A Comparison of Two Peripheral Displays for Monitoring Email: Measuring Usability, Awareness, and Distraction

Author names and affiliations have been removed for anonymous submission

ABSTRACT

Email overloading is an ongoing problem for information workers. One critical phase of email management involves deciding whether to respond to the arrival of a new message. We present the design and evaluation of two displays intended to support this activity (a Ticker and a physically-based colored Orb). Both are peripheral displays, designed to sit at the periphery of a user's attention. Unlike previous displays, our most successful display was easily visible even off the desktop, and both our displays limit notifications to emails that pass through a filter, thus limiting notifications about spam and other less useful emails. Surprisingly, our animated Ticker was not distracting. Both displays were well liked and supported awareness. We also present a comparison of the relative merits of different methods for measuring usability, awareness, and distraction in the lab and the field. Our results suggest that a combination of techniques is most informative for peripheral display design.

Keywords: Email management, peripheral displays, priming, lab study, field study, evaluation

ACM Classification Keywords: D.2.2: User Interfaces; H.5.2: Evaluation/methodology. General Terms: Design, Experimentation, Human Factors

INTRODUCTION

Emails have become an integral part of our lives. As the number of important emails we receive increases, and the number of spam or junk emails increases, it becomes critical to know when *relevant* email arrives. Users monitor their email inboxes while doing other tasks [15], and they need tools that can make this monitoring as efficient and low-cost as possible.

Currently, when a new message arrives, people tend to immediately look at it to decide if it is important [15]. Unfortunately, most existing systems provide minimal information about newly arrived email. For example, Outlook^{TM}simply and unobtrusively displays an envelope-shaped icon indicating the arrival of a new message, but no additional information. While the peripheral nature of this notification is important (it is designed not to interrupt the user), the lack of additional information is a major flaw. Systems should provide information about the sender, subject, or other content of an email [15]. Otherwise, a user must switch tasks and applications to check on an email message. More importantly, systems should only provide this information when it is likely that an email is *relevant* to the user. Finally, systems should provide this information in a way that is accessible to users even when they are not staring at their monitors.

Figure 1: The Ticker and Orb displays used in our studies. **left** The Ticker, shown in a magnified callout, is located just above the windows taskbar. **right** the Orb is the pink globe at the right front of the monitor.

We addressed these issues by designing and comparing two peripheral displays, shown in Figure 1, that support email inbox monitoring without requiring task switching. Peripheral displays support awareness of an information source while allowing a user to continue work on a primary task. We used two displays, a Ticker display modeled on tickers found in the literature, and an Ambient Orb (a product of Ambient Devices). Our design efforts focused on how email notifications were best displayed in these pre-existing systems. Both displays were modified to display the specific information our users cared about: email from a small number of crucial people (an email's sender is an important factor in determining it's importance [15]). Additionally, we present an in-depth evaluation of the relative merits of these displays. A major challenge in the design of peripheral displays is optimizing the trade off between awareness (*e.g.* of an email's arrival) and distraction from a primary task due to the peripheral display [10]. Thus, our evaluation considers not only usability, but also distraction and awareness. We measured all three factors using several different methods, and are therefore able to provide a nuanced picture of the factors leading to success and failure for both displays. One particular surprise

SUBMITTED TO CHI 2004

was the fact that our animated ticker turned out to be far less distracting than past work suggested it should be, leading to a less successful display with respect to awareness.

A second contribution of our work is a comparison of the methods we used to measure these three factors. We present an analysis of relationship between these methods, and their pros and cons – particularly for measuring awareness and distraction in peripheral displays.

The next section discusses two aspects of previous work tied to our two contributions, email monitoring and evaluation of peripheral displays. The following section describes how the Ticker and Orb display email arrival information, and including the formative evaluation that led to those designs. This is followed by a discussion of the methods we applied in our summative evaluation and our five hypotheses (three are about our first contribution, two are about our second). Our results section presents the data for and against each hypothesis, while our discussion elaborates on unexpected results such as the fact that the ticker animation was not distracting. In conclusions and future work, we re-iterate our main results and contributions.

RELATED WORK

The first half of this section provides some background on the problem of email management, particularly relating to the management of newly arriving emails, and gives examples of several displays designed to handle that problem. Past systems did not filter emails and were mostly limited to the desktop. The second half of this section discusses what needs to be measured to understand whether a display should succeed, and how should it be measured. We argue that in addition to usability, awareness and distraction are key factors in the success of peripheral displays. We present techniques used in the past to study usability, awareness, and distraction in the lab and in the field.

Email management

Email, as one of the most successful tools in use today, has been the subject of much study. An excellent overview of several different studies of email use conducted over the last several decades can be found in [15]. The high level summary is that people get a *lot* of email, and are often overloaded by it. Venolia *et al.* extended this work by conducting further interviews (with 6 people) and a survey (of 400 respondents) about email usage [15]. They found that email management includes five tasks: Flow, Triage, Task Management, Archive, and Retrieve. *Flow* refers to the problem of keeping up with email *while conducting other tasks*. Thus, it is precisely the type of task that a peripheral display might support. In the other four tasks, email is always primary and attention-centric.

Past solutions for providing peripheral information about email have varied greatly in exactly what they show. For example, Smith and Hudson created a "nonspeech audio glance" that played upon the arrival of each new email. This audio was generated dynamically depending on the priority, sender, number/type of recipients, and some information about content [5]. McCrickard created a GUI peripheral display called Irwin, that showed time of arrival, sender, and some subject and body content graphically on a portion of the desktop [8]. Cadiz *et al.* also created a GUI peripheral display called Sideshow, that showed the number of unread emails, and information about the sender and subject of newly arrived emails [3]. Both Irwin and Sideshow also showed many other, unrelated information sources such as weather and traffic. Finally, AudioAura and Nomadic Radio are both wearable audio peripheral displays that provided information about emails [11, 13]. Audio Aura provided ongoing ambient information about the number of unread emails (but not the sender or subject) [11]. Nomadic Radio provided notifications as new emails arrived indicating their length [13]. Nomadic Radio also inferred message priority, and encoded information indicating whether a message was for a group, personal, timely, or important. Information about content was provided if the system determined that the user was not currently having a conversation and the message was important.

Among those tools that were tested, success varied. Users found the audio notifications overly distracting and had difficulty interpreting arrival time information in Irwin. Almost 2000 different people used Sideshow in the field, and by far the most popular feature was the email notifications. On a questionnaire sent to those users, the interface scored low on distraction, and high on awareness. AudioAura was generally successful at supporting peripheral awareness, but the usefulness of the email feature was not reported on. Finally, Nomadic Radio was briefly tested with a single user in the field. The user "managed to have casual discussions with others while hearing notifications" but "preferred turning off all audio during important meetings," indicating that the device was quite distracting.

While Sideshow was clearly the most successful tool, it was not available to users off the desktop. Neither of the mobile tools demonstrably solve the problem of providing email arrival awareness, given their user study results. Additionally, one flaw common to all of these tools, except Nomadic Radio, is their lack of support for filtering of unimportant emails. As the prevalence of spam increases, interruptions will become more burdensome because messages are not prioritized.

Evaluation of peripheral displays

Although we report above on evaluation results, it is important to understand the methods used to arrive at those results as well. Evaluation of peripheral displays is not necessarily a straightforward task. As with all inter-

faces, basic usability is an issue for peripheral displays. However, these displays are designed to operate on the periphery, while other tasks are primary. As a result, a realistic evaluation must measure how they function in the periphery, requiring study design in which users are conducting a task of some sort while monitoring the display. Additionally, it must test not only usability, but also awareness (how easily can the user monitor the information they present) and distraction (how much do they detract from the primary task). Studies that do not take awareness and distraction into account may fail to identify important issues with peripheral displays [10]. Below, we give examples of different techniques used to evaluate these factors in the lab and in the field, and then discuss key issues for measuring awareness, distraction and usability. We also discuss any special issues relating to the design of field and lab studies.

Similar tools to those used in traditional usability evaluations have been applied to the problem of studying peripheral displays, including large-scale *questionnaires* [3], dual-task *lab studies* [2, 1, 6] in which awareness and distraction were measured directly and quantitatively, Likert Scale questions asking small sets of users to *self-report* awareness and distraction levels [3, 10, 1], *implicit measures* of awareness such as priming techniques [14], *field studies* that can lead to a deeper understanding of display success [8, 3, 12], and *interviews* with display users [12]. For example, Cadiz *et al.* conducted a large field study, and a survey asking users to self-report distraction and awareness levels using a Likert Scale [3].

Questions about *awareness* are not standardized, but are typically phrased in terms of attention [6], the use of the periphery [11], or an "overall sense of information" [10]. When awareness is explored through interviews, as is typical in field studies [13], it may be better understood. *Distraction* is typically measured in lab studies in terms of response time and other directly observable properties of user behavior. However, in some instances, users have been asked to report levels of distraction themselves, where direct data was not available. This particularly makes sense in the field [3]. *Usability* can be measured using similar techniques to those used in traditional studies, if the display is run in the user's periphery. One attempt to test the usability of a peripheral display without doing this had mixed results [10].

DESIGN OF TWO DISPLAYS

Our designs focused on the needs of administrative assistants in managing email, who receive a very high volume of emails. In interviews with ten administrative assistants, they indicated that they check email frequently, and often felt obligated to check who each new email was from almost immediately after noticing a notification about its arrival, even though the new mail might turn out to be spam or of little importance.

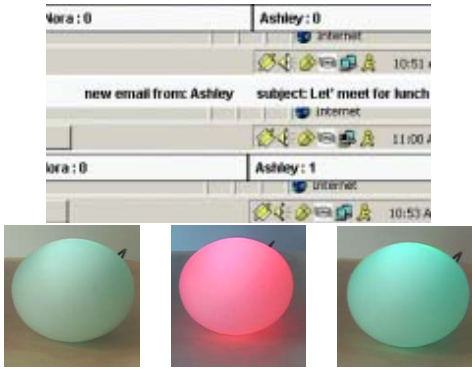

Figure 2: The Ticker (top) and Orb (bottom) displays used in our studies. First, each display is in its static state indicating zero unread emails from all monitored accounts. Second, an email has arrived from one person. Last, the display is back in its static state, indicating one unread email. The Ticker shows a sequence from top to bottom, and Orb from left to right.

Based on this information, we developed two different peripheral displays for showing information about arriving emails. Our displays were designed to monitor a person's IMAP account for email from up to five sets of email addresses, each associated with a name or nickname. We used two pre-existing displays, and modified them to display information about email arrivals. The first we used display was a Ticker, a common type of on-screen display that shows scrolling text (shown in Figure 2(a)) (other examples of tickers include [12, 9]). The second was a commercial display, a physical, frosted Orb (Ambient Orb) that sits on the user's desk and changes color in response to user input (see Figure 2(B)). In contrast to the ticker, the Orb displays information off the desktop and more abstractly.

We conducted a heuristic evaluation of both displays using heuristics specifically designed for peripheral displays [7]. The feedback we received caused us to minimize the amount of animation, flickering, blinking, and other distracting aspects of the displays. Additionally, we added the ability to associate multiple email addresses with each name or nickname, and the ability to get the subject of the unread messages from a single name by clicking on it. Based on our heuristic analysis and feedback from our pilot study, our final display designs were as follows:

Orb: For the Orb, the user associates a color with each set of addresses. Most of the time, the Orb shows a shade of cyan indicating the number of unread emails from up to five people combined, with lighter shades indicating more unread emails. When an email from a chosen person arrives, the Orb transitions to the color associated with that person for 10 seconds, and then transitions back into the cyan scale with a lighter shade because of the newly arrived unread message. Figure

2(bottom) shows this sequence. When an email from one of these people is read, the Orb simply updates the Cyan color to a darker shade.

Ticker: For the Ticker, the user associates a name with each set of addresses. Most of the time, the Ticker displays summary text: the total number of unread email from each of the five people, with no animation. For example, it might read "unread:3 John: 1 James: 2 Nancy: 0 Nora: 0 Ashley: 0." When an email arrives from one of those people, the Ticker begins scrolling at 7 characters per second, showing the name of the sender and the subject of the new email, and then reverting to the summary text (see Figure 2(top) for an example). A typical message will be shown for 25 seconds. When the Ticker is in the summary text mode, the users can click on a name to see more information about unread messages from that address or group of addresses.

SUMMATIVE EVALUATION

Because there is no consensus about the best approach for evaluating peripheral displays, we performed a series of different evaluations that included most of the techniques we found in previous work. For clarity, we split this section into a discussion of the techniques we used in the lab and the techniques we used in the field. Our lab study was a dual-task study in which we included measures of usability, distraction, and awareness. Our lab study methods included self-reporting of awareness and distraction; and objective records of performance on the primary task and on questionnaires about the peripheral display contents. Our field study included self-reporting of awareness, distraction, and usability; objective records of performance on questionnaires about the peripheral display contents; and qualitative information from interviews.

Lab Study

Our lab study was a dual task study. The primary task was to sort emails by saving them or removing them from a fake inbox that contained 1500 emails. The secondary task was to monitor the peripheral display. Participants were told that they would be asked questions about the peripheral display at the end of the study. We used a between subjects design, with half the subjects using an Orb (*Orb* condition), and the other half using a Ticker (*Ticker* condition). We ran a total of 26 participants in this study, divided equally between conditions. The participants were all college students within the ages of 18 and 23 and all of them had used email before. Orbs were placed to the right of the monitor, within 50 degrees of the user's focal vision. Tickers were located across the entire bottom of the screen, and took up 3% of the height of the monitor.

Participants were told to assume the role of a famous CEO who receives a lot of junk mail, but who also receives important emails from his/her three employees, Robert Chang, Lisa Brown and James Lewis and ten famous celebrities that all the participants were familiar with. The peripheral displays informed the user about emails from the three employees. To ensure that participants could remember the employees, they were trained until they could pass a simple memory test. Participants were asked to save the email if it was from one of the three employees or from one of the ten celebrities and to remove it otherwise. Fifteen new emails arrived at predetermined random intervals during the study. Emails were sorted from least recent to most recent, so new arrivals were only visible on the peripheral display, and not in the primary task inbox.

We gathered baseline data for 3 minutes, then started the displays, and continued for another 12 minutes. Participants were asked to remember as much information as possible from the peripheral display, as they would be given a quiz on the information later.

At the end of 12 minutes, we asked each participant a series of questions representing each of the methodologies we wished to explore. First, we asked each participant to self-report on awareness, answering questions such as "how often did you look at the display" and "how much attention did you pay to the peripheral display?". Second, we asked objective questions about how much information a participant had retained from the displays. These were questions such as: "How many new emails did you receive from James" and "who did you receive the most emails from during the first half of the study?" By asking general self-reporting questions before specific content questions, we hoped to minimize the impact that one type of question would have on answers to the next.

Field Study

Our field study included four participants, two using the Ticker and two using the Orb. Participants were all administrators in our department at our university. Participants were chosen based on their need to closely monitor email from a small number of people, and their having jobs that did not center entirely around email but included a significant amount of time spent using other desktop applications.

The field study lasted four weeks (one baseline week, two weeks with the displays present, followed by an additional baseline week). We collected data as follows: At six random intervals during the day, a pop-up window would appear on a participant's desktop, containing 9 questions about email and display awareness, similar to the questions asked of lab participants on their questionnaire. Preceding the appearance of the pop-up window, we shut down the Orb display by turning it black, and we shut down the Ticker display by turning it white and removing all text. A participant could respond to the questions, or ignore a pop-up window, in which case it

would disappear in one minute.

We also conducted a brief interview with our field study participants after they had used the displays for a week, to better understand display use patterns. At the end of the two weeks, we conducted a more detailed interview with each participant and also asked each participant to complete a questionnaire containing the same questions asked of our lab participants, as well as a series of usability questions based on the heuristics used in our formative heuristic evaluation.

Hypotheses

Our hypothesis can be split into two categories. The first (D) relates to the design of the displays, and the second (T) to the differences between techniques.

D1 The Orb is less distracting, and support a higher level of awareness, than would the Ticker.

D2 The Orb is more usable than the Ticker.

T1 The differences correlate across measures. For awareness, scores on self-reported awareness levels and scores on knowledge questions should correlate. For distraction, changes in primary task completion speed, and self-reporting should correlate.

T2 The level of awareness supported by a display in the lab correlates with the same in the field. Similarly, the level of distraction will correlate.

RESULTS

As stated above, we gathered several different kinds of data about awareness; distraction, and usability. To measure **awareness**, we used two types of questions: *self-reporting* questions in which users were asked to tell us how much attention they paid to a display, and *knowledge questions*, in which we tested how much information they had retained from the display objectively. To measure **distraction**, we used *self-reporting* questions similar to those used for awareness, and we measured changes in their primary task completion *speed* and *accuracy*. To measure **usability**, we asked participants to rate our displays against the heuristics used in our heuristic evaluation, and we interviewed them about their use of the displays. It should be noted that none of our results showed significant differences between the displays. Thus, we will not report on any t-tests. We use the convention (M=$X.XX$, SD=$Y.YY$) to report mean and standard deviation. Additionally, note that all of our Likert Scale questions were one a 5 point scale from 1 ("Not at all") to 5 ("Completely").

D1: Orb easier to monitor/less distracting

D1 states that the Orb supports a higher level of awareness, and is be less distracting, than the Ticker.

Awareness We measured awareness in two different ways. First, we asked the participants in the lab and the field to self-report their awareness level for the displays using a 5-point Likert Scale. In the lab, participants' self-reports of their awareness level for the *Orb* condition was slightly lower (M=2.85, SD=1.00) than that of the *Ticker* condition (M=3.00, SD=.80). In the field, both participants using the Orb reported that they were very aware of the display (4), while participants using the Ticker reported slightly lower values (2 and 3 out of 5).

Second, we tested participants in the lab and the field using knowledge questions about how many emails had actually arrived. In the field, participants in the *Ticker* condition scored higher on the knowledge questions (M=3.08, SD=1.32) than in the *Orb* condition (M=2.08, SD=1.55) condition, out of a maximum of 5 points. In the field, a flaw in our study design resulted in limited data: In most cases no emails had arrived because we only asked about the last 15 minutes. Among Orb users, there were 11 cases where the correct answer was not zero. We compared their score on the self-reporting questions with and without the display and found no significant difference. This is not surprising since they reported checking their inboxes as soon as a new email arrived both with and without the display, and thus know exactly from whom they had received emails.

Distraction Participants in the field and the lab were asked to self-report how distracting the displays were. Participants in the lab *Orb* condition reported being as distracted (M=2.23, SD=0.83) as those in the lab *Ticker* condition (M=2.23, SD=0.83). In the field, both participants using the Orb reported being not at all distracted, as did one Ticker user. This particular Ticker user actually requested that we make the display more distracting. She explained that she has grown accustomed to being interrupted by the nature of her job, and that she needed something flashier than simple scrolling to really catch her attention. The other Ticker user rated the display somewhat distracting (3).

We also tested distraction by measuring the change in speed and accuracy on the primary task (we could only do this in the lab, since the "primary task" in the field was unconstrained.) We used data from the second two minutes of the study, in which no displays were present, as a baseline (the first minute exhibited learning effects) We defined *speed* as (number of emails sorted) / (time in seconds). The speed of participants in the *Ticker* condition was reduced by a mean of 0.068 emails/second (SD=0.16), while the speed of *Orb* participants was reduced by 0.00 emails/second (SD=0.18) from the baseline to the second phase of the study. We calculated *accuracy* using the ratio of correct to total emails sorted. Once again, the difference between baseline and the actual study was calculated. The accuracy of participants in the *Ticker* condition decreased (-0.80%, SD=0.01) while than that of the *Orb* participants actually slightly increased (0.10%, SD=0.01).

D2: Orb more usable

Our second hypothesis was that the Orb is be more usable than the Ticker. Our results for this hypothesis are based on both questionnaires and interviews from the field study, since we did not ask lab participants any usability questions. We report below on overall usability, and several specific usability issues including how successful notifications and status were displayed (three users depended on the displays for notifications, while one monitored its status), how well they matched the original design heuristics (both displays scored highly on aesthetics and low on error prevention), and which one participants preferred (the Orbs were preferred).

Overall Usability In general, during their interviews, Orb participants seemed more excited about their displays than the Ticker participants. One reason was that they appreciated the Orb's aesthetics. Another was the benefit the Orb's visibility. It could convey information even when they are not working on the computer. Participants using the Orb could be walking around and talking to students and still notice the Orb on their desk changing colors. On our questionnaire, two participants (1 Orb and 1 Ticker) found the displays to be very useful (4) while the other two somewhat useful (3).

Three of the four participants thought the major flaw in the displays was that they did not support easy personalization of the addresses being monitored. This was a problem because they tended to work on short-term projects and depending on their projects, their list of contacts changed. However, they thought 5 people was the right number to monitor at a given time. One thought that we could improve the display more by expanding the filtering, perhaps notifying the users when they received replies. There was also one flaw with the Ticker that was not present in the Orb. Due to the location of the Ticker, certain parts of other computer applications could be covered.

Two of our participants were curious if we had plans to make our displays into a commercial product, as they were be interested in using it assuming that the aforementioned flaws were fixed. Another participant asked us to conduct a much longer study with her so that she could continue using the Orb. As she put it, "I have become attached [to the display]."

Notifications All participants found the notifications about new emails to be most effective, and all but one found the information shown about the number of unread messages least effective. When asked about the quality of support for monitoring the arrival of new emails, participants gave their displays an average rating of 3.75 out of 5 (median 4), but when asked about the quality of support for monitoring unread messages, one participant gave her display a rating of 4, and the other three participants gave their displays a rating of 1 out of 5.

The source of this disagreement in ratings is a difference in how participants used the display. One participant using our Ticker found the notifications about newly arrived messages to be too subtle. She kept track of emails solely by looking at the status line indicating number of unread emails from each address being monitored, and thus gave this feature a rating of 4. The other three participants told us that they checked their email inbox almost immediately after a notification. Therefore, the number of unread messages remained at 0 most of the time. Additionally, the participants using the Orb commented that the use of color intensity to represent the change in number of unread messages was difficult to perceive. The Ticker participant who did not make use of unread email status also commented that the status bar blended in too well with the background. She stopped noticing it after a few days of use.

Heuristics The participants were also given the list of heuristics relating to usability (the same heuristics we used in our formative heuristic evaluation [7]) and asked to rank how well the display matched the heuristics using a Likert Scale. In the case of the Orbs, the highest-rated heuristic was titled "aesthetics and pleasing design" (M=5)In our interviews, one Orb user commented that she enjoyed noticing the Orb changing color. "When you sit at a computer all day, reading email, anything to jazz it up...like oh, she emailed me!...just makes it more interesting." While the users of the Ticker also rated aesthetics highly (M=4.5), they rated the heuristic titled "match between system and real world" and "visibility of state" even higher (5). In contrast, these heuristics only received ratings of 2 and 3 respectively from Orb users. This probably reflects the fact that the Orb was more abstract and gave less overall information about new emails than did the Ticker.

In both cases, the displays fared worst on "error prevention and user control", with a mean of 2 (Orb) and 3 (Ticker). The low rating for error prevention and user control was caused by the fact that our system crashed when the IMAP service went down (this happened twice during the deployment, and additionally one Orb went down two other times due to unrelated problems). The displays did not alert users of the error, and it could take as long as half a day for the problem to be noticed.

Preference When asked whether or not she preferred our displays to what she used before our field study (a simplistic email notification system that showed no information about sender or subject referred to as "the popup window"), a Ticker user answered: "The popup window...is helpful, but it is also annoying because it pops up saying you have 3 new messages but sometimes all three of them are spam. It is almost a waste to check [my inbox]. But the display you guys did was better. Because I can see scrolling across I got a new message from so and so.... so it was helpful." An Orb user also

shared the same feeling: "I mostly ignored the Netscape popup after the department started receiving so much spam. After I received the Orb display, I completely ignored the Netscape popup."

Despite these positive comments, we wanted to know if our success was due solely to the fact that our displays only notified users about certain emails, or if there were other things about our display design that users liked. We asked users if they would still prefer our displays if the Netscape popup could filter emails. Orb participants said they still use the Orb due to the high visibility mentioned earlier. Among the Ticker users, one responded that the popup and Ticker would essentially be the same, whereas another participant responded that she would have preferred the popup because it would show all the information right away.

T1: Differences correlate across measures

In addition to our measures of how the displays performed, we were also interested in how our techniques performed. For hypothesis T1, we tested whether the different measures we used above would correlate. If they did, this would give an experimenter more freedom to use them interchangeably. However, our ability to test this hypothesis was limited by the fact that there was no significant difference between the displays' results. One thing we can say is that this lack of significant difference was consistent across all methods.

Despite this overall lack of significant difference, we felt that it was also important to see if there was a correlation between measures on a per-user basis. For example, we wanted to know if lab participants who reported a display as more distracting also were demonstrably slower at completing the primary task. On most measures, there was no correlation, but there was a weak correlation between the self-reported awareness ratings and how participants scored on the knowledge questions (Spearman's Rho=0.417, p=0.034).

T2: Differences correlate across lab/field

Hypothesis T2 measures the interchangeability of the lab and field study approaches. We hypothesized that the level of awareness supported by a display in the lab correlates with the same in the field. Similarly, the level of distraction would correlate.

Based on the results reported above, this hypothesis appears to be true in most cases. One big difference was the score of the Orb on awareness measures in the lab *vs.* the field. It was rated more highly, more consistently, in the field than in the lab. Qualitatively, participants using the Orb reported that it could be monitored even when they were not at their desks or interacting directly with their computers. This led them to give it high ratings on awareness, and was not something that participants in the lab could have been expected to notice.

DISCUSSION

Here, we group our results based on the two contributions of the paper: The hypotheses about the design of the displays, D1 and D2, indicate that the Orb was more successful and highlight the trade off between awareness and distraction. The hypotheses relating to the differences between evaluation techniques, T1 and T2, indicate that the techniques are *not* interchangeable.

Understanding the Designs: D1 and D2

D1: Orb easier to monitor/less distracting than Ticker In the lab, this hypothesis proved to be false: overall both displays got equivalent ratings (means between 2 and 3) for both awareness and distraction. One reason the displays fared so similarly that both the Orb *and* the Ticker performed very well on distraction, scoring low on all measures. In the field, the Orb was rated as very good at supporting awareness (4), and not at all distracting (1), by both participants. These results matched our hypothesis, since the Ticker was given an average rating of 2 for awareness and 2.5 for distraction.

D2: Orb more usable On usability, both displays scored quite highly. We found differences in how the displays were used by different participants, but cannot necessarily say that the Orb was more usable. However, the Orb was *preferred* for one important reason: It was visible throughout a participant's workspace, not just when she was in front of her computer.

There were three unexpected aspects to these results. First, our participants all checked their inboxes *very* frequently, and thus they were just as aware of *important* emails from before they began using our displays as they were after. What changed was the extent to which they were aware of *unimportant* emails. This is not something we measured, but qualitative evidence indicates that our participants appreciated the fact that they did not have to check their email as often with our displays.

Second, the animations in our ticker were much less distracting than expected. We designed our ticker to minimize distraction based in part on results from our formative studies, but we still expected the animations to be somewhat distracting, based on past work [9]. Surprisingly, the presence of the ticker had no significant effect on primary task completion or accuracy, and one user complained that the ticker was not distracting enough! Unsurprisingly, this led her to have problems with awareness of the notifications, emphasizing the connection between distraction and awareness. However, one positive result was her ability to make use of the ambient aspects of the display where notifications failed. This indicates the benefits of combining overall support for ongoing awareness with notifications.

Understanding the Methods: T1 and T2

T1: Differences correlate across measures This hypothesis was difficult to test because of a lack of differentiation between our two displays. However, there was a weak correlation (.417) between self-reported awareness and scores on the knowledge questions.

T2: Differences correlate across lab/field This hypothesis proved to be true in most cases. An exception was the Orb, which scored more highly in the field than in the lab on self-reported awareness.

Our ability to comment on the differences between our evaluation methods is limited by the fact that both displays scored very similarly on almost all measures. However, based on our experiences with these methods, we are able to make several observations:

Self-reporting and knowledge questions showed a weak correlation in the lab. Additionally, the display that ranked most consistently highly on awareness and low on distraction was also the display that participants preferred. This indicates that self-reporting may be a reasonable, low-cost technique to use for initial feedback. However, it appears that self-reporting is most useful when participants are able to use the display in a realistic setting, which suggests that even a short field study is better than a lab study when using this technique.

Both our objective measures of awareness (knowledge questions) and distraction (speed and accuracy) provided less useful information, overall, about awareness and distraction than we expected. This may be in part because the displays were both similarly successful on both measures. However, it is also partly because of the lack of realism inherent in our dual-task lab study.

Our survey of usability issues, and our interviews, were an important complement to our measures of awareness and distraction. They allowed us to explore the reasons for the answers participants gave us in more depth, and better understand the impact of awareness and distraction in our displays on overall usability.

CONCLUSIONS AND FUTURE WORK

In conclusion, we have presented the design and evaluation of two displays designed to support the monitoring of email from particular senders. While both displays were successful in the field, qualitative responses from field study participants indicate that one, the Orb, was a better design than the other. Our lab study gave more ambiguous results, as neither display was distracting (good), and neither was highly successful at supporting awareness. Our field study highlights the intricate relationship between awareness and distraction.

Our second contribution is a comparison of methods for testing awareness and distraction. Overall, self-reports of awareness and distraction, combined with interviews about actual display use, provided the most helpful information about display design. Although costly to run, a dual-task study with objective measures of awareness and distraction could also provide helpful information. For example, we confirmed that our displays were not at all distracting on objective measures in the lab.

In the future, we hope to develop a better understanding of the trade offs between evaluation techniques. This study did not include displays with enough variations to answer our hypotheses about techniques as definitively as we would like. By expanding our study, we can better understand the trade offs between different evaluation techniques, their limitations, and their advantages.

A second goal is to further expand our email monitoring tool by making it more customizable and expanding the sophistication of its filtering system to include a wider variety of information sources. We would also like to enhance its awareness of contextual cues and support a wider range of notifications based on a combination of email importance and current user goals.

REFERENCES

1. E. Arroyo and T. Selker. Arbitrating multimodal outputs: Using ambient displays as interruptions. *Proc. HCII'03*, 2003.
2. L. Bartram *et al.* Moving icons: Detection and distraction. *Proc. Interact'01*, 2001.
3. J. J. Cadiz, *et al.* Designing and deploying an information awareness interface. *Proc. CSCW'02*, pp. 314–323.
4. W.-L. Ho-Ching, *et al.* Can you see what I hear? the design and evaluation of a peripheral sound display for the deaf. *Proc. CHI'03*, 5(1):161–168.
5. S. E. Hudson and I. Smith. Electronic mail previews using non-speech audio. *Proc. CHI'96*, pp. 237–238.
6. L. Mamykina, *et al.* Time aura: Interfaces for pacing. *Proc. CHI'01*, pp. 144–151.
7. J. Mankoff, *et al.* Heuristic evaluation of ambient displays. *Proc. of CHI'03*, 5(1):169–176.
8. D. S. McCrickard. Maintaining information awareness with Irwin. *Proc. ED-MEDIA'99*, 1999.
9. D. S. McCrickard, *et al.* Establishing tradeoffs that leverage attention for utility: Empirically evaluating information display in notification systems. *IJHCS*, 8(5):547–582, 2003.
10. D. S. McCrickard, *et al.* A model for notification systems evaluation–assessing user goals for multitasking activity. *TOCHI*. To Appear.
11. E. Mynatt, *et al.* Designing Audio Aura. *Proc. CHI '98*. pp. 566–573.
12. S. Parsowith, *et al.* Tickertape: Notification and communication in a single line. *Proc. APCHI '98*, 1998.
13. N. Sawhney and C. Schmandt. Nomadic radio: Scaleable and contextual notification for wearable audio messaging. *Proc. CHI'99*, pp. 96–103.
14. D. Tan and M. Czerwinski. Information voyeurism: Social impact of physically large displays on information privacy. *Proc. CHI'03*, pp. 748–749, 2003.
15. G. Venolia, *et al.* Supporting email workflow. Technical Report MSR-TR-2001-88, Microsoft Research, 2001.