Initiations and Interruptions in a Spoken Dialog System

Leah Nicolich-Henkin and **Carolyn P. Rosé** and **Alan W Black**
Carnegie Mellon University
5000 Forbes Ave
Pittsburgh, PA 15213
{leah.nh, cprose, awb}@cs.cmu.edu

Abstract

Choosing an appropriate way for a spoken dialog system to initiate a conversation is a challenging problem, and, if done incorrectly, can negatively affect people's performance on other important tasks. We describe the results of a study in which participants play a game and are interrupted by spoken notifications in different styles. We compare people's perceptions of the notification styles, as well as their effect on task performance. The different notifications include manipulations of pre-notifications and information about the urgency of the task. We find that pre-notifications help people respond significantly faster to urgent tasks, and that 43% of people, more than in any other category, prefer a notification style in which the notification begins by stating the urgency of the task.

1 Introduction

As spoken dialog systems have improved, they have become an increasingly prominent part of our everyday lives. It is now common to interact with systems that not only perform a single, task-based function (e.g. booking an airplane flight (Bohus and Rudnicky, 2009)), but rather act as personified assistants across a range of domains, including answering questions, managing communication, and organizing schedules, as Apple Siri and Microsoft Cortana do. In particular, some dialog systems have taken up residence in our homes, acting as personal assistants, like Amazon Echo, or as the embodiment of a network of smart home devices (Oulasvirta et al., 2007). Most research has assumed that these systems are entirely user-initiated—that the system will always be responding to questions and requests from a person, rather than making its own. However, as these interactions become more natural and human-like, there are many situations in which the system will also have reason to initiate the dialog: for example, to notify someone about a time-sensitive event like taking medicine, or to start a conversation about planning dinner. In this paper we study how different wordings of initiations can change people's perceptions of the notifications and make it easier for them to manage interruptions. We find that pre-notifications and additional explicit information about urgency improve interruption management and are preferred by users. However, we also find that there is a large range of user preferences, and in particular people with greater working memory capacity have distinctly different preferences.

Allowing a system to begin a conversation raises questions about how the dialog can and should be initiated. Ideally, the system should be aware of the context of the users: where they are physically located, and what activities or interactions they are already engaged in. If it is acting as a single voice for multiple devices within the home, those devices may have competing goals, and it must decide which takes priority. Additionally, different people may have different preferences for how they want to interact with the system, including frequency, wording, and modality of interruptions, which the system should be able to accommodate. It is important that the system should not only be technically functional, but also be enjoyable to use. By taking these competing factors into consideration, it will provide a better user experience.

The factors mentioned above create a large and complex set of choices and possibilities. For the purposes of this paper we studied two variables—pre-notifications and urgency—that

Proceedings of the SIGDIAL 2016 Conference, pages 148–156,
Los Angeles, USA, 13-15 September 2016.

provide the potential to make very simple changes to a system which nonetheless have a large impact on people's preferences for and perception of notifications. As a language-based technology, a dialog system is uniquely situated to use linguistic strategies to accommodate different users and situations, so we focused specifically on changing the wording of the notifications to give increased amounts of warning and information at the cost of simplicity and directness.

The first variable we looked at was pre-notification. This was inspired by considering how people initiate conversations with each other. Schegloff (1968) introduces the idea that any two-person conversation typically begins with a "summons," which serves to propose the start of a conversation. The "summons" is followed by an "answer" from the other person, acknowledging participation in the conversation. Our first hypothesis is that by having the dialog system begin with a summons-style pre-notification it will feel more natural, so participants will find the interruption less annoying, and it will disrupt their task performance less.

Next, we consider how the urgency of a task affects performance and preference. A number of studies have found that people are more receptive to being interrupted if the notification is urgent (Vastenburg et al., 2008; Paul et al., 2011; Paul et al., 2015). Our second hypothesis is that participants will benefit even more from the pre-notification if it is not a generic phrase, but rather indicates the urgency of the task itself. This will allow participants to prepare themselves to complete the task as quickly as possible when necessary, but let them know that they can take their time when appropriate.

2 Related Work

There have been studies on the effects of wording in dialog systems and on different ways of providing notifications, but none of these studies have combined them to examine how linguistic style choices in a dialog system interact with and change people's perceptions of notifications.

Looking at the effects of wording in a spoken dialog system, Torrey et al. (2013) studied a system that helped people make cupcakes and found that using hedges and discourse markers made the system seem less commanding and more friendly. However, they focused on an in-progress dialog, rather than initiations and interruptions.

Interruptions, taking the form of notifications, have both benefits and risks. They can be extremely useful when reminding people of important tasks or appointments, and are particularly beneficial to people dealing with memory problems (McGee-Lennon et al., 2011). However, the danger of interruptions is that they decrease performance on the user's primary task by disrupting concentration, and, depending on the perceived worth of the notification, may also annoy or frustrate the user. Warnock et al. (2011) and Paul et al. (2015) find that all notifications cause errors in the task that is being interrupted. The ability to focus on a task and remember items despite interruptions is associated with the cognitive load of the task, which determines how much memory it requires; the working memory of the person, which measures how many items they can remember at once; and executive attention, which regulates which items they focus on (Engle, 2002). Because working memory is essential to interruption management, in this study we compare performance and perceptions across participants with different working memory capacities.

One way to address the problem of notification interruptions is by detecting natural breaking points, and interrupting during them (Hudson et al., 2003; Fogarty et al., 2005; Okoshi et al., 2015). However, this is a challenging task that relies on full knowledge or detection of the user's activities, and that may also raise privacy concerns. Changing only the delivery of the alert to make it less disruptive has a much lower barrier to entry and can be applied to a wide range of systems being designed to interact with users through dialog.

Certain types of notifications are less disruptive than others. McGee-Lennon et al. (2007) compare beeps, musical patterns, and speech-based notifications, finding that people perform slightly better with speech notifications, and that different people prefer different modalities of notification. Warnock et al. (2011) go further, also looking at notifications based on text, pictures, colors, iconic sounds, touch, and smell, where different variations must be associated with different tasks. However, although both the studies have speech as a notification option, they use only a single phrase type and do not consider the effect that different styles of speech may have on either performance or preference.

Style	Urgent notification	Non-urgent notification
base	The bathtub is overflowing.	The bathtub is dirty.
pre	Excuse me...the bathtub is overflowing.	Excuse me...the bathtub is dirty.
verbose	Urgent task...the bathtub is overflowing.	Whenever is convenient...the bathtub is dirty.

Table 1: Example notifications in each style and urgency level

A method of mitigating the negative effects of notifications is to first send a "pre-alert," as described by Andrews et al. (2009). They find that a pre-alert increases the speed with which the primary task is resumed after the interruption, and negates its disruptive effect. However, the types of alerts they compared were all non-linguistic, being either visual or consisting of a single tone.

Research has shown the benefits of pre-notifications, the relevance of urgency, the utility of language-based notifications, and the importance of wording choice to perceptions of a dialog system. These are all closely related concepts; however, unlike previous work, we incorporate all of them into a set of notifications which can be tested for their effect on users.

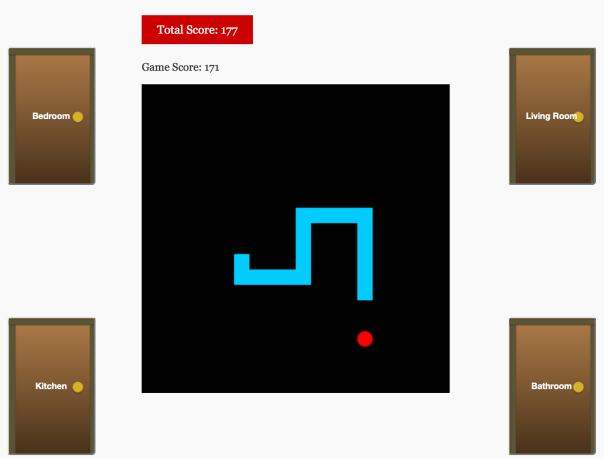

Figure 1: Screenshot of primary task screen

Figure 2: Screenshot of secondary task "Kitchen" screen

3 Experiment

As discussed above, there are many factors that can influence the perception and effect of notifications. To look at a well-defined space, we chose to study three factors: the urgency of the task, the presence of a pre-notification, and the prominence of urgency level in the notification. Participants played a browser-based game that involved going back and forth between primary and secondary tasks based on spoken notifications. By framing the activity as a game and giving players points for both types of tasks, players were encouraged to balance doing well in the game with responding to notifications. This models the real-world situation of balancing multiple tasks, some of which are prompted by a dialog system. The primary task took the form of a game of *Snake*, shown in Figure 1, and described in more detail in Section 3.2.2. In the secondary task, participants were periodically notified to go complete "household tasks." For example, in one task participants are told that the "toast is burning," and they must go click on the toast in the kitchen shown in Figure 2. Each participant went through three rounds with a different combination of variables in each round, so that we were able to both measure performance differences and get feedback on participants' perceptions of the different notifications. Examples of different notification types, described in more detail below, are displayed in Table 1.

3.1 Participants

The participants in the experiment were recruited using the Amazon Mechanical Turk crowdsourcing pool. We required participants to be located in an English-speaking country (US, Australia, Canada, Great Britain, or New Zealand), and to have a 95% or higher HIT approval rate. Additionally, participants had to sign a consent form stating they were at least 18 years old.

We conducted 206 sessions. Of those, we discarded the data of 31 because of issues including bugs in the game, people repeating the study, and people making no effort to play the game or

complete tasks. We sampled the remaining data to create a data set balanced between the 18 conditions described in Section 3.2.5, leaving a total of 144 participants in the study. Of these participants, 57.64% were male, and 41.67% were female. Their ages ranged from 20-69, with a mean of 34.31.

3.2 Procedure

The study procedure consisted of the following steps:

1. Forward digit span task
2. Primary task tutorial
3. *Snake* baseline session
4. Secondary task tutorial
5. Experimental manipulation (3 rounds)

3.2.1 Forward Digit Span task

Participants started by completing the Forward Digit Span task (Hunt et al., 1973), which we used as a metric for working memory, and is associated with attention. Working memory is closely related to people's ability to deal with interruptions, leading us to hypothesize that it would help distinguish different user groups with regard to their ability to manage interruptions. In the test, digits were presented visually to the participant at one-second intervals. Each digit was visible for half a second and was followed by a pause of another half second before the next digit was displayed. At the end of each sequence, the participants entered the sequence, as they remembered it, into a text box. Participants were presented with two different sequences of equal length, beginning with length 3. If they got at least one of the two correct, the sequence length was increased by one. When they got both wrong, the task ended and the longest length at which they got at least one of the two correct was considered the participant's "digit span score." The participants' scores had a mean of 7.37 and a standard deviation of 1.83.

3.2.2 Primary task tutorial

Following the digit span task, participants were given a brief tutorial on the primary task. This involved playing the computer game *Snake*, in which a player maneuvers a "snake" graphic around a box, trying to make it hit, or "eat," circles, without hitting the sides or itself. With each circle the snake eats, the player gets points equal to the length of the snake, and the snake gets longer, thus giving a greater reward to the more difficult situation of having a longer snake. Participants got a one-minute tutorial, presented by the same voice as used in the notifications, to familiarize them both with the game controls and the notification voice.

3.2.3 Snake baseline session

Participants played the game uninterrupted for one minute to get a baseline measurement of their skill at the game.

3.2.4 Secondary task tutorial

Next, participants were given a tutorial in the secondary task. Each task requires that the participants navigate to a different "room" in the house by clicking on a labeled door icon. This leads them to a different screen (e.g. the kitchen shown in Figure 2), where they must click on an item in the room, such as the television or the stove. Some tasks are urgent (they must be accomplished within 10 seconds), while others can be completed at any time during the game. Completing each of these tasks gives the participants 20 game points, incentivizing them to complete the task despite its potential disruption to the game.

3.2.5 Experimental manipulation

Each of the three rounds consisted of the participant playing *Snake* for 2 minutes, and being interrupted twice at 30 second intervals with notifications for secondary tasks. This simulated a person engaged in some activity in their home (represented by *Snake*) who is then interrupted by a spoken dialog notification system. To test our hypotheses that pre-notifications and additional urgency information would both be beneficial, in each round participants were given one of the three notification styles listed in Table 1.

The experiment comprises three independent variables. The first is notification style, with 3 different values, as discussed above. Second is urgency level, with two different values: urgent and non-urgent. Finally, room/task has three values: bedroom, bathroom, and kitchen. To control for the effects of different orders and combinations of variables, we conducted a 3 (notification style) $\times$ 3 (room/tasks) $\times$ 2 (urgency level) manipulation, with notification style and room/tasks as within-subject factors, and urgency as a between-subject (per-task) factor. In particular, we counterbalanced notification style and room/tasks using a 3×3 Latin Square, where each cell contained two

tasks in the same room, delivered with the same notification style, creating 3 separate conditions. Each participant experienced both urgency levels. However, we maintained a consistent urgency level within each room for each participant. We accomplished this by constructing all six possible sequences of three assignments of urgency level such that both urgency levels were in the sequence at least once. We then crossed the three conditions from the Latin Square with the six possible orderings of the between-subject factor. Thus, in total, there were 18 experimental settings.

The six types of notifications were designed to operationalize the variables of pre-notification, urgency level, and urgency information. In the first, *base*, the participant is given just the content of the notification (e.g. "The toast is burning"). In the second, *pre*, the participant is given a simple pre-notification ("Excuse me") followed by a pause, and then the content of the notification. In both the first and second conditions, the participant must determine based on the content of the notification whether it is urgent or not. In the third condition, *verbose* , the participant is given a pre-notification specifying the urgency of the notification (either "Urgent task" or "Whenever is convenient"), followed by a pause, and then the content of the notification. Each task object has an urgent and non-urgent version. These variations are represented in Table 1. To help participants identify separate notification types, each round was associated with a different room, including two unique tasks. This room/task pair was an additional manipulated variable.

3.3 Outcomes

In order to compare the notification types, for each notification type we measured how well the participant did in the primary task interrupted by those notifications, the number of secondary tasks completed, and the amount of time it took to complete tasks. In addition, we measured baseline performance on the primary task as a metric of individual skill.

In addition to quantitative measures, at the end of the study we also asked participants about their preferences. After participants completed the task, they were asked to identify the room associated with the notifications they liked the most, and the one associated with the notifications they liked the least. They were also instructed to give more detailed feedback about their preferences.

4 Results

In this section, we compare the outcomes of the study to our two hypotheses: first, that participants would perform better and prefer a system that begins with a pre-notification, and second, that participants would benefit even more from the pre-notification if it indicates the urgency of the task. To evaluate these interactions, we look first at performance on the primary task, second at performance on the secondary task, and finally at stated preference for different notification styles.

To answer the question of whether different types of notifications effect primary task performance, we analyzed the game score and number of game deaths, shown in Table 2. The gameplay was highly variable between individuals, so we compared game scores (Mean: 166.04, Standard deviation: 166.35) and number of game deaths (Mean: 7.16, Standard deviation: 5.31) using a repeated measures ANOVA across different notification styles. We also ranked each individual's rounds from one to three, and performed a chi-square test between notification style and rank. However, the only significant indicator of performance was the order of the rounds, with performance improving as people played more ($F(1, 143) = 23.80$, $p<.001$).

To answer the question of whether different types of notifications effect secondary task management, we analyzed the amount of time it took to complete tasks (Mean: 7547ms, Standard deviation: 7006ms) in different conditions, shown in Table 3. We found that whether the task was urgent or not has a significant effect on the completion time, validating our urgency manipulation ($F(1,422)=96.39$, $p<.0001$). For non-urgent tasks, the notification style did not appear to have an effect on completion time, but for urgent tasks it was highly significant ($F(2,208)=16.57$, $p<.0001$). A Tukey post-hoc test reveals that performance in *pre* and *verbose*, both of which have a pre-notification aspect to them, is virtually identical, but *base* is significantly slower. This shows that the presence of a pre-notification, regardless of the type, helps users manage interruptions to complete urgent tasks faster.

To answer the question of what effect different notification styles had on participants' perception of and preference for notifications, we per-

Table 2: Primary task performance

	F	df	P	
Game score by notification style	0.10	2,286	p = .905	
Game deaths by notification style	0.03	2,286	p = .972	
Game score by round order	23.80	1,143	p < .001	***
	χ^2	df	N	P
Rank of round by notification style	9.08	4	432	p = 0.059

Table 3: Secondary task completion times in different conditions

	F	df	P	
Urgent vs. non-urgent tasks	96.39	1,422	p < .0001	***
Non-urgent tasks with different notification styles	1.38	2,210	p = .254	
Urgent tasks with different notification styles	16.57	2,208	p < .0001	***

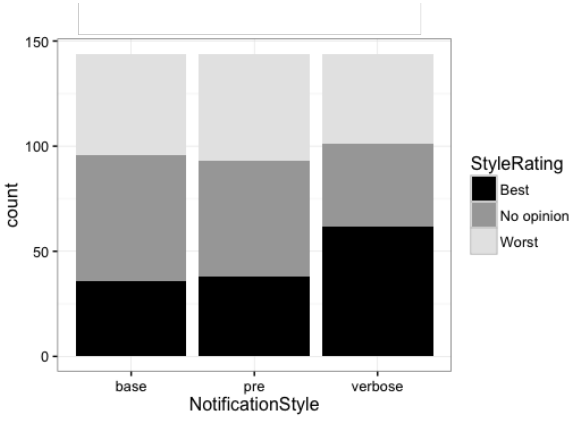

Figure 3: Preferences by Notification Style

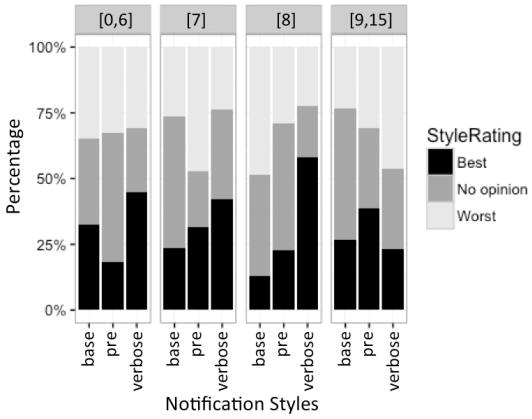

Figure 4: Preferences by Notification Style and Digit Span

formed a series of chi-square tests comparing different manipulations with the number of people who rated each best, worst, or not at all, as shown in Table 4, Here, the notification style was significant (χ^2(4, N=432)=14.61, p<.01), with the breakdown shown in Figure 3. A plurality of people liked *verbose* the most, with *base* and *pre* rated about the same. In looking at potential causes of variation, we also examined interactions between different components of the manipulation: the urgency of the task, the room containing the objects people clicked on in the task, and the order it was presented. Participants rated non-urgent tasks better, and also those in the bedroom worse (an unintentional effect of varying the tasks for different rounds was that the bedroom tasks were more difficult than others), but overall there were no significant interactions between these factors and notification style, displaying a stable main effect. The pre-notification and added urgency information were considered preferable across conditions.

These results show that having a pre-notification did not help participants on their main task, but it did help them complete urgent tasks more quickly, which is desirable. Adding urgency-related information to the pre-notification, as in the *verbose* style, did not affect task performance, but was a clear favorite across all different conditions.

Nonetheless, not all people liked the *verbose* style most, which raises the question of what factors determine a person's preferences. For the purpose of this study, we also looked at how three personal characteristics interact with notification preferences, shown in Table 5. Age and gender did not show a significant pattern, but digit span did. As depicted in Figure 4, across all groups there was a preference for the *verbose* style, except among people with the highest digit span scores.

Table 4: Interactions between round manipulation and preference distribution

Manipulated Variables	χ^2	df	N	P	
Notification style	14.61	4	432	$p < .01$	**
Urgency	18.95	2	432	$p < .001$	***
Room/task	13.26	4	432	$p < .05$	*
Order	2.11	4	432	p = .716	

Table 5: Interactions between personal characteristics and preference distribution

Personal Characteristics	χ^2	df	N	P	
Age (quartiles)	24.78	12	432	p = .016	
Gender	4.04	8	432	p = .854	
Digit span (quartiles)	28.99	12	432	$p < .005$	**

This group instead disliked *verbose* the most, and showed a slight preference for the *pre* style (χ^2 (12, N=432)=28.88, p<.005). As such, we can distinguish them as a distinct user group, with a different focus and different priorities. Given the difference in their preferences on notification style, we attempt to identify other factors that distinguish them as a group. However, we compared them to the rest of the participants using age, gender, baseline score, number of game deaths, and all the different round manipulation preferences, and they are not significantly different in any way.

To gain insight into user preferences, in addition to ranking their preferences, participants provided written comments about their favorite and least favorite rounds. We examined these to better understand what components of the round influenced their choice, and what they liked and disliked about the notifications themselves. The comments reflect what we see in the preference trends. Of the 270 comments, 23% focus only on the notification itself, while 35% focus only on the content and visual appearance of the rooms, 23% focus only on the urgency of the task, and 19% focus on other aspects of the system. When we include comments that mention multiple components of the study, 30% talk about the notification itself, while 37% talk about the room, and 30% talk about urgency. This includes a large amount of overlap, especially between notification and urgency, which are closely associated with each other.

Even though only 30% of people specifically mentioned the notification, of that 30%, the breakdown of urgencies and rooms they preferred mirrored that of the participants as a whole, suggesting that despite what people mentioned in their comments, everyone was motivated by similar factors. People who commented on notification style were most likely to say they liked *base* because it was simple and straightforward, but often complained that it didn't give them a pre-notification. For *pre*, some people said they liked it because it was polite, but even more people complained that the "excuse me" was "creepy", "annoying", or "unnecessary". Finally, for *verbose*, many people commented that it helped them to distinguish between different urgencies, but the most common complaint was the tone or wording of the message.

5 Discussion

Our hypothesis that pre-notifications would help people's performance was supported, but only in one condition, that of the urgent secondary task. This suggests that having a pre-notification does not help people manage interruptions to a primary task. However, it might be the case that the primary task, playing *Snake*, was ultimately not ideal for this evaluation because interruptions in general did not have as much of a negative impact on people's performance as we originally thought. Although participants were given a practice round, if they had a full round without any notifications at all, it would been easier for us to directly measure the performance effect of notifications, rather than just comparing different styles. Additionally, the best predictor of an individual's performance on any given round was the order of the round; in other words, people improved significantly at the game over the course of the study. Had game performance remained approximately the same over time, we might have seen a stronger effect of notification style.

Looking at preferences, we find that people who

scored higher on a test of working memory, as measured by digit span score, generally view the *verbose* style less favorably. The question remains, however, of what other factors determine people's preferences. For the purposes of this study we only looked at a few personal qualities, most of which were not good predictors of preference. If we gain more insight into people's traits, for example through additional demographics and basic personality questions, we may find other factors that affect their preferences.

The most interesting part of the extended feedback was that people frequently commented on the tone of the notification, either praising it for being "calm" and "friendly," or criticizing it for being "demanding" and "creepy." Additionally, we often assume that politeness is a positive, but in one case someone complained that the *verbose* notification was "too kind." We focused on the information contained in the notification, but subtle changes to wording or inflection can result in big changes in how language is perceived.

One shortcoming of this study was that it looked at people interacting with a system over a very short period of time. People's impressions of notifications may change over time as they become accustomed to the system. As they become used to being interrupted, the notifications may be less startling, but as people learn to predict what the system will say, they may value concise phrasing over added information. Our study lays a groundwork for what aspects of interruption people care about, but before making conclusions for use in a real-world system, it would be important to look at how people adjust and settle into patterns.

6 Conclusion

In general, the *verbose* style was both most preferred and best for people's performance. This suggests that people prefer a pre-notification before hearing the task, which enables them to know the urgency of the task and thus the type of reactions expected, and to prepare for the task. However, it's also important to note that many people did *not* like *verbose* the most, and even thought it was the worst. In particular, the difference in preference based on digit span score reveals that different types of people may have significantly different opinions. As we consider the design of spoken dialog systems, we should consider not only whether pre-notifications and urgency information can make interruptions more helpful and palatable, but also how we can accommodate a range of users. The conclusion we draw, then, is not that systems should contain urgency pre-notifications, but rather that they should have flexibility for different people to experiment with a range of initiation styles to choose the one they personally like the best.

Acknowledgments

This work is supported by Bosch Research and Technology Center North America.

References

Alyssa E Andrews, Raj M Ratwani, and J Gregory Trafton. 2009. The effect of alert type to an interruption on primary task resumption. In *Proceedings of the HFES Annual Meeting 2009*. Citeseer.

Dan Bohus and Alexander I Rudnicky. 2009. The ravenclaw dialog management framework: Architecture and systems. *Computer Speech & Language*, 23(3):332–361.

Randall W Engle. 2002. Working memory capacity as executive attention. *Current directions in psychological science*, 11(1):19–23.

James Fogarty, Scott E Hudson, Christopher G Atkeson, Daniel Avrahami, Jodi Forlizzi, Sara Kiesler, Johnny C Lee, and Jie Yang. 2005. Predicting human interruptibility with sensors. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 12(1):119–146.

Scott Hudson, James Fogarty, Christopher Atkeson, Daniel Avrahami, Jodi Forlizzi, Sara Kiesler, Johnny Lee, and Jie Yang. 2003. Predicting human interruptibility with sensors: a wizard of oz feasibility study. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 257–264. ACM.

Earl Hunt, Nancy Frost, and Clifford Lunneborg. 1973. Individual differences in cognition: A new approach to intelligence. *The psychology of learning and motivation*, 7:87–122.

Marilyn R McGee-Lennon, Maria Wolters, and Tony McBryan. 2007. Audio reminders in the home environment.

Marilyn Rose McGee-Lennon, Maria Klara Wolters, and Stephen Brewster. 2011. User-centred multimodal reminders for assistive living. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2105–2114. ACM.

Tadashi Okoshi, Julian Ramos, Hiroki Nozaki, Jin Nakazawa, Anind K Dey, and Hideyuki Tokuda.

2015. Reducing users' perceived mental effort due to interruptive notifications in multi-device mobile environments. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 475–486. ACM.

Antti Oulasvirta, K-P Engelbrecht, Anthony Jameson, and Sebastian Moller. 2007. Communication failures in the speech-based control of smart home systems. In *Intelligent Environments, 2007. IE 07. 3rd IET International Conference on*, pages 135–143. IET.

Celeste Lyn Paul, Anita Komlodi, and Wayne Lutters. 2011. Again?!! the emotional experience of social notification interruptions. In *Human-Computer Interaction–INTERACT 2011*, pages 471–478. Springer.

Celeste Lyn Paul, Anita Komlodi, and Wayne Lutters. 2015. Interruptive notifications in support of task management. *International Journal of Human-Computer Studies*, 79:20–34.

Emanuel A Schegloff. 1968. Sequencing in conversational openings. *American anthropologist*, 70(6):1075–1095.

Cristen Torrey, Susan Fussell, and Sara Kiesler. 2013. How a robot should give advice. In *Proceedings of the 8th ACM/IEEE international conference on Human-robot interaction*, pages 275–282. IEEE Press.

Martijn H Vastenburg, David V Keyson, and Huib De Ridder. 2008. Considerate home notification systems: a field study of acceptability of notifications in the home. *Personal and Ubiquitous Computing*, 12(8):555–566.

David Warnock, Marilyn R McGee-Lennon, and Stephen Brewster. 2011. The impact of unwanted multimodal notifications. In *Proceedings of the 13th international conference on multimodal interfaces*, pages 177–184. ACM.