Chapter 19
Evaluating Peripheral Displays

Tara Matthews, Gary Hsieh, and Jennifer Mankoff

Abstract Although peripheral displays have been a domain of inquiry for over a decade now, evaluation criteria and techniques for this area are still being created. Peripheral display evaluation is an acknowledged challenge in a field setting. This chapter first describes models and methods that have been tailored specifically to evaluating peripheral displays (measuring how well they achieve their goals). Then, we present evaluation criteria used in past evaluations of peripheral displays, ranging from issues such as learnability to distraction. After explaining how these criteria have been assessed in the past, we present a case study evaluation of two e-mail peripheral displays that demonstrates the pros and cons of various evaluation techniques.

19.1 Introduction

Although peripheral displays have been a domain of inquiry for over a decade now (Gaver et al., 1991; Weiser and Brown, 1996), evaluation criteria and techniques for this area are still being created. Peripheral display evaluation is an acknowledged challenge in a field setting (Carter et al., 2008; Mankoff et al., 2003). A user interface is a peripheral display if it is peripherally used (i.e., being used while multitasking and with low cognitive effort or interruption) (Matthews et al., 2007). Because peripheral use is important, criteria for peripheral display evaluation include awareness and distraction, which traditional desktop evaluation techniques do not emphasize. Gathering data about awareness and distraction is challenging. Awareness is difficult to evaluate because interactions with a peripheral display are often brief and changes in behavior may be small and unnoticeable. Distraction is difficult to measure without further distracting users. Various studies have explored how to gather data about peripheral displays in ways that take their unique usage constraints into consideration.

T. Matthews (✉)
IBM Research, Almaden Research Center, 650 Harry Road, San Jose, CA 95123, USA
e-mail: tlmatthe@us.ibm.com

P. Markopoulos et al. (eds.), *Awareness Systems*, Human-Computer Interaction Series,
DOI 10.1007/978-1-84882-477-5_19, © Springer-Verlag London Limited 2009

This chapter describes models and methods that have been tailored specifically to evaluating peripheral displays (measuring how well they achieve their goals). Then, we present evaluation criteria used in past evaluations of peripheral displays, ranging from issues such as learnability to distraction. After explaining how these criteria have been assessed in the past, we present a case study evaluation of two e-mail peripheral displays that demonstrates the pros and cons of various evaluation techniques.

19.2 Specialized Frameworks and Methods

Two evaluation frameworks (Matthews et al., 2007; McCrickard et al., 2003b) and two methods (Mankoff et al., 2003; Shami et al., 2005) have been created specifically for peripheral displays. The frameworks provide criteria that should be considered in peripheral display design and evaluation. The methods attempt to standardize questions that evaluators ask about their displays to gather data about important criteria. These frameworks and methods are discussed further in the next section to support our discussion of evaluation criteria.

19.2.1 Models

McCrickard et al. (2003b) present a design model for classifying different types of notification systems, and their definition of notification system includes peripheral displays. User goals are modeled based on the interruption, reaction, and comprehension caused by a system. The model can be used to suggest useful empirical and analytical evaluation metrics for tailoring usability evaluation methods. In particular, designers select target levels of *interruption*, *reaction*, and *comprehension* for their display and then evaluate it using these as criteria. They argue that key characteristics of a peripheral display evaluation are to (1) provide a realistic usage experience and (2) probe the use of the display according to trade-offs among interruption, reaction, and comprehension. As an example, McCrickard et al. present a survey they created for the Scope (Van Dantzich et al., 2002) with questions about the display's target level for each criterion.

Matthews et al. present an activity theory framework for evaluating peripheral displays (Matthews et al., 2007). As part of this framework, they discuss a set of evaluation criteria based on a literature survey, interviews with peripheral display creators, and an activity theory analysis of peripheral display use. The criteria are *appeal* (a user's qualitative enjoyment of a display), *awareness* (the amount of information shown by the display that people are able to register and use), *distraction* (the amount of attention the display attracts away from a user's primary action), *learnability* (the amount of time and effort required for users to operationalize their use of a peripheral display), and *effects of breakdowns* (how apparent breakdowns are to users and how easily users can recover from them). Given these criteria, the authors discuss peripheral display evaluation relative to design dimensions derived as part of the activity theory framework: scope (the number of activities supported), classes of activities supported (primary, secondary, or pending), and criticality (from low to high importance). Criteria will vary in importance and the practicality of evaluation methods will vary depending on a display's position along each design dimension. In general, as scope increases, so does the challenge of evaluating all criteria. When supporting primary and pending activities, displays tend to support a stable set of actions, making lab experiments more tenable. On the other hand, when supporting secondary activities, displays could be used in a variety of contexts, which may vary and change, making realistic usage difficult to simulate and necessitating an extended deployment. Finally, the criticality of information displayed changes the importance of evaluation criteria (e.g., for displays with *high criticality*, *awareness* is more important, while *aesthetics* are less important).

19.2.2 Methods

Two holistic methods tailored for peripheral displays attempt to standardize data gathering at early stages of peripheral display development. Shami et al. (2005) present context of use evaluation of peripheral displays (CUEPD), an evaluation method that relies on active user participation and emphasizes the experience of using peripheral displays. CUEPD captures the context of use through user scenario building, enactment, and reflection. Designers can use CUEPD when they have a working prototype to improve future designs. This new method attempts to increase realism in an in-lab experiment with scenarios collaboratively created by the designer and the user. It also provides guidance for evaluation criteria by suggesting survey question categories: noticeability, comprehension, relevance, division of attention, and engagement.

Mankoff et al. (2003) extended Neilsen's heuristic evaluation method by modifying the set of heuristics, based on a survey of ambient display designers, to reflect ambient display design goals. This modified method is meant for use in the early stages of design, suggesting usability goals to designers as they iterate. The ambient heuristics imply certain qualities of a usable ambient display that could lead to criteria for ambient display evaluation (e.g., the "peripherality of the display" heuristic implies that obtrusiveness is a metric).

Next we discuss common criteria used in peripheral display evaluations, and how the IRC model, the activity theory framework, the CUEPD, and the ambient heuristics relate. We also discuss specific examples of methods used in past studies to evaluate each criterion.

19.3 Evaluation Criteria

As mentioned in the Introduction, evaluation criteria represent a concrete way of measuring goals such as usability or low attention demand. Traditional graphical user interfaces tend to require focal attention to accomplish a set of predefined tasks.

Criteria for evaluating these interfaces are well established and many of them, such as user learnability, error visibility, usefulness, and user satisfaction, have been evaluated for peripheral displays as well. However, peripheral displays require a new set of criteria related to *attention issues* that are not usually measured in traditional interfaces. Peripheral displays tend not to be the focus of user attention, they are always used while multitasking, and there often is not a well-defined task being performed with them. In the following subsections we discuss two broad categories of criteria that are important in peripheral display evaluations: *traditional usability* criteria and criteria related to *attention issues*. We present past evaluations of peripheral displays that demonstrate the value of measuring both sets of criteria and how to effectively gather data about them.

19.3.1 Traditional Usability Criteria

The usability literature has developed a rich set of criteria for evaluating graphical user interfaces. Some of the most common usability criteria drawn from various Human–Computer Interaction and graphical user interface usability texts include the following (definitions below are from a survey of these usability texts; Seffah et al., 2006):

- *effectiveness*: the capability of the application to enable users to achieve specified tasks with accuracy and completeness (Booth, 1989; Brinck et al., 2002; Guillemette, 1995; ISO-9241-1, 1998; Shackel, 1991)
- *efficiency*: the capability of the application to enable users to expend appropriate amounts of resources in relation to the effectiveness achieved in a specific use context (Brinck et al., 2002; Constantine and Lockwood, 1999; Dumas and Redish, 1993; Hix and Hartson, 1993; ISO-9241-11, 1998; Nielsen, 1993; Preece et al., 1994; Schneiderman and Plaisant, 2004)
- *learnability*: the ease with which features needed for achieving particular goals can be mastered (Booth, 1989; Brinck et al., 2002; Constantine and Lockwood, 1999; Hix and Hartson, 1993; Nielsen, 1993; Preece et al., 1994; Schneiderman and Plaisant, 2004; Shackel, 1991)
- *memorability*: the degree to which the application's use can be remembered over time (Brinck et al., 2002; Constantine and Lockwood, 1999; Hix and Hartson, 1993; Nielsen, 1993; Schneiderman and Plaisant, 2004)
- *flexibility*: the degree to which the application can be tailored to suit a user's needs or preferences (Preece et al., 1994; Shackel, 1991)
- *errors*: the degree to which errors are avoidable, visible, and easy to recover from (Brinck et al., 2002; Constantine and Lockwood, 1999; Nielsen, 1993; Schneiderman and Plaisant, 2004)
- *usefulness*: the degree to which the application enables users to solve real problems in an acceptable way (Booth, 1989)
- *user satisfaction*: subjective responses from users about their feelings when using the application (Booth, 1989; Brinck et al., 2002; Constantine and Lockwood, 1999; Hix and Hartson, 1993; ISO-9241-11, 1998; Nielsen, 1993; Preece et al., 1994; Schneiderman and Plaisant, 2004; Shackel, 1991)

A number of these usability criteria are particularly important for peripheral displays. In this section, we highlight the traditional user interface criteria that have also frequently been evaluated for peripheral displays and the metrics used to gather data about them. These include *learnability*, *error visibility*, *usefulness*, and *user satisfaction*.

19.3.1.1 Learnability

Learnability is the amount of time and effort required for users to operationalize their use of a peripheral display (Matthews et al., 2007). Operationalize means *to accomplish a skilled level of use*, such that minimal attention is needed to use the display. Operational use is accomplished through extensive learning. If peripheral display designers can ease the learning process, their displays will be *used peripherally* more quickly (Matthews et al., 2007).

Van Dantzich et al. examined the learning process of simple, somewhat arbitrary visuals to convey a host of task information (e-mail, calendar, to dos, alerts) (Van Dantzich et al., 2002). This was important to their evaluation because the visuals chosen were not inherently meaningful or representative of the information they conveyed. They found that users learned to interpret the display in about an hour and enjoyed using the display. InfoCanvas is another display that uses nonintuitive information-to-visual mappings (Stasko et al., 2005). Creators asked users midway through a 1-month field study if they had learned to interpret the display without looking at a reminder sheet – all but one participant had. In an in-lab evaluation of IC2Hear, a sound awareness display for the deaf, participants were asked to rate how easy each display was to learn (Ho-Ching et al., 2003).

Other past peripheral display evaluation literature have focused on the design qualities that enable quick or easy operationalization, rather than on the user's learning process. For example, the ambient heuristics call for a "consistent and intuitive mapping," so that users spend less effort learning the mappings. The CUEPD survey asks if the user was able to understand information just by glancing at it, another indicator that information was easy to learn. Other evaluations have measured whether or not users *learned* to interpret a display (Skog et al., 2003), but not how long it took them or how challenging it was to learn.

It is important to evaluate that the learning process matches user expectations to bolster adoption. For example, an evaluation of sound displays for the deaf revealed that users disliked visualizations they thought were difficult to learn (Matthews et al., 2006c).

19.3.1.2 Error Visibility and Recovery

Error visibility refers to how apparent errors or breakdowns are to users and how easily users can recover from them. The visibility of errors is particularly important for peripheral displays because their updates tend to be subtle, infrequent, and not necessarily feedback based (i.e., the user often does not control the input and the display is not always reacting to the user's actions). Thus, users may not even notice breakdowns on a peripheral display.

Error visibility is often measured inadvertently when a display unexpectedly breaks down during an evaluation or deployment. Though not always considered before evaluations, errors can cause major problems for peripheral display users and evaluators. For example, in a field study we present later in this chapter, an e-mail peripheral display was not displaying anything for half a day before users noticed. This resulted in missing data for evaluators and less e-mail awareness for users. To avoid problems like this, the ambient heuristics state that "error prevention" is an important design consideration, since "users should be able to distinguish between an inactive display and a broken display" (Mankoff et al., 2003). Also, criteria from Matthews et al. include evaluating the "effects of breakdowns" (Matthews et al., 2007). We highlight the importance of considering this challenging aspect of peripheral display design, since it is often overlooked by designers and evaluators.

19.3.1.3 Usefulness

Usefulness is the degree to which the display provides value to the user. A common goal of peripheral displays is to convey information in a subtle, nondistracting way. However, a subtle display mechanism is at odds with conveying very important information; thus peripheral displays tend to show information of lower importance. The result is that compelling applications of peripheral displays are difficult to create. This is particularly problematic when peripheral displays share screen space with important task-related windows.

The ambient heuristics encourage designs that include "useful and relevant information." The CUEPD survey asks users about whether the display provides relevant, needed information (Shami et al., 2005). A number of studies have evaluated the usefulness of a display after participants had used it for a period of time. In a field study of the Sideshow display, Cadiz et al. (2002) asked users in surveys about usefulness and whether it was "worth giving up screen space to run Sideshow." Consolvo et al. (2004) found in a field study of the CareNet display (which provides adult children with awareness of their elderly parent's activities and environment) that it was useful to participants, having a positive impact on the elders' care. In a longitudinal field study of InfoCanvas, Stasko et al. emphasized the usefulness as a criterion, asking users several questions about it (Stasko et al., 2005). In general, usefulness tends to be measured using survey or interview feedback from participants who have used the display.

19.3.1.4 User Satisfaction

User satisfaction refers to a user's qualitative happiness or unhappiness with a display. All other criteria feed into a user's overall feelings about a display, hence this criteria is greatly affected by a user's priorities (i.e., which other criteria are most important to him or her) and is a general overview of a display's success.

Matthews et al. argue that user *appeal* is an important criterion. The CUEPD survey asks about a user's enjoyment using the display (Shami et al., 2005). Many peripheral display evaluations have asked about user satisfaction, such as "novelty and fun" and "summary impressions" (Stasko et al., 2005). In general, user satisfaction is typically measured through qualitative reports, such as surveys and interviews, following realistic usage.

Aesthetics is a design factor that can affect user satisfaction and has been prevalently discussed in peripheral display literature. A number of peripheral displays have emphasized aesthetics over intuitive designs (Dahley et al., 1998; Pedersen and Sokoler, 1997; Redström et al., 2000; Skog et al., 2003; Stasko et al., 2005). For these displays, it is important to evaluate aesthetic appeal to users. Accordingly, the ambient heuristics suggest the importance of an "aesthetic and pleasing design" and a "match between design of ambient display and environment" (Mankoff et al., 2003). Also, the CUEPD survey suggests gathering user feedback on the display's attractiveness (Shami et al., 2005).

19.3.2 Criteria Related to Attention Issues

Because peripheral displays are often used outside of a user's attentional focus, while multitasking, and in a nontask-driven manner, they require a new set of criteria related to *attention issues* that are not traditionally measured. A growing body of peripheral display evaluation literature has focused on two attention issues in particular: *awareness* and *distraction*. We discuss these issues in this section, along with examples of how they were measured in evaluations.

19.3.2.1 Awareness

Awareness refers to the amount of information shown by a display that people are able to register and use. It is a common criterion in most peripheral display evaluations. Past methods attempt to standardize questions about awareness. The CUEPD survey asks if users were "able to understand the information in the display" (Shami et al., 2005). The ambient heuristics prescribe that "useful and relevant information" is visible (Mankoff et al., 2003). The IRC model emphasizes questions about *reaction* and *comprehension*, which are similar to awareness (e.g., asking about a user's "overall sense of information") (McCrickard et al., 2003b). The activity theory framework also includes *awareness* as a criterion (Matthews et al., 2007).

Gaver et al. observed users as they monitor the status of a cola manufacturing process through the use of peripheral audio sounds (Gaver et al., 1991). Their observations and user reports both provided information about awareness in comparison to not using the sounds. More recent evaluations have asked users to specifically report on their awareness of displayed information, typically using Likert scales. For example, Mamykina et al. asked questions about attention (Mamykina et al., 2003), Mynatt et al. about the use of the periphery (Mynatt et al., 1998), Zhang et al. asked about awareness (Zhang et al., 2005), and Cadiz et al. asked about staying "aware of information that's critical for me to keep track of" (Cadiz et al., 2002). Consolvo et al. found through interviews following a field deployment that the CareNet display had a positive impact on elders' care and the caregivers' awareness of workload distributions.

Empirical evaluations have attempted to measure awareness through behavioral change. Arroyo and Selker asked participants to react to certain changes in a peripheral display and measured their level of awareness through their reaction speed (Arroyo and Selker, 2003). Dabbish and Kraut compared the timing of the messages sent by an "asker" to a "helper" for two different displays showing the helper's level of busyness (Dabbish and Kraut, 2004). In a task that involved using a peripheral display to manage multiple tasks, Matthews et al. gathered data about awareness using several metrics: the number of task switches, how quickly users resumed pending tasks when relevant information arrived, and the accuracy of task switches (e.g., did a user switch to e-mail when a spam message arrived?) (Matthews et al., 2006b). Ho-Ching et al. measured awareness of sound information by asking lab study participants to identify sounds as they were displayed (Ho-Ching et al., 2003).

19.3.2.2 Distraction

Distraction refers to the amount of attention a display attracts away from a user's primary task. It is another very common criterion measured in peripheral display evaluations. The ambient heuristics prescribe that a display "should be unobtrusive and remain so unless it requires the user's attention" and users "should notice an ambient display because of a change in the data it is presenting and not because its design clashes with its environment" (Mankoff et al., 2003). The CUEPD survey asks several questions about distraction (e.g., did the user notice the display? and was the user able to adequately focus on their primary task) (Shami et al., 2005). *Interruption* in the IRC model describes the event that causes a user to switch their focal attention to the notification, causing distraction from a primary task (McCrickard et al., 2003b). Finally, the activity theory framework discusses *distraction* as an important criterion (Matthews et al., 2007).

Distraction is often measured in lab studies in terms of directly observable properties of user behavior, such as changes in performance on a primary task (Arroyo and Selker, 2003; Ho-Ching et al., 2003; Matthews et al., 2006a), focal attention shifts to a secondary display (measured with eye tracking) (Dabbish and Kraut, 2004), and how often or how quickly a user switches to tasks about which the peripheral display conveys information (Matthews et al., 2006b). However, in some instances, users have been asked to report levels of distraction themselves, using Likert scale questions (Cadiz et al., 2002; McCrickard et al., 2003b; Zhang et al., 2005). This particularly makes sense in the field, where it is difficult to know exactly what the user's primary task is or to identify the cause of a change in performance on that task.

19.3.3 A Note About Design Mechanisms and Summary

In this section we have discussed criteria and methods for evaluating peripheral displays. Researchers also suggest various design mechanisms to accomplish many of the criteria discussed in this chapter. These include abstraction (Matthews et al., 2006b; Pedersen and Sokoler, 1997), glanceability (Matthews et al., 2006a; Van Dantzich et al., 2002), user customization (Stasko et al., 2005), sufficient information design, easy transition to more in-depth information (Mankoff et al., 2003), consistent visuals (Matthews et al., 2005; Van Dantzich et al., 2002), and many more. For example, glanceable visuals enable users to be more aware of information on a peripheral display with less distraction, which leads to greater user satisfaction (Matthews et al., 2006b).

In summary of this section, we first described models and methods that have been tailored specifically for peripheral displays. The models provide criteria that should be considered in peripheral display design and evaluation. The methods attempt to standardize questions that evaluators ask about their displays to gather data about important criteria. Next, we discussed two broad categories of criteria that are important in peripheral display evaluations: *traditional usability* criteria and criteria related to *attention issues*. We presented past evaluations of peripheral displays that demonstrate both the value of measuring these criteria and how to effectively gather data about them. In the next section, we present a case study evaluation of two e-mail peripheral displays that demonstrate the pros and cons of various evaluation techniques for evaluating some of these common criteria.

19.4 Case Study: Two E-Mail Display Evaluations

Here we present an example peripheral display evaluation process, intended to demonstrate the pros and cons of several in-lab and field evaluation techniques for gathering data about *awareness, distraction, learnability*, *error visibility*, *usefulness*, and *user satisfaction*. We start by describing the evaluated e-mail peripheral displays, which convey information about new e-mail messages (a graphical Ticker and a physical, colored Orb). We then discuss the formative and summative evaluation techniques used to evaluate and improve the displays. The evaluations gather data about the criteria mentioned previously, highlighting their importance for understanding the impact of a peripheral display. Finally, we compare the data yielded by the different evaluation techniques and discuss their pros and cons when applied to peripheral displays.

19.4.1 Display Designs to Improve E-Mail Awareness

Our studies focus on the e-mail domain, which can benefit greatly from peripheral displays. People are often distracted by e-mail, which can harm their productivity (Czerwinski et al., 2004). At the same time, e-mail is an important work tool that often requires regular monitoring. Knowing whether a new e-mail is important enough to interrupt the current task or can be ignored could significantly improve a user's ability to maintain task flow and resume tasks at opportune times (Matthews et al., 2006b). A past study showed that knowing which group a sender belongs to (e.g., coworker or family) is an important factor in deciding when to read a message (Dabbish et al., 2005). Our displays show e-mail sender group information in a glanceable way, to help users quickly and easily maintain awareness of new e-mail.

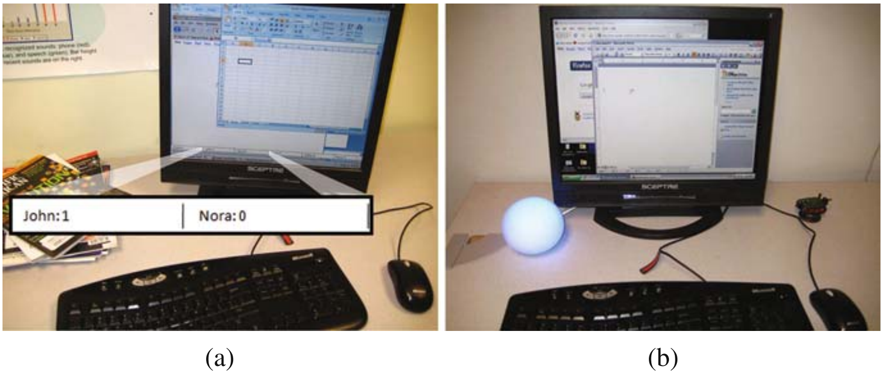


Fig. 19.1 The Ticker and Orb displays. **(a)** The Ticker, shown in a magnified callout, is located just above the windows taskbar. **(b)** The Orb is the pink globe to the right of the monitor

Our displays focused on the needs of administrative assistants managing e-mail since they receive many messages a day. We interviewed 10 administrative assistants, who indicated that they check e-mail frequently and often felt obligated to check who each new e-mail was from immediately upon notification, even though the new message might be spam or of little importance. Knowing which new messages are important could improve the overall productivity of the administrative assistants.

Peripheral displays may be a good solution for this problem. They could allow the assistants to focus on other tasks, while quickly and easily maintaining an awareness of their e-mail inboxes. To test this hypothesis, we developed two different peripheral displays for showing information about arriving e-mail. Our displays monitored a person's IMAP account for e-mail from up to five sets of e-mail addresses, each associated with a name or a nickname. We used two preexisting displays and modified them to display information about e-mail arrivals. The first display was a Ticker, a common type of on-screen display that shows scrolling text (shown in Fig. 19.1a). We chose a Ticker because they are common (many news channels use Tickers to show headlines, for example) and have been studied in the past (e.g., Maglio and Campbell, 2000; McCrickard et al., 2003a; Parsowith et al., 1998). The second was a commercial display called the Ambient Orb – a physical, frosted orb that sits on the user's desk and changes color in response to some input (see Fig. 19.1b). The Orb nicely complements the graphical, text-based Ticker, displaying information off the desktop and more abstractly (i.e., with color rather than a textual name and subject).

19.4.2 Formative Evaluation: Heuristic Evaluation

Formative evaluation is typically conducted during the design stage or early in the development stage. Fewer participants are typically needed than summative evaluations, since quick, iterative design cycles are valuable. Unfortunately, formative evaluation techniques used for traditional interfaces have focus primarily on the usability of graphical user interfaces intended for use in focal tasks. Therefore, most of them would not be able to provide feedback regarding awareness and distraction – two important criteria for evaluating peripheral displays. For this case study, a heuristic evaluation designed specifically for peripheral displays was used (Mankoff et al., 2003).

We conducted a heuristic evaluation of two versions of both the Ticker and the Orb with six graduate students who are all knowledgeable about peripheral display design. The evaluators were given a list of heuristics specifically designed for peripheral displays and asked to evaluate all four displays using the heuristic evaluation method, as described in Mankoff et al. (2003). In particular, evaluators were given descriptions and images of the displays in two scenarios. First was a notification version of each display that would change only when a new e-mail arrived from any of up to five people. Second was an ambient version that would constantly cycle through information about the current e-mails pending from each of up to five people. Evaluators were asked to indicate which version of each display they preferred.

19.4.2.1 Results

Five out of the six evaluators preferred the notification versions of the two displays. The notification versions were favored for two reasons: (1) there was a clear notification when a new e-mail arrived and (2) it was believed to be less distracting than the ambient version (which constantly cycled between information regarding the five people). Minimizing distraction was very important to our evaluators, who also suggested that we reduce the amount of animation, flickering, blinking, and other distracting aspects of the displays.

One evaluators suggested that we display the name or the nickname of the person who sent the e-mail, instead of the e-mail account from where the e-mail was sent, because knowing the sender is typically more important than knowing the specific e-mail address used. Nicknames enabled us to provide additional features, such as associating multiple e-mail addresses with one group nickname.

For the Ticker, some evaluators suggested we add an option to re-read an e-mail's subject if the user happened to miss it. In response, we enabled users to see the newest e-mail's subject line at any time by clicking on the Ticker.

19.4.2.2 Design Iteration

Based on our heuristic analysis, our final display designs were as follows:

- **Orb**. For the Orb, the user associated a color with each set of addresses. Most of the time, the Orb showed a shade of cyan indicating the number of unread e-mails from up to five people or (groups of people) combined, with lighter shades (increased brightness) indicating more unread e-mails. When an e-mail from a chosen person arrived, the Orb would transition to the color associated with that person for 10 s and then transition back into the cyan scale with a brighter shade

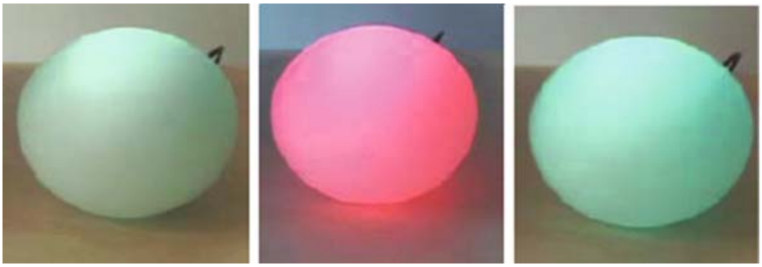


Fig. 19.2 Orb (***top***) and Ticker (***bottom***) sequence shown illustrates **(a)** a display with no unread e-mails, **(b)** the display showing the arrival of a new e-mail from Ashley, titled "Let's meet for lunch.", and **(c)** the display showing one unread e-mail

because of the new, unread message. Figure 19.2 (top) shows this sequence: the Orb transitions to red indicate a new message from Ashley has arrived and then transitions to a brighter cyan indicate there is one additional unread message in the inbox. When an e-mail from one of these people is read, the Orb shows a dimmer shade of cyan.

- **Ticker**. For the Ticker, the user associated a name with each set of addresses. Most of the time, the Ticker is not scrolling and displays only the summary text: the total number of unread e-mails from each of the five people or groups, with no animation. For example, it might read "unread: 3 John: 1 James: 2 Nancy: 0 Nora: 0 Ashley: 0." When an e-mail arrived from one of those people, the Ticker would begin to scroll at 7 characters per second, showing the name of the sender group and the subject of the new e-mail. Then it would revert to the summary text view (see the bottom of Fig. 19.2 for an example). Messages were shown for 25 s. When the Ticker was displaying summary text, users could click on a name to see message information.

19.4.3 Summative Evaluation

Summative evaluations are typically conducted when a display is fully functional. 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 and for the purpose of comparing different techniques, we split this section into a discussion of techniques we used in the lab and techniques we used in the field. After presenting the details of how each experiment was run, we summarize the techniques used to measure each of the criteria (see Tables 19.1 and 19.2). In the lab, we used a dual-task experiment and self-reports to evaluate the levels of awareness and distraction the displays caused. In our field study, we gathered self-report measures of awareness, distraction, learnability, error visibility, usefulness and user satisfaction via questionnaires and interviews and an objective measure of awareness (performance on knowledge questions about the peripheral display contents).

Table 19.1 Techniques used to measure awareness (Aw), distraction (D), error visibility (E), learnability (L), usefulness (U) and user satisfaction (S) in our lab study

	Aw	D	E	L	U	S
System log (primary task speed)		X				
System log (primary task accuracy)		X				
Knowledge questions	X					
Self-reports	X					

Table 19.2 Techniques used to measure awareness (Aw), distraction (D), error visibility (E), learnability (L), usefulness (U), and user satisfaction (S) in our field study

	Aw	D	E	L	U	S
Pre- and post-questionnaires	X	X	X	X	X	X
Interviews	X	X	X	X	X	X
ESM pop-ups (self-reports)	X	X				
ESM pop-ups (knowledge questions)	X					
System log (behavioral change)	X					

19.4.3.1 Lab Study

In this section, we first discuss the dual-task experiment setup and then describe in detail how each criterion was measured in this setup. Awareness was measured with knowledge questions and self-reports, while distraction was measured by system logs, focusing on performance changes in the primary task resulting from display use (see a summary of techniques used in Table 19.1).

Our lab study was a dual-task experiment. The primary task was to sort e-mail messages by saving them or removing them from a fake inbox that contained 1500 messages. The secondary task was to monitor the peripheral display. We chose our primary and secondary tasks such that the peripheral display would help to support the primary task. This helped to create a situation where the participant had a motive for monitoring the secondary display, while allowing us to reduce the extent to which we explicitly drew the user's attention to it. We chose to maximize realistic use by not interrupting the user at the time of a new update, but the trade-off was that users may not have recalled information when asked later.

We used a between-subjects design, with half the subjects using the Orb (Orb condition) and the other half using the Ticker (Ticker condition). We ran a total of 26 participants, divided equally between conditions. The participants were all college students between the ages of 18 and 23 and all of them had used e-mail before. The Orb was placed to the right of the monitor, within 50° of the user's focal vision. The Ticker was located across the entire bottom of the screen and took up 3% of the monitor's height.

The study was designed to represent a real-world situation where the participants needed to remove spam from their inbox. Participants were told to assume the role of a CEO of a major corporation. To make the study realistic, the CEO received three types of e-mail messages: junk mail (majority of the e-mails), important work-related messages from three of his/her employees (Robert Chang, Lisa Brown, and James Lewis) and social messages from 10 close friends who were famous celebrities (e.g., Arnold Schwarzenegger). Easily recognizable celebrity names were used to make it easier for the participants to recall the names of the friends. To ensure that participants could remember the employees' names, they were trained until they could pass a simple memory test. We considered allowing the participants to customize the employees' names for the study; however, we decided to use preset names to strengthen the control and reduce variability. The peripheral displays informed the user about messages from the three employees. Three employees, as opposed to the maximum five that the display supports, were used to decrease the challenge of memorizing the names. Participants were asked to save all e-mail from any of the three employees or from any of the 10 celebrities and to remove e-mail otherwise. Fifteen new e-mail messages arrived at predetermined, nonuniform intervals during the study. E-mail messages were sorted from least to most recent, so new arrivals were visible only on the peripheral display and not in the primary task inbox.

Participants performed the primary task for 3 min, allowing us to gather some baseline data. Then we started the peripheral display and gathered data for another 12 min. 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 min, we asked each participant a series of questions relating to the evaluation criteria. 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 that tested how much information a participant had retained from the displays, such as: "How many new e-mails did you receive from James?" and "From whom did you receive the most e-mails during the first half of the study?" There were a total of five knowledge questions. By asking general self-report questions before specific content questions, we hoped to minimize the impact that one type of question would have on answers to the next.

Like previous dual-task lab studies, we gathered data about distraction by measuring primary task accuracy and speed. We compared the 3 min of baseline performance data to the 12 min of dual-task performance. Accuracy was measured as the change in the percentage of correctly sorted e-mails, while speed was measured as the change in speed of sorting.

19.4.3.2 Field Study

In this section, we start by discussing the A–B–A' field study setup and then describe in detail the five different techniques used in the field and which criteria each technique allowed us to measure. Questionnaires and interviews were used to measure all of the six criteria. Experience-sampled self-report questions were used for both awareness and distraction, and experience-sampled knowledge questions and system logs were used to measure the level of awareness only (see a summary of techniques used in Table 19.2).

Our field study utilized an A–B–A’ format, where there was a week-long baseline, two weeks with the displays present, and an additional baseline week. This structure allowed us to obtain and compare measurements before, during, and after participants had used our displays. We conducted a brief interview with each of our field study participants after he or she had used his/her display for a week during phase B to better understand display-use patterns. At the end of the display usage period (phase B), we conducted a more detailed interview with each participant and also asked each participant to complete a questionnaire.

Our field study included four participants, two using the Ticker and two using the Orb. Participants were all administrators in a department at our university. Participants were chosen based on their need to closely monitor e-mail from a small number of people and their having jobs that included a significant amount of time spent using other desktop applications. All of the participants customized the display to show new messages from five people of importance to them (no participants chose to create groups of e-mail senders).

We used an experience sampling method (ESM) during phase B. This technique was very challenging to design, raising many issues: when should we ask questions, how should we administer questions, what kinds of questions should we ask, and what questions would help us determine if participants were aware of displayed information. If experience sampling were not designed correctly, participants could get extremely frustrated with the study and/or not provide any useful feedback.

Since our users were in front of a computer most of the day, we decided to ask the ESM questions using pop-up windows. Six sets of questions were asked at random times during the day. We tried to keep questions simple to save the user time. We asked them to select a number on a Likert scale (1–5), rather than asking for a textual answer to each question. A participant could respond to the questions, or ignore the pop-up window, in which case it would disappear in 1 min. We used a low-resolution webcam to take a snapshot of the environment when the pop-ups appeared to provide additional context information (e.g., was the participant at his/her desk, meeting with someone, and not present). During a pilot of the field study, we were surprised to discover that users would use their e-mail browser to help them answer our questions. For this reason, we instructed users not to use their desktop e-mail reader to find the answers to questions.

We structured the pop-up process as follows: 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. Next, to reduce the chance that users retained an after image of the display, we showed many different colors (this also functioned as an attention-getting mechanism). Next, we asked distractor questions to help clear working memory, such as “What is the current temperature in the office?” Following this, we asked awareness questions, such as “How often did you look at the display?” and “How much attention did you pay to the peripheral display?” Next, participants were asked how distracting the display had been. Finally, participants were asked six knowledge questions to test how much information they had retained over time from the display. Examples of specific knowledge questions that we asked the participants included “How many new emails did you receive from *person A* during the past 15 min?" and "Who did you receive the most emails from during the past 15 min?"

Throughout the peripheral display usage period, we logged e-mail activities related to the e-mail sender groups being monitored with our display. Specifically, we logged when an e-mail from one of the five selected e-mail sender groups was added, read, or deleted. We did not log any e-mail subject lines for privacy reasons; therefore, our data only indicate whether or not a message was added/read/deleted and not which message it was.

19.4.3.3 Results

In presenting our results, we focus on measurable differences between displays and between techniques, across each of the evaluation criteria.

Awareness

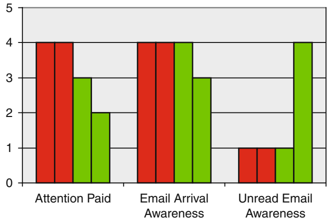
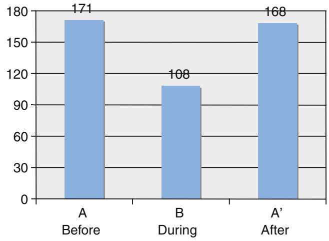
Awareness was measured using self-reports and knowledge questions in both the lab and the field. In the field, we also measured it with a post-usage questionnaire, interviews, and behavioral change (i.e., logged e-mail response time). The results for awareness are summarized in Table 19.3. In general, performance differences in the field could not be tested for significance because we had only two subjects in each condition. Instead we report average accuracy on ESM questions for all four users while the displays were deployed.

In the lab, we found that the Ticker led to better recall of the information being displayed than the Orb ($M_{\text{Ticker}} = 3.2$, $M_{\text{Orb}} = 1.9$, $t(24) = -2.19$, $p = 0.038$). The average self-reported awareness of 2.9 ($\sigma = 0.9$) was the same for both displays.

In the field, we learned from self-reports (on a Likert scale of 1–5) that both displays enabled good awareness of e-mail arrivals ($M = 3.75$), but not of unread messages ($M = 1.75$). Interestingly, one of the four users (a Ticker user) reported that the display was effective for providing a summary of unread messages (rating the display 4 out of 5) and moderately effective for providing an awareness of new message arrivals (rating the display 3 out of 5).

The source of this disagreement in ratings was further clarified by our interviews. The three participants who rated the displays *high* on awareness of e-mail arrivals and *low* on awareness of unread e-mails (two Orb users and one Ticker user) told us that they checked their e-mail inbox almost immediately after each notification. Therefore, the number of unread messages remained at 0 most of the time. Orb users commented that color intensity (which represented the change in the number of unread messages) was difficult to perceive. The Ticker user commented that the status bar blended in too well with the background. She stopped noticing it after a few days of use. The last participant, who rated the displays *high* on awareness of unread e-mails and moderately high on awareness of e-mail arrivals, reported that notifications about newly arrived messages on her Ticker were too subtle. Unlike the others, she kept track of e-mails solely by looking at the status line indicating the number of unread e-mails from each address being monitored.

Table 19.3 Summary of awareness measures and results

	Results
Lab knowledge questions	Ticker enabled better accuracy. $M_{\text{Ticker}} = 3.2$, $M_{\text{Orb}} = 1.9$, $t(24) = -2.19$, $p = 0.038$
Lab self-reports	Difference not significant: same rating
Field pre- and post-questionnaires	 Good awareness of e-mail arrivals (3.75), but not of unread messages (1.75). Red is Orb and green is Ticker
Field interviews	Three users depended on the displays for notifications of new arrivals; one monitored the number of unread messages
Field ESM pop-ups (self-reports)	Difference not significant
Field ESM pop-ups (knowledge questions)	Difference not significant
Field system log (behavioral change)	 No difference between displays; mean time to open e-mail is significantly lower for both displays during use, than before ($F(67) = 6.71$ $p = 0.012$) or after ($F(67) = 6.39$ $p = 0.014$)

Our second measure of awareness in the field was knowledge questions about how many e-mails had actually arrived. However, a flaw in our study design resulted in limited data. Because we asked about only the last 15 min, in most cases no e-mails from any of the five people of interest had arrived, even though e-mail from other senders may have arrived. For example, among Orb users, there were only 11 cases where someone of interest had sent e-mail in the last 15 min. For the small number of cases available to compare, there was no difference in performance between Orb and Ticker users.

Our last measure of awareness in the field came from the system log. We recorded how long it took from when a message arrived until a participant opened it. This analysis included only messages from any of the five preselected people shown by the displays. Since our participants checked their e-mails fairly often, for our analysis, we excluded any e-mail that took more than 5 min to open, interpreting that as an indication that the participant was unavailable. The *time to open* metric was analyzed using a repeated measure mixed-model analysis of variance, with periods (A–B–A') and participants. Participants, nested within display type, were modeled as random effects. Mean *time to open* was not significantly different between displays ($F(3)<0.001$ $p = 0.994$); however, there was a difference between period A and B and B and A' ($M_{\text{A}} = 171, M_{\text{B}} = 108, M_{\text{A'}} = 168, F(67) = 6.71$ $p = 0.012$, and $F(67) = 6.39$ $p = 0.014$, respectively). This indicates that the presence of the displays increased the speed with which participants opened an e-mail after it arrived and thus that the presence of the displays increased awareness.

Distraction

In the lab, data about distraction were measured via changes in speed and accuracy, and with self-reports. In the field, distraction was measured with self-reports and interviews. The results for distraction are summarized in Table 19.4.

We found no statistically significant differences in distraction in the lab when comparing between baseline data and data from when the displays were in use. The second and third minutes of the study, during which time no displays were present, were used as the baseline (the first minute exhibited learning effects). Though not statistically significant, the average speed *decreased* slightly (by 7%) for the Ticker and *remained about the same* for the Orb. Accuracy, in turn, *increased* slightly (by 9%) for the Ticker and *remained about the same* for the Orb. We defined speed as (number of e-mails sorted)/(time in seconds). We calculated accuracy using the ratio of correct e-mails sorted to total e-mails sorted.

Table 19.4 Summary of distraction measures and results

	Results
Lab system log (primary task speed)	Difference not significant: Ticker speed decreased slightly
Lab system log (primary task accuracy)	Difference not significant: Ticker accuracy increased slightly
Field pre- and post-questionnaires	Three users reported they were not distracted at all; one Ticker user reported being somewhat distracting
Field interviews	Three users thought their displays were not distracting; one wanted it to be more distracting. Interviews revealed that the displays reduced the extent to which e-mail arrival distracted users by providing quick awareness of important versus unimportant messages
Field ESM pop-ups (self-reports)	Difference not significant, all rated display to be not distracting

In self-reports, participants in the lab reported being slightly distracted by both displays ($M_{\text{Ticker}} = 2.38$, $M_{\text{Orb}} = 2.31$). In the field, both participants using the Orb reported not being distracted at all, as did one Ticker user. In fact, during our interviews, one participant actually requested that we make the display *more* distracting. She explained that she had grown accustomed to being interrupted by the nature of her job and that she needed something flashier than simple scrolling to capture her attention. The other Ticker user rated the display slightly distracting.

Our interviews of field study participants also revealed an unexpected side effect of our displays. Two participants reported that without the peripheral display, they would check e-mail every time their e-mail program notified them of a new arrival. However, with the peripheral display, they tended to ignore their e-mail program's notifications. This meant that they were not task switching every time a piece of spam or other unimportant e-mail arrived. Thus, the peripheral display reduced distraction as a *secondary* effect.

Learnability, Error Visibility, Usefulness, and User Satisfaction

Field study participants were given a list of peripheral display heuristics (the same heuristics we used in our formative heuristic evaluation, see Fig. 19.3) and asked to rank how well the display matched each heuristic using a Likert Scale. Some of these heuristics measure learnability, error visibility, usefulness, and user satisfaction. The participants were also given a separate set of questions to collect information regarding overall use.

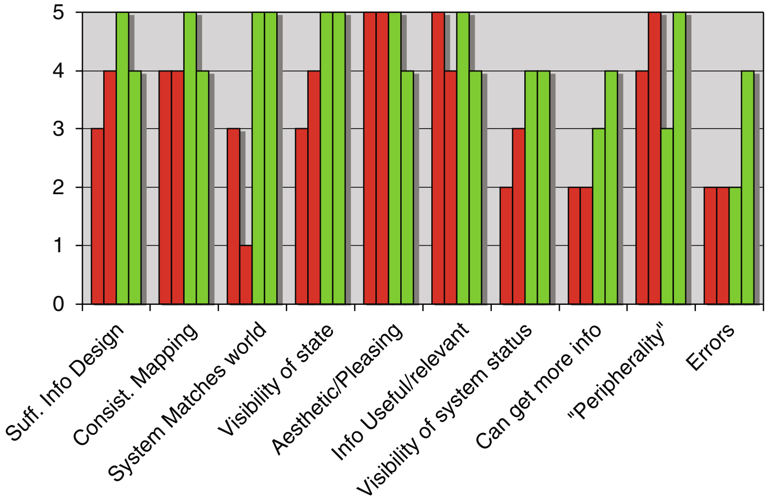


Fig. 19.3 Orb (*top*) and Ticker (*bottom*) sequence shown illustrates **(a)** a dim cyan display with no unread e-mails, **(b)** the display turning bright red, showing the arrival of a new e-mail from Ashley, titled "Let's meet for lunch.", and **(c)** the display turning a brighter shade of cyan than before the new message had arrived

Learnability. In general, our participants rated the displays highly for being easy to understand and having consistent and intuitive mappings.

Error Visibility. For both displays, the worst rating given was on "error prevention and user control," with ratings of 2, 2 (Orb) and 2, 4 (Ticker). These low ratings were caused by the fact that our system crashed whenever the users' 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 their breakdowns, and it took as long as half a day for a problem to be noticed, both because participants were not always at their desks (e.g., over lunch) and because they did not always receive e-mail from the five people being monitored frequently. There was one flaw with the Ticker that was not present in the Orb: other computer applications could occlude it.

Usefulness. Both displays rated very highly on "relevance and usefulness of the information." On our questionnaire, two participants (1 Orb and 1 Ticker) found the displays to be very useful (4 out of 5), while the other two somewhat useful (3 out of 5). Despite our participants' positive comments about our displays, we wanted to know if their success was due solely to the fact that they only notified users about certain messages from e-mail senders or if there were other aspects of their design that users liked. We asked users if they would still prefer our displays if standard e-mail notification solutions (e.g., pop-up windows provided by Outlook) could filter e-mails. Orb participants said they would still use the Orb due to its high visibility throughout their offices. Among the Ticker users, one responded that the pop-up and Ticker would essentially be the same, whereas another participant responded that she would have preferred the pop-up because it would show all the information in a single view rather than scrolling to reveal the information.

User satisfaction. While the Orb provided a more abstract representation of unread e-mail information than did the Ticker, and was rated lower on the "match between system and real-world" heuristic, users enjoyed using both displays and rated them very highly on "aesthetics and pleasing design."

During their interviews, the Orb participants seemed more excited about their displays than were the Ticker participants. One reason was that they appreciated the Orb's aesthetics. Another reason was the benefit of the Orb's visibility. It could convey information even when they were not working on the computer. Participants using the Orb could be walking around and talking to people and still notice color changes on the Orb. 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."

Two of our participants were curious if we planned to make our displays into a commercial product, as they would be interested in using it. 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)."

19.4.3.4 Discussion

This case study highlights the importance of evaluating peripheral displays using multiple methods to gather data about the six evaluation criteria: awareness, distraction, learnability, error visibility, usefulness, and user satisfaction. In particular, multiple methods provided either redundant information or in some cases, one method filled in missing data from other methods (e.g., while lab study results did not lead to significant differences in distraction measures, interview results revealed that users found the displays reducing their distraction caused by e-mail). Gathering data about the six evaluation criteria enabled us to understand more completely the affect of our displays on users.

Discussion of Heuristic Versus Summative Evaluations

In many ways, it is incorrect to compare a heuristic evaluation with summative evaluation techniques. First, there is no clear point of comparison; we used the techniques at two different stages of the display development. Second, the purposes of the evaluations are different. A heuristic evaluation does not replace a summative evaluation, but rather it is an inexpensive evaluation technique that provides initial design feedback for quick iterations. However, we can still examine the results of the heuristic evaluation and gain a sense of what kinds of problems it was unable to uncover about peripheral displays.

The heuristic evaluation was unable to suggest a correct level of distraction. It resulted in a general suggestion to minimize the amount of animation, flickering, blinking, but it was not clear what types of updates would appropriately balance distraction and awareness. This is illustrated by the fact that one field study participant wanted the updates to be more distracting to bolster her awareness.

Another set of problems discovered from the field study that was not apparent from our heuristic evaluation relates to the error visibility and recovery criterion. In our field evaluation, we found that the displays did not provide feedback regarding breakdowns. While heuristic evaluators could have foreseen such potential problems, since the list of heuristics does include an “error prevention and user control” heuristic, it is hard to predict those problems without seeing the actual system and without using it as a peripheral display. This denotes a shortcoming with using heuristic evaluation and suggests a potential area of improvement in heuristic evaluation procedures to accommodate for these types of problems.

Discussion of Lab Versus Field Evaluations

Lab studies are designed to test specific aspect of peripheral displays in a controlled setting. To evaluate awareness caused by peripheral displays, correct responses to knowledge questions and self-reports can be measured. To evaluate distraction caused by peripheral displays, accuracy and speed on the primary task and self-reports can be used. Had the lab study generated significant results, it would have been capable of providing concrete feedback comparing the awareness and distraction of the displays, which is both hard and costly to do in the field. It is challenging to reach significant results in the field because of a multitude of uncontrolled environmental factors. However, having more users might have helped in our case.

The only conclusive evidence from our lab study was that the Orb users had lower awareness than the Ticker users. This finding contradicts our findings in the field regarding awareness. While low participant numbers prevent us from finding significant differences, our interviews do suggest that Orb users were just as capable as Ticker users of noticing incoming messages. It is not clear what caused this discrepancy. We speculate it may have been related to learnability: the orb may have taken longer to learn and so it did not perform well until it had been used for a longer period of time in the field study.

Discussion of Field Techniques

Our use of redundant measures in our field study enabled better triangulation of data, which led to a deeper understanding of our displays' usage. For example, ESM popups provided a general sense that the displays did not support awareness of the number of unread e-mails well, but our final questionnaire helped us refine this result by showing that the displays were better at conveying information about e-mail arrival than about the number of unread messages. Interviews provided clarifying information: most of our users did not use the displays to monitor the number of unread e-mails because that number was almost always zero. None of these measures alone would have given nearly as complete a picture of how the displays performed.

How well did the methods we employed in the field study perform individually? Pre- and post-study questionnaires and interviews were comparable techniques used to collect qualitative feedback regarding overall use. 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. ESM self-reports supplemented the questionnaires and interviews by surveying the participants in situ. While the ESM self-reports could potentially provide more realistic results, they interrupt participants. The system log measure of awareness allowed us to obtain statistically significant results regarding behavioral change, but it would not act well as a stand-alone measure because it does not explain why the behavior changed. The least successful of the techniques was our objective knowledge questions measuring awareness. We believe our knowledge questions were unsuccessful due to the problems described above with not timing them to be administered after new messages had arrived. If used correctly, knowledge questions could still be potentially very useful for gathering data about user awareness.

Based on our experiences, we recommend an A–B–A' format for peripheral display field evaluations, though the length of each segment could change depending on factors specific to the experiment being run. For example, the length of the A–B–A' phases could be adjusted based on issues such as expected adoption and learning curves. The A–B–A' format would allow for the evaluators to notice behavioral changes, as we did in this study with system logs.

When using experience sampling to evaluate peripheral displays, we believe that evaluators will have to make decisions about sampling frequency and question contents that are specific to their display's usage context. However, we recommend not requiring answers of participants, to reduce the burden if they are busy. We also learned that using pop-up windows for experience sampling can be effective. Though pop-ups are commonly perceived as annoying, most of our participants did not find them irritating except when they were extremely busy. In fact, all of our participants stated that they would not mind participating again in a similar study. The one participant who found the pop-ups to be extremely annoying mentioned that they were at least designed in a way that “it is like a friendly thing that you’d kind of want to interact with.”

The field techniques presented in this case study have limitations. In particular, the evaluation was designed for people primarily using a desktop computer. For public displays which may be viewed in various contexts (e.g., while mobile), administering ESM questions is a problem. A potential solution is to ask users questions through their mobile devices as they walk past a display. Although the complexity of evaluation increases for off-the-desktop displays, we believe that ESM is a valuable technique for gathering data about awareness and distraction in the field.

19.4.4 Open Questions

Because realism is important to peripheral display evaluation, it is important to improve our ability to gather empirical data in the most realistic settings possible. The field study presented as part of our case study begins to address this need by demonstrating the effectiveness of experience sampling methods for gathering quantitative data about peripheral displays in situ. However, it remains an open question how to improve our ability to gather quantitative data in the field. Another open question is how to minimize attention that is unrealistically drawn to a display by an evaluation, something that makes existing field methods, such as experience sampling questionnaires and diaries, difficult to use.

19.5 Conclusion

Peripheral display evaluation is challenging, especially since it requires that criteria related to *attention* be examined in addition to more traditional usability criteria. Attentional criteria include *awareness* and *distraction*, which are difficult to measure due to their often unobservable nature and to the disruption caused by common evaluation techniques. To address these challenges, two evaluation frameworks (Matthews et al., 2007; McCrickard et al., 2003b) and two methods (Mankoff et al., 2003; Shami et al., 2005) have been created specifically for peripheral displays. The frameworks provide criteria that should be considered in peripheral display design and evaluation. The methods attempt to standardize questions that evaluators ask about their displays to gather data about important criteria. In addition, a growing body of peripheral display evaluation literature has highlighted a number of evaluation criteria: *awareness, distraction, learnability*, *error visibility*, *usefulness*, and *user satisfaction.*

We presented a case study evaluation of two e-mail peripheral displays both to demonstrate the use of various evaluation techniques for evaluating these common criteria and to highlight some pros and cons of various evaluation techniques. In the case study, we showed that heuristic evaluation and lab studies can provide important insights about the peripheral displays in a cost-effective manner. However, without actual field use, it is hard to predict the effects of the errors and to determine the right balance between awareness and distraction. In terms of techniques used in the field, questionnaire and interviews provided the most information, enabling us to explore our findings in depth. While the other field techniques tested (ESM and log analysis) can potentially be very useful, it is imperative to make slight modifications to customize them for the specific displays being evaluated (e.g., for ESM, question content and sampling frequency need to be determined).

References

Arroyo E, Selker T (2003) Arbitrating multimodal outputs: Using ambient displays as interruptions. *Proceedings of 10th International Conference on Human–Computer Interaction (HCI International)*, 591–595. Lawrence Erlbaum Associates, Crete, Greece.

Booth, P. (1989). *An Introduction to Human–Computer Interaction*. Lawrence Erlbaum Associates, London.

Brinck T, Gergle D, Wood SD (2002) *Designing Web Sites that Work: Usability for the Web*. Morgan Kaufmann, San Francisco.

Cadiz JJ, Venolia G, Jancke G, Gupta A (2002) Designing and deploying an information awareness interface. *Proceedings of the ACM Conference on Computer Supported Cooperative Work (CSCW)*, 314–323. ACM Press, New York.

Carter S, Mankoff J, Klemmer S, Matthews T (2008) Exiting the cleanroom: On ecological validity and ubiquitous computing. *Human–Computer Interaction Journal 23*(1):47–99.

Consolvo S, Roessler P, Shelton BE (2004) The CareNet Display: Lessons learned from an in home evaluation of an ambient display. *Proceedings of the 6th International Conference on Ubiquitous Computing (UbiComp)*, 1–17. Springer, Berlin.

Constantine LL, Lockwood LAD (1999) *Software for Use: A Practical Guide to the Models and Methods of Usage-Centred Design*. Addison-Wesley, New York.

Czerwinski M, Horvitz E, Wilhite S (2004) A diary study of task switching and interruptions. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*, 175–182. ACM Press, New York.

Dabbish L, Kraut RE (2004) Controlling interruptions: Awareness displays and social motivation for coordination. *Proceedings of the 2004 ACM Conference on Computer Supported Cooperative Work (CSCW)*. ACM Press, New York.

Dabbish L, Kraut RE, Fussell S, Kiesler S (2005) Understanding email use: Predicting action on a message. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*, 691–700. ACM Press, New York.

Dahley A, Wisneski C, Ishii H (1998) Water lamp and pinwheels: Ambient projection of digital information into architectural space. *Extended Abstracts of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*, 269–270. ACM Press, New York.

Dumas JS, Redish JC (1993) *A Practical Guide to Usability Testing*. Ablex Publishing Co, Norwood, NJ.

Gaver WW, Smith RB, O′shea T (1991) Effective sounds in complex systems: the ARKOLA simulation. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*, 85–90. ACM Press, New York.

Guillemette RA (1995) The evaluation of usability in interactive information systems. In Carey JM (ed.), *Human Factors in Information Systems: Emerging Theoretical Bases* (207–221). Ablex Publishing Co, Norwood, NJ.

Hix D, Hartson HR (1993) *Developing User Interfaces: Ensuring Usability Through Product & Process*. John Wiley, New York.

Ho-Ching FW-L, Mankoff J, Landay JA (2003) Can you see what I hear? The design and evaluation of a peripheral sound display for the deaf. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*, 161–168. ACM Press, New York.

Iso-9241-11 International Standards Organization (1998) *Guidance on Usability*. (Report ISO 9241–11).

Maglio PP, Campbell CS (2000) Tradeoffs in displaying peripheral information. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*, 241–248. ACM Press, New York.

Mamykina L, Mynatt E, Terry M (2003) Time aura: Interfaces for pacing. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*, 144–151. ACM Press, New York.

Mankoff J, Dey AK, Hsieh G, Kientz J, Lederer S, Ames M (2003) Heuristic evaluation of ambient displays. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*, 169–176. ACM Press, New York.

Matthews T, Blais D, Shick A, Mankoff J, Forlizzi J, Rohrbach S, Klatzky R (2006a) *Evaluating Glanceable Visuals for Multitasking*. (Technical Report EECS-2006-173). U.C. Berkeley.

Matthews T, Czerwinski M, Robertson G, Tan D (2006b) Clipping Lists and Change Borders: Improving multitasking efficiency with peripheral information design. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*, 989–998. ACM Press, New York.

Matthews T, Fong J, Ho-Ching FW, Mankoff J (2006c) Evaluating non-speech sound visualizations for the deaf. *Behaviour & Information Technology*, 25(4):333–351.

Matthews T, Forlizzi J, Rohrbach S (2005) *Designing Glanceable Peripheral Displays*. EECS-2006-113. U.C. Berkeley.

Matthews T, Rattenbury T, Carter S (2007) Defining, designing, and evaluating peripheral displays: An analysis using activity theory. *Human–Computer Interaction Journal*, 22(1):221–261.

McCrickard DS, Catrambone R, Chewar CM, Stasko JT (2003a) Establishing tradeoffs that leverage attention for utility: Empirically evaluating information display in notification systems. *International Journal of Human–Computer Studies*, 8(5):547–582.

McCrickard DS, Chewar CM, Somervell JP, Ndiwalana A (2003b) A model for notification systems evaluation – Assessing user goals for multitasking activity. *ACM Transactions on Computer-Human Interaction*, 10(4):312–338.

Mynatt ED, Back M, Want R, Baer M, Ellis JB (1998) Designing audio aura. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*, 566–573. ACM Press, New York.

Nielsen J (1993) *Usability Engineering*. Academic Press, London.

Parsowith S, Fitzpatrick G, Kaplan S, Segall B, Boot J (1998) Tickertape: Notification and communication in a single line. *Proceedings of the Third Asian Pacific Computer and Human interaction (APCHI)*, 139–144. IEEE Computer Society.

Pedersen ER, Sokoler T (1997) AROMA: Abstract representation of presence supporting mutual awareness. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*, 51–58. ACM Press, New York.

Preece J, Rogers Y, Sharp H, Benyon D, Holland S, Carey T (1994) *Human Computer Interaction*. Addison-Wesley, Wokingham.

Redström J, Skog T, Hallnäs L (2000) Informative art: Using amplified artworks as information displays. *Proceedings of the Conference on Designing Augmented Reality Environments (DARE)*, 103–114. ACM Press, New York.

Schneiderman B, Plaisant C (2004) *Designing the User Interface*. Addison-Wesley, Reading, MA.

Seffah A, Donyaee M, Kline RB, Padda HK (2006). Usability measurements and metrics: A consolidated model. *Software Qual Journal*, 14:159–178.

Shackel B (1991) Usability—Context, framework, definition, design and evaluation. In Shackel B, Richardson S (eds.), *Human Factors for Informatics Usability* (21–38). University Press, Cambridge.

Shami NS, Leshed G, Klein D (2005) Context of use evaluation of peripheral displays. *Proceedings of the Tenth IFIP TC13 International Conference on Human Computer Interaction (INTERACT)*, 579–587. Springer, Berlin.

Skog T, Ljungblad S, Holmquist LE (2003) Between aesthetics and utility: Designing ambient information visualizations. *IEEE Symposium on Information Visualization (INFOVIS)*, 233–240. IEEE Computer Society, Seattle, WA.

Stasko J, Mccolgin D, Miller T, Plaue C, Pousman Z (2005) *Evaluating the InfoCanvas Peripheral Awareness System: A Longitudinal,* In Situ *Study*. (Technical Report GIT-GVU-05-08). Georgia Institute of Technology, Atlanta, GA.

Van Dantzich M, Robbins D, Horvitz E, Czerwinski M (2002) Scope: Providing awareness of multiple notifications at a glance. *Proceedings of the Working Conference on Advanced Visual Interfaces (AVI)*. ACM Press, New York.

Weiser M, Brown JS (1996) Designing calm technology. *PowerGrid Journal*, 1(1).

Zhang L, Tu N, Vronay D (2005) Info-Lotus: A peripheral visualization for email notification. *Extended Abstracts of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*, 1901–1904. ACM Press, New York.