by C1 (message urgency) and C2 (user engagement in activities)

	general acceptability	
day 1	Q8=1.57C1 - 0.71C2 + 2.83	R^2=.84, $F_{(2,74)}$=187.0, p<.001
day 2	Q8=1.57C1 - 0.54C2 + 2.65	R^2=.83, $F_{(2,79)}$=198.6, p<.001
day 3	Q8=1.63C1 - 0.51C2 + 2.71	R^2=.86, $F_{(2,69)}$=215.2, p<.001
overall	Q8=1.58C1 - 0.60C2 + 2.74	R^2=.84, $F_{(2,228)}$=604.3, p<.001

In measuring the effect of mental activity load on acceptability, the design of the questionnaire was crucial. In the pilot study [22], notification messages were shown *before* users were asked to rate their mental activity load. Also, the presentation of messages was varied; new messages were signaled using alternately a shrill and a soft sound. In that study, we were unable to measure the effect of mental activity load, probably because the notification message biased people in rating their mental activity load.

In the present experiment participants had to rate their activities before the notification message was shown, the presentation style was fixed, and participants were reminded to consider both context and the notification message when assessing acceptability. Whereas in the pilot study no correlation was found between Q8 and C2, in the present study a significant negative correlation (overall: r=-.326, n=231, p<.005, day 1 only: r=-.262, n=77, p<.05) was found. The redesign of both the questionnaire and the reminder enabled measuring the relation between engagement in activities and general acceptability.

6.4 Methodological issues

In general, user ratings in short-term user studies with prototypical technology can be influenced by several artifacts that result from the nature of the study. These artifacts could – in this specific case- be solved by using a realistic system with real messages for a longer period of time. Short-term user studies with prototypical technology, such as the present study, may however guide the development of systems that can be used for longitudinal studies in terms of problem understanding and model construction.

Artifacts that might have influenced the results of the present study include:

- The number of notifications was set to an average of two messages per hour. A realistic notification system would activate messages based on their availability, which might lead

to many notifications in a short time span. Consequently, an oversupply of notifications might result in lower acceptability ratings.

- Notifications were only given in the vicinity of the messaging system. Therefore, the range of activities in which the user could have been engaged at the moment of interruption was by definition limited. This may have reduced the influence of user engagement on acceptability.
- Although asked to treat all messages as authentic messages, participants knew the notifications were artificial. The lack of authenticity in the user feedback could lead to different ratings of acceptability.

7 Conclusion

The present study demonstrates that message urgency is the primary indicator of acceptability of notifications. Existing studies on interruptibility and notification systems tend to focus on the state of the user and on the effect of interruptions on task performance; the effect of message urgency is generally not studied. The present study reveals that in the home setting user state and context are secondary predictors of acceptability of notifications. A cost-benefit approach towards predicting acceptability, in which acceptability is based on the value of the notification message and the cost of interrupting the user, is shown to be a workable approach.

At first sight, it might seem logical to discard context-aware systems, and focus on prediction mechanisms for perceived message urgency. It therefore seems important to study how contextual cues can help predict the perceived urgency of messages. For example, when people are on the phone, they tend to be highly engaged in their activity, and consequently the perceived urgency of messages tends to be lower. Low-urgent and medium-urgent messages could then be postponed till after the conversation.

In the present study, the acceptability of notifications was examined in the home living context. In measuring the effect of notifications in a realistic environment using realistic user activities, a major challenge is to avoid influencing user behavior. Rather than studying acceptability of notifications in an artificial lab environment using artificial user activities, participants in the present study could do whatever they usually did, and they could experience the notifications in a realistic and natural setting. The questionnaire design proved to be essential for measuring acceptability; by asking participants to rate the user state before showing the notification message, the effect of the message on perceived user engagement in activities could be assessed. Furthermore, the study shows the need for longitudinal user

studies, since changes in user experiences –due to for example product novelty- cannot be captured in short, one day experiments.

In conclusion, for a considerate notification system, highly urgent messages are easy to manage; these can be presented immediately. The real challenge is to present low and medium urgent messages in an acceptable manner. For these not-so-urgent messages, perceived urgency and acceptability are related to user state and context. A system that is aware of the actual user engagement and that can predict perceived message urgency for individual users, will be able to reduce unwanted interruptions and thereby improve acceptability.

8 Future work

A major challenge in the development of future aware home notification systems will be to predict the subjective urgency of messages. While in the present study subjective measures were asked directly to the users, an automated system will have to base predictions on objective measures. Subjective message urgency might be related to the message (message structure, phraseology, relationship between messages), the context (user activities, state of the environment) and the user (user values, user state). Additional user studies are needed to capture subjective message urgency and to create personalized prediction models. Studies on mobile interruptibility have shown that profiles for prototypical users can shorten the learning time of a notification system [16]; the use of prototypical user profiles for urgency prediction could be studied in a home environment.

The present study was restricted to short-term acceptability. A home notification system could also consider long-term effects when assessing acceptability of interruptions. As an example, think of prevention of repetitive strain injury (RSI). To prevent RSI, a typist should pause regularly to remove tension. The short-term acceptability of interruptions in the primary typing task tends to be very low; people do not like to pause typing. However, in the long run the pauses prevent RSI, resulting in a high acceptability of the pauses. Similarly, a home notification system could for instance induce interruptions in order to reduce stress levels.

Given a system which utilizes the level of message urgency to manage notifications, one could consider using different ambient displays for messages depending upon the classified level of urgency [24]. A system could display all low urgency messages via a non-intrusive interface in the background, for example a display next to a door. The high urgency messages could be communicated via an attention-demanding alert. The medium urgency messages could then be classified by an intelligent system in order to select the best interface and

intrusion level. Studies are planned to investigate the effect of presentation on the acceptability of notifications. Ideally, these studies should be conducted in a realistic setting with real messages over a longer period of time.

Acknowledgements The authors thank the participants for their input, and their colleagues and the ID-StudioLab for their advice and support.

9 References

1. Hartog, FTH den, Baken, NHG, Keyson, DV, Kwaaitaal, JJB, & Snijders, WAM (2004) Tackling the complexity of residential gateways in an unbundling value chain. In Proc of 15th IEE ISSLS, Edinburgh, Schotland, March 21-24, 2004: 1-10
2. McCrickard DS, Chewar CM (2003) Attuning notification design to user goals and attention costs. Comm of ACM 46(3): 67-72
3. McCrickard DS, Catrambone R, Chewara CM, Stasko JT (2003) Establishing tradeoffs that leverage attention for utility: empirically evaluating information display in notification systems. Int J Hum–Comput Stud 58: 547-582
4. McFarlane DC (1998) Interruption of People in Human-Computer Interaction. PhD Thesis, George Washington University
5. McFarlane DC (1999) Coordinating the Interruption of People in Human-Computer Interaction. In Proc of INTERACT'99. IOS Press, The Netherlands: 295-303
6. Interruptions in Human-Computer Interaction, http://interruptions.net/literature.htm
7. McFarlane DC, Latorella KA (2002) The Scope and Importance of Human Interruption in Human-Computer Interaction Design. Human-Comp Interaction 17:1-61
8. Horvitz E (1999) Principles of Mixed-Initiative User Interfaces. In: CHI'99 conference proceedings, 159-166
9. Vertegaal R (2003) Attentive user interfaces: Introduction. Comm of ACM 46(3): 30-33
10. Horvitz E, Kadie C, Paek T, Hovel D (2003) Models of Attention in Computing and Communication: From Principles to Applications. Communication of the ACM 46(3): 52-59
11. Gibbs WW (2005) Considerate Computing. Scientific American 292(1): 54-61
12. Maglio, PP, Campbell CS (2000) Tradeoffs in Displaying Peripheral Information. CHI Letters 2(1): 241-248
13. Miyata Y, Norman DA (1986) Psychological issues in support of multiple activities. In User-Centered System Design, Norman DA and Draper SW (eds), Lawrence Erlbaum Associates, Hillsdale, NJ: 265-284
14. Ho J, Intille SS (2005) Using Context-Aware Computing to Reduce the Perceived Burden of Interruptions from Mobile Devices. CHI 2005, Portland, Oregon, USA: 909-918
15. Horvitz E, Apacible J (2003) Learning and Reasoning about Interruption. ICMI '03: 20-27
16. Kern N, Schiele B (2006) Towards personalized mobile interruptibility estimation. In Proc of the 2nd International Workshop on Location- and Context-Awareness
17. Hudson SE, et al. (2003) Predicting Human Interruptibility with Sensors: A Wizard of Oz Feasibility Study. CHI Letters 5(1): 257-264

18. Tran QT, Calcaterra G, Mynatt ED (2005) Cook's Collage: Déjà vu Display for a Home Kitchen, in Home-Oriented Informatics and Telematics. In Proc of the IFIP WG 9.3 HOIT Conference: 15-32
19. Picard R (2000) Affective Computing. MIT Press
20. Haines V, Michell V, Cooper C, Maguire M (2006) Probing user values in the home environment within a technology driven Smart Home project. PUC (in press), DOI 10.1007/s00779-006-0075-5
21. Howard S, Kjeldskov J, Skov MB (2006) Pervasive computing in the domestic space. PUC (in press), DOI 10.1007/s00779-006-0081-8
22. Vastenburg MH, Keyson DV, Ridder H de (2004) Interrupting People at Home. In Proc of the 2004 IEEE Int Conf on Systems, Man and Cybernetics. Madison, Wisconsin, USA: Omnipress
23. SPSS 14.0, SPSS Inc, www.spss.com
24. Vastenburg MH, Ross PR, Keyson DV (2007) A user experience-base approach to home atmosphere control. UAIS (in press), DOI 10.1007/s10209-006-0065-5