Available online at www.sciencedirect.com
SCIENCE DIRECT
ELSEVIER
Int. J. Human-Computer Studies 58 (2003) 547–582
International Journal of Human-Computer Studies
www.elsevier.com/locate/ijhcs

Establishing tradeoffs that leverage attention for utility: empirically evaluating information display in notification systems

D. Scott McCrickard[a,*], Richard Catrambone[b], C.M. Chewar[a], John T. Stasko[c]

[a] *Department of Computer Science, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061, USA*
[b] *Department of Psychology, Georgia Institute of Technology, Atlanta, GA 30318, USA*
[c] *College of Computing and GVU Center, Georgia Institute of Technology, Atlanta, GA 30318, USA*

Received 1 May 2002; accepted 3 September 2002

Abstract

Designing and evaluating notification systems represents an emerging challenge in the study of human–computer interaction. Users rely on notification systems to present potentially interruptive information in an efficient and effective manner to enable appropriate reaction and comprehension. Little is known about the effects of these systems on ongoing computer tasks. As the research community strives to understand information design suitable for opposing usage goals, few existing efforts lend themselves to extensibility.

However, three often conflicting design objectives are interruption to primary tasks, reaction to specific notifications, and comprehension of information over time. Based on these competing parameters, we propose a unifying research theme for the field that defines success in notification systems design as achieving the desirable balance between attention and utility. This paradigm distinguishes notification systems research from traditional HCI by centering on the limitations of the human attention system.

In a series of experiments that demonstrate this research approach and investigate use of animated text in secondary displays, we describe two empirical investigations focused on the three critical parameters during a browsing task. The first experiment compares tickering, blasting, and fading text, finding that tickering text is best for supporting deeper comprehension, fading best facilitates reaction, and, compared to the control condition, none of the animated displays are interruptive to the browsing task. The second experiment investigates fading and tickering animation in greater detail with similar tasks—at two

*Corresponding author. Tel.: +1-540-231-6698; fax: +1-540-231-6075.
E-mail address: mccricks@cs.vt.edu (D.S. McCrickard).

1071-5819/03/$ - see front matter
doi:10.1016/S1071-5819(03)00022-3

different speeds and sizes. Here, we found smaller displays allowed better reaction but were more interruptive, while slower displays provides increased comprehension. Overall, the slow fade appears to be the best secondary display animation type tested. Focusing research and user studies within this field on critical parameters such as interruption, reaction, and comprehension will increase cohesion among design and evaluation efforts for notification systems.

Keywords: Secondary display; Animation; Critical parameters; Empirical evaluation

1. Introduction

Ongoing advances in technology make it possible and easy to access information on virtually any topic and people naturally wish to stay informed about information of interest. For instance, office workers may want to stay appraised of the weather outside, the traffic situation for the ride home, stock performance, or their favorite team's current score. Such information is often not related to the primary task at hand, or is at best tangentially related, but the gain in knowledge can help in the planning of future tasks, interacting with others socially and professionally, and completing simple tasks in a timely manner. While people may want to maintain awareness of certain information, or perhaps even monitor it intermittently, such increased knowledge ideally should cause minimal distraction to their primary work or task. In examples like the ones listed previously, rarely does the primary task involve keeping track of weather, stock, or sports information. Rather, some primary task typically dominates the attention of the user, with attention diverted as minimally as possible to allow the user to maintain awareness at a desired level.

A variety of information communication devices have been developed to help people maintain a sense of casual awareness of interesting information. These *notification systems* attempt to deliver current, important information through a variety of platforms and modes in an efficient and effective manner. The benefits of notification systems are numerous, including rapid availability of important information, access to nearly instantaneous communication, and heightened awareness of the availability of personal contacts. Visually implemented notification systems are characterized as presenting information with some type of display, reflecting the current state of a corresponding information resource. While such a system display may be usually thought of as a computer monitor, it could also be a large screen or wearable display, or even a real world object, such as Mark Weiser's dangling string representation of network traffic (Weiser and Brown, 1996).

Common, classic desktop notification systems include instant messaging systems, status programs, and news and stock tickers. However, over the years, there have been many more intriguing examples of notification systems from the research community. These include the Peepholes system for monitoring the presence and activities of colleagues and the Notification Collage that uses images to reflect local items of interest (Greenberg, 1996; Greenberg and Rounding, 2001).

The important issue in notification systems research for the human–computer interaction domain is the display of information. Visual implementations of notifications that typically are not a user's main attention priority are called *secondary displays*. Users willingly sacrifice brief interruptions from their primary task to achieve view additional information of interest in a secondary display. Although secondary display design goals are generally the same as those of other notification systems (considering primary task interruption while enabling reaction to and comprehension of information), they often must conform to other design considerations as well. For instance, secondary displays are usually permitted very little screen real estate while still attempting to convey a fairly large amount of information. Unlike a typical computer interface, a secondary display portrays information that is intended to be perceived and interpreted in a quick glance or through a series of glances, rather than a longer period of a user's full attention.

Computer users have long used secondary displays like clocks, email alert tools, and system load monitors, suggesting that people may be willing to tolerate an interruption if the information presented proves to add utility through appropriate, timely reaction or long-term comprehension. While demand for these types of displays appears to be increasing, questions remain regarding the effects of visual notifications on ongoing computing tasks. They are often perceived as distracting, but the degree to which they distract a user is not well understood.

The notification systems that use these displays must provide greater benefit through information gain than irritation through distraction to other tasks.

This presents an important distinction between notification systems and traditional HCI research:

> The success of notification systems hinges on accurately supporting attention allocation between tasks while simultaneously evaluating utility through access to additional information.

This design paradigm provides a unifying theme for notification systems research that is quite different and more specific than other interface study.

Notification systems research should focus on exploring the balance between interruption, reaction, and comprehension design objective parameters, each supported differently by various forms of information representation within these secondary displays. If tradeoffs can be determined for information design options across platforms and information types, then various usage scenarios can be reliably, and perhaps automatically, supported with optimal presentation features.

Interruption caused by the reallocation of attention from a primary task to a notification is clearly an important issue. Some notification systems are designed to attract attention and compel other activities, thus by their very nature they must interrupt users from some primary task and minimize delay in attending to a notification. However, many systems are intended to preserve as much primary task attention and performance as possible. Understanding how and when to best accomplish both is important to this field.

Reaction to a specific secondary information cue while performing a primary task is another important goal of a notification system. As the urgency of the primary and secondary tasks varies, the desired reaction and reaction time may vary as well. Understanding how to invoke urgent or gradual reactions and corresponding interaction through information design is a second challenge to notification systems research and evaluation.

Comprehension of information presented in secondary displays over a period of time may also support important facets of notification systems design requirements. Comprehension usually requires higher level cognitive processing, assimilation, and integration necessary to move information from a user's working memory to long term memory, and can be achieved or can fail independent of short term recognition of and reaction to information presence. Understanding how information encoding effects memorability and comprehension, while inducing interruption or facilitating reaction, is pivotal for supporting this likely requirement for secondary displays.

While demand for innovative and useful application of the latest technology continues to drive development, there are few guidelines to help designers determine aspects of a display that make it distracting to primary tasks or more suitable for communicating secondary information in certain situations. Although establishing guidelines based on observed and expected behavior is an important approach, lab-based empirical studies are also desirable in identifying how people react to specific features of displays. In a controlled but realistic environment, a researcher can study and compare differences between non-application-specific features and cues for information presentation. This allows establishment of guidelines and tradeoffs for information design possibilities that best support system development requirements and emerging ideas. However, for this body of knowledge to achieve any influence or utility for practitioners, notification systems empirical studies should conform to common themes, use common metrics, and report findings with common language. Focusing studies on these parameters will help achieve this external validity.

As a research community, we need to achieve this focus now, so this paper is designed to assist current notification systems researchers and newcomers to achieve greater context regarding our research challenges and approach. Section 2 outlines related work in each of the three notification systems design parameters mentioned previously: *interruption*, *reaction*, and *comprehension*.

Here, we also discuss some recent efforts at establishing or validating general guidelines applicable to visual notification systems, as well as imperatives for well-conducted empirical studies. Section 3 describes a series of empirical studies conducted to investigate information design tradeoffs resulting from use of animation of text in a secondary display. The results of the experiments and implications for design are discussed in Section 4. Section 5 provides a summary of our work and speculates on future directions for continued efforts in this field.

2. Related work

In this section, we provide a brief survey of some prior work relating to each of the three design objectives, in pursuit of the proposed unifying attention-utility theme for notification systems research. We begin by discussing specific studies that introduce key concepts and relate to each of the three parameters: interruption, reaction, and comprehension.

2.1. Design objectives for notification systems

Before we turn our consideration entirely toward empirical study of notification systems, we look at a few other important studies that have advanced understanding of notification systems information design. However, these individual studies perhaps were not designed to provide the extendibility critical for sequential research addressing an emerging design paradigm. Framing such studies around commonly accepted critical parameters provides a solution to this problem and motivates our research approach. Toward this end, we group discussion of related work by each of the three design objectives.

2.1.1. Interruption

User goals and usage scenarios for notification systems often have some requirement regarding the interruption of primary tasks. In the context of notification systems study, we define *interruption* as an event within the notification system prompting transition of attention focus from a primary task to a notification. User goals and usage scenarios for notification systems often have some requirement for or against primary task interruption. Some situations require that a notification system not intrusively disrupt user attention devoted to a main task, while other situations explicitly call for notification prompted task-switching. Still other usage scenarios allow some interruption to the primary task in order to accommodate acquisition of secondary information, but seek to minimize interruption before unacceptable primary task performance degradation occurs. User-initiated interruption are sometimes incorporated into a supporting design, allowing a user to choice when to attend to a notification based on natural breaks in a primary task. Depending on an application's interruption objective, the result of interruption can either be an unwanted *distraction* from an important task or *attraction* to valued information.

Certainly negative effects of interruption on other tasks caused by the presence of a changing notification display are a concern. Much of the prior work on distraction in notification systems considers secondary displays for in-vehicle information systems, where distraction from the primary vehicle control task can be harmful or fatal. Guidelines established in these areas suggest defining limited numbers and types of interactions with the displays, restricting the amount that displays change, and limiting the time that a display is present. (Ballas et al., 1992; Green, 1999; Lee et al., 1999; Tufano et al., 1996; Sheridan, 1991). In most desktop computer usage situations, when the consequences of distraction are not life-threatening, annoyance threshold seems to determine the amount of distraction that is acceptable, although research suggests that performance on an interrupted task will suffer for longer than simply the time required to perform the secondary task (Bailey et al., 2001). However, as ubiquitous notification displays and devices increase in popularity and are coupled with more attention-intensive primary tasks, understanding how to satisfy this design objective becomes increasingly more important.

Recent work by McFarlane presents additional applicable background for understanding aspects interruption and provides results of an empirical study evaluating four design implementations to coordinate interruptions (immediate, negotiated, mediated, or scheduled) (McFarlane, 2002). The tentative guidelines he established, which are particularly useful for supporting a user-initiated interruption design, exhibit design goal tradeoffs among the coordination methods. Negotiation-based interruption coordination appears to be best for many cases. Additionally, he introduces a taxonomy describing eight major dimensions of interruptions (McFarlane, 1998).

Horvitz's models and inference procedures present some hope for this design objective, an imperative driven by his belief that human attention is the most valuable commodity in HCI (Horvitz et al., 1999; Horvitz, 1999). These models are designed to improve notification utility by considering cost of user interruption and introducing notification presentation appropriately. To support this type of emerging notification adaptivity, we must be as certain as possible about comparative interruption properties of information design attributes so that they can be properly mapped to interruption levels. However, selection of information design for a notification system that is driven by inferred suitability of interruption will likely have impacts on the two other design objectives (reaction and comprehension) and affect overall system utility.

2.1.2. Reaction

A complementary yet competing measure to interruption is the rapid and accurate response to important information provided by notification systems, which we refer to as *reaction*. Generally, notification systems present cues intended to inform the user of information of interest. With respect to facilitating reaction, the ideal presentation of information minimizes unnecessary interruption while enabling a user to recognize changes in information state. As requirements for minimal interruption to a primary task become more important in user goals, requirements are also likely to increase for properties that allow new information to be detected with short, quick glances. Therefore, important in understanding reaction to notifications is research in preattentive processing that considers how information can be highlighted using colors, shapes, and motion, such that the information can be assimilated in a single glance (Enns and Rensink, 1991; Healey et al., 1996; Healey and Enns, 1999; Bartram, 1998, 2001; Bartram et al., 2001). Specifically, the research of Bartram considers the effectiveness in using motion cues to enable signal detection, identification, and reaction. This work examined the speed and accuracy with which motion cues can draw a key-pressing reaction, relative to other visual attributes like color and shape. The findings showed that for this purpose, motion cues outperform static representations in displays in the periphery of the screen.

Since existing desktop notification systems often employ constantly moving textual displays, it is particularly important to understand reaction tendencies associated with differences in textual motion implementations. Early work used moving, changing text as a method for presenting information in hands-off displays, for example, for people with disabilities or in work environments that require them to use their hands in other ways. Some of the earliest evaluations of constantly changing textual displays examined the perceptibility and readability of rapid serial visual presentations (RSVPs) of letters, strings, and words. For example, Foster found that participants could correctly identify about four out of six words in a sentence when rapidly presented a word at a time in a single visual location (Foster, 1970). Duchnicky and Kolers performed a series of experiments examining the readability of text scrolled on visual display terminals as a function of window size (Duchnicky and Kolers, 1983). They found that larger displays typically led to faster performance on reading tasks. These types of studies investigated rapid reaction to information, yet they did not consider more in-depth understanding of it.

2.1.3. Comprehension

While rapid and accurate reaction to an informational cue is important in many situations, often it is also vital to use a notification system with the goal of remembering and making sense of the information they convey at a later time.

We refer to this as *comprehension*. Users of notification systems may desire high levels of comprehension over time, even though they are unwilling to accept primary task interruption and only will devote occasional glances or limited interaction periods toward this objective.

Again, we consider research relating to textual motion as an initial example for studying relative comprehension of secondary display information. Juola found that comprehension of information was comparable when presented as RSVPs and in multi-line paragraph format (Juola et al., 1982). A study led by Granaas found that in scrolled displays, larger jumps (four to ten characters) led to better comprehension than smaller jumps (one to two characters) (Granaas et al., 1984). Kang and Muter, in comparing a tickering effect to a non-animated RSVP effect, found no difference in comprehension for a reading task (Kang and Muter, 1989). Our own recent efforts have focused on evaluation of various attributes (position, area, and color) in a secondary displays for supporting information extraction and comprehension as part of tasks requiring detection, estimation-ratioing or estimation-compare (Chewar et al., 2002). We found that the three attributes are significantly different in enabling comprehension at various levels of primary task degradation.

Studies like these have advanced our understanding of how comprehension can be best achieved with different forms of information design, but efforts comparing information representation that are most applicable to notification systems should capture tradeoffs among all three proposed critical parameters.

2.1.4. Tradeoffs among objectives

In studying the usability of various peripheral or secondary display information representations, many researchers focus only on information gained, without measuring the changes in primary task performance caused by these displays.

Similarly, some studies create an unrealistic testing environment given that the distinction between reaction and comprehension is unclear. While reaction and comprehension tasks are often closely related, the two objectives may imply differences to notification system information design. For example, individuals generally may want to know the status (thus achieving comprehension) of a stock quote, but during critical times they may want to quickly invoke some action (exhibiting a reaction) at a specific trading price. Similarly, when keeping track of the score of a ball game, users may desire to be aware that one team is far ahead of the other or that it is early in the game, but if the score remains close near the end of the game they want to focus attention on the game and react to specific changes. Cadiz's Awareness Monitor system semantically and functionally emphasizes the blending of these concepts, using a series of effects (including tickers) to address both reaction and comprehension issues (Cadiz et al., 2000). Gaining notification information from radios, cuckoo clocks, and television could lead to either reaction or comprehension and must be appropriately assessed.

One type of notification system for which users often must balance interruption, reaction, and comprehension is instant messaging (IM). Research on the effects of IM notifications on desktop computer tasks found that IM typically was disruptive to primary tasks, particularly so for fast, stimulus-driven search tasks and cognitively intense primary tasks (Cutrell et al., 2001). However, IM, like many notification systems, does not use smooth animation in its updates, possibly exacerbating the distraction. Maglio and Campbell performed a series of dual-task experiments to examine the tradeoffs in displaying information using animated textual displays (Maglio and Campbell, 2000). Participants performed a series of primary tasks where they were asked to edit a document. Simultaneously, a continually scrolling, start-and-stop scrolling, or fading display would show information. They concluded that continually scrolling displays are more distracting than displays that start and stop, but information in both is remembered equally well. Scrolling direction does not seem to effect performance, and additional cues that are auditory have a more negative impact than additional visual cues.

2.2. Evaluation of notification systems

Other recent efforts have sought to establish methods and frameworks for showing a wide range of data types in a peripheral manner, and should be readily usable for notification systems information design. Evaluated displays use visual presentation methods to show news, weather, sports, personal data, and other information in a small portion of the desktop. The "What's Happening?" system employed smooth animation of both text and graphics to show a wide variety of information types in an effort to build local community (Zhao and Stasko, 2000). Irwin used graphical encodings and changing text to notify users of new email, Usenet news, and changes to web sites (McCrickard, 1999). Microsoft's Sideshow, a sidebar multimedia awareness system, was developed to reflect information on email, software bugs, and much more (Cadiz et al., 2001). Sideshow system designers developed a number of guiding principles in building Sideshow, including the use of smooth animation, minimal use of audio, and easy transition to more in-depth sources of information. However, neither effort employed or validated design guidelines that were evaluated in a manner that allows far reaching applicability. Although efforts like these could have enormous potential for the advancement of notification systems science as a demonstration of empirically proven guidelines, they are limited to implementation reports that only provide cursory hints at effective information design.

We believe that as an emerging research area, the field of notification systems will quickly provide valuable, compounding results if a solid empirical model is embraced by its researchers for evaluation and reporting of systems and information design. This is particularly important since notification systems research already seems to lag behind its practitioners, evidenced as new systems are deployed with little fundamental research to support design decisions of new applications. To guide our advocacy of empirical study, we turn to a few influential works from other research areas within computer science and HCI, particularly Crowder's imperative for reporting of mathematical software experiments (Crowder et al., 1979) and Gray and Saltzman's plea for better design of experiments comparing usability evaluation methods (UEMs) (Gray and Saltzman, 1998). Crowder's work was introduced to provide authors and referees a set of suggestions that would produce better documentation of research contributions. Although he specifically targets reporting of computational and algorithm studies, his suggestions on experimental design and reporting standards are largely applicable to this context. Similarly, Gray and Saltzman motivate their treatise as a prevention for bad research: "for HCI practitioners, making choices based on misleading or erroneous claims can be detrimental—compromising the quality and integrity of the evaluation, incurring unnecessary costs, or undermining the practioner's credibility within the design team." Although their work provides a critique of UEM studies, it is orchestrated in a manner that supplies a concise, well-illustrated review of components required for experimental validity. Recent work by Chin also stresses the need for solid empirical studies within the user models and user-adapted systems research area (Chin, 2001).

In the section that follows, we discuss two empirical studies of information design features for notification systems that would be used with a text reading primary task. We designed and reported these experiments particularly mindful of our attention-utility theme and the three proposed critical parameters for notification systems: interruption, reaction, and comprehension.

3. Experiments

In this section, we describe two empirical studies examining animation of text within a desktop computer notification system—specifically, displays that use scrolling and fading effects to communicate information. These effects are used frequently on Web pages, in applications, and even in crawling displays at the bottom of many news and sports television channels. However, there are few guidelines that dictate when these animations are most appropriate for user goals relating to interruption, reaction, and comprehension.

Animation is a natural fit for secondary displays. Because secondary displays on the desktop are typically small, this often translates into some use of animation to cycle through items of interest via scrolling or fading techniques. However, a natural confound of using animation as a secondary display aid is that constant change in a visual field is a strong perceptive attention draw for humans (Ware et al., 1992). Consequently, animation naturally appears to distract people from their primary focus. People may be willing to accept a small amount of movement in the visual field as long as it serves some useful purpose (such as raising awareness or assisting in monitoring) and does not unduly distract them from their primary work focus. While Maglio and Campbell conducted studies on animated textual displays, they employed a 5-pixel jump when scrolling, creating a jerky effect that may have resulted in unnecessary distraction (Maglio and Campbell, 2000). However, our pilot-testing showed that a smoother ticker updating a pixel at a time may introduce less distraction and possibly affect the other design objectives (reaction and comprehension) as well.

We were specifically interested in determining whether or not levels of interruption, reaction, and comprehension in a dual-task system exhibit significant differences when any of three animation types are used in the secondary display: a smooth ticker, a fading display that gradually fades between pieces of information, and a RSVP-style "blast" that places items in the display without smooth animation. To examine these comparative properties of animated displays, two empirical evaluations were conducted. In both, participants were asked to complete a series of browsing tasks that while simultaneously watching a notification system, in the form of a secondary display, showing constantly changing news, weather, stock, and sports information.

The secondary display provided a continuous channel of information that was meant to be *non-interruptive* to the browsing task, but allow participants to monitor changes and *react* to specific items (monitoring tasks), as well as to gain *comprehension* necessary to answer longer-term knowledge-gain questions (awareness questions).

3.1. Experiment platform

Each experiment was conducted on Sun Sparcstation 2 workstations connected to a 15-in monitor with an optical mouse. Participants were run in small groups, one participant per computer. Experiment procedures were explained to participants verbally and again on the computer using examples. The experiments consisted of several rounds (six in the first experiment, eight in the second), each consisting of four browsing tasks, two monitoring activities, and up to five awareness questions. Therefore, a possible total of 24 or 36 browsing observations were recorded for each participant. The layout of the information on the computer screen is in Fig. 1. Independent variables, animation settings, and introduction of test conditions are unique to each experiment and is discussed later. Motivations for our experimental choices follow.

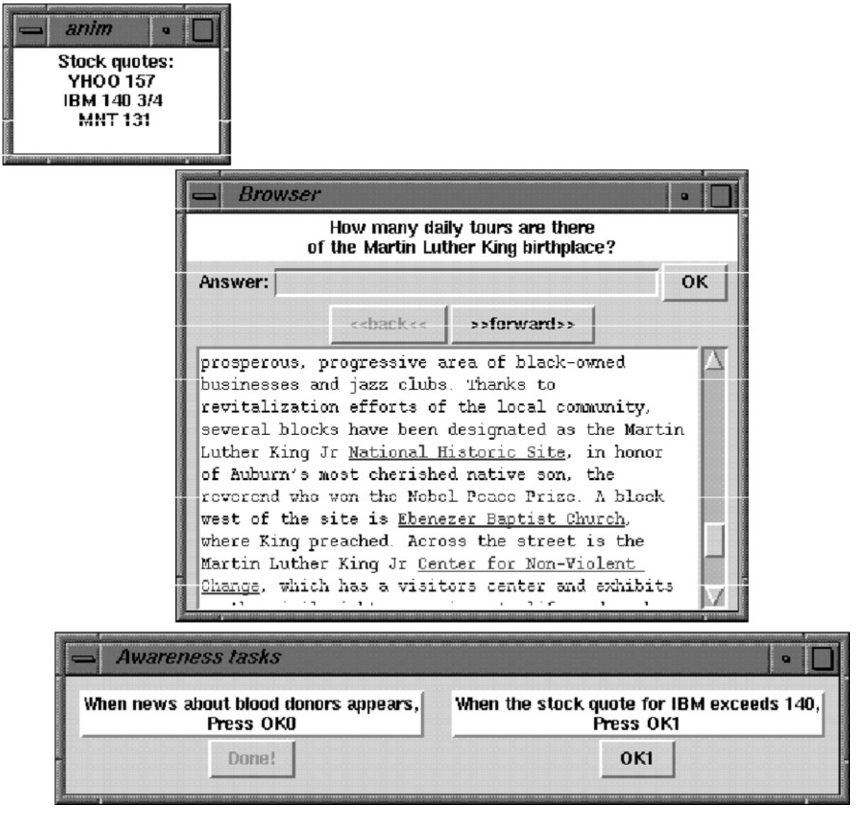

Fig. 1. Layout of the experimental environment experienced by participants. At the center is the browser used in the experiment. At the top of the screen is a secondary display that cyclically showed information with three different animation types. At the bottom is the area used for monitoring activity interaction.

3.1.1. Primary task

We selected the hypertext browsing task as the primary task because of its similarity with tasks that many people perform as part of their daily computer activities, during which they are likely to use an animated information display as a notification system. In performing the browsing tasks, participants used a simple browser and hypertext pages. The browser consisted of a textual information area containing a number of condensed pages from World Wide Web sites. The text-only information area contained highlighted, underlined links that pulled up other pages when clicked with the mouse. The participants navigated the information space by clicking on the links and by using the forward and back buttons. The browsing tasks were non-trivial: the participants had to read and navigate through a hypertext space to find certain information in the pages, enter it into a box connected with the browser, and press a button to continue. To minimize the typing required, all solutions to browsing task questions were numerical (for example, “In what year was Mount Rushmore carved?”). If an incorrect answer was entered, the interface beeped and the participant had to continue working on the problem until the correct answer was entered. When the correct answer was entered, the participant could proceed to the next browsing task.

3.1.2. Secondary tasks

While performing the browsing tasks, the participants used information in the secondary display to complete a set of monitoring activities and to answer a series of awareness questions. We felt that a simple question-and-answer session at the end of the primary task would not realistically simulate the concern that an individual may have for secondary information, or allow understanding of animation effects in simultaneous resource consumption, so we asked participants to react as soon as they saw certain information in the secondary display and try to remember all of the information they saw.

The secondary display cyclically showed instances of different types of information, such as sports score, a stock quote, and a weather report. Participants were asked to press a button when the information in the secondary display matches some criteria (for example, "When the temperature drops below 35, press OK1."). Each round included two such monitoring activities. If the button was pressed at the correct time (that is, after the information was presented), it was greyed out to alert the participant that the task had been completed successfully. If the button was pressed too soon, the interface beeped, and the button remained active.

At the end of each round, the participants were asked awareness questions that required recall of information shown in the secondary display. The questions were multiple-answer four-choice questions that addressed both content and temporal issues.

The first question in each set listed four types of information and asked participants to choose those that had been displayed. If they correctly recalled seeing information, later questions addressed specific details (such as which news stories appeared, which stock quotes constantly increased, or which sports team scored the most points). Therefore, if a participant correctly noted that news headlines had been displayed, later questions would present a list of headlines and ask the participant to select the ones that had appeared.

All of the information was fictional but realistic, and no attempt was made to intentionally deceive the participants with slightly different information (for example, participants were not asked to differentiate between a stock quote that constantly or intermittently increased).

3.1.3. Dual-task situation

Task goals between the two tasks were unrelated, and neither task contained any data-link constraints. We expected that participants would use a discrete split of their full attention to attend to both visual, intramodal tasks, and we did not place any indication of priority or urgency on any aspect of the experiment. Since both tasks required a moderate level of cognitive processing, we did not expect participants to exhibit task automaticity.

3.2. Data collection and evaluation

To compare performance among groups, times for all browsing tasks and monitoring activities and answers to the post-round awareness questions were collected. The results were analysed to determine whether differences in certain measures occurred for participants in different conditions (participants using different types of secondary displays in the first experiment, and participants using different sizes and speeds in the second).

3.2.1. Interruption metrics

We collected several measurements of primary task performance to identify comparative levels of interruption introduced by each animation type. The *browsing time* was the time from which the browsing task and browser information appeared on the screen to the time when the participant typed in the correct answer and pressed the OK button. The order in which browsing tasks were presented was held constant for all participants in each experiment. *Incorrect answers* were the number of answering attempts made before the correct answer was entered. *Bad link selections* captured the number of times a participant pressed the "Back" button to return to a previous page. The final primary task measurement was obtained at the end of the experiment; participants answered several questions relating to their preferences of animation types. One question asked about the amount of *perceived intrusiveness* to browsing caused by each animation type.

3.2.2. Reaction and comprehension metrics

We are interested in two general aspects of secondary task performance—reaction to information presence (monitoring) and comprehension of displayed information (awareness). Monitoring effectiveness can be measured by how long a participant takes to recognize the presence of information that he or she is searching for; *monitoring latency* captures the difference between the time that the information is present within the cyclic display and when the participant acknowledges it by pressing a button. For awareness, we use multiple performance measurements to capture animation type differences and similarities. When participants accurately report viewing kinds of information (such as local or world news, weather, or stock quotes) that were actually displayed during a round, this reflects a basic awareness hit, so *basic awareness hit rate* expresses the ratio of actual hits to possible hits. Conversely, when participants report viewing kinds of information that were not displayed, the *basic awareness false alarm rate* increases. Awareness measurements for more detailed information are similar, except we make a basic awareness hit a precondition for testing more detailed awareness because comprehension of information is difficult to imagine when the presence of the information is not recognized. *Detailed awareness rate* reflects successful answer choice selections—capturing recognition of correct and incorrect choices. *Detailed awareness false alarm rate* shows instances of participant comprehension confidence where there should not have been any (since the information was not properly understood). Finally, participant responses to the second end-of-experiment question express relative opinions about *perceived ease of awareness* for the three types of secondary displays.

3.3. Experiment 1—method

The first experiment compared relative performance when using fading, tickering, and blasting displays as well as when no secondary display was present. We focused on isolating three comparison factors with respect to alternate secondary display representations: performance changes in a browsing task, speed in identifying and reacting to notification information, and ability to remember information that appeared in a secondary display. Seventy undergraduate students participated in this experiment for class credit. The participants performed six rounds of browsing tasks, monitoring activities, and awareness questions.

3.3.1. Animation settings

In each round, participants completed four browsing tasks and two monitoring activities using either a fade, ticker, or a blast animation. The speed with which the information was displayed corresponded to the mean speeds for each device selected by the participants in a previous study (McCrickard, 2000b). While this resulted in different rates of information display for the animations, we felt it was a more realistic and ecologically valid measure of how people would use them. The ticker continually shifted horizontally one pixel every 50 ms, while the fade and blast updated their entire contents every 2 s. The fade required 500 ms to fade between items, while the blast updated instantaneously.

3.3.2. Presentation of test conditions

As a base case, one group of participants ($n = 15$) did not have any animations present at any time and as such performed only the browsing tasks. For the other groups, all participants experienced all animations, with orders based on a Latin square design (browse–fade–ticker ($n = 17$), fade–ticker–browse ($n = 17$), or ticker–browse–fade ($n = 21$)). A different animation was used for each of the first three rounds with the order repeated on the last three.

3.3.3. Data validation and screening

Prior to analysis of the data, we screened each of the 70 participants' performance results of all 24 browsing tasks, enforcing expected experimental conditions. As a baseline definition for attentive performance, we required completion of each browsing task within 5 min and correct answering of the browsing objective within five attempts. We excluded data associated with a browsing task if the participant failed to meet either of these conditions. Out of the 1680 possible browsing task results, 1629 or nearly 97%, fell within this criterion and formed the analysis set. This screening caused some differences and additional observations from previous analysis (McCrickard et al., 2001), but is thought to be essential in maintaining the empirical study's internal validity.

3.4. Experiment 1—results

Analysis of the data reveals several performance strengths and weaknesses for each of the three secondary display text animation types—ticker, fade, and blast. No single animation type is superior for all aspects of simultaneously minimizing distraction, facilitating reaction to changing information, and enabling information comprehension. This implies that display designers must prioritize importance of these design objectives, and select or eliminate animation types accordingly. To support such decisions, we have analysed aspects of each design objective and can often identify a single animation type allowing significantly better (or worse) performance according to each metric. Furthermore, analysis of filtered data, capturing instances of optimal performance under key metrics, has allowed deeper understanding of performance trends resulting from the use of each animation type. This section presents details of these analyses according to potential design parameter—grouped as either primary task, secondary task or overall dual-task system objectives.

3.4.1. Primary task performance

The primary task in this experiment is the browsing task—requiring participants to search for an answer to a question within hyperlinked pages of a simulated browser. During this task, there are several measurable aspects of participant performance which capture aspects of unwanted interruption (distraction) levels: *browse time*, *incorrect answers*, *bad link selections*, and *perceived intrusiveness*. Testing of primary task performance varies between participant groups only in the presence (control group did not have one) or type (ticker, fade, or blast) of an animated secondary display.

Impact on browsing speed. Browse times for tasks during which participant attention was split between secondary ticker, fade, or blast displays were not significantly different from each other. Likewise, control condition browsing tasks, with no secondary display, were not completed significantly faster than tasks with any single type of animated display ($F(3, 1625) = 1.93$, $MSE = 2460.4$, $p = 0.12$). Confidence in the difference between control and blast was strongest ($z(346) = 0.82$, $p = 0.41$). The control group was also not significantly faster than the mean task completion time ($z(346) = 1.06$, $p = 0.29$). As expected, mean completion times show the control group had the fastest task completion times; the ticker and fade follow, while the tasks with the blast present were the slowest. Fig. 2 shows mean completion times with 95% confidence intervals for display type and control groups, in relation to the population mean. Note that the actual differences in average completion times vary by only 4 s in a 60 s frame of reference.

Impact on browsing comprehension. Analyzing the number of incorrect answers supplied before entering the correct answer to the browsing task shows a significant effect due to animation type or lack of secondary display presence ($F(3, 1625) = 5.58$, $MSE = 0.54$, $p < 0.01$). Specifically, questions asked under the control condition were answered with less incorrect attempts than on average ($z(346) = 2.90$, $p < 0.01$). This significant difference is primarily due to the difference

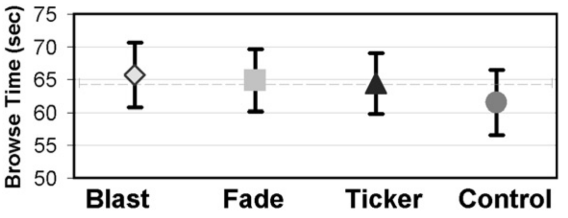

Fig. 2. Mean completion times for display type and control groups.

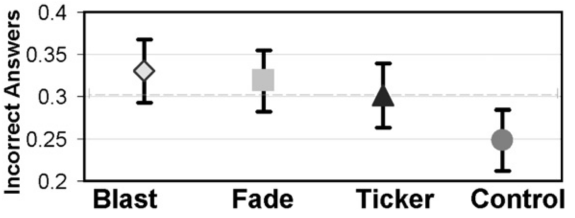

Fig. 3. Relationships between results by type.

between the control condition and the blast display ($z(346) = 2.18$, $p = 0.03$). However the differences between the control condition and fade, as well as the control and ticker, are not significant ($z(346) = 1.89$, $p = 0.06$ and $z(346) = 1.40$, $p = 0.16$). Examination of means for minimal numbers of incorrect answers produces an identical ranking as fastest browse times: control, ticker, fade, and then blast were most effective. The relationships between results by each type, with 95% confidence intervals, are shown in Fig. 3. The meaning of actual differences in incorrect attempts is also slight—participants were most likely to have no incorrect answers, even for tasks during which the blast was present.

Impact on link selection accuracy. Unlike the previous two primary task performance measures, average numbers of bad link selections were not minimized with the control condition. Instead, tasks performed with the presence of the ticker resulted in the lowest average of bad links selected; the fade display average was next lowest, and the blast and control conditions were nearly equal. None of the conditions were statistically different from each other ($F(3, 1625) = 0.28$, $MSE = 8.28$, $p = 0.84$). It is also important to note that, again, the differences in actual data were only slight—under any of the four conditions participants averaged between 1.5 and 2 selections of non-useful links.

3.4.2. Secondary task performance

Non-control participants ($n = 55$) performed their primary task (browsing for an answer to a question) while a secondary display was present. During each round, these participants were given the same secondary task, for which they were required to recognize specified information in their secondary display. Although the secondary tasks are the same for all participants in a given round, these tasks are supported by different types of animation within the display (blast, fade, or ticker) for each of three participant groups. After every other round, secondary display animation types rotate to the next group of participants, ensuring that learning effort is equally introduced for each type.

Reaction time for recognizing a key information state. Results of the monitoring latency analysis show relevant and significant differences between the animation types ($F(2,618) = 9.53$, $MSE = 2818$, $p < 0.01$). Notably, monitoring latency is only one of two metrics (the other is monitoring false alarm rate) by which the ticker rates significantly worse than at least one other animation type. Here, it is evident that participants required significantly more time to notice information introduced by the ticker than by the blast ($z(852) = 2.84$, $p < 0.01$) or fade displays ($z(852) = 2.49$, $p = 0.01$). Examination of the data shows the mean latency for ticker monitoring was 54.3 s, while monitoring with the blast had an average 33.6 s delay, 35.5 s with the fade. Therefore, the average cost of using a ticker is an additional 65% of the time participants take to notice information in a blast display—since this time period is on the order of a half minute, this difference could be quite relevant. Fig. 4 shows these differences between monitoring latency for the three animation types.

Memorability of secondary information. Recalling that a high basic awareness hit rate indicates high participant tendency in recognizing presence of information categories within the secondary display, differences between all three animation types are apparent. The ticker enables significantly better basic hit rates, compared to blast ($z(165) = 4.57$, $p < 0.01$) and fade displays ($z(165) = 8.11$, $p < 0.01$).

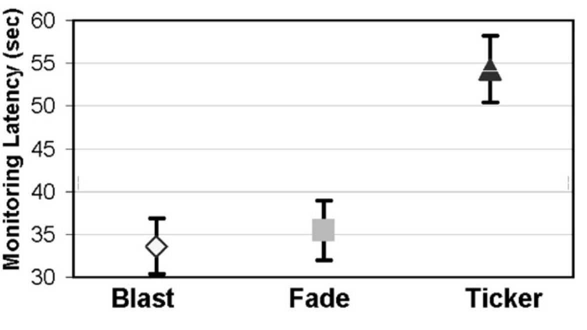

Fig. 4. Differences in monitoring latency.

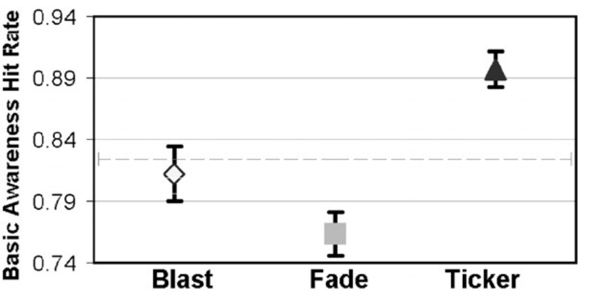

Fig. 5. Awareness hit rate differences.

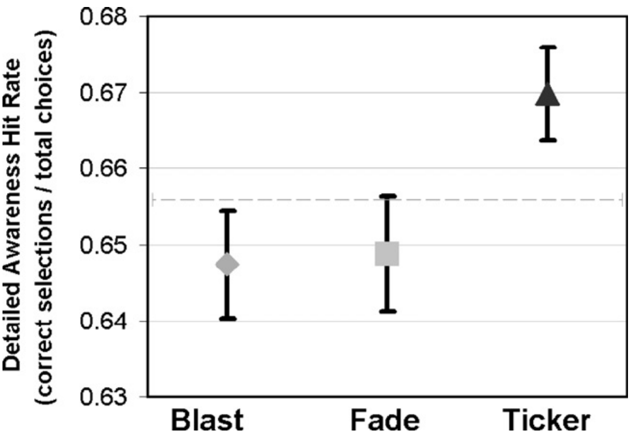

Fig. 6. Comparison of detailed awareness results.

Additionally, blast displays are significantly better by this measure than fade displays ($z(165) = 2.40$, $p = 0.02$). Most importantly, these differences are relevant as well—mean hit rates with ticker displays can be expected to fall within 88–91%, while rates using blast and fade displays can only be expected to fall between 79–83% and 74–78%. These differences can be clearly seen in Fig. 5, since the 95% confidence intervals for any pair do not overlap.

When participants indicate seeing information categories which were actually not resent, this increases the basic awareness false alarm rate associated with the secondary display animation type. This metric is the second of two by which the ticker exhibited significantly worse performance compared to the other animation types. Basic awareness false alarm rates were minimized with the fade displays—we can be most confident in the difference between fade and ticker ($z(330) = 3.17$, $p < 0.01$), but the blast displays were also significantly better than the ticker ($z(330) = 2.16$, $p = 0.03$). There is no significant difference between the blast and fade animations. However, relevance of even the significant differences may be marginal at best, since the 95% confidence intervals for rates of all three animation types span only a 13 point range.

Comprehension of secondary information. Once a participant accurately reports seeing a particular information category throughout a round in the secondary display, his or her comprehension of that information is further tested. Understanding which stock prices were consistently rising or falling—while performing the browsing task—is an example of maintaining secondary information awareness. The ticker animation enables significantly better detailed awareness than the blast ($z(1072) = 3.32$, $p < 0.01$) and the fade animation ($z(1008) = 3.04$, $p = 0.03$). There is no significant difference between the blast and the fade ($z(1008) = 0.19$, $p = 0.85$). Fig. 6 illustrates these differences; however, examination of mean values or the y-axis scale shows little relevance for these differences. At the very most, tickers can only be expected to yield a 4% increase in detailed awareness levels—a slight improvement to comprehension that already includes approximately two-thirds of the secondary information presented. Differences in detailed awareness false alarm rate follow suit—while there are significant differences between animation types, the differences are so miniscule they are irrelevant. However, unlike the relatively poor performance of tickers in minimizing basic awareness false alarm rate, they excel at minimizing detailed awareness false alarm rate. Ticker rates are significantly lower than blast rates ($z(1072) = 4.28$, $p < 0.01$) and fade rates ($z(1008) = 3.42$, $p < 0.01$), and again here, there is no significant difference between the blast and fade animation types.

3.4.3. Dual-task performance

We have discussed various assessments of primary and secondary task performance, but actual design scenarios may require animation used in a secondary display to effectively and simultaneously support both parts of the dual-task system. For instance, in some cases, minimizing distraction to a non-critical primary task may not be as important as presenting constantly changing secondary information that can be quickly identified. In other cases, the primary task may be critical and largely uninterruptible, although a long-term understanding (perhaps throughout the duration of an afternoon) of secondary information is desired. While the numerous combinations of design priorities and somewhat limited generalizability of this study restricts the utility of an exhaustive dual-task analysis, we gain some relative indications of dual-task effectiveness for the three animation types based on the final participant survey question and a filtering analysis technique, both which are described below.

Tradeoffs resulting from ideal primary task performance. A snapshot of dual-task performance under likely system design goals can be achieved by filtering the data to include only task instances reflecting desired performance results. This allows comparison of the animation types across the other metrics to anticipate likely tradeoffs in cases of otherwise ideal performance. Certainly, such "what if" analyses cannot be extrapolated into causal inferences and predictions of future performance, but it is sometimes useful to capture significant and nonsignificant filtering effects. To define the pool of data representing optimal dual-task performance, we set the following conditions: participant browse time on a given trial must be less than one standard-error from each question's mean browse time and the correct browsing answer must be supplied within two attempts. In order to consider monitoring and awareness data for a given round and end-of-equipment survey results, we required that the participant complete at least one of the four browsing tasks under these optimal conditions. Interestingly enough by itself, 102 sets of results qualified (out of 1680 possible, or approximately the top 6%)—corresponding secondary animation types were: 28 blast, 20 fade, 25 ticker, and 29 instances of browsing tasks with no secondary display (control condition). The set of optimal trials was produced by 41 of the 70 participants. These small samples were compared to each other and the set of "all trials" using the same measurements discussed above.

Several notable differences existed between the optimal performance trials and all trials. We were surprised to see that by every measurement, except detailed awareness false alarm rate, animation types exhibited significant differences.

Although incorrect answer attempts was a filtering criteria for the optimal performance group, we observed that tasks completed during blast display and no display (control) rounds had a significantly lower incorrect answer rate than rounds with the ticker or fade. In fact, the correct answer was supplied on the first try for all blast and control rounds within this group. There is some indication that bad link selections were affected by the presence of a secondary display, since tasks performed under the control condition resulted in fewer instances of bad selections ($t(28) = 1.84$, $p = 0.08$). While the basic awareness hit rate was significantly lower with the fade display for all trials, the optimal performance group had the most predictably low rates with the blast display (confidence intervals range from 0.66 to 0.78)—the ticker still resulted in the highest hit rates (confidence intervals of 0.83 to 0.92). A final difference between the set of all trials and optimal case results was animation effectiveness for basic awareness false alarm rates. While the all trials sample exhibited lower false alarm rates with fade displays, the optimal cases achieved significantly better results (lower monitoring false alarm rates) with the ticker compared to the blast ($z(50) = 2.85$, $p < 0.01$). Although there were not significant differences between the ticker and fade displays ($z(54) = 0.13$, $p = 0.89$), fade did allow better false alarm rates than the blast ($z(50) = 2.06$, $p = 0.04$). Other results from optimal trials agree with the findings from the all trials analysis.

3.5. Experiment 1—discussion

Table 1 summarizes the findings presented for both groups, according to all primary, secondary, and dual-task measurements discussed above, as well as three post-test questions about perceived intrusiveness, ease of maintaining awareness, and expected frequency of use. Recommendations made (positive and negative) are based on significant findings ($p < 0.05$) from inferential statistical analysis. Differences in recommendations for the optimal performance group are noted separately. A few general observations can be made from these results about the presence of secondary displays and the three animation types (ticker, blast, and fade).

No impact on browsing speed. Perhaps the most noteworthy finding is the lack of impact the presence of a secondary display had on the primary task. Out of three measurements of primary task performance interruption, only one showed significantly stronger performance for the control group (no secondary display). Browsing times for the control group were not significantly faster, and these participants chose about the same numbers of bad hypertext links during their browsing experience. However, answers to browsing tasks without a secondary display were answered correctly in fewer attempts. This indicates that, although presence of a secondary display may affect primary task concentration somewhat, focus is quickly recovered and the task is completed in similar reading tasks.

Tradeoffs between reaction, memorability, and comprehension. Since secondary display animation types are not significantly different in the amount of measured distraction caused to the primary task (although we observed that user opinion favors ticker as being less intrusive), perhaps the three animation types can be best judged according to secondary display objectives of facilitating reaction to changing

Table 1
Recommendations for various notification systems design objectives, based on significant differences observed in the experiment. "Optimal Trial Only" column indicates result findings for the filtered optimal performance group

Ticker vs. Fade vs. Blast animation	Recommended	Not recommended	Optimal trials only
Primary task	—	—	—
Browse time	No secondary display	—	Recommended—blast
Incorrect answers	—	—	—
Perceived intrusiveness	Ticker	Blast	(Same)
Secondary task			
Monitoring latency	Blast, then fade	Ticker	
Basic awareness hit rate	Ticker	Fade	Not recommended—blast
Basic awareness false alarm rate	Fade	Ticker	Recommended—ticker. Not recommended—blast
Detailed awareness hit rate	Ticker	Blast	—
Detailed awareness false alarm rate	Ticker	Blast	(Same)
Perceived ease of awareness	Ticker	Blast	(Same)
Dual-task			
Expected frequency of use	Ticker	Blast	(Same)

information (monitoring) and enabling information comprehension over a longer period of time (awareness). Clearly, the ticker seems to be the better choice for maintaining basic and detailed awareness. However, we see a weakness in the ticker at enabling quick recognition of changing information, or minimizing monitoring latency, which is better facilitated by either the blast or fade. Results are contradictory for all trials and the optimal trial sample as to the effectiveness of the ticker for preventing basic awareness false alarms (reporting to see categories of information which were not actually displayed), although from this we can infer that expert browsers may be able to monitor tickers more accurately.

Slow tickers tested here appear to support minimization of primary task interruption, reaction and comprehension facilitation to changing secondary information, as well as user satisfaction for ease of use and expected use frequency. This makes them particularly well suited to support various design objectives of dual-task systems. Our test of filtering data to capture instances of optimal performance supports this conclusion, since tickers exhibited the same distinctions over blast and fade animations. However, selection of an animation type for a secondary display should be guided by Table 1, ensuring that the most important design objectives are best supported.

3.6. *Experiment 2—method*

The previous experiment suggested that there are differences in performance when using the fade and ticker displays. In a follow-up experiment, we wanted to explore whether certain factors, namely display size and animation speed, impacted performance in any way. Perhaps making the display area larger would result in faster recognition times and allow the awareness questions to be answered with greater accuracy, or perhaps a slower speed would be less distracting, resulting in lower times on the browsing tasks.

Since the blast display resulted in performance similar to the fade display and was consistently rated as the least favorite display by participants in the first experiment, it was not used in the second experiment. Ninety-one undergraduates participated in this experiment for class credit. The materials and procedure were similar to the ones used in the previous experiment with the differences described here.

The number of rounds was increased from six to eight, since it was determined that participants would still be able to complete the experiment within the requested hour. The awareness question introduction criteria also differed from the first experiment. In the first experiment, the first question asked participants to select the types of information that were displayed, then for each case where the participant stated correctly that a type of information was displayed, two additional questions were asked about that information, the first relating to content and the second relating to order. In this experiment, each participant answered all five of the questions. This change seemed reasonable since a cue such as a word or phrase in a question can aid retrieval from memory. Of course, this experiment also included three different implementations for the two types of animation tested. Testing combined variable assumptions seemed to also be an ecologically valid assumption.

3.6.1. *Animation implementation settings*

The participants were presented with a secondary display having one of three animation implementation characteristics: normal size and speed, normal size but slow speed, or small size but normal speed. The *normal* displays were used as the comparison point for the small and slow displays. Normal displays used large display areas and fast speeds, though both well within the ranges of sizes and speeds selected by participants in a previous study (McCrickard, 2000a). Both the fade and ticker had a width of 1180 pixels (about 160 characters) with a height of one line. This size was chosen because it fits nicely along the top or bottom of the screen and because it is large enough to hold long streams of information (such as news headlines and weather bulletins) in their entirety. The ticker speed was at the upper range of the possible speeds for the platform, one pixel per 20 ms. The fade cycle step had a 100ms delay between each of five steps with a 3 s delay before the next fade.

The *small* display used a smaller area but the same speed as the normal display. The fade and ticker width was more than halved to 840 pixels (about 70 characters), small enough to fit above a single terminal window. This reduction in size meant that most streams of information could not be shown in their entirety. The *slow* display was the same size as the normal display, but slower. The speed was chosen to be at the slow end of the range selected by participants in the previous study. The ticker updated at a rate of one pixel every 140 ms. The fade updated one shade every 150 ms with a delay of 9 s before the next fade. The size for the widgets was the same as in the normal display.

3.6.2. Presentation of test conditions

A between-subjects size and speed condition was added to the within-subjects design testing the normal implementation conditions used in the first experiment. Participants alternated between using fade and ticker every other round, with one group starting with a fade, and the other group starting with a ticker. Therefore, half of the participants a round animation order of fade ($\times$ 2), ticker ($\times$ 2), fade ($\times$ 2), ticker ($\times$ 2), and the other half saw ticker ($\times$ 2), fade ($\times$ 2), ticker ($\times$ 2), fade ($\times$ 2). In summary, there were six groups of participants differentiated by animation implementation (normal, slow, or small) and starting animation (fade or ticker). Each group had 15 participants except the slow fade-first group with 16 participants.

3.6.3. Data validation and screening

To prepare the data from the experiment for analysis, we used the same screening criteria as discussed for Experiment 1, again ensuring the study's internal validity. That is, we required each browsing task to be completed within 5 min and each browsing question to be answered within five attempts. Browsing tasks which did not conform to these conditions were eliminated from further consideration and analysis. Twenty-four browsing tasks were performed by the 91 participants, providing a total of 2184 possible trial results. However, 116 trials (5.3%) failed to meet one or both of the screening criteria, and were excluded from the analyses.

3.7. Experiment 2—results

This analysis was conducted in a very similar fashion to that of Experiment 1: we examine various aspects of primary, secondary, and dual-task performance for nine different categories or types of secondary displays. Since we tested versions of ticker and fade animations presented normally or at slow rates or within a smaller-sized window (see above for specifics), we are able to compare each of the three animation implementations (i.e., normal, slow, and small) for tickers and fades, as well as normal, slow, and small displays regardless of animation type. This allows possible distinction between animation implementations or general display settings for each potential design goal. A single table summarizing all findings is presented at the end of this section (Table 2).

3.7.1. Primary task performance

The same four primary task measurements used in Experiment 1 and discussed earlier (see Section 3.2) serve as the metrics here: *browse time*, *incorrect answers*, *bad link selections*, and *perceived intrusiveness*. Of these four measures, only two show some differences between animation types and display settings—browse time and incorrect answers. Instances of bad link selections and participant perceptions of browsing intrusiveness did not differentiate any of the six types of secondary displays from each other.

Impact on browsing speed. The results from the browse time analysis present the most discriminating insights into this experiment's primary task performance ($F(5, 2906) = 2.21$, $MSE = 14770$, $p = 0.05$). Fig. 7 illustrates these differences well,

Table 2
Animation implementation recommendations by notification system goal, based on significant findings from the second experiment

Ticker vs. Fade animations, using slow, small or normal implementations	Recommended	Not recommended	Comments
Primary task			
Browse time	Slow ticker, or any slow	Small ticker, or any small	
Incorrect answers	Small fade ($p = 0.065$)	Small ticker ($p = 0.071$)	
Bad link selections	—	Small ticker ($p = 0.062$)	
Perceived intrusiveness	—	—	No differences
Secondary task			
Monitoring latency	Any fade	Normal or slow ticker	
Basic awareness hit rate	Small ticker	Slow ticker	
Basic awareness false alarm rater	Slow ticker	Any ticker, or small fade	
Detailed awareness hit rate	Slow fade	Any ticker, small fade	
Detailed awareness false alarm rate	Slow fade	Small fade	
Perceived ease of awareness	—	—	No differences
Dual-task			
Expected frequency of use	—	—	No differences

showing the 95% confidence intervals for each sample. Visual inspection allows identification of significant differences, which are sample pairs that do not have overlapping confidence intervals. Overall, slow display rates allow significantly faster browsing completion times ($z(944) = 2.24$, $p = 0.03$), while all small displays (ticker and fade displays considered together) had significantly slower times. Although this trend can be observed for both ticker and fade displays individually according to mean completion times, only the tickers exhibit significant browse time differences between slow and small display settings ($z(452) = 2.06$, $p = 0.04$). Mean performance was longer for tasks with the normal ticker compared to rounds with the slow setting, but not significantly. Since the actual differences in the data reflect variations of up to 20 s within a 2 min timeframe, distinction between slow and small display settings may have high relevance for an actual application.

Comparing numbers of incorrect answers to browsing tasks supplied by participants using the various types of displays shows a few marginally significant differences, as well as some interesting trends. Perhaps the most remarkable aspect of Fig. 8, which depicts the means and confidence intervals of wrong answer attempts, is that the implementation versions of the secondary displays for fades and tickers have entirely different effectiveness orderings. For instance, with fade displays, small implementations best minimize incorrect answer instances, followed by normal, and then slow versions. However, tickers have equally good results with slow or normal implementations, and the small tickers perform worst in terms of incorrect answers.

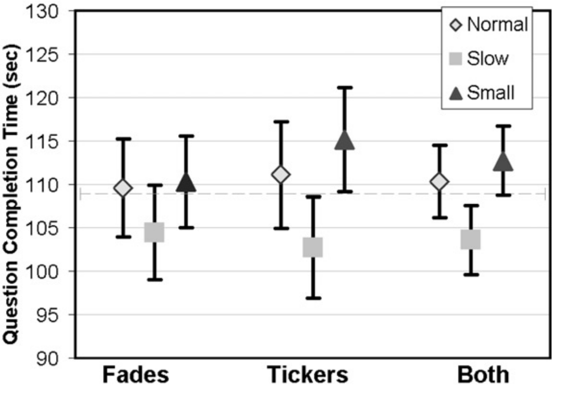

Fig. 7. Browse time results.

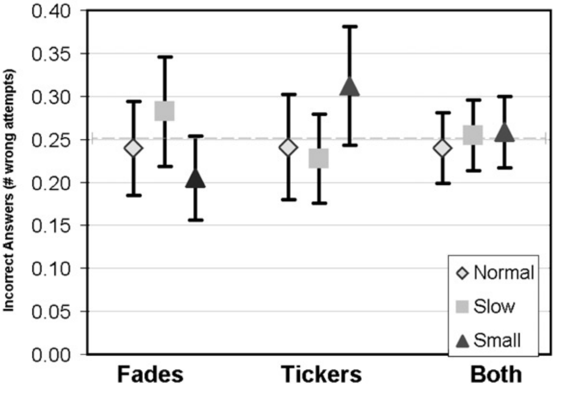

Fig. 8. Means and confidence intervals of wrong answer attempts.

The difference between small fades and small tickers is marginally significant ($z(452) = 1.78$, $p = 0.08$), and both are also marginally different from the population mean ($z(492) = 1.85$, $p = 0.07$ for the superior small fade and $z(432) = 1.82$, $p = 0.07$ for the inferior small ticker). Most interestingly, the same general graphical pattern present within Fig. 8 is repeated in Fig. 9, which shows the various amounts of bad link selections, however, there are no significant effects ($F(5, 2906) = 0.70$, $MSE = 10.18$, $p = 0.62$). Here, only the small ticker seems distinct from the other secondary display types, with its near non-inclusion of the population mean; however, there is only a marginally significant difference between small tickers and the population mean with respect to bad link selections ($z(492) = 1.87$, $p = 0.06$). For both measurements (incorrect answers and bad link selections), the relevance of these marginally significant differences appears to be low, since all secondary display implementations under both measures are well within the same whole number values of the metric.

Preference indifference. While participant survey answers distinguish the animation types in the first experiment, subjective ratings of the fade and ticker animations were largely indiscernible in this experiment. Confidence intervals for perceived intrusiveness all fell within 0.55 and −0.69, and always included zero, which was the middle response value available to the participants. Survey question responses for perceived ease of maintaining awareness and expected frequency of use produced similarly unexciting results, indicating strong preference indifference for all nine implementations. For this reason, no additional analysis of these question results is presented.

3.7.2. Secondary task performance

As in the first experiment, primary task performance in the second experiment tended not to differ between animation types or implementation options. Considering the lack of differences apparent though the survey questions, this elevates the importance of the secondary task metrics somewhat. We used the same measurements of secondary task performance as in the analysis of Experiment 1, which are introduced in Section 3.2: *monitoring latency*, *basic awareness hit rate*,

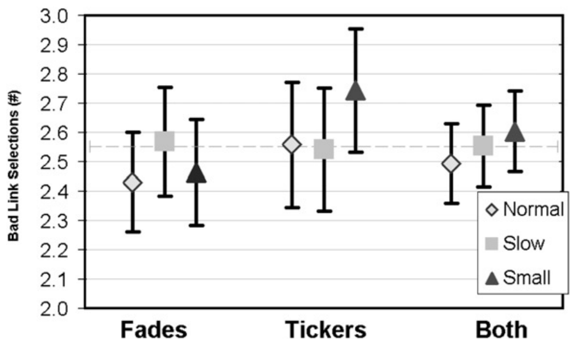

Fig. 9. Bad link selections.

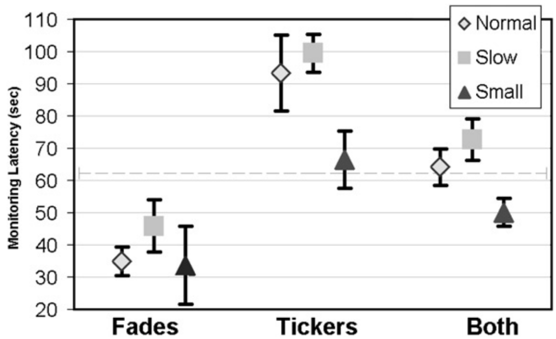

Fig. 10. Recognition times.

basic awareness false alarm rate, *detailed awareness rate*, and *detailed awareness false alarm rate*. This part of the analysis shows many differences or potential design tradeoffs for the various animation type implementations.

Reaction time for recognizing a key information state. Monitoring latency, or the time difference between the presentation of certain information and the participant's acknowledgment of seeing it, exhibits significant differences due to animation implementation ($F(5, 1289) = 79.82$, $MSE = 2032$, $p < 0.01$). Specific differences are between the fade and ticker displays, echoing the results found in the first experiment. Fade animation enabled much quicker recognition of information presence within a changing display than tickers; this is quite apparent in Fig. 10. For instance, in comparing the normal fade and ticker implementations, mean latency with the fade displays was about one-third that with ticker displays—a highly relevant difference on nearly any scale. The difference between all of the fade implementations are significantly better compared to all three of the ticker implementations ($p < 0.01$ for all but slow fade vs. small ticker, which is $z(240) = 2.84$, $p < 0.01$), yet the three fade implementations are indistinguishable from each other. The small ticker stands out from the other two ticker implementations (which show no difference from each other), allowing significantly lower latency compared to normal ($z(240) = 2.61$, $p = 0.01$) or slow ($z(240) = 3.09$, $p < 0.01$). Curiously, small displays in general show significantly lower monitoring latency than either normal ($z(480) = 2.71$, $p = 0.01$) or slow ($z(480) = 4.04$, $p < 0.01$) display versions.

Memorability of secondary information. Analysis of basic awareness hit rates also show significant differences between the several pairs of secondary display implementations, however, hit rates with all display types were within about eight points, centering just below 0.90. Therefore, unless an application is extremely sensitive to hit rate imperfection, this measure lacks relevance. As Fig. 11 shows, the three fade displays exhibit no differences from each other, even in amount of participant hit rate variance, which we can approximate from the relative length of the confidence intervals. Ticker implementations show differences between pairs:

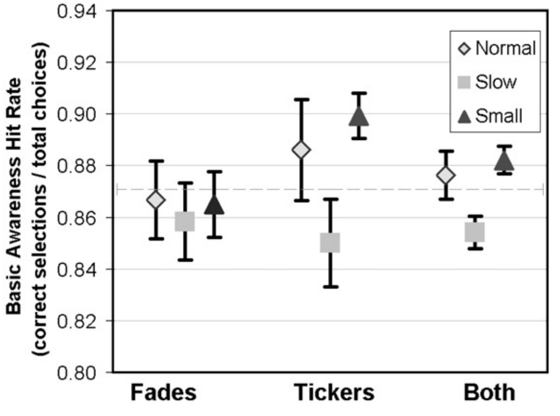

Fig. 11. Means and confidence intervals for awareness hit rates.

slow tickers had significantly lower hit rates than small tickers ($z(120) = 3.73$, $p < 0.01$) and the difference between slow and normal tickers was marginal ($z(120) = 1.94$, $p = 0.05$). Additionally, we can say slow displays had significantly lower basic awareness hit rates than the population mean ($z(120) = 3.03$, $p < 0.01$), and small displays had significantly higher rates ($z(120) = 3.64$, $p < 0.01$).

Both basic and detailed awareness measurements of secondary task false alarm rates, which reflect participant reports of seeing information which was not actually present, show only that all animation implementations produce similar performance. There were significant differences: small fade displays consistently yielded higher false alarm rates while slow implementations proved to be more effective—tickers for basic awareness and fades for greater detail. However, particularly for the detailed awareness false alarm rates, examination of the scale of changes reveals very slight and irrelevant variations between the animation types. All basic awareness false alarm rates fell within a 15-point span, centering at about 0.30; individual condition confidence intervals spanned between three and ten points. Detailed awareness false alarm rates were all between 0.11 and 0.08, usually spanning only two points each.

Comprehension of secondary information. Analysis of detailed awareness hit rates would not be noteworthy, except for the radical difference in the slow fade display. Although there are other significant differences between other secondary displays types, such as between the normal and small fade, the normal and slow ticker, and the marginally significant difference between the normal and small tickers, certainly no other pairwise comparison approaches one including the slow fade in terms of difference certainty (always $p < 0.01$) or relevance (allowing a 20-point improvement in awareness). This can be clearly seen in Fig. 12, which shows the various means and confidence intervals for the detailed awareness hit results.

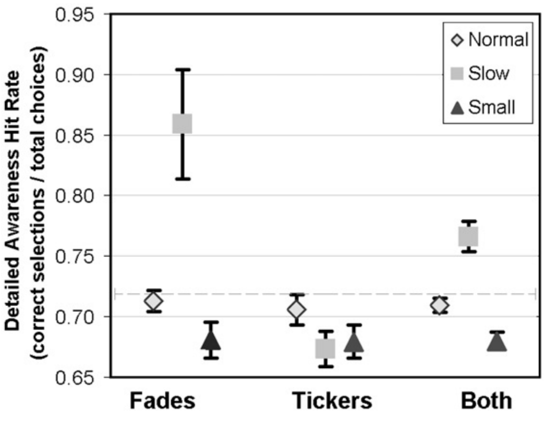

Fig. 12. Comprehension comparison.

3.8. Experiment 2—discussion

Significant differences ($p<0.05$, except where noted) found between the secondary display animation implementations for each of the primary and secondary task measurements are summarized in Table 2. From this table and the results presented above, there are several relevant findings which impact general understanding of the normal, small, and slow implementations of ticker and fade animation. These findings indicate that the slow fade implementation may be the overall best secondary display animation type tested.

Tradeoffs between reaction and comprehension. When we considered the relevance of the significant differences found with the various measurements, it appears that only the monitoring latency and detailed awareness hit rates provide solid discriminating data. No differences were found with primary task browse times, and the other measurements and the subjective survey questions did not demonstrate meaningful differences. However, even when these measurements exhibiting low relevance are considered in aggregate, the slow fade (and slow conditions in general) does not perform significantly worse than other implementations. This is not true for slow or small tickers or small fade displays. Focusing on the relevant results, all of the fade animation implementations have significantly lower monitoring latency results than the versions of the ticker, but neither the normal, small, nor slow fades can be differentiated from each other by this measure. However, the slow fade demonstrated significant superiority in the hits on detailed levels of awareness of information in the secondary displays.

These results are very important when considered together with the results of Experiment 1—there the ticker enabled significantly higher awareness than either the blast or the fade displays, and all displays had detailed awareness hit levels between 0.64 and 0.68. If this experiment had tested only normal and small displays, this result may have been replicated. However, testing of the slow fade implementation refinement reveals a much more suitable comprehension-enabling device, reversing the general understanding gleaned from Experiment 1. Additional comparisons between the results of the two experiments, to include instances of replication, are presented in the following section.

4. General discussion of findings

This section discusses the aggregation of findings from the two experiments. This includes a summary of result replication between experiments, comments about result generalizability, and an iteration of contributions made from this effort. The goal of the empirical evaluations was to explore the balance between interruption, reaction, and comprehension using secondary animated displays in dual-task situation with a browsing task. Neither tickering, fading, nor blasting secondary displays significantly interrupted users from a primary task (compared with the control group without a secondary display), yet could effectively communicate information. The type of animation had various performance impacts on monitoring activities and awareness questions. The second experiment showed that changes in size and speed also could impact performance on monitoring activities and awareness questions. In both experiments, eight measurements were used to assess the impact of animation types and/or secondary display implementations on primary, secondary, and dual-task performance. Additionally, three subjective questions captured participant opinions about the animation types, in order to test differences in perceived effectiveness.

4.1. Result replication between experiments

Since the two experiments both included fade and ticker secondary displays, we expected there to be instances of result replication. In all cases except the detailed awareness measurements and the subjective survey questions, findings replication occurred for both the ticker and fade. One change made in the experimental design between the two experiments may account for the non-replication of detailed awareness findings: Experiment 1 participants who did not demonstrate basic awareness of secondary information were not presented with detailed awareness questions; in Experiment 2, all participants were asked detailed awareness questions. This change was made to see if detailed questions (and answer choices) could stimulate memory and affect performance differences—this appears to have significantly helped recall of information in fade displays, but not affected ticker awareness at all. Findings from the survey questions were also different between the two experiments. Experiment 1 participants subjectively rated the ticker higher than the blast and fade animations on all three opinion questions. However, considering that participants in the second experiment never saw blast animation as part of the experiment, their frame of reference could have been different enough to explain the lack of significant differences in Experiment 2 answers.

4.2. General application of results

One frequently mentioned limitation to empirical evaluations is that they cannot often be generalized enough to accurately represent real-world situations. For example, in the experiments, participants used the animated display for less than an hour. However, there are many real-world situations where monitoring tasks would reasonably extend across several hours—perhaps to monitor the traffic or weather throughout the workday, or to track breaking news. Another potential limitation is that the homogeneous population selected to perform the studies consisted of undergraduate students. This is perhaps the ideal group to examine if one hopes to obtain positive results, since they are part of the computer and video-game generation that is generally comfortable with animated displays. Actual users of animation in secondary displays may not resemble this group in terms of perceptual tendencies, so results may not apply or differences could be more apparent. In these experiments, we used a browsing task as the primary task, since the text reading and search properties are common to many computing activities. However, we would not expect these results to generalize to dissimilar primary tasks, especially those that include mixed modality of information presentation or those that include intensive cognitive processing. Since we observe that animation effectiveness could vary greatly with slight differences in implementation (such as size or speed), application specific implementation variations (which could also include relative position to the primary task, use of color, information type, etc.) may also affect secondary display animation performance.

4.3. Recommendations for secondary display animation

The results from our studies have added to our understanding of animation tradeoffs. The following recommendations can be derived from the results of the experiments presented here:

- **Animation can be used in secondary displays** with minimal negative impact on certain primary tasks. While other work (Maglio and Campbell, 2000) seems to suggest otherwise, both experiments supported this claim. The difference may result from a primary task that is less cognitively demanding and smoother, slower animations.
- **In-place displays such as fade and blast are better than motion-based displays like ticker for rapid identification of items.** Participants had lower monitoring latency when using the fade and blast than when using the ticker. This seems to extend prior results indicating that moving text is more difficult to read than static text (Sekey and Tietz, 1982; Granaas et al., 1984).
- **Motion-based displays such as a ticker are better than in-place animations for comprehension and memorability, except when a cue is provided.** While in-place displays aid rapid identification, on the awareness questions participants who used the ticker obtained better detailed awareness than those who used the blast and the fade. This suggests that if it is essential to remember specific details of monitored information, and application of detailed awareness relies on first demonstrating basic awareness (rather than being cued), a motion-based display should be used.
- **Smaller displays result in faster identification of changing information.** This may be related to the amount of information that a viewer can read in a glance. Larger displays may make it difficult to obtain desired information.
- **Slow fade animation provides best all-around support for a notification task in a secondary display used with a browsing task.** Results from the second experiment show no deficiencies for slow fade animation under any of the primary or secondary task metrics. Although other animation implementations may allow better performance under different and specific design objectives, slow fade minimizes performance tradeoffs best.

These contributions enhance understanding of animation tradeoffs according to notification system design objectives. This research also furthers the body of knowledge relating to the attention-utility theme and the intrinsic disparity between user wants and needs with respect to information monitoring tasks.

5. Conclusions

The results from our experiments as well as the growing demand for constantly changing information notification suggest that people are willing and able to divide their attention to include use of secondary displays to add utility to their overall tasks. While the specific findings included in this work contribute to current understanding of design costs and benefits resulting from secondary display animation types, we feel our work exemplifies a framework that links empirical studies to proposed critical parameters for notification systems research, which we hope will apply to future information design evaluation efforts. As Chin implored his research community to improve design and reporting of empirical efforts, he mentioned that within his field during the previous nine years only about one third of the articles included any type of evaluation (Chin, 2001). The notification systems community would benefit from a more rapid research program observation and adjustment. Reports of studies that allow unified and complementary advancements of our science are critical for early progress.

The contributions made by this paper for the notification systems research community are as follows:

- Presented a theme for the research area which describes differences from other HCI research and provides a general definition for design success.
- Proposed three critical parameters of interruption, reaction, and comprehension, which should guide usability evaluations for information design in notification systems.
- Provided an empirical study on text-based animation demonstrating an extensible, replicable study that increased our understanding of tradeoffs in the proposed critical parameters for secondary display information design.

We have established a solid empirical and analytical methodology that can be repeated with implementation variations made to any aspects of the primary and secondary tasks. For example, although a browsing task serves well in the study we conducted, we have used simple games and other attention-demanding primary tasks in similar experiments.

Despite a notification system's implementation options within the design space we propose, we believe it can be evaluated with an empirical study designed, conducted, and analysed similar to those presented in this work.

This effort provides a starting point for additional discussion and research efforts. Although we suggest a theme for the notification systems field and a statement of fundamental design objectives, we hope that others will build on these ideas to provide an expanded vision, which will lead to the recognition of critical parameters among the research community. There is much work to do in understanding the three parameters we propose—interruption, reaction, and comprehension. While each objective has been considered individually by many previous studies, tradeoffs occur between the characteristics, and we believe it is rare that any given notification system will have specific, albeit fluctuating, requirements for any less than two of these three design objectives. Understanding how they systematically impact each other throughout the design space requires comprehensive study, but would contribute enormously to satisfying the attention-utility theme. Specifically, these are some important questions that should be addressed toward this goal of determining linkages and causal relationships between interruption, reaction, and comprehension:

- Are rapid reactions to cues always accompanied by significant interruption to primary tasks?
- Does information design optimized for attention attraction always prompt reaction?
- Is comprehension level directly related to total time spent attending to secondary information, regardless of attention period duration, or are short, quick glances different than less frequent, more concentrated observations?

Implicit in this task is the investigation of all parts of the design space, which should be formally articulated by a taxonomy describing important variation in system characteristics, the dual-task situation, and the relationship between the primary and secondary tasks. For example, notification context may vary by level of importance in relation to the primary task—implying instances where interruption is desired and welcomed, as well as scenarios where primary task distraction is intolerable. Both cases require suitable notification system information design, but studies must be conducted to understand effective representations. Crowder states that a standard set of test problems should be used for experimentation within his field (Crowder et al., 1979), certainly benefiting construct validity and expediting research progress. Perhaps such a notion can apply to empirical study of notification systems as well.

The vast domain of information design possibilities provides many directions for future work as well. In actual systems, notification information varies in complexity, length, and data type (nominal, ordinal, quantitative) and information extraction tasks can involve simple detection, estimation, association, or even complete memorization. Information encoding can be accomplished with different attributes (i.e., color, shape, or position), implementations of attributes (i.e., various color hues or levels of luminescence), and with alternate modalities. Although work has been done to establish guidelines for encoding information in a user's focus for data types and extraction tasks, similar questions must be investigated specifically for notification systems residing beyond the focus. Related work has shown that focal guidelines do not necessarily extend to secondary displays presenting passive notification information (Chewar et al., 2002), providing an imperative for further investigation.

Lab-based empirical work is only one method that can extend our knowledge of how to present these various kinds of continuously changing information across a spectrum of possible user needs. Other usability evaluation methods, to include analytic approaches such as walkthroughs or heuristics, as well as ethnographic considerations in field studies or surveys, each provide additional insight. However, unless studies target commonly accepted problems and parameters, the notification systems research field will be slow to achieve relevance. While we propose some initial ideas, the community will benefit most from increased discussion of alternate possibilities and refinements. Additional studies like those we describe in this paper, tied to critical parameters, will certainly add to the growing understanding of notification system design tradeoffs, optimizing usability of future systems.

Acknowledgements

This work was supported in part by both the National Science Foundation under award IIS-0118685 and by Air Force Office of Scientific Research Grant F49620-98-1-0362.

References

Bailey, B.P., Konstan, J.A., Carlis, J.V., 2001. The effects of interruptions on task performance, annoyance, and anxiety in the user interface. In: Proceedings of the IFIP TC.13 International Conference on Human–Computer Interaction (INTERACT 2001). Tokyo, Japan, pp. 593–601.

Ballas, J.A., Heitmeyer, C.L., Perez, M.A., 1992. Evaluating two aspects of direct manipulation in advanced cockpits. In: Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI '92). Monterey, CA, pp. 127–134.

Bartram, L., 1998. Enhancing visualizations with motion. In: Proceedings of the IEEE Symposium on Information Visualization (InfoVis '98). Raleigh, NC, pp. 13–16.

Bartram, L.R., 2001. Enhancing information visualization with motion. Ph.D. thesis, Simon Fraser University, Canada.

Bartram, L., Ware, C., Calvert, T., 2001. Moving icons: detection and distraction. In: Proceedings of the IFIP TC.13 International Conference on Human–Computer Interaction (INTERACT 2001). Tokyo, Japan.

Cadiz, J.J., Fussell, S.R., Kraut, R.E., Lerch, F.J., Scherlis, W.L., 2000. The awareness monitor: A coordination tool for asynchronous, distributed work teams. Available from www.cs.cmu.edu/kraut

Cadiz, J., Venolia, G.D., Jancke, G., Gupta, A., 2001. Sideshow: Providing peripheral awareness of important information. Tech. Rep. MSR-TR-2001-83, Microsoft Research, Collaboration, and Multimedia Group, September.

Chewar, C.M., McCrickard, D.S., Ndiwalana, A., North, C., Pryor, J., Tessendorf, D. 2002. Secondary task display attributes: Optimizing visualizations for cognitive task suitability and interference avoidance. In: Proceedings of the Joint Eurographics/IEEE Symposium on Visualization (VisSym '02). Barcelona, Spain.

Chin, D.N., 2001. Empirical evaluation of user models and user-adapted systems. User Modeling and User-Adapted Interaction 11, 181–194.

Crowder, H., Dembo, R.S., Mulvey, J.M., 1979. On reporting computational experiments with mathematical software. ACM Transactions on Mathematical Software 5 (2), 193–203.

Cutrell, E., Czerwinski, M., Horvitz, E., 2001. Notification, disruption, and memory: effects of messaging interruptions on memory and performance. In: Proceedings of the IFIP TC.13 International Conference on Human-Computer Interaction (INTERACT 2001). Tokyo, Japan, pp. 263–269.

Duchnicky, R.L., Kolers, P.A., 1983. Readability of text scrolled on visual display terminals as a function of window size. Human Factors 25 (6), 683–692.

Enns, J.T., Rensink, R.A., 1991. Preattentive recovery of three-dimensional orientation from line drawings. Psychological Review 98, 335–351.

Foster, K.I., 1970. Visual perception of rapidly presented word sequences of varying complexity. Perception and Psychophysics 8, 215–221.

Granaas, M.M., McKay, T.D., Laham, R.D., Hurt, L.D., Juola, J.F., 1984. Reading moving text on a CRT screen. Human Factors 26 (1), 97–104.

Gray, W.D., Saltzman, M.C., 1998. Damaged merchandise? A review of experiments that compare usability evaluation methods. Human–Computer Interaction 13 (3), 203–261.

Green, P., 1999. The 15-second rule for driver information systems. In: Proceedings of the ITS America Ninth Annual Meeting. Washington, DC (CD–ROM).

Greenberg, S., 1996. Peepholes: Low cost awareness of one's community. In: Conference Companion for the ACM Conference on Human Factors in Computing Systems (CHI '96). Vancouver, BC, pp. 206–207.

Greenberg, S., Rounding, M., 2001. The notification collage: posting information to public and personal displays. In: Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI '01). Seattle, WA, pp. 515–521.

Healey, C.G., Enns, J.T., 1999. Large datasets at a glance: combining textures and colors in scientific visualization. IEEE Transactions on Visualization and Computer Graphics 5 (2), 145–167.

Healey, C.G., Booth, K.S., Enns, J.T., 1996. High-speed visual estimation using preattentive processing. ACM Transactions on Human Computer Interaction 3 (2), 107–135.

Horvitz, E., 1999. Principles of mixed-initiative user interfaces. In: Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI '99). Pittsburgh, PA, pp. 159–166.

Horvitz, E., Jacobs, A., Hovel, D., 1999. Attention-sensitive alerting. In: Conference on Uncertainty and Artificial Intelligence (UAI '99). Stockholm, Sweden, pp. 305–313.

Juola, J.F., Ward, N.J., McNamara, T., 1982. Visual search and reading of rapid serial presentations of letter strings, words, and text. Journal of Experimental Psychology: General 111 (2), 208–227.

Kang, T.J., Muter, P., 1989. Reading dynamically displayed text. Behaviour and Information Technology 8 (1), 33–42.

Lee, J.D., Carney, C., Casey, S.M., Campbell, J.L., 1999. In-vehicle display icons and other information elements: Preliminary assessment of visual symbols. Tech. Rep. FHWA-RD-99-196, United States Federal Highway Administration, December.

Maglio, P.P., Campbell, C.S., 2000. Tradeoffs in displaying peripheral information. In: Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI 2000). The Hague, the Netherlands.

McCrickard, D.S., 1999. Maintaining information awareness with Irwin. In: Proceedings of the World Conference on Educational Multimedia/Hypermedia and Educational Telecommunications (ED-MEDIA '99). Seattle, WA.

McCrickard, D.S., 2000a. Agentk: A toolkit for enhancing agent interfaces; login: The Magazine of USENIX and SAGE 25, 4 (July).

McCrickard, D.S., 2000b. Maintaining information awareness in a dynamic environment: Assessing animation as a communication mechanism. Ph.D. thesis, Georgia Institute of Technology, Atlanta GA.

McCrickard, D.S., Catrambone, R., Stasko, J.T., 2001. Evaluating animation in the periphery as a mechanism for maintaining awareness. In: Proceedings of the IFIP TC.13 International Conference on Human–Computer Interaction (INTERACT 2001). Tokyo, Japan, pp. 148–156.

McFarlane, D.C., 1998. Interruption of people in human–computer interaction. Ph.D. thesis, George Washington University, Washington DC.

McFarlane, D.C., 2002. Comparison of four primary methods for coordinating the interruption of people in human–computer interaction. Human–Computer Interaction 17, 3.

Sekey, A., Tietz, J., 1982. Text display by saccadic scrolling. Visible Language 16, 62–76.

Sheridan, T., 1991. Human factors of driver-vehicle interaction in the IVHS environment. Tech. Rep. DOT HS 807 837, National Technical Information Service, Spring_eld, MA.

Tufano, D., Knee, H., Spelt, P., 1996. In-vehicle signing functions and systems concepts. In: Proceedings of the 29th International Symposium on Automotive Technology and Automation (ISATA) Dedicated Conference on Global Deployment of Advanced Transportation Telematics/ITS. Florence, Italy, pp. 97–104.

Ware, C., Bonner, J., Cater, R., Knight, W., 1992. Simple animation as a human interrupt. International Journal of Human–Computer Interaction 4 (4), 341–348.

Weiser, M., Brown, J.S., 1996. Designing calm technology. PowerGrid Journal 1 (1).

Zhao, Q.A., Stasko, J.T., 2000. What's Happening? The community awareness application. In: Conference Companion of the ACM Conference on Human Factors in Computing Systems (CHI '00). The Hague, the Netherlands, pp. 253–254.