An Investigation of the Effects of Matching Attentional Draw with Utility in Computer-Based Interruption

by

Jennifer Shari Gluck

B.Sc., Queen's University, 2004

A THESIS SUBMITTED IN PARTIAL FULFILMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

Master of Science

in

The Faculty of Graduate Studies

(Computer Science)

The University Of British Columbia

August 2006

Abstract

We all experience interruption in our daily lives when something causes a break in our actions, activities, or concentration. The number of channels through which we may interrupt each other has multiplied with the advent of communication and information technology, beginning with the telephone and increasing with email and instant messaging systems. Moreover, technology itself has become a source of interruption through calendar systems, software update reminders, and even battery monitor warnings. Interruption has become pervasive to the point where it is overwhelming. Consequently, research in the Human-Computer Interaction literature has focused largely on the negative effects of interruption. Yet, the fact that we continue to propagate and tolerate computer-based interruption suggests that there is some value associated with it. In this thesis, we explore how interruption can be harnessed for beneficial means by empirically investigating a design guideline that may help to mitigate negative effects: matching the amount of attention attracted by an interruption's notification signal to the usefulness of the interruption content.

In three controlled studies, we investigated the effects of matching attentional draw of notification to interruption utility in terms of annoyance, benefit, workload, and performance. Study 1 examined notification signals in terms of their detection times and established a set of three significantly different notification signals along the spectrum of attentional draw. Study 2 was an initial investigation of matching these different signals to interruptions with different levels of utility. In our final study we compared our strategy of matching attentional draw and utility to the status quo of static notification methods. Our results indicate that interfaces that matched attentional draw to utility were associated with decreased annoyance and an increased perception of benefit compared to interfaces that used a static level of attentional draw. These and other secondary results are discussed, along with design implications and directions for future work. The research presented is an initial step towards understanding and exploiting the benefits of matching attentional draw of notification to the utility of interruption content.

Contents

List of Tables

List of Figures

Acknowledgements

First and foremost, I would like to thank my wonderful supervisor Dr. Joanna McGrenere for her inspired guidance and insight. I am also grateful for the invaluable contribution made by PhD candidate Andrea Bunt, who acted unofficially as co-supervisor. Working with Joanna and Andrea has been both an honour and a pleasure. I have learned an incredible amount from these two brilliant women and I have the utmost respect for the affable manner in which they dedicate themselves to their research. I would also like to thank my second reader Dr. Gail Murphy for her remarkably expeditious feedback. Thanks are also due to Dr. Lyn Bartram for contributing valuable advice early in my research.

This experience has taught me that, analogous to the ancient African proverb that it takes a village to raise a child, a supportive community of graduate students is essential in the creation of a thesis. I am grateful to my peers for alternately providing indispensable advice and essential distraction. In particular, I would like to thank Karen Parker and David Sprague for their continual support in both work and play.

Finally, I dedicate this thesis to my parents, Lisa and Tom Gluck. They are my support system and personal cheering squad whose love and tireless encouragement inspire all my achievements.

JENNIFER GLUCK
The University of British Columbia
August 2006

Chapter 1

Introduction

We all experience interruption in our daily lives when something causes a break in our actions, activities, or concentration. In the past, external sources of interruption were largely limited to direct human interaction, for instance, a coworker stopping by your office to chat. Sources of interruption have multiplied with the advent of communication technology, beginning with the telephone and increasing with mobile phones and pagers. Further advances in information technology have brought email and instant messaging (IM) to the masses. As the number of channels through which we may contact one another has increased, so has the ease with which we can interrupt. Once upon a time there was a certain amount of effort required to interrupt someone: we had to walk down the hall, or at least pick up the phone. Now we can interrupt each other with the touch of a button, and we often do. Moreover, technology itself has become a source of interruption, for instance, through calendar systems, software update reminders, and even battery monitor warnings. Interruption has become pervasive to the point where it is overwhelming.

This ubiquity of interruption in modern life has motivated a considerable amount of research in the field of Human-Computer Interaction (HCI), much of which has focused on the negative effects of interruption. Widely-employed systems such as email-alerting and IM software are often implicated as disruptive interruption offenders [5, 24, 30], yet their rampant popularity testifies to their usefulness. The fact that we continue to propagate and tolerate computer-based interruption suggests that there is some value associated with it. In this thesis, we explore how interruption can be designed to promote a positive user experience.

The term *interruption* is deceptive in its apparent simplicity. It in fact encompasses a multifaceted research domain, and so we begin by discussing the dimensions of interruption that arise in our work. The two components of computer-based interruption that we focus on in this research are notification and content. Notification concerns the way in which the interruption is presented to a user. McFarlane discusses this as the *method of expression* while Latorella uses the term *annunciation stimulus* [37]. In this thesis we refer to this component as a *notification signal.* Specifically, we consider *signals* as graphical events used to alert users to some change in an interface display [8].

The second component we address is interruption content, which may be examined in terms of both relevance and utility. These two dimensions are different and yet related. Relevance considers how pertinent the content is to the recipient of the interruption. Content may be relevant to the primary task at hand, but also may be relevant to some secondary task. Moreover, content unrelated to any particular task may still be personally relevant to the human user. Utility, on the other hand, defines how useful, important, or urgent the interruption content is to the recipient [26, 31]. Relevance is a component of utility but does not define it. For instance, interruption content that is highly utilitarian must be relevant to the user in some way; however, it is possible for relevant content to have low utility.

It is common in the research literature to classify interruption systems in terms of a method of *coordination* [36]. Coordination encompasses notification (described above) as well as the timing of onset of each notification signal. The method of coordination also delineates the amount of control users are given in dealing with interruption, for instance, whether a user must respond to a particular interruption at the time of onset or if there is the option to postpone the response to a later moment.

With a basis now in place for understanding the dimensions of interruption considered in this research (i.e., notification, content, utility, relevance, and coordination), we now examine how interruption has been treated in the HCI literature. As already mentioned, the research literature predominantly casts interruption in a very negative light. Studies commonly point to interruption as a cause of lowered performance on primary task [15, 22] and a threat to the emotional state of users. Bailey, Konstan, and Carlis [7] reported increases in user anxiety and annoyance as well as perceived difficulty of completing a primary task as a result of interruption. Latorella's Interruption

Management Stage Model [37] defined the four general effects of interruption as *diversion*, *distraction*, *disturbance*, and *disruption*. This focus on the detrimental nature of interruption leaves its potential value ignored or forgotten.

A small number of researchers [43, 45] recognize and emphasize positive aspects of interruption. While these researchers urge the community to see the potential benefit of reminders, notifications, suggestions, warnings, and alerts, there is a lack of understanding of how to exploit these interruptions for positive purposes while minimizing their disruptive properties. Meanwhile, negative aspects of interruption remain the focus, as studies that are regarded as foundational to the topic of interruption in HCI examine interruptions that are relevant to neither the primary task nor the user. For instance, when studying methods for coordination of interruption, McFarlane [36] employed a video game as the primary task while the interrupting task was an unrelated shape-matching task. This type of irrelevant interruption stands no chance of being perceived as beneficial by the user. In reality, interruption content is often relevant to either a user's primary task or to his or her life or job more generally. Thus, interruption researchers should consider potentially positive aspects of interruption by studying relevant interruptions.

We have already noted that the massive popularity of communication technologies such as email and IM implies that interruption does provide some benefit. Innovative interruption-based systems also have a real potential to improve how users experience computer software. Recommender [9, 32] and mixed-initiative [18, 27, 44] systems offer value by assisting users in a context-sensitive, interactive manner. Proactive information and recommender systems such as the FXPAL Bar [9] aim to enhance resource discovery while limiting information overload. Mixed-initiative approaches to interface customization such as FlexExcel [44] and the Adaptive Bar [18] offer a solution to software complexity by helping users to engage in effective customization of graphical interfaces. Both types of systems aim to aid users by making, in real time, context-sensitive suggestions that have the potential to reduce the amount of time necessary to perform a task, but their success hinges on the ability of users to perceive interruptions positively. It is essential that such systems present interruptions diplomatically so that users neither ignore suggestions nor are driven by annoyance to stop using the system, as was the case of the anthropomorphic office assistant we all love to hate: Microsoft's ill-fated Clippy [2, 49]. Billsus, Hilbert, and Maynes-Aminzade [9] observe that the key problem with recommender systems is that their notification methods are often either too subtle or too obtrusive, depending on context. Consequently, investigating techniques for emphasizing the beneficial aspects of interruption is a worthwhile research endeavour.

A design guideline put forward by Obermayer and Nugent [40] and advocated by McFarlane and Latorella [37] could help to promote positive interruption. The guideline recommends making the method of expression of an interruption - specifically, the "level of attention-getting" of a notification signal - relative to the utility of the interruption content. Obermayer and Nugent refer to this strategy as "multi-level attention-getting" [40]. According to this strategy, interruptions that are highly important are presented with high attentional draw and are thus noticed immediately. Unimportant interruptions are presented more subtly using notification signals with very low attentional draw, to be noticed by users only during a natural break from the task at hand. Attentional draw for interruptions with utilities between these endpoints is scaled accordingly. In this manner, users are only truly interrupted from a task when it is important to do so. If this technique can effectively reduce the negative effects commonly associated with interruption, then positive aspects of interruption may become more conspicuous.

In work considered to be the foundational source of information in the interruption literature, McFarlane and Latorella [37] argue that Obermayer and Nugent's design guidance is simplistic: alone, it cannot solve the disruptive aspects of interruption. While McFarlane and Latorella do advocate matching attentional draw and utility, it is but one item in their long list of design recommendations. More disconcerting is the fact that no commercially available interruption system has adopted the strategy. While we suspect that the value of the guideline has been underestimated in both the literature and the industry, empirical investigation of the multi-level attention-getting strategy is absent in the literature. We felt it worthwhile to examine the effects of matching utility and attentional draw to determine if this strategy alone can in fact help to ease the disruptive elements of interruption. If positive effects are significant, interaction designers may be more strongly encouraged to incorporate the strategy into interruption interfaces. Thus, this thesis comprises an empirical investigation of Obermayer and Nugent's guideline by examining the effects of matching attentional draw and utility.

1.1 Definition of Attentional Draw

For the purposes of our research, we define attentional draw (AD) as the time elapsed between when an interruption is presented and when the user first notices its presence. The low end of the AD spectrum corresponds to a large amount of time to notice the interruption; the high end corresponds to a very short amount of time to notice the interruption.

1.2 Scope of Utility

In order to study the impact of matching the attentional draw of an interruption signal with the utility of its content, it is first necessary to delineate the scope of the interruption utility. In this thesis, we define utility as relevance to the primary task. Our interrupting task comprises context-sensitive hints designed to help subjects perform a primary task, but subjects decide if and when to utilize each hint. In this manner, we effectively emulate a mixed-initiative system.

1.3 Contributions

This thesis documents work done to examine the effects of matching attentional draw of notification to interruption utility in terms of annoyance, perceived benefit, workload, and performance. We sought to prove that a correct match between AD and utility decreases annoyance and increases perceived benefit associated with the interrupting application without negatively impacting performance or workload. Our findings indicate that the matching strategy does result in decreased annoyance and increased benefit compared to a strategy that employs static attentional draw, with neither a positive nor a negative effect on workload or performance. This research also establishes a set of three significantly different notification signals along the spectrum of attentional draw.

Our research interests did not include how to appraise the utility of interruption content computationally. Instead, we selected a primary task for which we could generate hints with three objective levels of utility (very helpful, somewhat helpful, and not helpful).

This research arose from the need to reduce the negative effects associated with interruption as a means to facilitating positive perception of interruption. Because we emulated a mixed-initiative context, we expect our findings to apply most readily to mixed-initiative and recommender systems, but the results will likely also apply to other interruption systems.

1.4 Overview

This thesis comprises three studies that were designed to evaluate the effect of matching attentional draw of notification to interruption utility. Previous work relevant to this research is summarized in Chapter 2. In order to begin our investigation, we required a set of notification signals with different levels of AD. Chapter 3 discusses Study 1, which was designed to investigate notification signals in terms of their attentional draw and in which we established a set of three notification signals whose mean detection times were significantly different from one another. Several methodological decisions were made in the design of the two subsequent studies that investigated the effect of matching these different signals to interruptions with different levels of utility. Chapter 4 discusses several of these decisions, such as the choice of interrupting task and experimental conditions. Chapter 5 presents Study 2, which was our initial investigation of utility and attentional draw. Results of this between-subjects experiment allowed us to pare down the number of conditions in order to capitalize on the power of a within-subjects design in Study 3, which is discussed in Chapter 6. In that study we compared a strategy that matched attentional draw to utility, a strategy that used static attentional draw, and a control condition that did not interrupt. The latter two studies measured annoyance and perceived benefit associated with the interrupting system, as well as cognitive workload and primary task performance. Qualitative feedback was also collected, using questionnaires and follow-up interviews, to gain insight into additional components such as preference and perception of the notification signals and interruption utility. Finally, Chapter 7 discusses directions for future work and concludes this thesis.

Substantial portions of this thesis appear in a conference paper submission jointly authored with Andrea Bunt and Joanna McGrenere.

Chapter 2

Related Work

In this chapter we review literature relevant to our research in the area of interruption. We begin by providing a general background on methods for coordinating interruptions and presenting notification. We also present systems that have implemented the matching strategy we investigate in our research and discuss other strategies reported in the literature for improving how users experience interruption. Throughout, we discuss how previous techniques for investigating and improving interruption compare to and support our work.

In the following sections, we simplify our examination of interruption to the context of computer systems. For a more in-depth examination of interruption, refer to McFarlane's Taxonomy of Human Interruption [37].

2.1 Coordination of Interruption

Four methods for coordinating user interruption in HCI have been documented in the literature [36, 37]: *immediate*, *negotiated*, *mediated*, and *scheduled*. In *immediate* coordination, an interruption is presented to the user as soon as it occurs and in a manner that requires the user to address the interruption immediately. *Negotiated* coordination refers to a system that presents an interruption as soon as it occurs, but that supports negotiation with the user in order to give the user control over when or whether to deal with the interruption. *Mediated* coordination employs an agent that interrupts indirectly by requesting interaction with the user through some sort of personal broker, which in turn determines when and how to present the interruption to the user. *Scheduled* coordination restricts interruption presentation to a prearranged schedule, such as once every 15 minutes.

We employ a hybrid of negotiated and mediated coordination in the studies presented in this thesis. Our strategy is negotiated in that each interruption is presented without delay and the user decides when to address it. Our approach is mediated in the sense that the system decides how to present the interruption.

2.2 Notification

Little research has examined the attentional draw associated with how an interruption is presented. Ware, Bonner, Knight, and Cater's [47] preliminary research into moving icons as interruptions motivated the foundational Moticon work by Bartram, Ware, and Calvert [8]. The Moticon research studied the perceptual properties of visual motion applied to notification in terms of detection and distraction. In a series of experiments, the authors found that icons with simple motion (termed *moticons*) outperformed colour and shape changes in attracting attention, but that different kinds of motions were associated with differing levels of distraction. Slow linear motion was found to most effectively balance detection speed with low distraction and irritation. Our research builds upon this base in terms of investigating perceptual properties such as motion, colour, and blink according to attentional draw, using new signals and tasks. We also base much of the experimental design methodology of our first study on Bartram et al.'s research.

In the context of multimodal interruption, Arroyo, Selker, and Stouffs [3] examined the disruptiveness of five modes of notification: heat, smell, sound, vibration, and light. The interruptions - which had no content - were presented while subjects performed a computer-based reading comprehension task, and subjects were asked to acknowledge each interruption by clicking on an icon. At the end of the study, subjects rank ordered each interruption modality according to perceived disruptiveness. Smell and vibration were rated most disruptive and second-most disruptive, respectively. The authors speculatively attributed the disruptiveness associated with these modes to their novelty. Unfortunately, the authors did not report on detection times.

More recently, Robertson, Lawrance, and Burnett [42] compared "high-intensity" and "low-intensity" *negotiated* interruptions in the context of end-user debugging. Interruptions notified subjects of information relevant to the task of debugging a spreadsheet. The authors discussed the difference in intensity in terms of amount of mental stimulation or *arousal*. This is similar in spirit to our definition of attentional draw, though the study made no attempt to quantify or measure the notion of intensity. High-intensity interruptions differed from low-intensity interruptions in terms of size, colour and blink. Findings indicated that high-intensity notification impaired subjects' ability to learn debugging features and was associated with lowered effectiveness in both debugging and judgment of ability to debug. Some subjects were also observed to engage in counter-productive activity in order to terminate a high-intensity notification signal. The authors concluded that interruptions that are very intense (i.e., have high attentional draw) should be avoided. However, intensity was not linked to utility of interruption content. Furthermore, the study did not measure annoyance, perceived benefit, or workload.

2.3 Timing of Onset

Substantial effort has been made to investigate the effect of timing of interruption onset. Reseachers in this area theorize that presenting an interruption at an ideal moment - and postponing the interruption if the moment is inopportune - can help to mitigate negative effects. The research literature points to a number of different approaches to determining the ideal moment of interruption. Chen and Vertegaal [12] employed physiological sensors to detect attentional state, on which they based the decision of when to interrupt.

Fogarty, Hudson, and Lai [19] and Ho and Intille [26] have focused on using simple sensors to model interruptibility from an environmental context. Fogarty et al. utilized sensors such as microphones, magnetic switches (to detect door position and phone use), and motion sensors, as well as keyboard and mouse logging, to model interruptibility. Ho and Intille used accelerometers to detect activity transitions, finding that interruptions delivered during an activity transition were received more positively than those delivered at random moments.

Adamczyk and Bailey [1] and Iqbal and Bailey [29] studied the use of mental workload to determine opportune interruption moments during task execution. The earlier work [1] found that interruptions occuring during different moments in a task sequence impacted user emotional state (i.e., annoyance, frustration and respect) differently. Later work [29] investigated how characteristics of task structure can be used to predict the cost of interruption at subtask boundaries.

Our research is similar to Adamczyk and Bailey's work [1] in terms of measuring the effect of interruption on workload and annoyance. However, Adamczyk and Bailey examined the effect of manipulating the timing of an interruption (i.e., at different moments within task sequence), whereas we manipulate the notification style of interruption. The authors used an *immediate* coordination method with highly intrusive notification: a full-screen modal popup box. We employ a hybrid of *negotiated* and *mediated* coordination and vary attentional draw from low to high, but even our highest level of AD is less intrusive than the Adamczyk and Bailey's notification style. Furthermore, the interrupting tasks in Adamczyk and Bailey's study were unrelated to the primary tasks, while our interruptions are directly relevant.

We argue that controlling the timing of interruption is not the only viable strategy for alleviating distraction and annoyance. It is also important to investigate strategies that do not delay delivery of important messages. Postponing an interruption until a more opportune moment (e.g., within the task execution model) may sometimes be a mistake, since the interruption might very well change how the user completes the task. While researching the timing of interruption is important, it is also necessary to investigate how to present interruptions to users.

2.4 Utility and Relevance

In the previous chapter we discussed relevance and utility as separate but related dimensions of interruption content: relevance is a component of utility but does not necessarily define it. In this section we discuss previous work in both of these areas.

2.4.1 Utility of Interruption

The ability to assess the utility of an interruption is a fundamental component of multi-level attention-getting. Appraisal of utility of interruption content has been investigated by a number of researchers. Horvitz, Koch, and Apacible [28] harnessed machine learning to generate models for inferring the "cost" of interrupting a user. Cost was assessed as a user's willingness to pay in dollars to avoid a particular interruption at a particular moment. While users performed their everyday computer-based tasks, the system interrupted intermittently to ask users whether they were busy or not at that moment. Models of interruptibility were then built based on user responses as well as factors such as users' computer activity and meeting status, location, time of day, and a conversation detection agent. The generated models were intended as input to a mediating agent that decides if, when, and how to relay interruptions. Speculatively, low cost may indicate that an interruption has high utility. On the other hand, low cost may merely indicate that the user does not mind being interrupted at a particular moment, regardless of utility. However, it is feasible to imagine the construction of models of utility using a similar experiment in which the training system asks users to rate the utility of typical interruption content.

Gievska and Sibert [21] similarly employed machine learning to develop an interruption model that incorporates utility as *relevance* and *urgency*. The authors intend for their model to recommend appropriate timing of an interruption, but we note that a variation of the system could be used to isolate utility for the purposes of selecting a level of notification AD.

In the context of IM, Avrahami and Hudson [5] developed and tested statistical models to predict user responsiveness to an incoming message - that is, they predicted whether a user is likely to respond to a message within a certain amount of time. The models were based on IM events such as starting and stopping the IM client, sending messages, opening and closing message windows, and changes in online status, as well as on desktop events such as key presses, mouse events, and window activity. Results indicated that the models could predict whether a user would respond to a message within 30 seconds, 1, 2, 5, and 10 minutes, with as much as 90% accuracy. The authors suggest that their models may be used to increase the salience of important incoming messages when responsiveness is predicted to be low.

In a departure from computational assessment of utility, White and Zhang [48] investigated a system where users, as opposed to machines, judged the utility of messages in a sender-initiated email notification system. In the study, both senders and recipients reviewed their shared communications history and, for each email, indicated how soon the recipient needed to read the message following delivery (i.e., immediately, within an hour, or by the end of the day). Findings indicated that the strategy might not be viable because senders underestimated the immediacy of the message in 27% of cases. However, the authors admit that their sample size of seven sender-recipient pairs was small and that further investigation is warranted.

In our research, we do not assess interruption utility computationally. Instead, we create interruption content to conform to predefined levels of utility. However, this related work shows promising signs that computational appraisal of utility may be possible in the future. When utility can be assesed computationally, the multi-level attention-getting strategy will be more attainable.

2.4.2 Relevance of Interruption

Very little research has investigated the impact of interruptions with varying degrees of relevance of to the user. The main exception is work by Czerwinski, Cutrell, and Horvitz [17], which examined the effect of relevance on disruption in instant messaging. The experiment investigated the effects of IM interruption during different phases of task execution under two levels of relevance. A component of the primary task was to classify web sites into categories based on quality of graphic design. *Relevant* IMs told subjects the category of the current website, while *irrelevant* IMs conveyed some useless factoid about the website. Results indicated that time spent on an interruption and time taken to resume the primary task after an interruption was longer for irrelevant than relevant interruption content. The study did not consider qualitative aspects of disruption, for instance, annoyance or perceived benefit. Furthermore, only one (unidentified) notification method was employed.

There is also unpublished work on relevance that appears to be somewhat related to our research. The work [41], which is referenced in research by Robertson et al. [43], implies that interruptions that are unrelated to a primary task cause higher user annoyance than interruptions that are related. However, a detailed reading of the unpublished manuscript leaves many question unanswered.

2.5 Mixed-Initiative and Recommender Systems

Our research was partly motivated by the need for effective notification methods in recommender [9, 32] and mixed-initiative [18, 27, 44] systems. These systems are well suited to a multi-level attention-getting strategy because the utility of their interruption content is often inherent. Billsus et al.'s FXPAL Bar [9] recommendation algorithm, for instance, assesses the level of relevancy of each recommendation it makes. Flex-Excel [44] extends the user interface of a common spreadsheet application, providing adaptive suggestions for defining new menu entries or short-cut keys for frequently-used functions. Similarly, Debevc, Meyer, Donlagic, and Svecko's Adaptive Bar [18] is a modification of the customizable toolbar supplied in Microsoft Word that suggests additions or deletions of items on the toolbar based on a history and frequency of use. These and similar mixed-initiative systems could infer utility of suggestions from existing function frequency and recency measures as well as from estimates of how much time a suggested customization could save users [10].

2.6 Peripheral Awareness Systems

A substantial amount of interruption research has focused on peripheral awareness systems [13, 33, 35], addressing utility, attention, and presentation. Research in awareness systems has been fruitful, and recommendations exist for presenting peripheral data in such a way that when data in a secondary awareness task becomes relevant, it will "grab" user attention.

Peripheral awareness research most similar to our work is Matthews, Dey, Mankoff, Carter, and Rattenbury's Peripheral Displays Toolkit (PTK) [34], which provides tools for managing attention in peripheral displays. The authors define five *notification levels* that are equivalent to what we refer to as utility levels: differences in information importance. Similarly, their *transitions* are comparable to our notification signals. The PTK library included *notification map* components to match transitions and notification levels. To demonstrate the toolkit's capabilities, the authors used the PTK to create a display designed to give users a sense of activity level in a remote location. A camera sensed and analyzed activity in an office environment using image differencing. An abstraction of this activity was displayed in remote location using a commercial Ambient Orb^{TM}, which changed colour slowly (i.e., with low AD) or rapidly (i.e., with high AD) according to the level of activity sensed by the camera.

Awareness systems that cater to this type of continuous divided-attention task situation employ a stream of continuous content. The type of interruption discussed up to now, on the other hand, involves discrete moments of interruption, each containing its own specific content. The differences between these two contexts is further evidenced in diverging definitions of utility. McCrickard, Catrambone, Chewr, and Stasko [35] discuss utility as the value provided by the peripheral system as a whole, while in this thesis we consider utility as the importance of the content of a particular interrupting message. Because of these differences, discrete interruption systems cannot necessarily employ guidelines and recommendations tailored to peripheral systems. Further research is required to construct similar guidelines for discrete interruption.

2.7 Matching Attentional Draw with Utility

A handful of systems described in the literature have heeded Obermayer and Nugent's design guidance to match AD to utility. Avrahami and Hudson's QnA IM Client [4] used two levels of AD and two levels of utility. The system increased the salience of incoming messages hypothesized to deserve immediate attention based on message content. Regular non-urgent messages were presented using an unidentified notification method native to the Trillian Pro IM client. When an IM was identified as a question or an answer to a previously identified question, this potentially urgent message was presented with non-modal popup box that indicated, "[Sender] is asking a question," "[Sender] might be answering your question," or "[Sender] might be replying with a question." The system was not formally evaluated, but preliminary feedback seemed promising.

Billsus et al.'s FXPAL Bar recommender system [9] similarly employed two levels of AD and utility. Recommendations that were judged to be of "exceptionally high quality" were displayed using a notification method similar to the "toaster popup" used in Microsoft Outlook 2003. All other recommendations were presented in a more subtle manner, by changing the colour of a button on the Microsoft Internet Explorer toolbar. The authors did not discuss the level of relevance at which recommendations were determined to be of "exceptionally high quality." A previous design of the system had employed only the toolbar-based notification. The authors used a questionnaire to elicit feedback from users who worked with the intial design for an undisclosed period of time. After the updated two-level AD approach had been in use for a month, the same users filled out a second questionnaire. Responses indicated that, in the updated system, users were more likely to access system content through the toaster popup than through the toolbar notification, and that user awareness of the system was higher in the two-level AD system than in the initial design.

In the most sophisticated use of utility and AD to date, Oberg and Notkin [39] developed a Pascal code debugger that used colour to alert users to the existence and location of errors. Employing a negotiated method of interruption, the system provided a full error report for a particular bug only upon request. Saturation of the colour alert communicated the age of an error. Importance of errors was also indicated: colouring of important errors (e.g., type mismatches) got darker more quickly than colouring of less important errors (e.g., undeclared variables). Oberg and Notkin did not intend their work as a validation study, and so did not compare their design with other methods of notification. They indicate, however, that anecdotal evidence endorses the usefulness of the technique.

None of these efforts endeavored to verify Obermayer and Nugent's guideline. Although the multi-level attention-getting strategy was exercised, the goal of these works was not to study its efficacy, but to develop a novel system. Our goal, on the other hand, is explicitly to examine the effects of matching attentional draw and utility through a series of empirical evaluations.

Chapter 3

Study 1

In our first experiment we investigated notification *signals* in terms of their attentional draw (AD). *Signals* are graphical events used to alert users to some change in an interface display [8]. We define AD as the time elapsed between when the interruption is presented and when the user first notices its presence. Our goal was to identify between three and five signals whose mean detection times were significantly different from one another. In subsequent experiments, we investigated the effect of matching these different signals to interruptions with differing levels of utility. Because most existing interruption research and interrupting applications (e.g., email notification, IM software, and calendar applications) utilize the visual field, we focused on visual notifications rather than using other modalities such as haptics or audition.

We leveraged related work in the design of this experiment, basing our trial design and one of our primary tasks on Moticon research by Bartram et al. [8]. Taking into account well-understood properties in both psychology and information visualization literature such as size, colour, motion, and location, we designed 10 signals and then carried out an experiment to determine which signals generated the greatest spread of detection times.

3.1 Primary Task

In order to ensure that the notification signals defined in this experiment remained valid in our subsequent experiments, it was crucial that we employed the same primary task across all studies. We selected this primary task according to two key requirements: (1) the need to be able to generate interruptions with an objective measure of utility; and (2) the need to involve concentration such that moving from the primary task to the interrupting task required a cognitive context switch. Thus, we sought a primary task that involved considerable cognitive workload and provided some amount of engagement and motivation to perform well.

In order to increase the generalizability of our work, we also investigated detection times using a primary task with lower workload. The low-workload task allowed us to gauge reaction times when subjects did not need to be pulled out of heavy concentration.

Finally, we required both primary tasks to be mouse-driven because one of the signals we examined involved augmenting the cursor.

3.1.1 High-Workload Task: Memory Game

A computer-based version of the game Memory satisfied the task requirements outlined above. This traditional game involves a set of picture cards made up of pairs of matching cards. Initially all cards are face down. Players try to match all of the cards as quickly as possible, turning over only two cards at a time. When an attempt is unsuccessful, cards are returned to the face-down position. When a match is found, the cards remain face up. In our implementation, when a subject found all of the matches on the board before the end of a block, the board was reset with a different deck of cards and the matching task continued. Four different decks of cards were used (each with a different set of pictures), after which subjects returned to the first deck. The deck size was 64 cards (32 pairs). The large number of cards ensured that the task required considerable concentration and thus provided a high workload. The game-like nature of the task also ensured that subjects were engaged in the task. Figure 3.1 shows a screen capture of the Memory Game task.

3.1.2 Low-Workload Task: Simple Editor

Our low-workload task was based on the simple editing task in Bartram et al.'s Moticon research [8]. A large non-scrollable editing window contained a 20×20 table of numbers from zero to nine. Subjects had to find all of the zeros in the table (80 in total) and replace them with ones by left-clicking with the mouse on the table entry. When

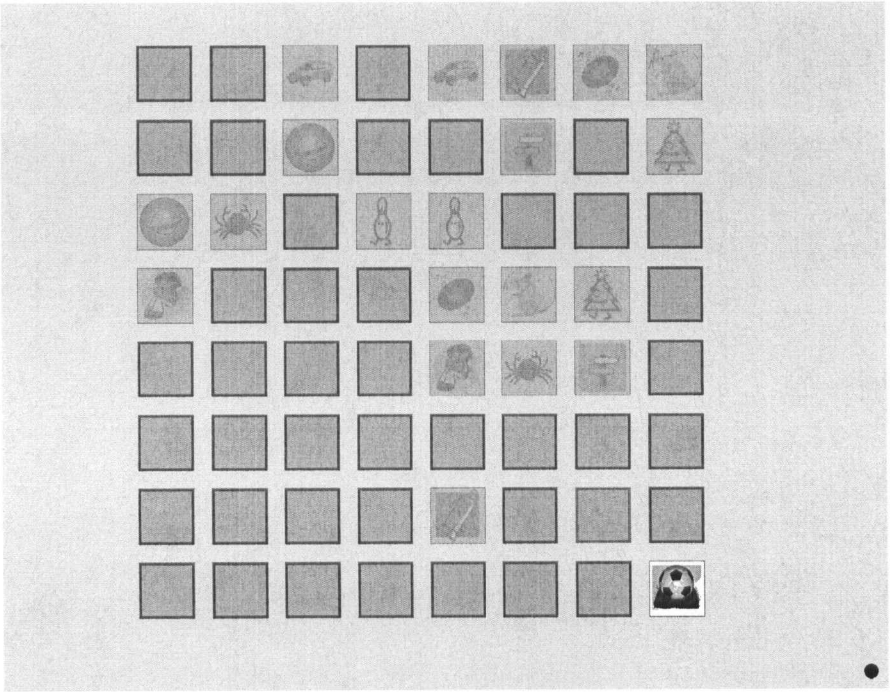
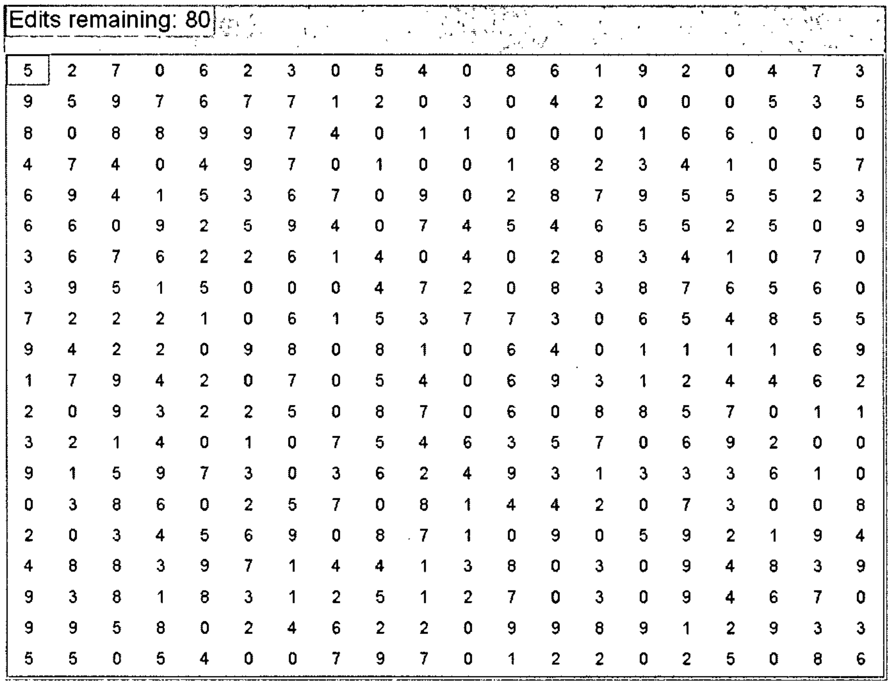


Figure 3.1: Screenshot of the Memory Game task. The soccer ball in the bottom right corner is the currently selected card. The greyed-out cards have already been matched. All other cards have yet to be matched.



Edits remaining: 80

5	2	7	0	6	2	3	0	5	4	0	8	6	1	9	2	0	4	7	3
9	5	9	7	6	7	7	1	2	0	3	0	4	2	0	0	0	5	3	5
8	0	8	8	9	9	7	4	0	1	1	0	0	0	1	6	6	0	0	0
4	7	4	0	4	9	7	0	1	0	0	1	8	2	3	4	1	0	5	7
6	9	4	1	5	3	6	7	0	9	0	2	8	7	9	5	5	5	2	3
6	6	0	9	2	5	9	4	0	7	4	5	4	6	5	5	2	5	0	9
3	6	7	6	2	2	6	1	4	0	4	0	2	8	3	4	1	0	7	0
3	9	5	1	5	0	0	0	4	7	2	0	8	3	8	7	6	5	6	0
7	2	2	2	1	0	6	1	5	3	7	7	3	0	6	5	4	8	5	5
9	4	2	2	0	9	8	0	8	1	0	6	4	0	1	1	1	1	6	9
1	7	9	4	2	0	7	0	5	4	0	6	9	3	1	2	4	4	6	2
2	0	9	3	2	2	5	0	8	7	0	6	0	8	8	5	7	0	1	1
3	2	1	4	0	1	0	7	5	4	6	3	5	7	0	6	9	2	0	0
9	1	5	9	7	3	0	3	6	2	4	9	3	1	3	3	3	6	1	0
0	3	8	6	0	2	5	7	0	8	1	4	4	2	0	7	3	0	0	8
2	0	3	4	5	6	9	0	8	7	1	0	9	0	5	9	2	1	9	4
4	8	8	3	9	7	1	4	4	1	3	8	0	3	0	9	4	8	3	9
9	3	8	1	8	3	1	2	5	1	2	7	0	3	0	9	4	6	7	0
9	9	5	8	0	2	4	6	2	2	0	9	9	8	9	1	2	9	3	3
5	5	0	5	4	0	0	7	9	7	0	1	2	2	0	2	5	0	8	6

Figure 3.2: Screenshot of the Simple Editor task. Subjects had to locate all zeros and replace them with ones. The base notification icon can be seen in the bottom right-hand corner.

a subject completed all necessary edits before the end of a block, the board was populated with new values and the editing task continued. A running counter in the upper left hand corner indicated the number of zeros remaining. Figure 3.2 shows a screen capture of our Simple Editor task.

In Bartram et al.'s version of this task, the editing window was small, the table was scrollable, and subjects could use arrow keys on the keyboard in addition to the mouse to navigate through the table. Our version of the task differed from Moticon task because Bartram et al.'s research focused on detection in the periphery: most of the screen real estate was utilized for icon detection, necessitating a small window for the primary task. The detection area with which we were concerned was meant to mimic the system tray area of the Windows OS, which uses only a very small amount of screen real estate. Similarly, the primary task was meant to simulate typical software tasks that may be performed on the desktop. Thus, we used a larger window which was not scrollable and took up most of the screen. Subjects did not have the option of using the keyboard for navigating or editing because we wanted subjects to engage with the system using only the mouse.

3.1.3 Training

After each task was introduced, subjects completed a training block to ensure that they understood the task. The training block for the Memory Game task used a smaller board that contained 16 cards, and the training block for the Simple Editor task used a smaller table (10×10 entries) that contained 10 zeros. The training block ended as soon as the subject had found all eight matches or all 10 zeros (depending on the task) and observed the board reset action.

3.2 Interruption Detection Task

While performing the primary tasks, subjects were asked to respond by pressing the space bar with their non-dominant hand whenever they noticed a notification signal.

3.2.1 Notification Signals

In order to establish between three and five signals for our subsequent experiments, we designed and studied a base set of 10 different signals. These notification signals were presented while subjects performed the two primary tasks. Signals were comprised of transformations applied to an icon that was present on the screen at all times. The base icon used was a blue circle with a diameter of 21 pixels (0.62cm) that was located in the bottom right-hand corner of the screen. The placement of this icon was meant to emulate the Windows OS system tray. The base icon can be seen in Figure 3.1.

We designed the notification signals across four categories that we hypothesized would span the spectrum of AD. Parameters such as colour change rates and movement velocities were based on the perception literature as well as informal piloting. Relevant literature is cited in our hypotheses about the detection time of these signals (Section 3.10). The signals, by category, were as follows.

Category A: Single State Change

Each of these signals consists of a single state change. If the user does not notice the change as it happens, only a polling action (i.e., looking directly at the icon) will result in detection.

FLAG (FG): A yellow exclamation mark appeared in the centre of the icon.

COLOUR (CR): The icon colour changed to yellow.

GROW (GR): The icon smoothly grew to 200% of its original size (41 pixels, 1.09cm diameter), centered on its origin. This grow action took place over 500 ms.

Category B: Continuous Slow State Change

These signals demonstrate continuous transformation. This ongoing activity is more likely to attract attention than the single change nature of Category A.

OSCILLATE (OS): The icon moved slowly up and down a path of 17 pixels (0.5cm) with sinusoidal motion. It took 1700 ms for the icon to complete one a full cycle (up and back down again).

SLOW ZOOM (SZ): The icon smoothly and continuously grew and shrank between 100% and 200% of its original size, centered on its origin. This continuous motion occurred at a slow velocity: it took 1500 ms for the icon to complete one full cycle of growing and shrinking.

SLOW BLINK (SB): The icon continuously flashed back and forth from blue to yellow. This continuous colour change occurred at a slow velocity: the icon changed colour every 1000ms.

Category C: Continuous Fast State Change

As in Category B, these signals demonstrate continuous transformation, but they are more likely to attract attention than Category B because the transformation occurs at a higher velocity than its corresponding cue in Category B.

BOUNCE (BC): The icon moved up and down a path of 17 pixels (0.5cm) with a bouncing motion. One bounce took 800 ms to complete, and each bounce occurred as soon as the previous bounce finished.

FAST ZOOM (FZ): The icon smoothly and continuously grew and shrank between 100% and 200% of its original size, centered on its origin. This continuous motion occurred at a high velocity: it took 780 ms for the icon to complete one full cycle of growing and shrinking.

FAST BLINK (FB): The icon continuously flashed back and forth from blue to yellow. This continuous colour change occurred at a high velocity: the icon changed colour every 300 ms.

Category D: Continuous Location Change

The signals in Categories A, B, and C leave the icon in its initial location in the periphery of the screen. FOLLOW, on the other hand, brings a copy of the icon into the fovial area.

FOLLOW (FL): A copy of the icon appeared beside the mouse cursor and continued to follow the cursor until detection occurred or the trial timed out.

A visual representation of these signals can be found in Figure 3.3.

3.2.2 Block Design

Our experimental setup was designed so that subjects would *be interrupted* rather than *wait for an interruption*. Similarly to Bartram et al. [8], we introduced variation in interruption onset times in two ways. First, signal onset occurred at a random point for each trial between 5 and 20 seconds after the trial started. The signal was presented until it was detected or until the trial timed out after 30 seconds. A trial began immediately after the previous trial ended (i.e., after either signal detection or timeout). The structure of a trial is illustrated in Figure 3.4.

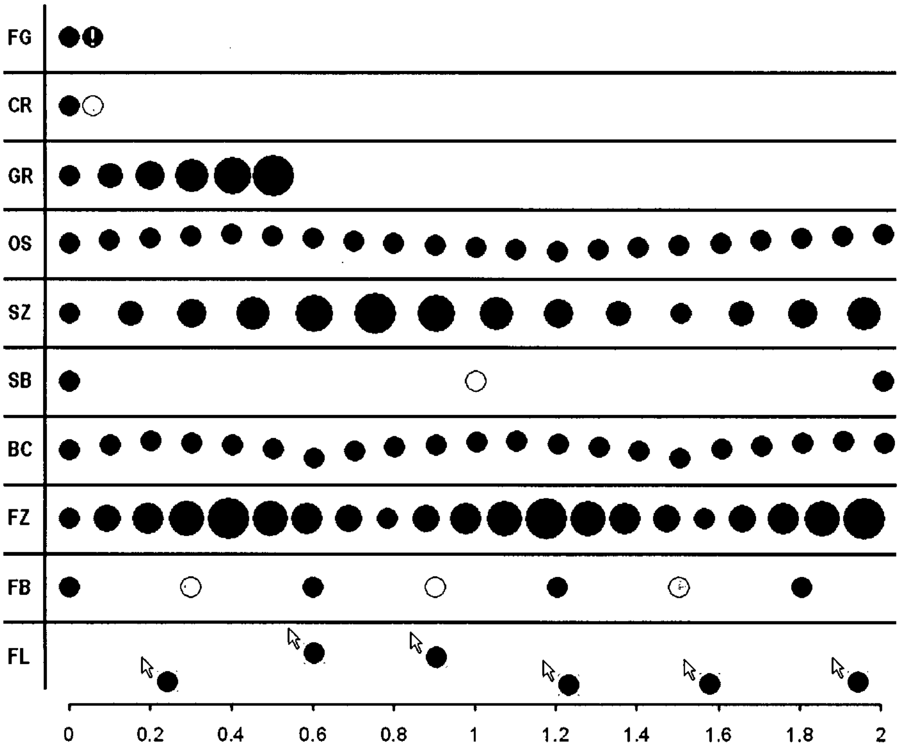


Figure 3.3: A visual representation of each of the 10 notification signals during the first 2 seconds of notification. Single state change signals are illustrated only up to the point in time when they cease to change.

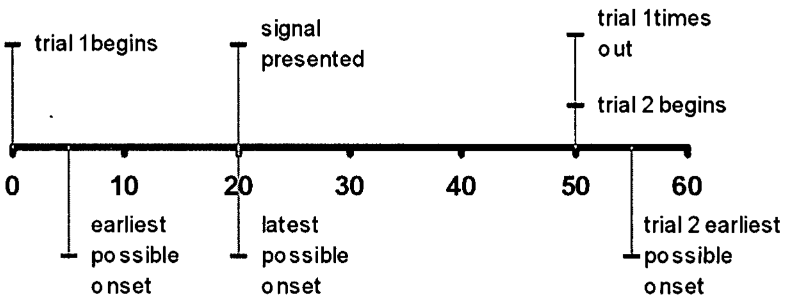


Figure 3.4: Structure of a trial, in seconds. In this trial, the signal was not detected and so a timeout occurred.

Second, we inserted a number of “dummy” cases in which no signal was presented. For each replication of the 10 signals we included three dummy slots, resulting in 13 potential slots for interruption. Thus, in 23% of the slots nothing happened. A block contained two replications of each signal and six dummy slots, for a total of 26 potential trial slots with 20 actual interruption trials. The ordering of signal presentation and the placement of the dummy slots were randomized within a block independently for each subject. Blocks were repeated three times for each of the two primary tasks, totaling 120 trials per subject.

3.2.3 Training

At the start of the experiment, subjects were given a training block that demonstrated all 10 signals in order to ensure that they were familiar with all of the signals before the experiment began.

3.3 Duration

On average, subjects took between 8 and 10 minutes to complete each of the six blocks. The duration varied with how quickly the subject detected the signals because the start time of each trial was based on the end time of the previous trial. This allowed us to strictly control the time between interruptions, as well as to replicate the design of Experiment 1 in Bartram et al.'s Moticon research [8]. The duration of the entire experiment (trials, instructions, surveys, and detection task) was approximately 90 minutes.

3.4 Motivation

Motivating the intended cognitive split between primary task and the secondary signal detection task was a challenge. Our definition of intended behaviour was somewhat fuzzy: we wanted subjects to focus mainly on the primary task and to have a milder interest in the interrupting task. This was intended to mimic common user behaviour where, for example, the user is engaged in writing a document but is also open to glancing at email as it arrives. Although we wanted subjects to concentrate mainly on the primary task, some of the low-AD signals (i.e., FLAG and OSCILLATE) could only be detected via polling (i.e., looking directly at the icon). We motivated realistic and consistent behaviour by carefully using the words "primary" and "secondary" in the instructions given to the subjects. Ultimately, we realized that we could not fully control polling behaviour. Instead, we utilized the exit questionnaire and informal interview to track the detection strategy used by each subject.

To further motivate subjects to focus on the primary task but also devote some attention to the detection task, subjects were told that an extra $10 would be provided to the 1/3 of the subjects who achieved the best performance. Subjects were told that their comprehensive scores would be largely based on scores for the primary tasks but would also take into account detection of the notification signals.[1] The explanation of scoring was deliberately vague so that participants would not try to fit their performance to the specifics of the scoring system. The 1/3 ratio was chosen to encourage subjects to believe they had a reasonable chance of being paid the extra money.

See Appendix B.1 for the exact wording used to instruct the subjects.

[1] Scores for each block were calculated by dividing the total number of matches (for the Memory Game task) or edits (for the Simple Editor task) by the block duration. The comprehensive score for each subject was the average of the block scores.

3.5 Apparatus

The experiment was conducted on a system running Windows XP with a 3GHz Pentium 4 processor, 1.0 GB RAM, an nVidia GEForce 6800 GT video card, and a 19 inch monitor configured at a resolution of 1280×1024. The experimental software, including all notifications signals, was fully automated and was coded in Java 1.5.0.

3.6 Participants

Twelve subjects (1 female) between 18 and 39 years of age participated in the study and were compensated $15 for their participation. All subjects had normal colour vision, were right-handed, and were recruited using an online experiment management system accessed by students at the University of British Columbia.

3.7 Design

The experiment used a within-subjects 2 × 10 × 3 (primary task × notification signal × block) design. There were also two orders of presentation of the primary task, a between-subjects control variable introduced to minimize order effects.

3.8 Procedure

The procedure was as follows. (1) A questionnaire was used to obtain information on user demographics. (2) A signal training block demonstrated all 10 notification signals. (3) For each of the two primary tasks, verbal instructions and a training session ensured that each subject understood the primary task. Subjects then performed three blocks for each task. A block ended after all interruption trials had occurred. Each block took approximately 8 to 10 minutes to complete, and there was a 2-minute break in between the blocks. There was also a 2-minute break between the two primary task conditions. (4) A questionnaire was used to collect annoyance rankings for the notification signals. Brief, informal interviews were also conducted with some of the subjects when it was necessary to obtain clarification on questionnaire responses.

All the study instruments including questionnaires and the exact wording of the instructions given to participants can be found in Appedix B. The background questionnaire used to collect user demographic information is given in Appendix A.

3.9 Measures

Our main dependent variable was detection time. This measure was capped at 30 seconds when the notification signal timed out. We also report on the number of timeouts and false detections. A false detection occurred when a subject pressed the spacebar while no signal was present.

Annoyance measures were collected via a questionnaire at the end of the study. Subjects were asked to indicate how annoying each signal was in two ways. First, subjects rated the annoyance of each of the 10 signals on a 5-point Likert scale where 1 indicated low annoyance and 5 indicated high annoyance. In order to guard against the case where a subject rated all signals equally, we also asked subjects to rank the three most annoying and three least annoying signals. Annoyance was defined as "To make slightly angry; to pester or harass; to disturb or irritate."

As mentioned in Section 3.4, we also tracked detection strategy.

3.10 Hypotheses

H1: Categories B, C, and D will have faster detection rates than Category A.

i.e. Continuous changes will dominate single state changes.

The blinking and moving targets in Group B are persistent, unlike the instantaneous icon changes in Group A which, according to Ware [46, 47], rapidly fade from attention and are likely to be missed unless they are explicitly monitored by subjects. Bartram et al. [8] and Baecker and Small [6] concur that motion and blinking are more noticable than single state changes.

H2: Category C will have faster detection rates than Category B.

i.e. Speed will be the dominant factor.

The FAST ZOOM and FAST BLINK cues in Group C are the same as SLOW ZOOM and SLOW BLINK cues in Group B except that the zooming and blinking take place at a higher velocity. Ware et al. [47] tell us that motion at a higher velocity results in faster detection times. By extension, intuition tells us that an icon blinking at a higher velocity will result in faster detection time than an icon blinking at a slower velocity.

H3: Category D will have faster detection rates than Category C.

i.e. Fovial location will dominate periphery location

Fast blinking and high-velocity motion in the periphery are both known to be near the fast end of the detection rate spectrum [23, 47]. However, we hypothesize that placing an icon directly into the fovial area will result in even faster detection rates. When icons are in the periphery, the goal is to attract the user's attention to that location. Detection time should be faster if the computer does the work instead of the user by placing the icon on the fovial area instead of forcing an eye movement that moves the fovial area to the peripheral icon location.

H4: All detection rates will be faster during the Simple Editor task than during the Memory task.

We expect that subjects will detect the icon changes faster when they are working on a primary task that has low workload than when they are working on a primary task that has high workload.

3.11 Results

Data for four trials was discarded due to logging corruption. In order to salvage the remaining data from the affected subjects, we averaged means across the two replications of signals in each block.

A series of 2 (task) by 10 (notification signal) by 3 (block) ANOVAs were per-

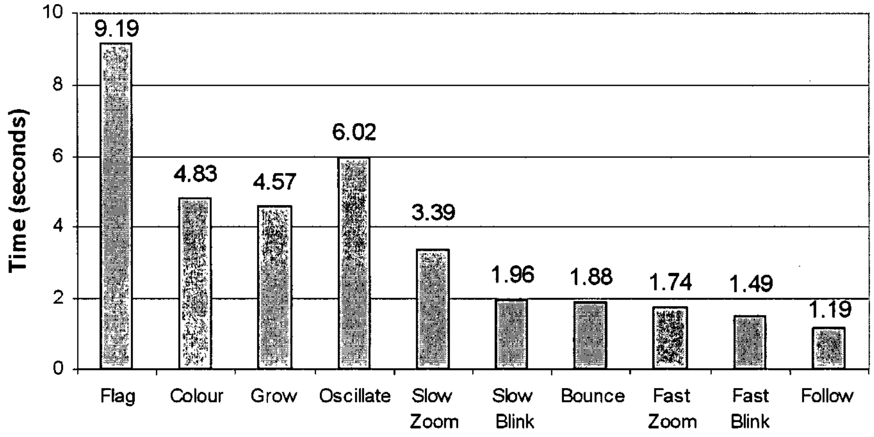


Figure 3.5: Mean detection times by signal ($N = 12$).

formed.[2] The Greenhouse-Geisser adjustment was used for non-spherical data, and the Bonferroni adjustment for post-hoc pair-wise comparisons. Along with statistical significance, we report partial eta-squared (η^2), a measure of effect size. To interpret this value, .01 is a small effect size, .06 is medium, and .14 is large [14].

3.11.1 Detection Times by Signal

Detection times for each signal are summarized in Figure 3.5. As expected, there was a very large main effect of signal on detection time ($F(2.632, 28.954) = 14.204, p < .001, \eta^2 = 0.564$). Figure 3.6 summarizes the significant pairwise comparisons.

The goal of this study was to discover a set of signals with mean detection times that are significantly different from one another. Pairwise comparisons show that one subset of the signals fulfills this requirement: a three-way significant comparison between FLAG (FG), SLOW ZOOM (SZ), and FOLLOW (FL). The detection time for FLAG was significantly slower than both SLOW ZOOM ($p = .036$) and FOLLOW ($p = .015$), and the detection time for SLOW ZOOM was significantly slower than

[2] A 2 (task) by 10 (signal) by 3 (block) by 2 (presentation order) ANOVA showed no significant main ($F(1,10) = .001, p = .982, \eta^2 < .001$) or interaction effects of presentation order, so in all subsequent analysis we examine only the effects of task, notification signal, and block.

	FG	CR	GR	OS	SZ	SB	BC	FZ	FB	FL
FG	▨									
CR		▨								
GR			▨							
OS				▨						
SZ	■				▨					
SB	■		■	■		▨				
BC	■			■			▨			
FZ	■		■	■				▨		
FB	■		■	■					▨	
FL	■		■	■	■					▨

Figure 3.6: Pairwise comparisons of detection times: a square indicates that the row signal had a faster detection time than the column signal. Significance is at the .05 level ($N = 12$).

FOLLOW ($p = .044$). There were no significant comparisons of four or more signals.

3.11.2 Detection Times by Signal Category

We also analyzed the signal data using the categories defined in Section 3.2.1 to test our hypotheses. Figure 3.7 shows detection times by category. There was a significant main effect of signal category ($F(1.292, 14.208) = 28.059, p < .001, \eta^2 = 0.718$). Pairwise comparisons showed that, consistent with H1 and H2, detection rates for Category A were significantly slower than for Category B ($p = .005$), Category C ($p = .001$), and Category D ($p = .001$), and detection rates for Category B were significantly slower than for Category C ($p = .002$) and Category D ($p = .003$). Although the mean detection time for Category D (1.19 s) was faster than Category C (1.70 s), the difference was not statistically significant and so we cannot accept H3.

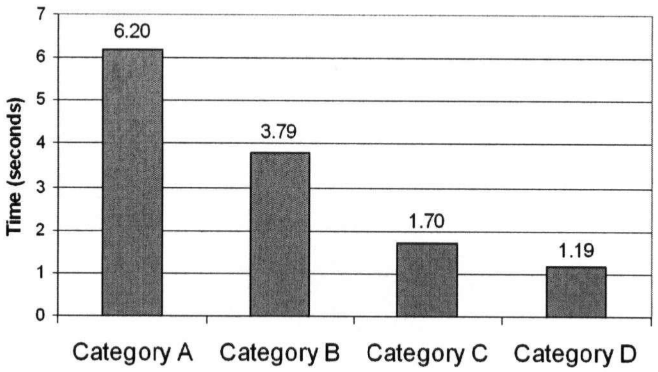


Figure 3.7: Mean detection times by category ($N = 12$).

3.11.3 Effect of Task

Counter to H4, there was no significant main effect of task $(F(1,11)) = .781, p = .396, \eta^2 = .066)$. Although there was a significant interaction effect of task and signal $(F(3.918, 43.103) = 2.676, p = .045, \eta^2 = 0.196)$, we found no consistent pattern across signal and task. Figure 3.8 shows that there is no distinct pattern: some of the means are higher for the Memory Game task while others are higher for the Editor Task. However, when we look only at the 3 signals identified above, the effect of task is not a large concern: paired-samples t-tests showed that detection times were not significantly different between tasks for FLAG ($p = .084$), SLOW ZOOM ($p = .479$) or FOLLOW ($p = .231$). Thus, the three identified signals were robust to tasks with varying cognitive workload.

3.11.4 Timeouts and False Detections

Timeout rates are reported in Figure 3.9. There was a significant main effect of signal $(F(2.542, 27.961) = 3.630, p = .031, \eta^2 = .248)$, but no post-hoc pairwise comparisons were significant. Timeout rates were low across the board, especially for SLOW ZOOM and FOLLOW (0.35% and 0%, respectively), and FLAG had a timeout rate of only 4.55%.

The mean number of false detections in each block per participant ranged from 0.17 to 1.50, and on average was less than one per block. This very low number indicated

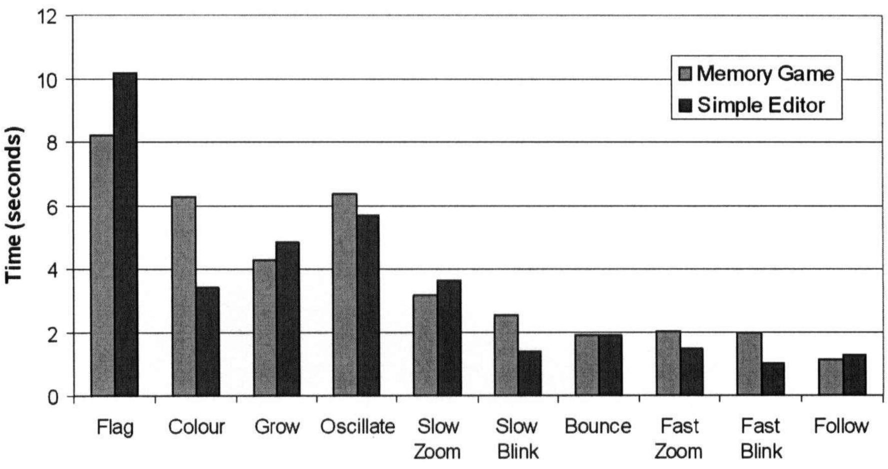


Figure 3.8: Mean detection times by signal and task ($N = 12$).

that false detections were not a concern.

3.11.5 Learning Effects

There was a clear learning effect on detection times ($F(2,22) = 6.150, p = .008, \eta^2 = 0.359$), number of timeouts ($F(2,22) = 4.142, p = .030, \eta^2 = 0.274$), and performance ($F(2,22) = 5.725, p = .010, \eta2 = 0.342$). Subjects detected the notification signals more quickly ($p = .042$), missed fewer signals (borderline: $p = .056$), and achieved higher scores ($p = .038$) between the first and third blocks. The fact that subjects improved at both performing the primary task and detecting the signals over time is not surprising.

3.11.6 Detection Strategy

Based on subjects' written and verbal descriptions of their detection strategies in the exit survey and informal interview, we defined three categories of detection:

1. Mostly peripheral vision (five subjects)

   Subjects in this category indicated that they relied mainly on their peripheral

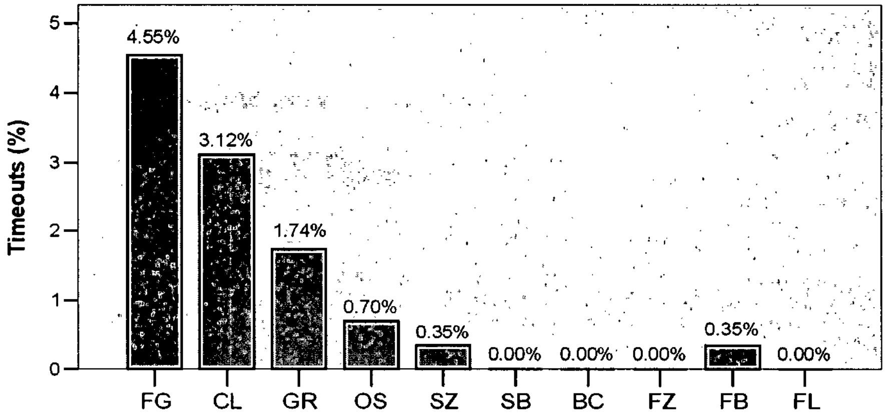


Figure 3.9: Mean timeout rate by signal ($N = 12$). Note the scale is from 0 to 5%.

vision to catch the icon changes. The subjects either did not poll or indicated that they polled "every once in awhile," or "not often at all." We define this category to describe subjects whose answers indicated that they devoted roughly 0-10% of effort to polling.

2. Mostly explicit polling of the icon (three subjects)

   Subjects in this category indicated that they regularly polled the icon (i.e., deliberately looked directly at the icon) in order to monitor icon changes. Many of these subjects revealed that they were able to detect most of the icon changes using peripheral vision, with the notable exception of the FLAG signal. Because they were unable to reliably detect FLAG using peripheral vision, these subjects polled regularly. We defined this category to be roughly 75-100% of effort devoted to polling.

3. A mixture of peripheral vision and polling (four subjects)

   A number of subjects indicated that they relied on a mixture of polling and peripheral vision to detect the icon changes. We defined this category to be roughly 11-74% of effort devoted to polling.

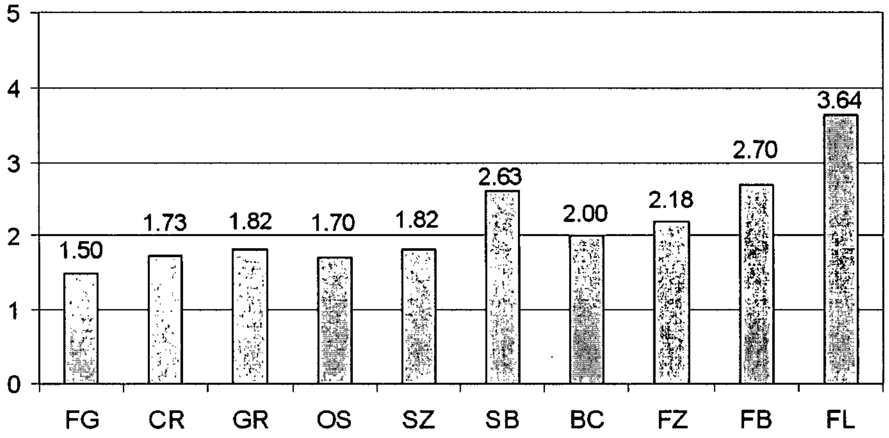


Figure 3.10: Annoyance ratings, by signal, on a 5-point Likert scale where five is most annoying ($N = 11$).

We were concerned that the diversity in strategy could affect detection times. A 2 (task) by 3 (block) by 3 (strategy: between-subjects) ANOVA did not reveal any significant main effect of strategy ($F(2,9) = 1.552, p = 0.264, \eta^2 = 0.256$) on detection times. However, the large effect size and low power (.248) indicate that an effect might have been present but we did not have enough power to detect it. Thus, our result is inconclusive. Future work might employ eye-tracking to address this issue more carefully.

3.11.7 Self-Reported Measures

Results of the self-reported measures are summarized in Figure 3.10, Figure 3.11, and Figure 3.12.[3] A one-way ANOVA on the annoyance ratings revealed a statistically significant main effect of signal ($(F(9,9)) = 3.285, p = .002, \eta^2 = .243$). Pairwise comparisons showed that the FOLLOW signal was more annoying than FLAG ($p = .006$), COLOUR ($p = .008$), GROW ($p = .016$), OSCILLATE ($p = .010$), and SLOW ZOOM ($p = .016$).

[3] Due to his misunderstanding of the questionnaire, data for one subject was excluded.

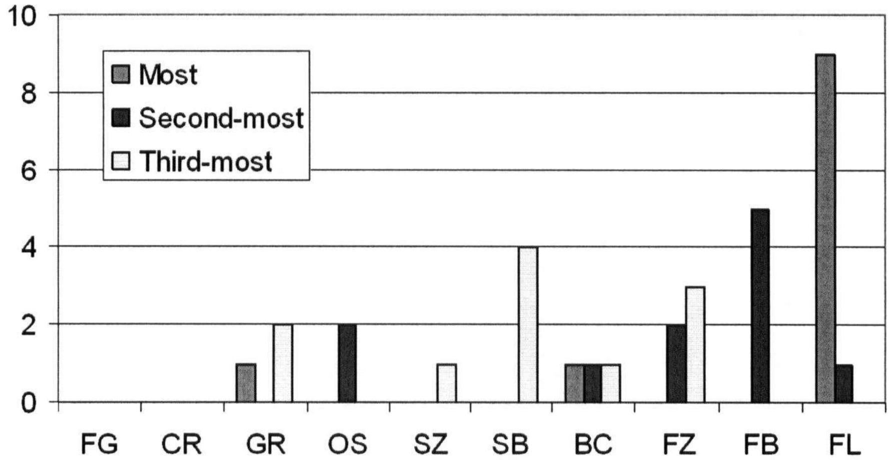


Figure 3.11: Frequency distribution of "Most Annoying" rankings (N=11).

Figure 3.11 shows that FOLLOW was clearly the most annoying of the signals. From Figure 3.12, we can see that FLAG was the least annoying signal.

3.11.8 Summary of Results

We summarize our results according to our hypotheses:

H1 supported. Groups B, C, and D had faster detection rates than Group A, i.e. continuous changes dominated single state changes.

H2 supported. Group C had faster detection rates than Group, i.e. speed was the dominant factor.

H3 not supported. Group D did not have faster detection rates than Group C, i.e. fovial location did not dominate periphery location.

H4 not supported. Detection rates were not faster in the Simple Editor task than in the Memory task.

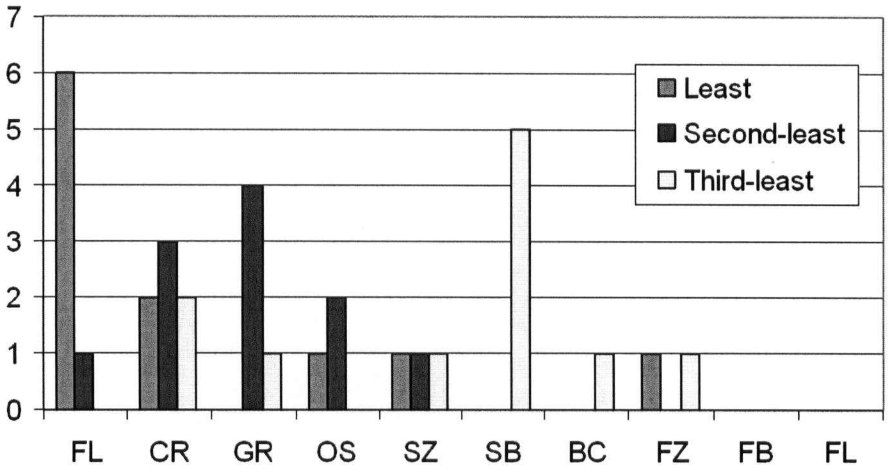


Figure 3.12: Frequency distribution of "Least Annoying" rankings ($N = 11$).

3.12 Limitations

We recognize that there was considerable variation in the amount of polling. This could have been due to the artificial lab setting, insufficiently precise instructions on polling behaviour, or simply individual differences. It may only be possible to see true polling behaviour in a less artificial setting (i.e., one in which subjects are legitimately motivated to anticipate interruptions). Even in such a setting, however, there would likely be considerable variation between users as well as context dependencies.

3.13 Discussion and Conclusions

The goal of this study was to identify a set of three to five statistically significantly different signals. Our results revealed one subset of the signals that met this requirement: FLAG (mean detect time: 9.193 s), SLOW ZOOM (mean detect time: 3.385 s), and FOLLOW (mean detect time: 1.187 s). These signals were in categories A, B, and D, respectively. We were not concerned about the number of missed interruptions since timeout rates were low. In Studies 2 and 3, we used FLAG as the signal with low AD, SLOW ZOOM as the signal with medium AD and FOLLOW as the signal with high

AD.

The negligible effect of task on the selected subset of signals suggests that the mean detection times for these signals generalized across primary task workload, indicating that this set of signals may be utilized successfully in future research.

Our main motivation in performing this experiment was the selection of signals to be used in subsequent studies, but we did not plan to base this selection heavily on the subjective measures of annoyance. Initially we had viewed the annoyance ratings as a backup in case we had multiple high-attentional-draw candidate signals from which to choose, but the qualitative results generally support the detection time results. We were also curious to record some baseline measurements with which to compare results from our second study.

Results of the self-reported measures indicate that FLAG was considered the least annoying of the signals, while FOLLOW was the most annoying signal. We were very interested to determine in subsequent experiments if matching utility and attentional draw is able to mitigate the high level of annoyance associated with the FOLLOW signal.

Chapter 4

Experimental Approach for Studies 2 and 3

We conducted two controlled experiments in order to examine the effects of matching the attentional draw associated with an interruption to the utility of its content. The two studies were similarly structured and we document the core experimental approach taken in this chapter. The specific motivation and experimental design of each study along with other methodological differences are presented in Chapter 5 (Study Two) and Chapter 6 (Study Three).

4.1 Scope of Utility

In order to study the impact of matching the attentional draw of an interruption signal with the utility of its content, it was first necessary to delineate the scope of the interruption utility. We considered defining utility as relevance to (a) a primary task, and (b) to the user in general, as are typically delivered via personal systems such as IM, email, or calendar software. Option (b) is very difficult to simulate in a lab environment, requiring knowledge about the personal lives of individual subjects (e.g., work, family, and living situations) and relying on a certain degree of subject suspension of disbelief and role playing (i.e., "Pretend you have a child and it is important if the day care centre contacts you," or, "Imagine you are waiting for a very important email," etc). Moreover, it is not clear that defining utility as general relevance to users would provide extra insight above utility as relevance to the task. Thus, in our experiments, interruption utility was relevant solely to the primary task.

4.2 Number of Levels of Utility and Attentional Draw

We employed three levels of AD and three levels of utility in our studies. We had initially planned to use more levels of AD than levels of utility in order to examine the effects of using different levels of AD with the same level of utility. For example, this type of study could utilize the following design:

- 3 levels of utility: $U1, U2, U3$
- 4 levels of AD: $A1, A2, A3, A4$
- 4 conditions:
    1. $U1/A1$, $U2/A2$, $U3/A3$
    2. $U1/A1$, $U2/A2$, $U3/A4$
    3. $U1/A1$, $U2/A3$, $U3/A4$
    4. $U1/A2$, $U2/A3$, $U3/A4$

Results from this type of study could offer interface recommendations in terms of ideal levels and thresholds by answering questions such as, "Is there a level of AD that is so high it will be perceived as annoying no matter how high the utility?" and, "Is there such a thing as AD that is 'too low?' Will users get frustrated if presentation format is too subtle, even for the minimum utility level?" However, we realized that these goals were too ambitious for an initial experiment. We decided to focus on our primary goal using a more simplified experimental design, leaving the examination of ideal levels of AD for future work. Thus, we worked with an equal number of levels of utility and attentional draw (three of each).

4.3 Primary Task

Our primary task was the Memory Game used in Study 1. As mentioned in Chapter 3, the selection of this task was based on two key requirements: (1) the need to be able to generate interruptions with an objective measure of utility; and (2) the need to involve concentration such that moving from the primary task to the interrupting task requires a cognitive context switch.

4.4 Interrupting Task

The interrupting task was comprised of context-sensitive hints and comments, many of which aimed to aid the subject in playing the game. As we have already mentioned, an interruption typically consists of two components: notification and content. In our studies, the content was a hint. The notification signal indicated the availability of a hint. Once subjects noticed the notification, they could view the hint by clicking on the icon located in the lower right-hand corner of the screen. Each hint was associated with a particular level of utility. Study 1 provided us with a set of three notification signals. Because we decided to use an equal number of signals and utility levels, we defined three levels of utility: *low* (not helpful), *medium* (somewhat helpful), and *high* (very helpful). Subjects saw equal numbers of each of the three types of hints in all interruption conditions. This interrupting task was designed to emulate a mixed-initiative system: the enhanced Memory Game interface offered, in real time, context-sensitive suggestions with the potential to reduce the amount of time necessary to perform the task at hand, but it was up to the user to decide if and when to utilize each suggestion.

4.4.1 Notification Signals

The time scale in Figure 4.1 shows our spectrum of attentional draw according to the three presentation formats we selected in Study 1.

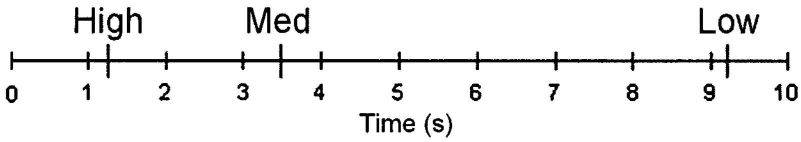


Figure 4.1: Time scale showing the three notification signals along the spectrum of attentional draw.

As a reminder, FLAG, SLOW ZOOM, and FOLLOW were selected for our low, medium, and high AD notification signals, respectively. See Section 3.2.1 for details about these signals.

4.4.2 Hint Utility

One of our main goals was to investigate the perceived benefit of matching utility with the attentional draw. To achieve *perceived* benefit, however, we felt that it was necessary for the interruption system to *actually* be beneficial. Thus, our interruptions were fashioned to boost performance on the primary task.

Hints were defined differently for the two studies and will be discussed in each respective study chapters.

Relationship between Hint Utility, Performance and Workload

Hart and Lowell [25] define workload as the cost incurred by a user to achieve a particular level of performance. That is, workload is proportional to cost and inversely proportional to performance. Interruption requires extra effort from the user to switch between primary and interruption tasks and thus increases cost to the user. If there is no compensatory increase in performance, workload goes up. This is the case when an interruption is unrelated or yields little performance benefit to the main task. If the interruption content increases performance on the primary task, however, there is a potential to actually reduce workload. In our studies, we expected interruption to boost task performance enough to mitigate the increase in cost such that workload under interruption was no higher than workload in the no-interruption condition.

4.4.3 Structure of an Interruption

As in Study 1, we implemented an interruption timeout of 30 seconds. If a subject did not respond to the notification signal within this amount of time, the notification stopped and the subject missed the hint. Piloting revealed an upper bound of 6 seconds on the amount of time a subject would need to attend to one of the hints. In order to prevent subjects from feeling overwhelmed by the interruptions, we also implemented a hard lower bound of 10 seconds between when a subject finished with a hint and when the next notification signal began. Interruption onset was again varied; however, to ensure that all blocks were identical in length for all subjects regardless of signal detection times, an interruption occurred every 65 seconds plus or minus a random number between 1 and 10 seconds. Thus, interruptions were at least 45 seconds and at most 85 seconds apart, depending on the random onset. Figure 4.2 illustrates the structure of an interruption trial.

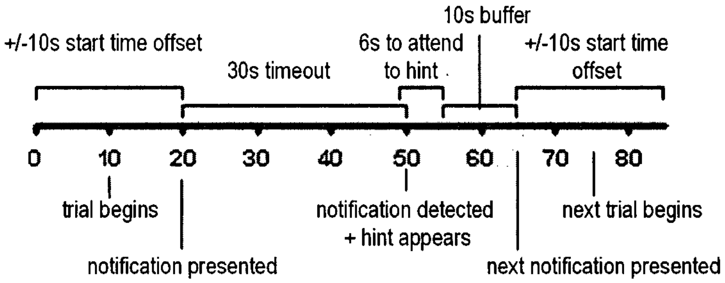


Figure 4.2: Structure of a trial, in seconds. In this trial, the notification signal was detected at the last possible moment before a timeout could occur. Latest possible onset for the first trial and earliest possible onset for the second trial create the case where interruptions are closest together (45s).

In Study 2, similarly to Study 1, we increased variation in interruption onset by including a number of "dummy" cases in which no signal was presented (see Section 3.2.2). The inclusion of these extra slots lowered the average interruption frequency from 65 seconds to 80 seconds.

In Study 3, we did not include the dummy slots, leaving the average interruption frequency at 65 seconds. This was done in order to minimize block duration and increase the frequency of interruption. Piloting indicated that the onset variation provided enough variance in interruption timing such that subjects would be interrupted rather than wait for an interruption.

Interruption frequency in previous work [1, 2, 7, 15, 16, 17, 20, 29, 36, 41] ranged from 3 seconds to 5 minutes with an average of 2 minutes, but with the exception of the primary task. In Study 2 we chose an interruption frequency of 80 seconds, which is slightly above the average in the research literature. This was done because our interruptions were relevant to the primary task and we needed sufficient opportunity for the high utility hints to boost performance and the low ones to annoy. Results of Study 2 suggested that despite this effort, annoyance was rather low, and so we futher increased interruption frequency in Study 3 to 65 seconds.

Previous work examining relevancy of interruption [17] utilized three to four replications for each type of interruption. In Study 2, because we used a between-subjects design, fatigue was not a concern and so we had the luxury of using a large number of replications (8) to give subjects opportunity to discover the relationship between notification signal and hint utility in the Match condition. In Study 3, we balanced maximizing the number of replications of each hint utility (i.e., to give subjects the opportunity to comprehend the coordination of notification signal and hint utility) and minimizing the duration of game play to avoid excessive subject fatigue.

4.5 Conditions

In order to study the effects of matching attentional draw to utility, we compared annoyance, perceived benefit, workload, and performance across four conditions: Match, Static, Random, and Control. Study 2 included all four conditions, and Study 3 included all conditions but Random.

4.5.1 Match

This condition represented a system where the AD of the notification was matched with its corresponding level of utility. We matched low utility hints with the low AD signal (FLAG), medium utility hints with the medium AD signal (SLOW ZOOM), and high utility hints with the high AD signal (FOLLOW).

4.5.2 Static

This condition was designed to emulate current practices, where all notification takes the same form. Billsus et al. [9] state that the key problem with proactive information systems - a problem we believe generalizes to most interfaces that interrupt - is that notification is either too subtle or too obtrusive. We attempted to avoid this problem by employing the notification signal with medium AD (SLOW ZOOM).

4.5.3 Random

This condition was meant to emulate the worst-case scenario of a system that tries to match AD to utility but incorrectly assesses utility. Here, the type of notification signal was randomly selected for each hint, regardless of utility.

4.5.4 Control

We included an interruption-free Control condition in order to establish baseline workload and performance measures.

4.6 Experimental Design Issues

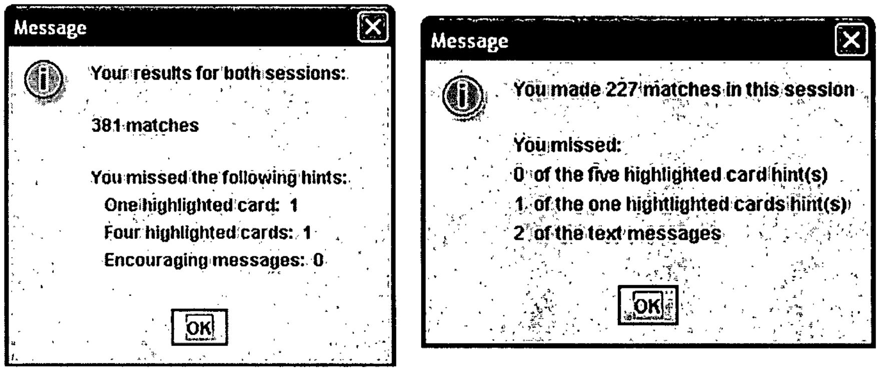
Designing a controlled lab experiment to examine annoyance and perceived benefit in interruption presented unique challenges. Here, we highlight two important challenges.

4.6.1 Motivation

It was necessary to motivate subjects to utilize the hints. Subjects were told that an extra $10 would be provided to the 1/3 of the subjects who found the highest number of matches during the experiment. The goal was to encourage subjects to maximize their performance, thereby motivating them to use the hints if they recognized that doing so would help them to achieve higher scores. The 1/3 ratio was chosen to encourage subjects to believe they had a reasonable chance of being paid the extra money.

4.6.2 Sources of Annoyance

An irrelevant or poorly-timed interruption is an obvious cause of annoyance. Another possible cause of annoyance to a subject is the retrospective knowledge that she missed a hint that would have boosted her score. In an equivalent real world context, when a user misses an interruption that is highly important, it usually follows that she later finds out about the missed message and experiences annoyance, even aggravation. To elicit this type of annoyance, we required subjects to know when they had missed high utility interruptions. To this end, at the end of the Match, Static and Random conditions, subjects were informed of the number and types of hints that were missed during that condition. Figure 4.3 shows the dialog boxes used to convey this information.



Message

Your results for both sessions:

381 matches

You missed the following hints:
One highlighted card: 1
Four highlighted cards: 1
Encouraging messages: 0

Message

You made 227 matches in this session

You missed:
0 of the five highlighted card hint(s)
1 of the one hightlighted cards hint(s)
2 of the text messages

Figure 4.3: End-of-session dialog boxes listing missed hint information. The box on the left was used in Study 2. The box on the right was used in Study 3. The only differencees between the two are the descriptions of the hints.

4.7 Measures

Our main dependant measures were annoyance, perceived benefit, workload, and performance. Performance was measured as the number of matches made in each condition. The remaining three measures were self-reported through questionnaires. We also measured detection times for the notification signals and the number of timeouts.

We used the NASA-TLX scales [25], a standardized instrument for assessing various dimensions of workload. In Study 2, the dimensions used were mental demand, temporal demand, effort, perceived performance, and frustration. In Study 3, we also included physical demand. Ratings were measured on a 20-point scale. In Study 2 we graded these ratings from 1 to 20, while in Study 3 we graded from 5 to 100; however, the scales presented to subjects were identical in both studies.[4] Definitions for each of the dimensions can be found in the questionnaires in Appendix D.

Perceived benefit and annoyance were assessed through additional questions we added to the TLX, in a manner similar to [1], where subjects rated statements from low to high on a 20-point scale. The statements rated were as follows:

perceived benefit: "To what extent did your performance benefit from the hints?"

general annoyance: "How annoyed (i.e. pestered, harassed, disturbed or irritated) did you feel during the task in general?"

interruption annoyance: "How annoyed (i.e. pestered, harassed, disturbed or irritated) were you by the notifications and hints in particular?"

With respect to annoyance, piloting indicated that good performance on the game tended to mitigate annoyance specific to the interruptions. This phenomenon resulted in low annoyance ratings even when subjects later admitted to feeling considerable annoyance caused by the interruptions. Therefore, we defined two measures of annoyance: one related to the task in general, and one specific to the interruptions, as given above.

The questionnaires also elicited fatigue ratings on a 5-point Likert scale. Subjects were asked to respond to the statement, "I felt fatigued during this session."

Secondary qualitative measures were dependent on the individual study and are discussed in each respective study chapter.

[4] Grading on a scale of 100 is necessary for certain additional TLX assessments that we considered but did not end up persuing in Study 3.

4.8 Apparatus

Both studies were conducted on a system running Windows XP with a 3GHz Pentium 4 processor, 1.0 GB RAM, an nVidia GEForce 6800 GT video card, and a 19 inch monitor configured at a resolution of 1280 × 1024. The experimental systems, including all notifications signals, were fully automated and were coded in Java 1.5.0.

4.9 Summary

The experimental approach outlined in this chapter provides the foundation for our second and third controlled lab experiments. We have outlined our primary and interrupting tasks, as well as the conditions we employed to examine the matching of attentional draw and interruption utility.

Chapter 5

Study 2

Study 2 was our initial investigation of utility and attentional draw. This chapter presents the study design and results, as well as a discussion motivating Study 3 (Chapter 6).

5.1 Methodology

The methodology used for this study is based on the core experimental approach documented in the previous chapter. Only additions and clarifications are highlighted here.

5.1.1 Conditions

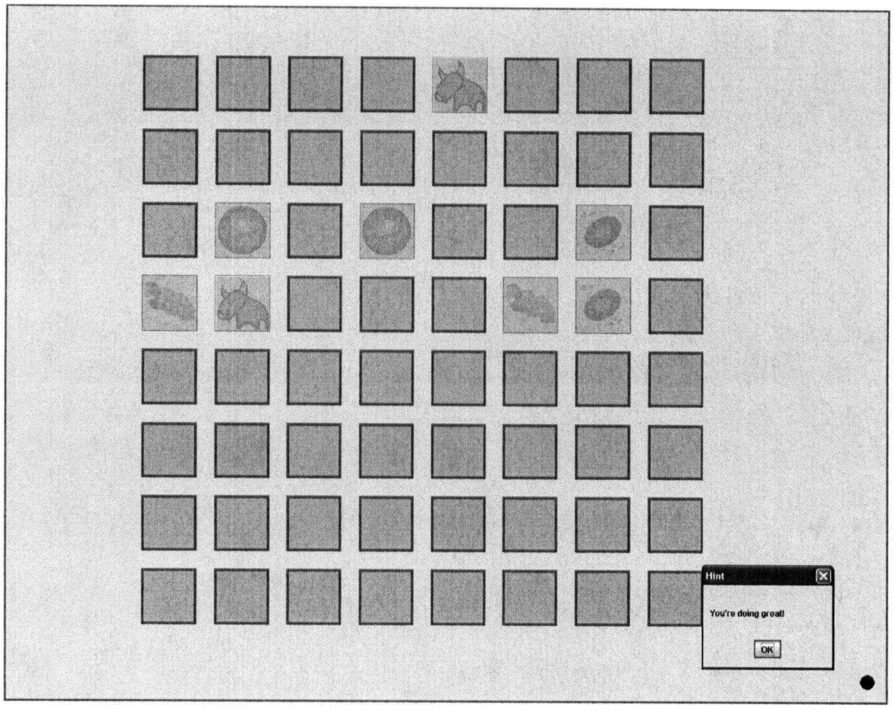
In this study we investigated all four of the conditions defined in the previous chapter: Match, Static, Random, and Control.

5.1.2 Hints and Utility

As mentioned in the previous chapter, our goal was to create hints that would help subjects to perform the primary task. Because the hints were helpful we assumed that they would boost performance but we did not calculate a precise estimate. We required hints with three levels of utility: not helpful (low), somewhat helpful (medium), and very helpful (high).

Low-Utility Hint

Hints with the lowest utility did not provide the user with any assistance in finding a match. Instead, this type of interruption always showed a popup box with the encour-



You're doing great!

Figure 5.1: Screenshot of a low-utility hint. The popup box reads, "You're doing great!.

aging message, "You're doing great!" Subjects were required to dismiss the popup box by clicking on "OK" before they could continue playing the game. Figure 5.1 shows a screen caputure of a low-utility hint.

Medium-Utility Hint

Medium-utility hints cut down the search space. This type of hint turned over one card and highlighted four other cards, one of which was the match for the selected card. A popup box instructed the subject, "The match for the selected card is one of the cards currently highlighted." Subjects were required to dismiss the popup box by clicking on "OK" before they could continue playing the game. The cards remained highlighted until the subject clicked one of them. Figure 5.2 shows a screen capture of a medium-utility hint.

High-Utility Hint

High-utility hints were guaranteed to help the subject make a match. This type of hint turned over one card and highlighted its match in yellow. A popup box instructed the subject, "The match for the selected card is highlighted." Subjects were required to dismiss the popup box by clicking on "OK" before they could continue playing the game. The card remained highlighted until the subject clicked on it. Figure 5.3 shows a screen caputure of a high-utility hint.

5.1.3 Frequency of Interruption and Block Design

Our design used a total of eight replications of each of the three types of hints per condition, with an average interruption frequency of 80 seconds. We presented 12 interruptions per block in two 16 minute blocks. Thus, subjects in the interruption conditions (Match, Static, and Random) each saw a total of 24 interruptions. Hint order was randomized independently in each block for each subject. Interruption timing is explained in Section 4.4.3

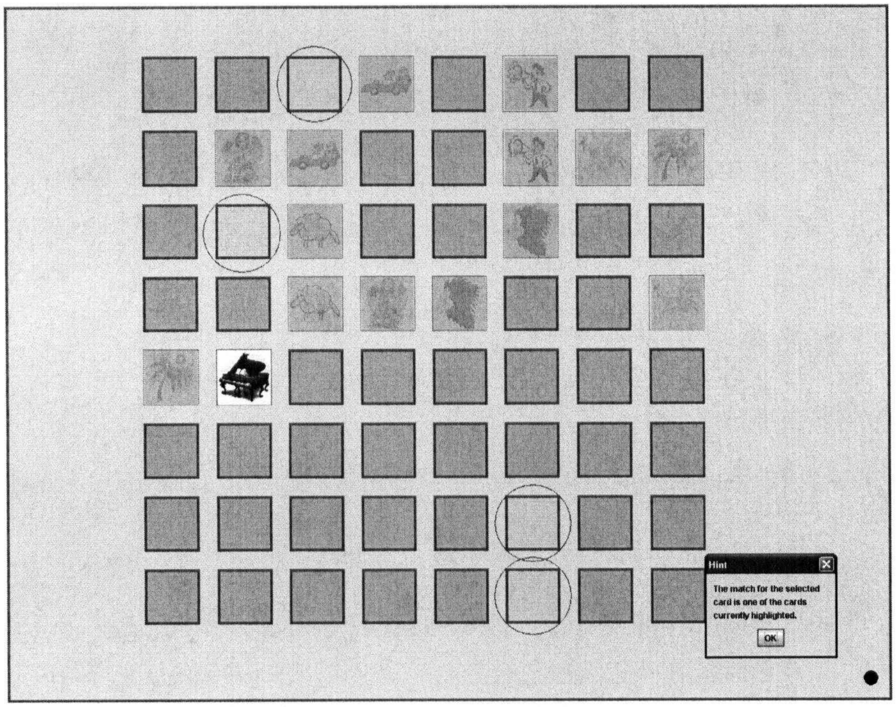
Similarly to Study 1, we included variation in interruption onset by including a number of "dummy" cases in which no signal was presented. There were four dummy slots overall (two per block). Thus, in 14% of the slots nothing happened.

The Control condition consisted of two 16 minute blocks with no interruptions.

5.1.4 Experimental Design

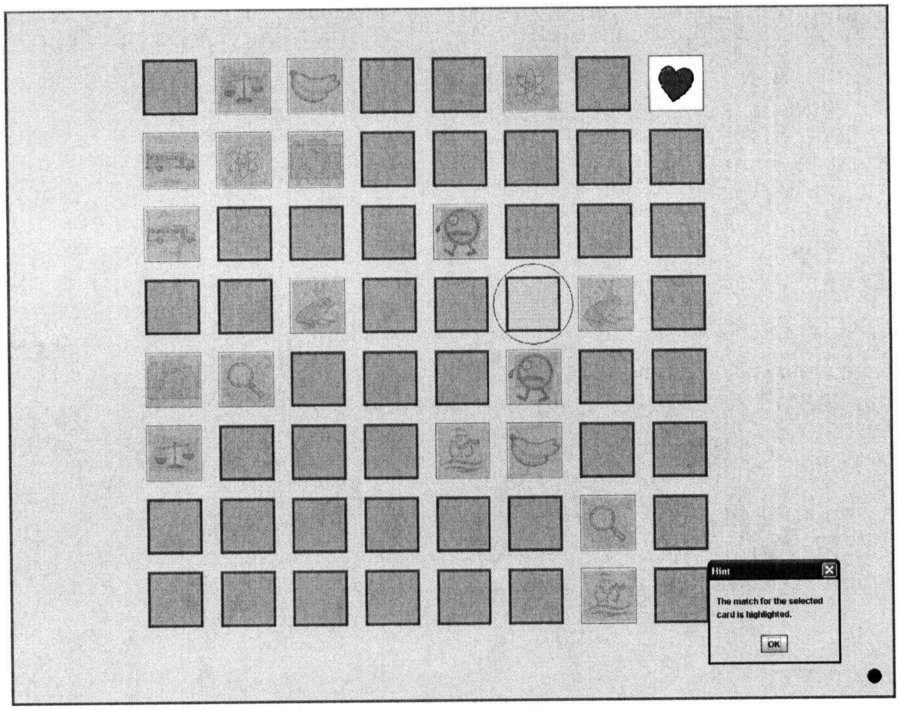
The experiment used a four level (level 1 = Match, level 2 = Static, level 3 = Random, level 4 = Control) between-subjects design, where levels 1, 2 and 3 were nested with three hint utilities, and levels 1 and 3 were also nested with three notification signals.

Initially, we had intended to use a within-subjects design in order to capitalize on its increased power and to allow for comparison amongst the three interruption schemes. In order to create a study short enough to avoid excessive subject fatigue (i.e., two



The match for the selected card is one of the cards currently highlighted.

Figure 5.2: Screenshot of a medium-utility hint. The piano in the middle-left portion of the screen is the currently selected card. The four cards highlighted in yellow are circled here for ease of viewing in greyscale. One of these is the matching piano card. The popup box reads, "The match for the selected card is one of the cards currently highlighted. The greyed-out cards have already been matched. All other cards have yet to be matched.



The match for the selected card is highlighted.

Figure 5.3: Screenshot of a high-utility hint. The heart in the top right-hand corner is the currently selected card. The card highlighted in yellow (circled) is the matching card. The popup box reads, "The match for the selected card is highlighted." The greyed-out cards have already been matched. All other cards have yet to be matched.

hours or less), however, we would have been limited to one block per condition. Piloting showed a number of problems with our intended design. First, fatigue was a major concern because pilot subjects experienced fatigue and rejuvenation at unexpected moments throughout the four blocks. Second, this design allowed for only four replications of each interruption utility per condition. Piloting revealed that this small number of replications did not allow subjects to detect the differences between the conditions. Yet, it was crucial to our research that subjects be able to differentiate between the Match and Random conditions. Furthermore, even with counterbalancing, this design risked the occurrence of negative transfer between conditions, e.g. subjects in the Random condition first might get so annoyed in this condition that they decide to ignore all notifications in subsequent conditions. For these reasons, we opted for a between-subjects design.

5.2 Participants

Forty subjects (26 female) between 18 and 39 years of age participated in the study and were compensated $10 for their participation. Thirty-eight were right-handed and all had normal colour vision. Subjects were recruited using the same online system as in Study 1.

5.3 Procedure

The experiment was designed to fit in a single one hour session. The procedure was as follows. (1) A questionnaire was used to obtain information on user demographics. (2) A training session ensured that each subject understood the Memory game interface (see Section 3.1.3). (3) A hint training block ensured that subjects in the Match, Static and Random conditions were familiar with all three notification signals and all three hint types. (4) Subjects performed two blocks of the same condition. At the end of both blocks, a dialogue box listed the total number of matches made. In the Match, Static and Random conditions the number of hints missed for each hint type was also displayed (see Figure 4.3). Subjects were given a 2-minute break in between the blocks. (5) After completing the second block, subjects filled out a survey that measured workload and fatigue. Annoyance and perceived benefit were included in the survey in the Match, Static and Random conditions. (6) A structured interview was conducted to understand subject perception of the notifications and hints, and strategies for their usage.

All the study instruments including questionnaires and interviews as well as the exact wording of the instructions given to participant can be found in Appedix C. The background questionnaire used to collect user demographics is given in Appendix A.

5.4 Measures

In addition to the measures outlined in Section 4.7, secondary measures were collected in a structured interview. We gathered subject perception of how the notifications and hints affected performance and documented strategies for their use. We also took this opportunity to determine if subjects in the Match and Random conditions comprehended any relationship between the notification signals and the hints.

5.5 Hypotheses

H1: Interruption annoyance is lower in the Match condition than in the Static and Random conditions.

H2: Perceived benefit is higher in the Match condition than in the Static and Random conditions.

H3: Workload in the Match condition is no different from, if not lower than, all other conditions.

H4: Performance is higher in the Match condition than in all other conditions.

H1 and **H2** are relevant only to the Match, Static and Random conditions. **H3** and **H4** concern all four conditions.

5.6 Results

Data for three outlier subjects was removed from the analysis (two from the Random condition and one from the Static condition). Outliers were subjects whose number of missed hints was more than two standard deviations from the mean. We defined outliers in this manner to ensure that subjects saw a sufficient number of interruptions to have "experienced"' the conditions. In the Static condition we counted the total number of hints missed, regardless of notification signal ($M = 6.63$). In the Match condition we considered only the number of high attentional draw hints ($M = 0.53$), because we anticipated that subjects who deciphered the signal-utility relationship might reasonably ignore low- and medium-utility hints. We considered outliers in the Random condition in the same manner as in the Match condition, in case subjects mistakenly assumed a signal-utility relationship and ignored low and medium attentional draw notifications.

Statistical adjustment strategies were identical to those employed in Study 1, and we again report effect sizes.

5.6.1 Annoyance and Benefit

To test H1 and H2, a one-way ANOVA with 3 levels (Random, Static, Match) was performed for the interruption annoyance and benefit ratings. Results for annoyance and benefit are illustrated in Figure 5.4 and Figure 5.5, respectively.

There was a statistically significant main effect of condition on annoyance ($F(2,24) = 3.903, p = .034, \eta^2 = .245$), where annoyance was significantly higher in the Random condition than in the Match condition ($p = .032$). Figure 5.4 shows that annoyance ratings in the Static condition fell in between ratings for Random and Match, but not with statistical significance. Compared to the Random and Static conditions, there was relatively little variance in the annoyance ratings for the Match condition.

No significant main effect of condition on perceived benefit was present ($F(2,24) = 2.202, p = .132, \eta^2 = .155$). However, the effect size was large while power was not high ($.405$). Trends are obvious in the boxplot (Figure 5.5): perceived benefit was very low in the Random condition, and seemed to be roughly the same in the Static and Match conditions. There was, however, greater variation in perception of benefit in the

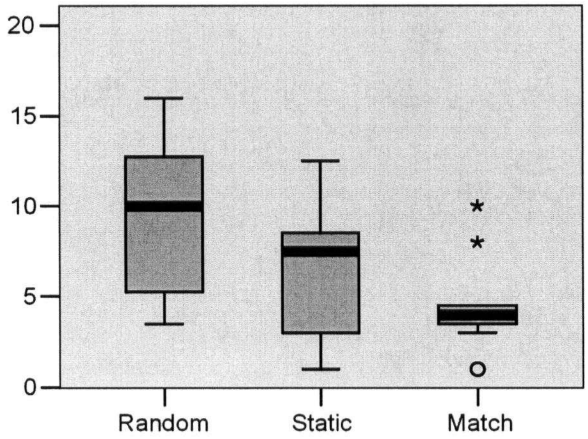


Figure 5.4: Boxplot of interruption annoyance rating by condition (scale: 1-20, where 20 indicates highest annoyance) ($N = 27$).

Static condition than in the other two conditions.

5.6.2 Workload

To test H3, a series of one-way ANOVAs with 4 levels (Control, Random, Static, Match) were performed on the workload measures. The ANOVA results are summarized in Table 5.1. There was a significant main effect of condition on temporal demand ($F(3,33) = 2.974, p = .046, \eta^2 = .213$), where temporal demand was greater in the Static condition than in the Control condition, with borderline significance ($p = .053$). Figure 5.6 shows the temporal demand results by condition. No other significant effects were found. However, the effect size for mental demand was large ($.172$) while power was not high ($.526$). The boxplot in Figure 5.7 indicates that mental demand may have been lower in the Control condition than in the interrupting conditions.

5.6.3 Performance

To test H4, a 4 (condition: Control, Random, Static, Match) by 2 (block) ANOVA was calculated for the number of matches found. A significant main effect of block

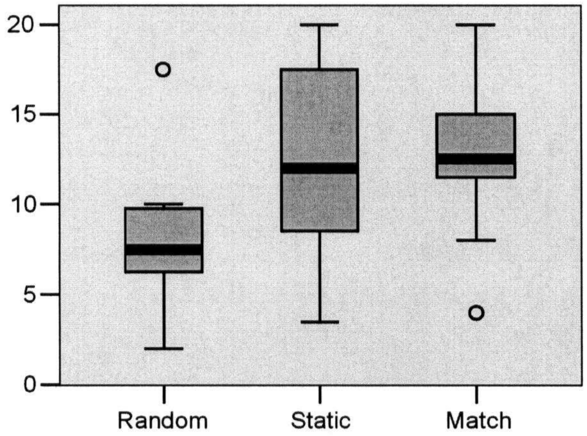


Figure 5.5: Boxplot of perceived benefit rating by condition (scale: 1-20, where 20 indicates highest benefit) ($N = 27$).

NASA-TLX Factor	**F(3,33)**	**p**	η^2	**power**
Mental Demand	2.290	.097	.172	.526
Temporal Demand	2.974	.046*	.213	.650
Effort	1.242	.310	.101	.301
Perceived Performance	.027	.994	.003	.054
Frustration	.607	.615	.052	.162

Table 5.1: Results of ANOVA on NASA-TLX workload measures ($N = 37$).

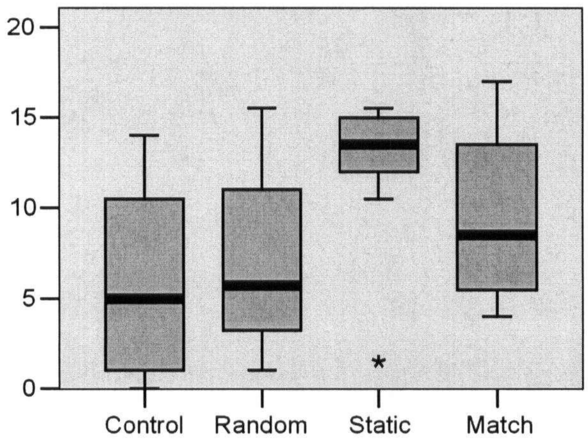


Figure 5.6: Boxplot of temporal demand rating, by condition (scale: 1-20, where 20 indicates highest temporal demand) ($N = 37$).

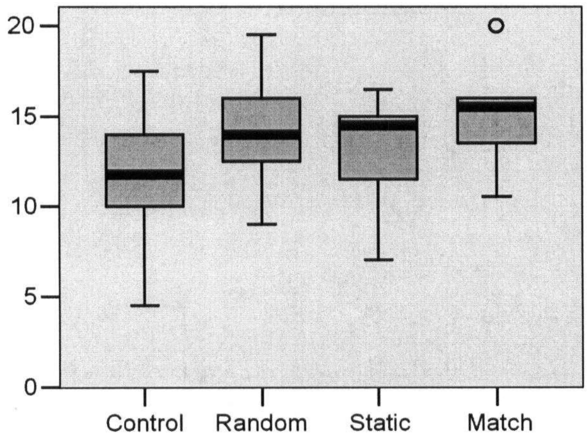


Figure 5.7: Boxplot of mental demand rating by condition (scale: 1-20, where 20 indicates highest mental demand) ($N = 37$).

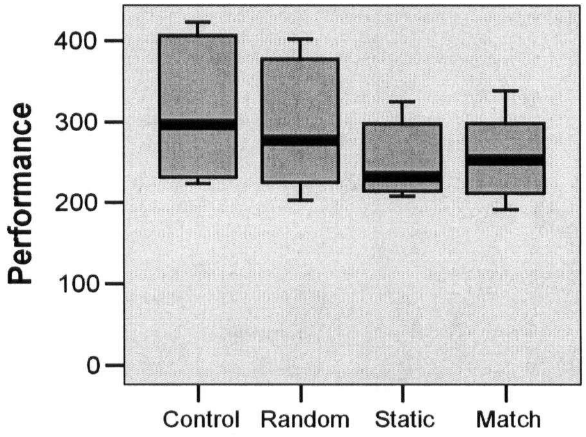


Figure 5.8: Boxplot of performance by condition ($N = 37$). Performance is measured as the number of matches made.

($F(1,33) = 51.807, p < .001, \eta^2 = .611$) indicated that a learning effect was present. No significant main effect of condition was present ($F(1,3) = 1.750, p = .176, \eta^2 = .137$). However, effect size was large while power was low ($.414$). The boxplot in Figure 5.8 indicates that performance may have been highest in the Control condition. The boxplot also reveals that variance in performance was quite high.

5.6.4 Qualitative Feedback

Written and verbal feedback on the low-utility hints helped to elucidate the annoyance and benefit ratings and indicate a greater disparity in perception between the Static and Match conditions. Table 5.2 summarizes two frequent sentiments expressed in the surveys and informal interviews. When asked what aspects of the notifications and hints annoyed subjects, 63% of subjects in the Random condition indicated that they were annoyed by the low-utility hints. In contrast, only 30% of subjects in the Match condition complained about the low-utility hints, and all 30% qualified the annoyance of these hints as being only "a little bit" annoying. Subjects in the Static condition were in the middle, with 44% of subjects annoyed by the low-utility hints.

Unsurprisingly, detrimental effects of the low-utility interruptions were felt most

Sentiment	Random	Static	Match
Low-utility hints were annoying	63%	44%	30%
Low-utility hints wasted time and hindered performance	88%	33%	20%

Table 5.2: Summary of self-reported annoyance sentiments ($N = 27$).

strongly in the Random condition (88%). In the Static condition, 33% of subjects complained that the low-utility hints wasted time and hindered performance. Of the 20% of Match subjects who had the same complaint, 10% indicated that the low-utility hints were only "a little bit of a waste of time." The other 10% ignored the low-utility hints after ascertaining the relationship between utility and notification signal, and so the detriment desisted.

Discussion with many of the subjects during the interview revealed that the medium-utility hints were more disruptive than we had intended. Many subjects had attempted to memorize all four potential match locations during this type of hint. Subjects recalled that this endeavour had broken their concentration and they forgot what they had been working on before the interruption occurred. Overall, 30% of subjects complained that all hints in general broke their concentration and interfered with memory. An unexpected reversal, some subjects were glad to see encouragement from the low-utility text messages.

Comprehension of the Attentional Draw-Utility Relationship

Of the 10 subjects who saw the Match condition, only three comprehended the relationship between the notification signals and the hint utilities. All three utilized this knowledge to ignore the low-utility hints. None of the subjects in the Random condition incorrectly surmised a relationship between the notification signals and the hints.

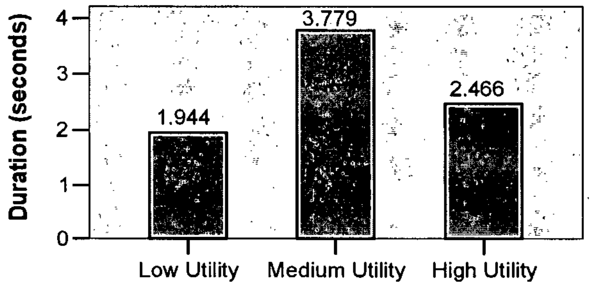


Figure 5.9: Hint duration times by utility ($N = 27$).

5.6.5 Secondary Quantitative Measures

Hint Duration

We examined hint duration in order to understand how long subjects took to process the interruption content. Hint duration was measured as the time between when a subject clicked on a notification signal and when the subject clicked on the first card after dismissing the hint popup box. Mean hint duration times by utility are displayed in Figure 5.9.

A 3 (utility) by 3 (condition: Random, Static, Match) by 2 (block) repeated measures ANOVA was run to investigate the amount of time subjects spent on the different types of hints. There was no main or interaction effect of condition. A significant main effect of utility ($F(2,42) = 67.948, p < .001, \eta^2 = .764$) showed that duration times were the highest for the medium-utility hint ($p < .001$ compared to both low and high utility), and lowest for the low-utility hint ($p = .002$ compared to high-utility). A significant main effect of block ($F(1,21) = 73.216, p < .001, \eta^2 = .777$) where duration times were lower in the second block showed that, unsurprisingly, subjects learned to deal with hints more efficiently by the second block.

Detection Time and Timeouts

We verified that our set of notification signals was still significantly different in terms of attentional draw. Mean detection times and timeout for each signal are presented

Notification Signal	Mean Detect Time (s)	Mean Timeouts
Flag	14.92	57.5%
Slow Zoom	8.74	28.6%
Follow	2.70	1.9%

Table 5.3: Detection times and timeout rates, by signal, for Match and Random conditions ($N = 18$).

in Table 5.3. We ran a 3 (notification signal) by 2 (block) by 2 (condition: Random, Match) repeated measures ANOVA on detection times and timeouts for the two conditions that used all three signals. As expected, there was a main effect of signal on both detection times ($F(1,11) = 70.992, p < .001, \eta^2 = .866$), and timeouts ($F(2,32) = 41.146, p < .001, \eta^2 = .720$). Differences were significant for all post-hoc pairwise comparisons for both measures. There were no main or interaction effects for block or condition. This confirms the findings from Study 1 that the three notification signals lay along the spectrum of attentional draw.

We also investigated detection times across the three interruption conditions for the medium AD signal only, in order to determine if seeing three types of signals versus only one type of signal affected detection of this signal. For instance, subjects may have polled more in the Match condition in order to detect the low AD signal and thus may have detected the medium AD signal more efficiently as a result. We examined detection of the medium AD signal across all three interruption conditions using a 3 (condition: Random, Match, Static) by 2 (block) ANOVA on detection times and timeouts. There were no main or interaction effects for either condition or block on detection times. Thus, subjects who saw only the medium AD signal in the Static condition did not detect it differently than subjects who saw all three types of signals in both the Match and Random conditions. There was a significant main effect of block on number of timeouts ($F(1,24) = 6.567, p = .017, \eta^2 = .215$), where there were fewer timeouts in the second block.

5.6.6 Summary of Results

H1 not supported. Interruption annoyance was lower in the Match condition than in the Random condition, but was no different from the Static condition.

H2 not supported. Perceived benefit was not significantly higher in the Match condition than in the Random and Static conditions.

H3 supported. Workload in the Match condition did not differ significantly from the other conditions.

H4 not supported. Performance did not differ significantly across the three conditions.

5.7 Discussion

Unfortunately, the study design had lower statistical power than we had anticipated. This was the result of combining a between-subjects design, reliance on subjective measures, and a task that had relatively high variance in individual performance. However, the results did reveal a number of interesting trends.

As expected, annoyance levels were significantly higher in the Random condition than in the Match condition. Although there was no statistically significant difference in annoyance between the Match and Static conditions, trends in both annoyance ratings and qualitative interview feedback indicate that, with more power, we would likely have seen such a difference. Low power was also a problem in the analysis of perceived benefit. Again, however, trends indicated that benefit was lowest in the Random condition. Further investigation of the Match and Static conditions was certainly merited.

These results were enough to convince that, in terms of annoyance and perceived benefit, the Random condition was significantly worse than the status quo of static notification. Removing this condition from consideration opened the possibility of a more powerful within-subjects design, which was used in Study 3 to investigate the remaining three conditions.

Results of task performance and qualitative feedback indicated that we did not succeed in creating beneficial hints. Without performance-boosting interruption content, we could not expect the workload detriment generally associated with interruption to be mitigated. This explains why trends pointed to higher mental and temporal demand in the interrupting conditions than in the control condition. This lack of actual benefit might also have interacted with our results for perceived benefit. Thus, further investigation necessitated a reformulation of the high and medium-utility hints to ensure that their content offered benefit to subjects by increasing performance. Hint-duration results and interview feedback indicated that the medium-utility hints were far more disruptive than had been intended, while low-utility hints were not disruptive enough. We were very careful to avoid repeating these mistakes when redesigning the hints for Study 3.

Chapter 6

Study 3

6.1 Methodology

The methodology used for this study is based on the core experimental approach documented in Chapter 4. Only additions and clarifications are highlighted here.

6.1.1 Conditions

In this study we investigated only three of the conditions defined in Chapter 4: Match, Static, and Control.

6.1.2 Hints and Utility

Hints were redesigned in this study according to the following objectives:

- High-utility hint
  - offers performance boost
  - always helpful
- Medium-utility hint
  - less disruptive than in Study 2
  - sometimes helpful, sometimes not
- Low-utility hint
  - more disruptive than in Study 2
  - never helpful

Based on subject performance in Study 2, we expected an average performance boost of 15% (approximately 22 extra matches per session) if subjects looked at all hints.

Low-Utility Hint

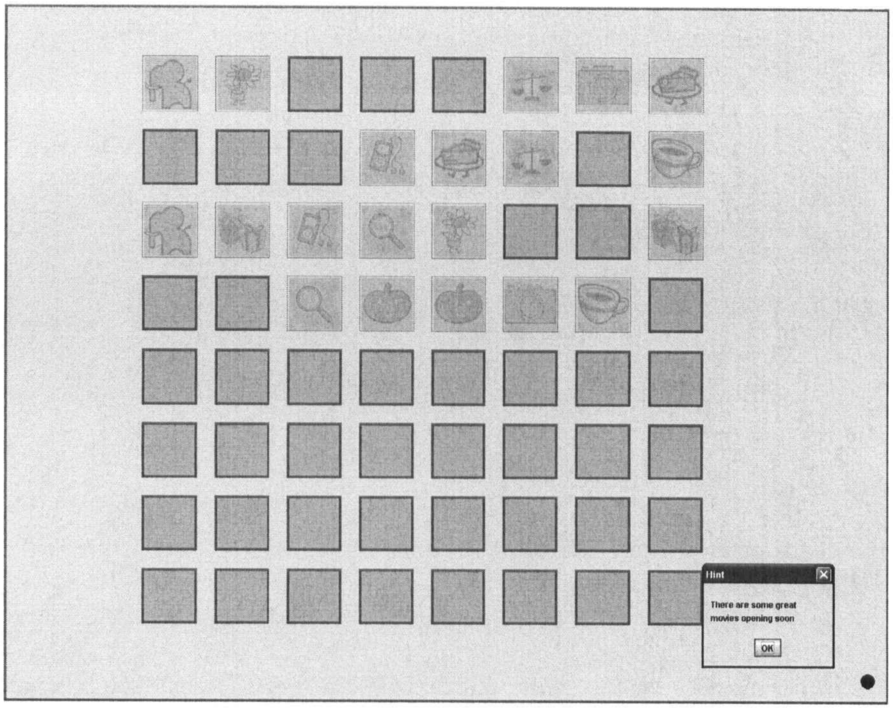
Hints with the lowest utility were not relevant to the game play and thus did not provide the user with any assistance in finding a match. Instead, this type of hint always showed a text message unrelated to the game using a pop-up box. Subjects were required to dismiss the popup box by clicking on "OK" before they could continue playing the game. In order to force subjects to spend more time on the irrelevant hints than in Study 2, a series of different text messages were used throughout each block. The messages were as follows:

- Nice weather we're having.
- Traffic was terrible this morning.
- Am I interrupting?
- There are some great movies opening soon.
- Stock prices are up.
- Did you watch the game last night?

Figure 6.1 shows a screen caputure of a low-utility hint.

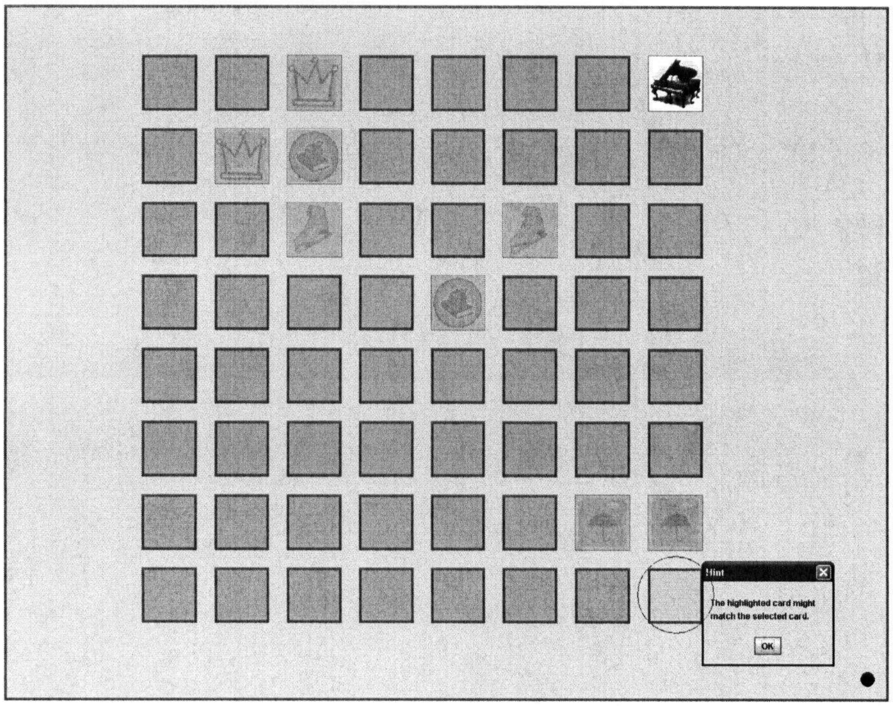
Medium-Utility Hint

A medium-utility hint turned over one card and highlighted a second card in yellow; 40% of the time, the highlighted card was the match for the selected card, while 60% of the time it was not. A popup box instructed the subject, "The highlighted card might match the selected card." Subjects were required to dismiss the popup box by clicking on "OK" before they could continue playing the game, and the card remained highlighted until the subject clicked on it.



There are some great movies opening soon

Figure 6.1: Screenshot of a low-utility hint. The popup box reads, "There are some great movies opening soon.



The highlighted card might match the selected card.

Figure 6.2: Screenshot of a medium-utility hint. The piano in the top right-hand corner is the currently selected card. The card highlighted in yellow (circled) in the bottom right-hand corner has a 40% chance of being of being the matching piano card. The popup box reads, "The highlighted card might match the selected card." The greyed-out cards have already been matched. All other cards have yet to be matched.

This type of hint was designed to be "somewhat helpful" and needed to be appreciably different from the high-utility hint. High-utility hints always helped while low-utility hints never helped. Had the medium-utility hints always helped, two thirds of the interruptions overall would have been helpful. We believe that is rare for a real life interruption system to be this pertinent. Our initial intention was to make this hint helpful 50% of the time; however, our use of an odd number of replications did not allow this. Thus, a 40/60 split was used. Figure 6.2 shows a screen capture of a medium-utility hint.

High-Utility Hint

A high-utility hint showed the location of five matches by highlighting 10 cards, using different colours to indicate the matched pairs. A popup box instructed the subject, "Five matches are highlighted." Subjects were required to dismiss the popup box by clicking on "OK" before they could continue playing the game. The cards remained highlighted until the subject clicked on each of them to uncover the five matches.

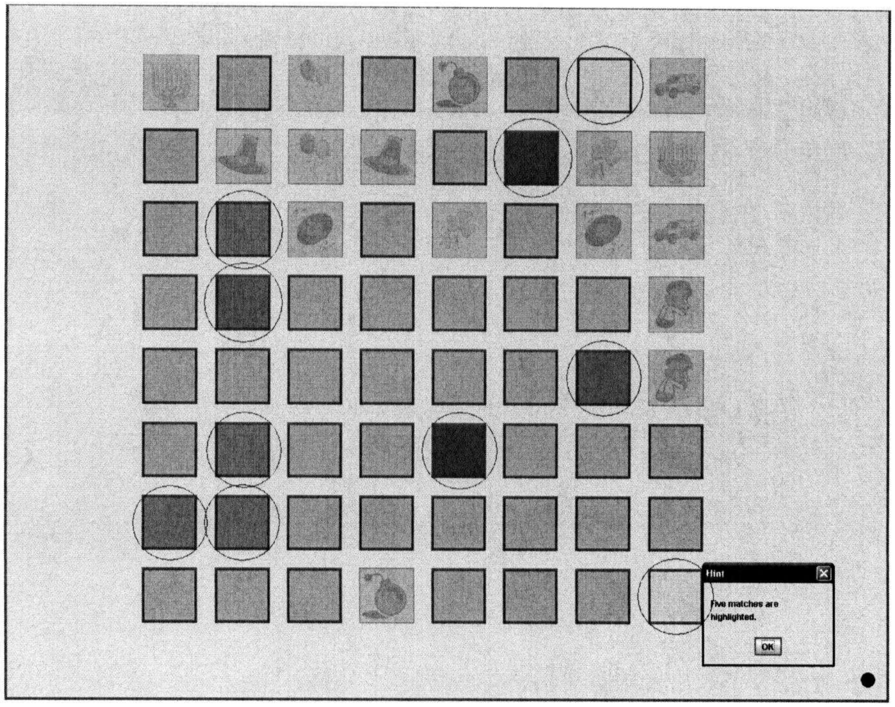
If there were fewer than 10 cards left on the board during this type of hint, the highlighted matches would carry over to the next board, i.e., if there were four cards left on the board, the subject saw hints for the last two pairs on the current board, and there were hints for three more matches on the new board after it reset. Figure 6.3 shows a screen capture of a high-utility-hint.

6.1.3 Frequency of Interruption and Block Design

Our design used five replications of each of the three types of hints with an average interruption frequency of 65 seconds. The 15 interruptions were presented in a 17 minute block, and hint order was randomized independently for each subject. Accordingly, the Control condition lasted 17 minutes but had no interruptions. Interruption timing is explained in Section 4.4.3

6.1.4 Experimental Design

The experiment used a 3 level (level 1 = Match, level 2 = Static, level 3 = Control) within-subjects design, where levels 1 and 2 were nested with three hint utilities, and level 1 was also nested with three notification signals. A within-subjects design was chosen for its increased power and because it allowed for comparative comments on the two interruption schemes. Order of condition presentation was fully counterbalanced. There were six orders of presentation of the conditions, a between-subjects control variable introduced to minimize order effects.



Five matches are highlighted.

Figure 6.3: Screenshot of a high-utility hint. Ten cards (circled) are highlighted using five different colours. The popup box reads, "Five matches are highlighted." The greyed-out cards have already been matched. All other cards have yet to be matched.

6.2 Participants

Twenty-four subjects (15 female) between 18 and 39 years of age participated in the study and were compensated $20 for their participation. Twenty-three were right-handed and all had normal colour vision. Subjects were recruited using the same online system as in Study 1, as well as through advertisements posted throughout the university campus.

6.3 Procedure

The experiment was designed to fit in a single two-hour session. The procedure was as follows. (1) A questionnaire was used to obtain information on user demographics. (2) A training session ensured that subjects understood the Memory game interface (see Section 3.1.3). (3) A hint training block ensured that each subject was familiar with all three notification signals and all three hint types. (4) Subjects performed each of the three conditions. At the end of each condition, a dialogue box listed the total number of matches made. In the Match and Static conditions the number of hints missed for each hint type was also displayed (see Figure 4.3). (5) After each condition, subjects filled out a survey that measured workload and fatigue in all conditions, as well as annoyance and perceived benefit in the Match and Static conditions. Six-minute breaks were given following the survey in the first two conditions. (6) A structured interview was conducted to collect condition preferences, as well as to understand subject perception of the notification and hints and strategies for their usage.

All the study instruments including questionnaires and interviews as well as the exact wording of the instructions given to participants can be found in Appedix D. The background questionnaire used to collect user demographics is given in Appendix A.

6.4 Measures

In addition to the measures outlined in Section 4.6, we conducted a structured interview where subjects rank ordered all three conditions according to overall preference.

Subjects were also asked if the hints were equally helpful in both the Match and Static conditions, or if one condition was more helpful than the other. Similarly, we asked if the hints hindered performance equally in both interruption conditions, or if there was greater hindrance in one or the other. We also documented subject perception of the notifications and hints, and strategies of their use.

6.5 Hypotheses

H1: Interruption annoyance is lower in the Match condition than in the Static condition.

H2: Perceived benefit is higher in the Match condition than in the Static condition.

H3: Workload in the Match condition is no different from, if not lower than, the other two conditions.

H4: Performance is higher in the Match condition than in the other two conditions.

H1 and **H2** are relevant only to the Match and Static conditions. **H3** and **H4** concern all three conditions.

6.6 Results

Data for four outlier subjects were removed from the analysis. As in Study 2, outliers were subjects whose number of missed hints was more than two standard deviations from the mean in either of the two interruption conditions. We defined outliers in this manner to ensure that subjects saw a sufficient number of interruptions to be able to perceive a difference between the two interruption conditions. In the Static condition we counted the total number of hints missed ($M = 2.92$), regardless of utility. In the Match condition we considered only the number of high-utility hints ($M = 0.33$), because we anticipated that subjects who deciphered the signal-utility relationship might reasonably ignore low- and medium-utility hints.

Dependent Variable	Match	Static
Interruption Annoyance	28.914	40.707
Perceived Benefit	71.288	59.722

Table 6.1: Mean ratings for interruption annoyance and benefit (scale: 5-100, where 100 indicates highest annoyance or benefit) ($N = 20$).

Statistical adjustment strategies were identical to those employed in Study 1 and Study 2, and we again report effect sizes.

6.6.1 Annoyance and Benefit

To test H1 and H2, a 2 (condition: Match, Static) by 2 (presentation order)[5] ANOVA was performed for annoyance and benefit ratings. Results for these ratings are summarized in Table 6.1. As hypothesized, annoyance was significantly lower in the Match condition than in the Static condition ($F(1,18) = 5.239, p = .034, \eta^2 = .225$). Likewise, perceived benefit was significantly higher in the Match condition than in the Static condition ($F(1,18) = 5.074, p = .037, \eta^2 = .220$). No effect of presentation order was found.

In addition to interruption annoyance, we also examined general annoyance across all three conditions. Mean general annoyance ratings are presented in Figure 6.4. To examine these ratings across all three conditions, a 3 (condition: Match, Static, Control) by 6 (presentation order) ANOVA was performed. A main effect of condition ($F(2,28) = 2.788, p = .079, \eta^2 = .166$) approached significance with a large effect size. These results suggest that the Static condition may have been more annoying than the Match condition in terms of general annoyance as well as annoyance specific to the interruptions. Furthermore, general annoyance did not differ largely between the Con-

[5] When comparing only the Match and Static conditions, we present ANOVA results based on two presentation orders: (1) Match before Static and (2) Static before Match. We also performed this analysis based on all six presentation orders in case the order of the Control condition affected the annoyance and benefit ratings. Values were slightly different, but there were no differences in whether or not the results reached significance for any of the analyses.

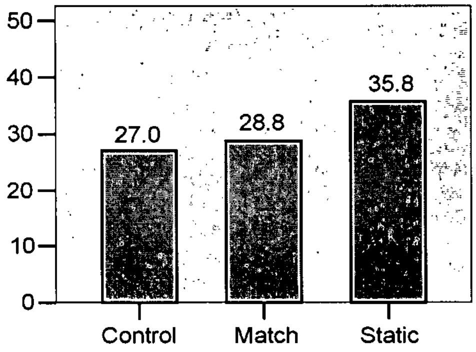


Figure 6.4: Mean general annoyance ratings by condition (scale: 5-100, where 100 indicates highest annoyance) ($N = 20$). Note the scale on this figure is from 0 to 50.

trol and Match conditions, despite consensus in the literature that interruptions cause undue annoyance (e.g., [7]).

6.6.2 Workload and Performance

H3 and H4 pertained to all three conditions. To test these hypotheses, a 3 (condition: Match, Static, Control) by 6 (presentation order) ANOVA was performed for workload measures and performance.

Workload

Results for the NASA-TLX workload measures are summarized in Table 6.2. There were no significant differences among the three conditions for any of the NASA-TLX workload measures, and no effect of order was present. This was consistent with our hypothesis H3 in which we speculated that workload would be no worse in the Match condition than in the other conditions.

Performance

Performance results are presented in Figure 6.5. Counter to our hypothesis H4, there was no significant effect of condition on performance ($F(2,28) = .812, p = .454, \eta^2 =$

NASA-TLX Factor	**F(2,28)**	**p**	η^2
Mental Demand	.057	.945	.004
Physical Demand	2.335	.115	.143
Temporal Demand	1.069	.357	.071
Effort	.118	.889	.008
Perceived Performance	1.347	.276	.088
Frustration	.381	.687	.027

Table 6.2: Results of ANOVA on NASA-TLX workload measures ($N = 20$).

.055). However, a main effect of presentation order ($F(5,14) = 2.720, p = .064, \eta^2 = .493$) and an interaction effect of condition and presentation order ($F(10,28) = 2.035, p = .068, \eta^2 = .421$) both approached significance with large effect sizes. These effects are illustrated in Figure 6.6 and Figure 6.7, respectively. However, because individual differences were large and there was sparse data in each cell of the design, no clear trends were evident, as can be seen in the graphs.

Not unexpectedly, we observed a borderline significant learning effect on performance ($F(2,38) = 3.171, p = .053, \eta^2 = .143$). There was also a significant effect of block on the self-reported fatigue measure ($F(2,38) = 5.327, p = .009, \eta^2 = .219$), where subjects were more fatigued in the third block than the first ($p = .017$).

6.6.3 Interview Results

Preference, helpfulness and hindrance

We calculated the Chi-square statistic for preference, helpfulness, and hindrance responses to determine if actual frequencies were significantly different from the case in which all frequencies are equal. A summary of the results is shown in Table 6.3. Chi-square was significant for all of the measures. Consistent with our annoyance and benefit findings, the majority of subjects preferred the Match condition, finding it to be more helpful than the Static condition. The majority of subjects also found that interruptions in the Static condition hindered performance more than interruptions in the Match condition.

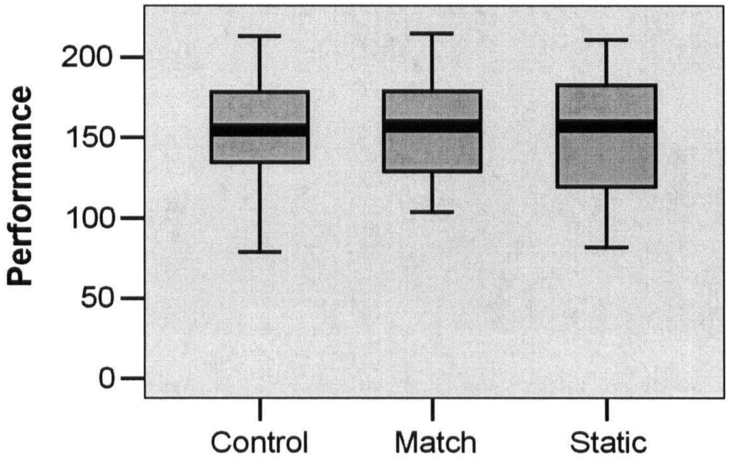


Figure 6.5: Boxplot of performance by condition ($N = 20$). Performance is measured as the number of matches made.

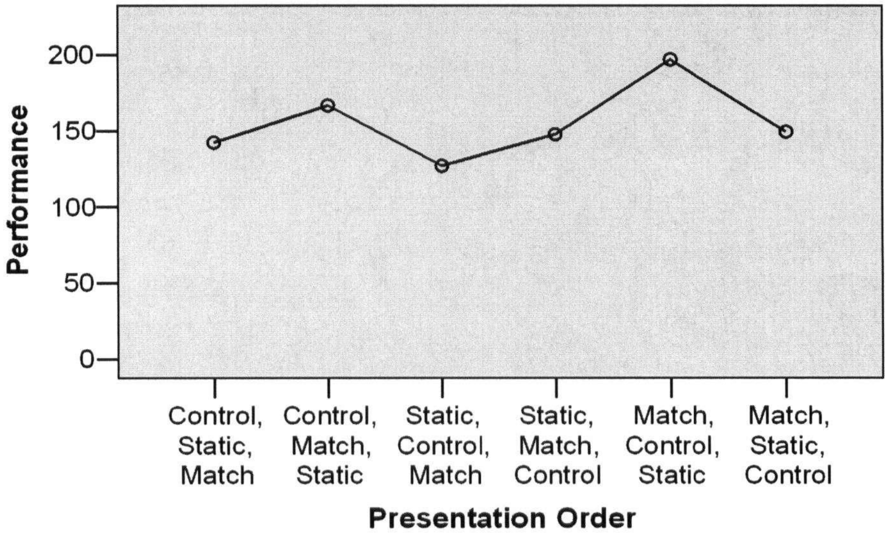


Figure 6.6: Performance by presentation order ($N = 20$).

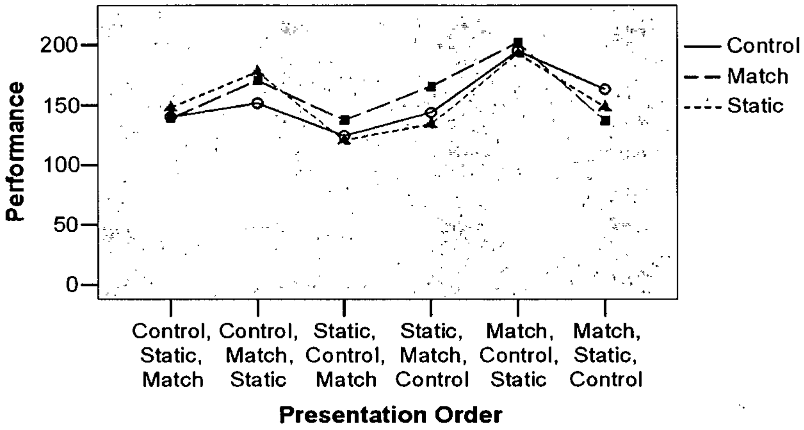


Figure 6.7: Interaction effect of condition and presentation order on performance ($N =$ 20).

Dependent Variable	Match	Static	Control	df	Chi-square	p
Preferred overall	15	3	2	2	15.70*	<.001
Dependent Variable	**Match**	**Static**	**Equal for Match & Static**	**df**	**Chi-square**	**p**
More helpful	12	3	5	2	6.70*	.035
More hindering	2	11	7	2	6.10*	.047

Table 6.3: Chi-square statistic for qualitative results. The top dependent variable is compared across all three conditions. The bottom two dependent variables was compared only across the Match and Static conditions ($N = 20$).

Comprehension of the Attentional Draw-Utility Relationship

In terms of understanding the relationship between the hints and the degree of AD in the Match condition, the interviews revealed that 25% of subjects made no comprehension of the relationship. The relationship between the high AD notification signal and the high-utility hints was comprehended by 45% of subjects, while 40% of subjects comprehended the "medium" relationship, and 70% of subjects comprehended the "low" relationship. Overall, 40% of subjects understood all three relationships and all of these subjects preferred the Match condition. In terms of strategies of hint usage, 40% of subjects utilized their relationship knowledge to ignore low-utility hints. This type of learned behaviour was anticipated.

Perception of hints and notification signals

All subjects perceived the high-utility hints to be helpful, while 30% thought the medium-utility hints were helpful, and no subjects found the low-utility hints to be helpful. In terms of hindrance, 80% of subjects responded that the low-utility hints hindered performance, 50% of subjects said that the medium-utility hints hindered, and only one subject (5%) thought that the high-utility hints hindered performance. These results indicate that, in contrast to Study 2, subjects did perceive the hints along the intended spectrum of utility.

Furthermore, the interruption conditions shaped how subjects perceived the different types of hints. In surveys distributed following each condition, we asked what aspects of the notifications and hints annoyed subjects during that condition. In the Static condition, 85% of subjects indicated that they were annoyed by the low-utility hints. In the Match condition, only 60% of subjects admitted to being annoyed by the low-utility hints. This included 20% who stated that annoyance associated with low-utility hints lowered significantly - if not ceased - once they began purposely to ignore these hints.

The matching of AD and utility also seemed to colour subject perception of the notification signals. After the structured portion of the interview, subjects were asked if they had any additional thoughts they wanted to share about the three signals and 65%

of subjects volunteered comments involving affective perception of the signals. These comments revealed a positive perception of the high-utility (FOLLOW) notification signal: 35% of subjects spontaneously remarked that that they "liked" or "loved" the signal, noting that it was "hard to miss," because, "you didn't have to look away from what you were doing." Astute subjects (10%) mentioned that they were glad this signal was associated with the high-utility hint because it was the easiest to see.

The low-utility (FLAG) notification signal was received less favourably: 30% of subjects complained that it was "hard to see without looking [directly] at it," and that was a "bad thing if you want[ed] to notice the hints." These complaints were voiced by subjects who either did not comprehend the relationship between utility and AD (15%), or who did comprehend the relationship but continued to monitor and view the low AD signal because they did not completely trust the perceived correlation (15%). On the other hand, another subset of subjects (15%) - those who comprehended and trusted the relationship - appreciated the subtlety of the FLAG signal because it was easy to ignore. The remaining 55% gave no opinion about FLAG.

The advantages of coordinating AD and utility were best summarized by two subjects. One said of the high-utility signal, "If [the hint] is useful, it's better that it's presented like this, but I wouldn't want to get the [low-utility hint] this way." Another subject remarked that the low-utility signal was "least able to pull my attention away from where it was, which was fine because they [sic] seemed to correlate with the least useful hints, [and] so I allowed myself to ignore it."

6.6.4 Secondary Quantitative Measures

Detection Time and Timeouts

Mean notification signal detection times and mean timeout rates by signal are illustrated in Figure 6.8 and Figure 6.9, respectively. We separated the medium-AD signal (SLOW ZOOM) between the Static and Match conditions because subjects may have responded to these signals differently depending on whether or not there was a relationship between AD and utility. To ensure that our notification signals were still significantly different in terms of attentional draw, a 4 (notification signal: low AD, medium

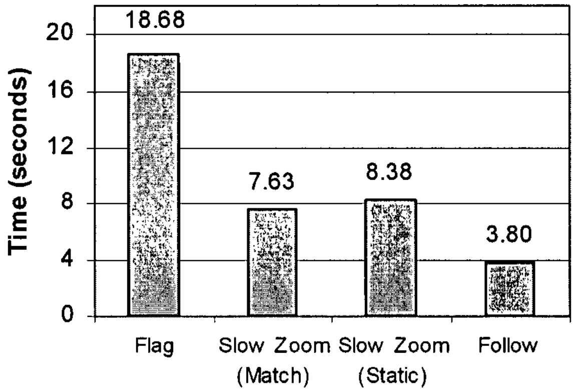


Figure 6.8: Signal detection times ($N = 20$).

Signal (i)	**Signal (j)**	**Mean Diff. (i − j)**	**Std. Error**	**p**
Flag	SZ (Match)	11.047*	1.794	<.001
Flag	SZ (Static)	10.297*	2.325	.002
Flag	Follow	14.878*	2.088	<.001
SZ (Match)	Follow	3.831*	1.172	.026
SZ (Static)	Follow	4.580*	1.468	.036
SZ (Match)	SZ (Static)	-0.749	1.353	1.00

Table 6.4: Pairwise comparisons for detection times ($N = 20$).

AD (Static), medium AD (Match), high AD) by 2 (order of presentation) ANOVA was performed on detection times and timeout rates. There was a significant main effect of notification signal on detection time ($F(2.056, 37.005) = 26.473, p < .001, \eta^2 = .595$), where detection times for the low, medium, and high signals were all statistically significantly different, but there was no difference between the medium signal in the Match and Static conditions. This is consistent with our findings from Study 1 and Study 2. Table 6.4 summarizes the post-hoc pairwise comparisons.

There was also a significant main effect of notification signal on timeout rate ($F(1.609, 28.957) = 16.008, p < .001, \eta^2 = .471$), where the FLAG signal had more

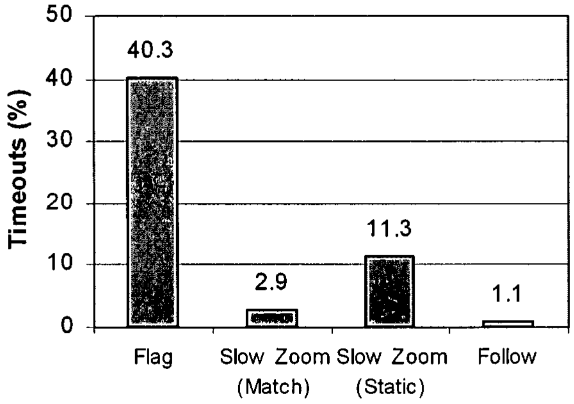


Figure 6.9: Mean timeout rates, by signal ($N = 20$). Note the scale is from 0 to 50%.

timeouts than all other signals ($p = .001$, $p = .031$, $p < .001$, compared to SLOW ZOOM Match, SLOW ZOOM Static, and FOLLOW respectively). The high timeout rate for FLAG (40%) was not unexpected since many subjects purposely ignored the signal (see Section 6.6.3). There was no main or interaction effect of presentation order on detection time or timeout rate.

Hint Duration

As in Study 2, we examined hint duration in order to understand how long subjects took to process the interruption content. Hint duration was again measured as the time between when a subject clicked on a notification signal and when the subject clicked on the first card after dismissing the hint popup box. Mean detection times by hint utility are displayed in Figure 6.10. We ran a 3 (utility) by 2 (condition: Match, Static) by 2 (presentation order) ANOVA to ensure that medium-utility hints were not disproportionately disruptive, as in Study 2. There was a significant main effect of utility ($F(2,36) = 6.839, p = .003, \eta^2 = .275$), where subjects spent less time on low-utility hints than on medium- ($p = .036$) or high- ($p = .026$) utility hints. Note, however, that the difference was on the order of only 300ms. The fact that the medium-utility hints did not take more time to deal with than the high-utility hints, combined with

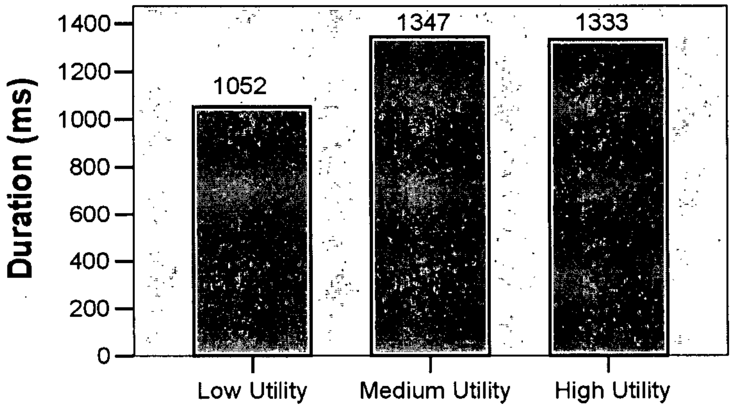


Figure 6.10: Hint duration times by utility ($N = 20$).

the results on qualitative perception of hint utility (see Section 6.6.3) indicates that our intended utility scale was achieved in this study.

Statistical analysis also revealed two interaction effects: one between condition and presentation order ($F(1,18) = 7.137, p = .016, \eta^2 = .284$), and another between condition and utility ($F(2,36) = 9.959, p < .001, \eta^2 = .356$). Figure 6.11 shows a graphical representation of the interaction between condition and presentation order. The graph suggests that hint duration was lower in the second condition the user saw, regardless of which condition it was. However, paired-samples t-tests reveal that the difference (i.e., hint duration was lower in the second condition) was statistically significant only in the Match-Static order ($p = .002$).

Figure 6.12 shows a graphical representation of the interaction between condition and utility. Paired-samples t-tests confirm that subjects dealt with low-utility hints more efficiently in the Match condition than in the Static condition (842 ms versus 1257 ms, $p = .022$), and with high-utility hints more efficiently in the Static condition than in the Match condition (1117 ms versus 1257 ms, $p = .002$). The difference in low-utility hint duration likely had to do with subject knowledge of the relationship between notification signal and utility: most subjects knew that the hint would be a text message, and so didn't bother reading the text message before dismissing the popup.

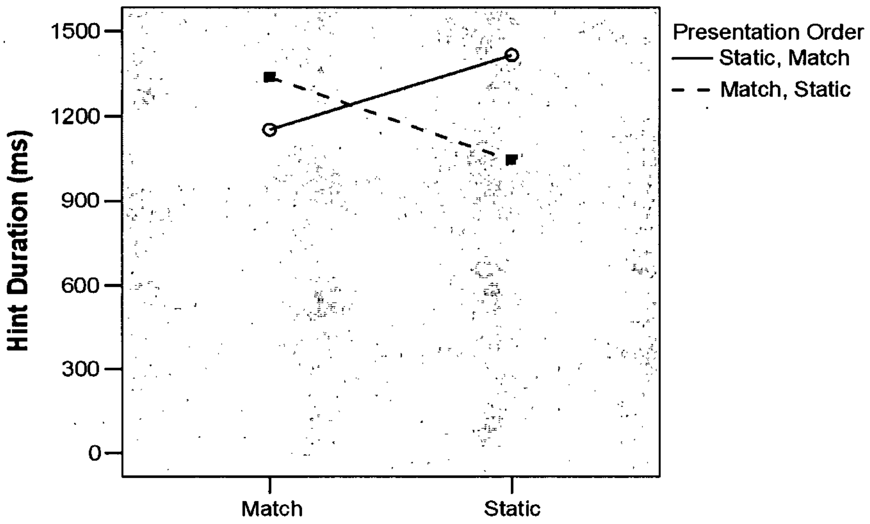


Figure 6.11: Interaction effect of condition and presentation order on hint duration ($N = 20$).

We cannot account for the duration differences for the high-utility hint.

6.6.5 Summary of Results

H1 supported. Interruption annoyance was lower in the Match condition than in the Static condition.

H2 supported. Perceived benefit was higher in the Match condition than in the Static condition.

H3 supported. Workload did not differ significantly across the three conditions.

H4 not supported. Performance did not differ significantly across the three conditions.

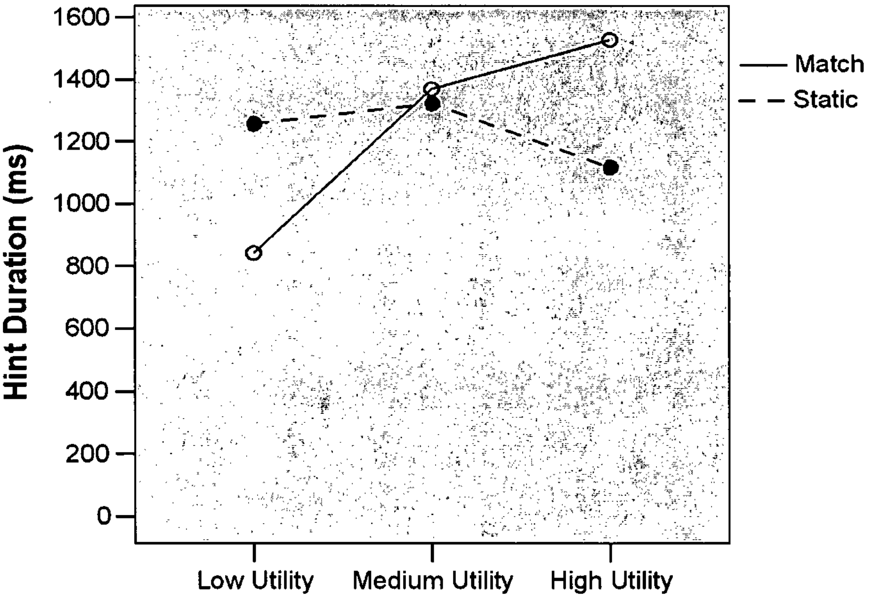


Figure 6.12: Interaction effect of condition and utility on hint duration ($N = 20$).

6.7 Discussion

6.7.1 Perception of Notification Signals

The differences between Study 1 and Study 3 in terms of qualitative feedback on the notification signals highlight the importance of context in interruption systems. In Study 1, where notifications were irrelevant to the task, the signal with highest AD was perceived by subjects to be the most annoying (82%), while the signal with lowest AD was ranked as least annoying (55%).

When utility became a factor in Study 3, perceptions reversed. The signal with high AD fell into favour with subjects (35%) who realized that its content improved their performance on the primary task. Conversely, the low AD signal drew mixed reviews: subjects who either did not comprehend the relationship between utility and AD, or who did comprehend but did not trust it, complained that the FLAG signal was difficult to detect (30%). In contrast, subjects who trusted the relationship seemed pleased that the low-utility hints were less disruptive and easily ignored (15%). This attitude characterizes the expected affective response to an interruption system where the relationship between AD and utility is explicitly known to users.

These results highlight the significance of Billsus et al.'s [9] observation about current static notification methods being alternatively too subtle and too obtrusive, depending on context. Interruption is most detrimental when important interruptions are too subtle and unimportant interruptions are too obtrusive. As our study shows, when the utility of an interruption is known to the system, an interrupting system that uses multiple levels of AD is perceived in a more favourable light than one that collapses AD across the board using a medium-level signal.

6.7.2 Performance and Workload

Understanding the performance impacts of helpful interruptions was not the primary goal of our study; however, we had hoped that our matched interruption presentation strategy would yield small performance gains in addition to improved annoyance and benefit perception. It seems that the help offered by the hints was enough to mitigate the additional effort and distraction associated with interruption (evidenced by the workload results), but was not enough to significantly boost performance above the interruption-free control condition. Unfortunately, fatigue, learning, presentation order, and interaction effects made it impossible to interpret the performance results.

Note that when interruptions are not directly related to a specific primary task, performance gains are not expected consequences of ideally matched interruptions. Future work is required to determine if performance gains can be achieved when interruptions are specific to the primary task. For instance, finding that effective interface customization can increase performance, Bunt, Conati, and McGrenere [11] suggest that adaptive support could help users customize effectively. Multi-level attention-getting may be applied to this domain by using high attentional draw to present suggestions for customization that the system is certain will save the user a lot of time, and using low attentional draw to present suggestions that may not save much time. It is worthwhile to determine whether an interruption-based mixed-initiative approach to interface customization that employs multiple levels of attentional draw in this manner can boost performance associated with the customized interface.

Although neither performance nor workload varied across the conditions, annoyance and perceived benefit responses were significantly better in the Match condition. The use of multiple notification signals did not increase workload, and the majority of subjects (75%) preferred the Match condition. Perhaps if the hints had elicited a performance boost, our self-reported measures would have been even stronger. The results of our research form a persuasive argument for the matching AD with utility.

6.7.3 Generalizability

Our research examined three levels of utility and an equal number of levels of AD. This use of three levels was motivated by the findings of Study 1, and also distinguishes our work from previous research [17, 42]. Our results show promise for the strategy of matching utility and AD in interruption, but also raise questions about how our work generalizes to real-world contexts where interruptions have a wide range of utilities.

Further study is necessary to understand the tradeoffs between increasing the set of notification signals beyond three to permit a wider range of utilities to be conveyed, and the potential cognitive overload associated with having to interpret the meaning behind this increased set. We saw good results using three levels, but can a larger number of levels help users to distinguish more finely-grained utility levels? Conversely, is there a threshold beyond which distinguishing amongst too many levels of utility or too many different notification signals escalates workload and erodes performance? Whether or not users can manage an increased set of notification signals likely depends on the properties of those signals. In our study we used three blatantly discernible signals; however, a larger number of distinct notification signals may overload users. Thus, increasing the set of signals to convey additional utility levels likely requires notification signals that lie along a smooth scale of AD but are not distinctive. For instance, certain properties of a single signal are continuous (e.g., for a signal that uses motion, velocity) and could be manipulated to create varying levels of AD without unduly taxing the user by requiring recognition of distinct signals. In the motion example, users would not be expected to recognize differences in velocity; rather, faster velocities would simply grab user attention more quickly, and so users would be notified of important interruptions more effectively than less important interruptions.

Maximum AD threshold is another question: is there a level of AD so high that it will create disturbance, no matter how high the utility? Our results seem to indicate not, since our FOLLOW signal had very high AD and was still received favourably.

Finally, there is the question of generalizability of scope and context. We examined utility in the scope of a primary task. We hypothesize that our results could generalize to utility in the context of personally relevant interruptions, but further research is required to confirm this belief. As we noted earlier, however, determining the utility in such contexts is likely much more difficult.

6.7.4 Design Implications

Our work is intended to persuade interruption system developers to heed Obermayer and Nugent's design guideline to match the attentional draw of a notification to the interruption utility. Some have argued that this design guidance is too simplistic [37].

Our research suggests otherwise. In our work, identical interruptions were presented to subjects; our two interruption conditions differed only in terms of the level of AD associated with the signals used to notify subjects. Yet, subjects perceived the interruptions to have significantly different levels of benefit and annoyance across the two conditions. Thus, this relatively simple solution can in fact provide significant improvement over current methods of interruption with static notification signals. The value of Obermayer and Nugent's design guidance has clearly been underestimated by the research literature and the industry. Our results provide a strong argument for interface designers to begin harnessing AD to improve interruption systems, as long as some estimation of utility is available.

As discussed in the related work, systems capable of assessing utility do currently exist (e.g., mixed-initiative and recommender systems); auspiciously, these are the types of systems for which a positive perception of interruption is most crucial. Alternatively, when interruptions are human-generated, senders could designate utility. In terms of extending the strategy to diverse sources of interruption, our work motivates research into computationally appraising utility of arbitrary interruption content. Results from Study 2 indicate, however, that caution must be exercised when utility ratings are not reliable.

In our experiments, the relationship between AD and utility was not explicitly made known to users because we wanted to see if benefits could be perceived at an unconscious level. Even with limited exposure (15 interruptions in 17 minutes), 75% of subjects at least partially deciphered the relationship. Still, not all subjects fully deciphered the relationship; moreover, many did not trust the perceived relationship. Thus, systems that adopt the strategy of matching AD to utility should make the relationship known so that users can work with the system instead of fighting it; however, trust is likely to remain an issue for some users.

The use of multiple levels of AD may also benefit research systems that are currently concerned with timing of interruption (e.g., [20, 29]): when the system wants to interrupt but determines that the particular moment is inopportune, utilizing a notification signal with low AD could be an alternative to postponing the interruption.

Chapter 7

Conclusions and Future Work

The primary goal of the work presented in this thesis was to examine the effects of matching the attentional draw of interruption notification to the utility of the interruption content. While previous work has recommended this strategy, few researchers have heeded the guidance, and there has never been an empirical investigation of the potential benefits of multi-level attention-getting. We conducted three experiments to examine the potential benefits of matching interruption presentation in terms of annoyance, perceived benefit, workload, and performance. Our results indicate that interfaces that vary attentional draw with utility are associated with decreased annoyance and an increased perception of benefit compared to interfaces that use a static level of attentional draw. Our research also establishes a set of three significantly different notification signals along the spectrum of attentional draw.

7.1 Limitations

As with any lab experiment, our studies exercised a trade-off of realism and generalizability for increased precision [38]. The formulation of primary and interrupting task employed in the study as an emulation of mixed-initiative systems helped to maintain a degree of ecological validity. However, the degree of realism provided by this factor is tempered by a relatively small amount of exposure to each condition. The generalizability of the results of this study is likewise limited by the use of a narrow range of utility and attentional draw. Given these limitations, this work should be regarded as an initial step towards exploring the potential benefits of varying attentional draw with utility in interruption.

7.2 Future Work

Several possibilities for future studies arise from the results described in this thesis. As already mentioned, further study is necessary to create notification methods that can maximize the number signal-utility pairs without cognitively overloading users.

Furthermore, our results motivate research into computationally appraising the utility of arbitrary interruption content. If utility can be automatically assessed, the multilevel attention-getting strategy can be extended to diverse sources of interruption content. In contexts where interruptions are specific to the primary task, our hypotheses may be retested to determine if performance gains can be expected consequences of ideally matched interruptions.

Finally, a logical direction for continuing this work is to explore the effects of matching attentional draw to utility in a more naturalistic setting with an existing interruption system. Varying attentional draw with utility in a chat client where senders indicate the utility of their messages will allow us to investigate whether our results generalize to utility in the context of personally relevant interruptions. This type of field study would also allow us to retest our hypotheses in a context where the relationship between attentional draw and utility is explicitly known to users, motivation to look at interruptions is uncontrived, and where users can have in-situ exposure over time to the various interruption conditions.

7.3 Concluding Remarks

This work should be seen as an initial step towards understanding and exploiting the benefits of matching attentional draw of notification to the utility of interruption content. Results presented in this thesis demonstrate that, contrary to the argument that such a matching strategy alone is too simplistic to make any real difference, it can in fact provide significant improvement over current methods of interruption that employ static notification. Thus, this strategy can help to emphasize beneficial aspects of interruption. Before developing specific guidelines that will be appropriate for a wide range of applications, further work needs to be done in order to evaluate this strategy of interruption notification in more realistic contexts.

Bibliography

[1] P. D. Adamczyk and B. P. Bailey. If not now, when?: the effects of interruption at different moments within task execution. In *CHI '04: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 271–278. ACM Press, 2004.

[2] E. M. Altmann and J. G. Trafton. Task interruption: resumption lag and the role of cues. In *Proceedings of the 26th Annual Conference of the Cognitive Science Society (CogSci 2004)*, 2004.

[3] E. Arroyo, T. Selker, and A. Stouffs. Interruptions as multimodal outputs: which are the less disruptive? In *ICMI '02: Proceedings of the 4th IEEE International Conference on Multimodal Interfaces*, page 479. IEEE Computer Society, 2002.

[4] D. Avrahami and S. E. Hudson. QnA: augmenting an instant messaging client to balance user responsiveness and performance. In *CSCW '04: Proceedings of the 2004 ACM Conference on Computer Supported Cooperative Work*, pages 515–518. ACM Press, 2004.

[5] D. Avrahami and S. E. Hudson. Responsiveness in instant messaging: predictive models supporting inter-personal communication. In *CHI '06: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 731–740. ACM Press, 2006.

[6] R. Baecker and I. Small. Animation at the interface. In B. Laurel, editor, *The Art of Human-Computer Interface Design*, pages 97–115. Addison-Wesley, 1990.

[7] B. P. Bailey, J. A. Konstan, and J. V. Carlis. The effects of interruptions on task performance, annoyance, and anxiety in the user interface. In *INTERACT 2001 Conference Proceedings*, pages 593–601, 2001.

[8] L. Bartram, C. Ware, and T. Calvert. Moticons: detection, distraction and task. *International Journal of Human-Computer Studies*, 58(5):515–545, 2003.

[9] D. Billsus, D. M. Hilbert, and D. Maynes-Aminzade. Improving proactive information systems. In *IUI '05: Proceedings of the 10th international conference on Intelligent user interfaces*, pages 159–166. ACM Press, 2005.

[10] A. Bunt. User modelling to support user customization. In *User Modeling 2005: 10th International Conference, UM 2005*, pages 499–501, 2005.

[11] A. Bunt, C. Conati, and J. McGrenere. What role can adaptive support play in an adaptable system? In *IUI '04: Proceedings of the 9th International Conference on Intelligent User Interface*, pages 117–124. ACM Press, 2004.

[12] D. Chen and R. Vertegaal. Using mental load for managing interruptions in physiologically attentive user interfaces. In *CHI '04 Extended Abstracts on Human Factors in Computing Systems*, pages 1513–1516. ACM Press, 2004.

[13] C. M. Chewar, D. S. McCrickard, and A. G. Sutcliffe. Unpacking critical parameters for interface design: evaluating notification systems with the irc framework. In *DIS '04: Proceedings of the 2004 Conference on Designing Interactive Systems*, pages 279–288. ACM Press, 2004.

[14] J. Cohen. Eta-squared and partial eta-squared in communication science. *Human Communication Research*, 28:473–490, 1973.

[15] E. Cutrell, M. Czerwinski, and E. Horvitz. Notification, disruption, and memory: effects of messaging interruptions on memory and performance. In *Human-Computer Interaction - INTERACT 2001 Conference Proceedings*, pages 263–269, 2001.

[16] M. Czerwinski, E. Cutrell, and E. Horvitz. Instant messaging and interruption: Influence of task type on performance. In *Proceedings of OZCHI 2000*, pages 356–361, 2000.

[17] M. Czerwinski, E. Cutrell, and E. Horvitz. Instant messaging: effects of relevance and time. In *People and Computers XIV: Proceedings of HCI 2000, Vol. 2*, pages 71–76, 2000.

[18] M. Debevc, B. Meyer, D. Donlagic, and R. Svecko. Design and evaluation of an adaptive icon toolbar. *User Modeling and User-Adapted Interaction*, 6(1):1–21, 1996.

[19] J. Fogarty, S. E. Hudson, and J. Lai. Examining the robustness of sensor-based statistical models of human interruptibility. In *CHI '04: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 207–214. ACM Press, 2004.

[20] J. Fogarty, A. J. Ko, H. H. Aung, E. Golden, K. P. Tang, and S. E. Hudson. Examining task engagement in sensor-based statistical models of human interruptibility. In *CHI '05: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 331–340. ACM Press, 2005.

[21] S. Gievska and J. Sibert. Using task context variables for selecting the best timing for interrupting users. In *sOc-EUSAI '05: Proceedings of the 2005 Joint Conference on Smart Objects and Ambient Intelligence*, pages 171–176. ACM Press, 2005.

[22] T. Gillie and D. Broadbent. What makes interruptions disruptive? a study of length, similarity and complexity. *Psychological Research*, 50(4):243–250, 1989.

[23] D. A. Goldstein and J. C. Lamb. Moving icons as a human interrupt. *Human Factors*, 9(5):405–408, 1967.

[24] A. Gupta, R. Sharda, R. A. Greve, and M. Kamath. An exploratory analysis of email processing strategies. In *Proceedings of 35th Annual Meeting of the Decision Sciences Institute*, pages 7671–7676, 2004.

[25] S. G. Hart and L. E. Staveland. Development of NASA-TLX (Task Load Index): results of empirical and theoretical research. In P. Hancock and N. Meshkati, editors, *Advances in Human Psychology: Human Mental Workload*, pages 139–183. Elsevier Science, 1988.

[26] J. Ho and S. S. Intille. Using context-aware computing to reduce the perceived burden of interruptions from mobile devices. In *CHI '05: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 909–918. ACM Press, 2005.

[27] E. Horvitz. Principles of mixed-initiative user interfaces. In *CHI '99: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 159–166. ACM Press, 1999.

[28] E. Horvitz, P. Koch, and J. Apacible. Busybody: creating and fielding personalized models of the cost of interruption. In *CSCW '04: Proceedings of the 2004 ACM Conference on Computer Supported Cooperative Work*, pages 507–510. ACM Press, 2004.

[29] S. T. Iqbal and B. P. Bailey. Leveraging characteristics of task structure to predict the cost of interruption. In *CHI '06: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 741–750. ACM Press, 2006.

[30] T. Jackson, R. Dawson, and D. Wilson. The cost of email interruption. *Journal of Systems and Information Technology*, 5(1):81–92, 2001.

[31] N. Kern, S. Antifakos, B. Schiele, and A. Schwaninger. A model for human interruptability: experimental evaluation and automatic estimation from wearable sensors. In *ISWC '04: Proceedings of the Eighth International Symposium on Wearable Computers (ISWC'04)*, pages 158–165. IEEE Computer Society, 2004.

[32] F. Linton, D. Joy, H.-P. Schaefer, and A. Charron. Owl: A recommender system for organization-wide learning. *Educational Technology & Society*, 3(1), 2000.

[33] P. P. Maglio and C. S. Campbell. Tradeoffs in displaying peripheral information. In *CHI '00: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 241–248. ACM Press, 2000.

[34] T. Matthews, A. K. Dey, J. Mankoff, S. Carter, and T. Rattenbury. A toolkit for managing user attention in peripheral displays. In *UIST '04: Proceedings of the 17th Annual ACM Symposium on User Interface Software and Technology*, pages 247–256. ACM Press, 2004.

[35] D. S. McCrickard, R. Catrambone, C. M. Chewar, and J. T. Stasko. Establishing tradeoffs that leverage attention for utility: empirically evaluating information display in notification systems. *International Journal of Human-Computer Studies*, 58(5):547–582, 2003.

[36] D. C. McFarlane. Comparison of four primary methods for coordinating the interruption of people in human-computer interaction. *Human-Computer Interaction*, 17(1):63–139, 2002.

[37] D. C. McFarlane and K. A. Latorella. The scope and importance of human interruption in human-computer interaction design. *Human-Computer Interaction*, 17(1):1–61, 2002.

[38] J. E. McGrath. Methodology matters: doing research in the behavioral and social sciences. In R. M. Baecker, J. Grudin, W. Buxton, and S. Greenberg, editors, *Readings in Human-Computer Interaction: Towards the Year 2000*, pages 152–169. 1995.

[39] B. Oberg and D. Notkin. Error reporting with graduated color. *IEEE Software*, 9(6):33–38, 1992.

[40] R. W. Obermayer and W. A. Nugent. Human-computer interaction for alert warning and attention allocation systems of the multi-modal watchstation. In *Integrated Command Environments: Proceedings of SPIE, Vol. 4126*, pages 14–22, 2000.

[41] P. Pongched. A more complex model of relevancy in interruptions. Unpublished manuscript obtained through personal contact with author, 2003.

[42] T. J. Robertson, J. Lawrance, and M. Burnett. Impact of high-intensity negotiated-style interruptions on end-user debugging. *Journal of Visual Languages & Computing*, 17(2):187–202, 2006.

[43] T. J. Robertson, S. Prabhakararao, M. Burnett, C. Cook, J. R. Ruthruff, L. Beckwith, and A. Phalgune. Impact of interruption style on end-user debugging. In *CHI '04: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 287–294. ACM Press, 2004.

[44] C. G. Thomas and M. Krogsoeter. An adaptive environment for the user interface of excel. In *IUI '93: Proceedings of the 1st International Conference on Intelligent User Interfaces*, pages 123–130. ACM Press, 1993.

[45] M. Walji, J. Brixey, K. Johnson-Throop, and J. Zhang. A theoretical framework to understand and engineer persuasive interruptions. In *Proceedings of 26th Annual Meeting of the Cognitive Science Society (CogSci 2004)*, 2004.

[46] C. Ware. *Information visualization: perception for design.* Morgan Kaufmann Publishers Inc., 2000.

[47] C. Ware, J. Bonner, W. Knight, and R. Cater. Moving icons as a human interrupt. *International Journal of Human-Computer Interaction*, 4(4):341–348, 1992.

[48] G. White and L. Zhang. Sender-initiated email notification: using social judgment to minimize interruptions. Technical Report MSR-TR-2004-126, Microsoft Research, 2005.

[49] J. Xiao, J. Stasko, and R. Catrambone. An empirical study of the effect of agent competence on user performance and perception. In *AAMAS '04: Proceedings of the Third International Joint Conference on Autonomous Agents and Multiagent Systems*, pages 178–185. IEEE Computer Society, 2004.

Appendix A

Background Questionnaire

The following questionnaire was used to collect information on user demographics in all three of our studies.

Subject #: _______

Background Questionnaire

1. In what age group are you?

- □ 19 and under
- □ 20 - 29
- □ 30 - 39
- □ 40 - 49
- □ 50 - 59
- □ 60+

2. Gender:

- □ Male
- □ Female

3. Are you right-handed or left-handed?

- □ Right-handed
- □ Left-handed

4. How many hours a week on average do you use a computer (including work and non-work related activities)?

- □ < 1
- □ 1 - 5
- □ 5 - 10
- □ > 10

5. Which operating systems do you currently use on a regular basis (at least on a weekly basis)? Please tick all that apply.

- □ Windows (Microsoft)
  - □ Windows XP
  - □ Windows 2000
  - □ Windows ME
  - □ Windows 98
  - □ Windows 95
  - □ Other - please specify: ________________
- □ Mac (Apple)
  - □ OS X
  - □ OS 9 or lower
- □ Unix - specify window manager: ________________

Subject #: _______

6. In terms of your current occupation, how would you characterize yourself?

□ Writer
□ Administrative Assistant
□ Journalist
□ Secretary
□ Academic
□ Professional
□ Technical expert
□ Student - please specify your area of study: ________________
□ Designer
□ Administrator/Manager
□ Other - please specify: ________________

Appendix B

Study 1 Resources

B.1 Study 1 Instructions

Explanation of the Icon Change Detection Task

Over the course of this experiment you will be asked to perform two primary tasks, which will be explained later. You will work on each of the primary tasks 3 times in a row, with a short break in between each session.

While you are performing each task there will be a blue circular icon visible in the bottom right-hand corner of the screen. Throughout the experiment, this icon may change visibly. This icon may:

1. Have a small yellow exclamation point appear in the centre of the icon.

2. Change colour.

3. Grow to a larger size.

4. Move slowly up and down continuously

5. Bounce continuously

6. Grow and shrink continuously (slowly or quickly)

7. Change colour continuously (slowly or quickly)

8. Follow your mouse cursor

(Each icon change will be demonstrated on screen during the above explanation.)

As soon as you notice any of these icon changes, using your non-mouse hand, please press the space bar to indicate that you have noticed the change. Once you have pressed the space bar, the icon will return to its normal size and colour.

Your performance during the experiment will be recorded. If your score falls within the top third of participants, you will be paid an additional $10, so you should try to perform to the best of your ability. Scoring for each of the primary tasks will be explained when the task is introduced. Your overall score will be largely based on your scores for the two primary tasks, but will also take into account your detection of the icon changes.

Do you have any questions?

Explanation of the Editor Task

During this task you will see a table that contains numbers between 0 and 9. Clicking with the mouse on a number will change that number to a 1. Your task is to replace all of the 0s with 1s. At the top of the screen you will see how many 0s remain on the board. Once you have replaced all of the 0s on the board, the table will be reset with new values, and your editing task will continue. Please continue to perform this task until a popup box on the screen tells you to stop.

Your score on this task will be calculated as the total number of edits you make.

Please take a moment to practice this task by editing all the 0s on this board. (practise board)

Note that during the actual trial, the table will be larger and you will also be performing the secondary icon change detection task at the same time.

Do you have any questions?

You will now be asked to perform this task 3 times in a row, with a short break in between.

Explanation of the Memory Game Task

This task is a matching game. The game board consists of cards. Each card has a picture on the front, and every card on the board has a matching card that contains the same picture. At the start, all cards are face-down. When you use the mouse to click on a card, that card will "turn over" and you will see the picture on the card. You may turn over two cards at a time. If you find a match, the cards will remain face up. If you flip over two cards that do not match, both cards will automatically be turned back over. Your goal is to find all of the matches on the board. You have won the game when all the cards are face-up. Please continue to play the game until a popup box on the screen tells you to stop. If you win the game before this popup appears, the board will reset and you will begin a new game.

Your score on this task will be calculated as the total number of matches you find.

Please take a moment to practice this task by finding all of the matches on this board. (pratice board)

Note that during the actual trial, the game board will be larger (i.e. there will be more cards) and you will also be performing the secondary icon change detection task at the same time.

Do you have any questions?

You will now be asked to perform this task 3 times in a row, with a short break in between.

B.2 Study 1 Questionnaire

The following questionnaire was administered after subjects completed all six blocks.

Subject #: _______

Post-Experiment Questionnaire

To refresh your memory, you saw 10 different icon changes in the experiment:

Flag: A yellow exclamation mark appeared in the centre of the icon.
Yellow: The icon colour changed to yellow.
Grow: The icon smoothly grew from small to large.
Oscillate: The icon moved slowly up and down.
Bounce: The icon moved up and with a bouncing motion.
Slow zoom: The icon continuously grew and shrank at a slow velocity.
Fast zoom: The icon continuously grew and shrank at a fast velocity.
Slow blink: The icon slowly flashed back and forth from blue to yellow.
Fast blink: The icon quickly flashed back and forth from blue to yellow.
Follow: The icon followed the cursor as you moved it around the screen.

We define the term "annoy" using the following phrases:
To make slightly angry; to pester or harass; to disturb or irritate.

Part 1

By circling 1, 2, 3, 4, or 5 below, please indicate the degree to which to you felt annoyed **during the experiment** by each of the icon change types. If you do not recall seeing a particular icon change at all during the experiment, circle "did not notice."

	Very slightly or not at all	**A little**	**Moderately**	**Quite a bit**	**Extremely annoyed**	
Flag	1	2	3	4	5	didn't notice
Yellow	1	2	3	4	5	didn't notice
Grow	1	2	3	4	5	didn't notice
Oscillate	1	2	3	4	5	didn't notice
Bounce	1	2	3	4	5	didn't notice

Subject #: ______

	Very slightly or not at all	**A little**	**Moderately**	**Quite a bit**	**Extremely annoyed**	
Slow zoom	1	2	3	4	5	didn't notice
Fast zoom	1	2	3	4	5	didn't notice
Slow blink	1	2	3	4	5	didn't notice
Fast blink	1	2	3	4	5	didn't notice
Follow	1	2	3	4	5	didn't notice

Part 2

Please list the icon change types that you found to be the **most** annoying **during the experiment**:

Most annoying: ____________

Second-most annoying: ____________

Third-most annoying: ____________

Part 3

Please list the icon change types that you found to be the **least** annoying **during the experiment**:

Least annoying: ____________

Second-least annoying: ____________

Third-least annoying: ____________

Thank you for your participation!

Appendix C

Study 2 Resources

C.1 Study 2 Instructions

The following instructions were given to subjects in the Match and Random conditions. Subjects in the Control condition did not receive the instructions about the notifications and hints. Subjects in the Static condition did not receive the instructions that explained the FLAG and FOLLOW signals.

Explanation of the Memory Game Task

During this experiment you will be asked to play a card-matching game. I will begin by explaining how to play this game.

The game board consists of cards. Each card has a picture on the front, and every card on the board has one matching card that contains the same picture. At the start, all cards are face-down. When you use the mouse to click on a card, that card will "turn over" and you will see the picture on the card.

You may turn over two cards at a time. If you find a match, the cards will remain face up. If you flip over two cards that do not match, both cards will automatically be turned back over. Your goal is to find as many matches as possible. If you find all of the matches on the board before the end of a session, the board will reset and you will continue to play the game.

Please take a moment to practice this task by finding all of the matches on this board.

Note that during the actual trial, the game board will be larger (i.e. there will be more cards).

You will be asked to play the game for 2 sessions, each lasting 16 minutes. At the end of each session, a popup box will appear on the screen informing you that the session has ended. You will be given a short break in between the sessions.

Do you have any questions?

Explanation of the Notifications and Hints

Note that, while you are playing the game, there will be a blue circular icon visible in the bottom right-hand corner of the screen.

While you are playing the game, you may be provided with hints to help improve your game performance.

A hint has two stages: Stage 1 is a notification that a hint is available. This notification is presented as a visible change to the blue icon. There are three ways in which this icon may change. The three types of notification are (demonstrated on screen as I read):

1. A small yellow exclamation point appears in the centre of the icon.
2. The icon begins to grow and shrink continuously.
3. The icon begins to follow your mouse cursor around the screen.

Once you notice a notification, you may click on the icon to see the hint. This is the second stage of the hint. There are three kinds of hints (demonstrated on screen as I read):

1. A popup box with an encouraging message.
2. A popup box combined with one highlighted card that shows you the location of the match for the selected card. If you have a card selected at the time that you click on the notification, the hint will tell you about the match for that card. If you do not have a card selected at the time you click on the notification, a card will be randomly selected for you, and the hint will tell you about the match for that card.

3. A popup box combined with 4 highlighted cards. The match for the selected card is one of the highlighted cards. Again, if you have a card selected at the time that you click on the notification, the hint will tell you about the match for that card. If you do not have a card selected, one will be randomly selected for you.

You may dismiss the hint popup box by clicking on OK. Note that the cards will continue to be highlighted after you dismiss the popup box. They will stop being highlighted as soon as you click on one of the cards.

Things to note:

- Once you click on the icon during a notification, the notification will no longer be displayed (i.e. the icon will return to its normal size and colour).
- If you do not click on the notification within a certain amount of time, then notification will cease and you will no longer be able to access that particular hint.

Your performance during the game play will be recorded. Your score will be calculated as the total number of matches you find during both sessions. If your score falls within the top third of participants, you will be paid an additional $10, so you should try to perform to the best of your ability.

Do you have any questions?

C.2 Study 2 Interview

The following interview questions were administered to subjects in the Match, Static, and Random conditions. No interview was administered to subjects in the Control condition.

1. How did the notifications and hints affect your performance? (i.e. benefited vs. hindered)

2. (Look at Annoyance scales; if moderate-high) I see from your answer on the survey that your annoyance level was [moderate/quite high]. What was it that annoyed you?

3. Did you use any particular strategy for dealing with the notifications and hints? (Both in terms of detecting the notifications and deciding when to click on the icon to bring up the hints.)

4. Did you purposely ignore any of the hint notifications? Were some types harder to ignore than others?

5. Did you notice any relationship between the notification types and the helpfulness of the hint? If so, what was it?

6. Do you have any other comments about this experiment that you would like to share?

C.3 Study 2 Questionnaires

The first questionnaire presented was administered to subjects in the Match, Static and Random conditions. The questionnaire that follows was administered to subjects in the Control condition.

Subject #: _____

Post-Experiment Survey

With respect to **both sessions** of game play, please answer the following questions by marking an 'X' along the scale beside the corresponding question.

How much mental and perceptual activity was required to play the game and attend to the hints (e.g., thinking, remembering, looking, searching, deciding, etc.)?	**MENTAL DEMAND** **Low** **High**
How annoyed (i.e. pestered, harassed, disturbed or irritated) did you feel **during the task in general**?	**GENERAL ANNOYANCE** **Low** **High**
How annoyed (i.e. pestered, harassed, disturbed or irritated) were you **by the notifications and hints in particular**?	**NOTIFICATION/HINT ANNOYANCE** **Low** **High**
How much time pressure did you feel due to the rate or pace at which the tasks or task elements occurred?	**TEMPORAL DEMAND** **Low** **High**
How hard did you have to work (mentally and physically) to accomplish your level of performance?	**EFFORT** **Low** **High**
How successful do you think you were in accomplishing the goals of the task set by the experimenter (or yourself)?	**PERFORMANCE** **Poor** **Good**
How insecure, discouraged, irritated, stressed and annoyed versus secure, gratified, content, relaxed and complacent did you feel during the task?	**FRUSTRATION** **Low** **High**
To what extent did your performance benefit from the hints? Did you appreciate the assistance they provided?	**BENEFIT** **Low** **High**

Subject #: _____

Please indicate the extent to which you agree or disagree with the following statements (circle one):

I was motivated to look at the hints.	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
I felt fatigued during the sessions.	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree

During the **first** session, how many times did it ***seem*** like you were notified that a hint was available? (circle one)

1-3	4-6	7-9	10-12	13-15	16-18	19+

During the **second** session, how many times did it ***seem*** like you were notified that a hint was available? (circle one)

1-3	4-6	7-9	10-12	13-15	16-18	19+

Subject #: _____

Post-Experiment Questionnaire

With respect to **both sessions** of game play, please answer the following questions by marking an 'X' along the scale beside the corresponding question.

How much mental and perceptual activity was required to play the game and attend to the hints (e.g., thinking, remembering, looking, searching, deciding, etc.)?

MENTAL DEMAND

Low **High**

How annoyed (i.e. pestered, harassed, disturbed or irritated) did you feel **during the task in general**?

ANNOYANCE

Low **High**

How much time pressure did you feel due to the rate or pace at which the tasks or task elements occurred?

TEMPORAL DEMAND

Low **High**

How hard did you have to work (mentally and physically) to accomplish your level of performance?

EFFORT

Low **High**

How successful do you think you were in accomplishing the goals of the task set by the experimenter (or yourself)?

PERFORMANCE

Poor **Good**

How insecure, discouraged, irritated, stressed and annoyed versus secure, gratified, content, relaxed and complacent did you feel during the task?

FRUSTRATION

Low **High**

Appendix D

Study 3 Resources

D.1 Study 3 Instructions

Explanation of the Memory Game Task

During this experiment you will be asked to play a card-matching game. I will begin by explaining how to play this game.

The game board consists of cards. Each card has a picture on the front, and every card on the board has one matching card that has the same picture. At the start, all cards are face-down. When you use the mouse to click on a card, that card will "turn over" and you will see the picture on the card.

You may turn over two cards at a time. If you find a match, the cards will remain face up. If you turn over two cards that do not match, both cards will automatically be turned back over. Your goal is to find as many matches as possible. If you find all of the matches on the board before the end of a session, the board will reset and you will continue to play the game.

Please take a moment to practice this task by finding all of the matches on this board.(practise board)

Note that during the actual trial, the game board will be larger (i.e. there will be more cards).

Do you have any questions?

You will be asked to play this game for 3 sessions. Each session will last 17 minutes. At the end of each session, a popup box will appear on the screen to tell you that the session has ended and you will stop playing. After each session, you will fill out a survey and then you will get to take a break.

Explanation of the Notifications and Hints

Note that, while you are playing the game, there will be a blue circular icon visible in the bottom right-hand corner of the screen.

While you are playing the game, you may be provided with hints to help improve your game performance.

A hint has two stages: The first is a notification that a hint is available. This notification is presented as a visible change to the blue icon. There are three ways in which this icon may change, and so there are three types of notifications (demonstrated on screen as I read):

1. A small yellow exclamation point appears in the centre of the icon.

2. The icon begins to grow and shrink continuously.

3. A copy of the icon begins to follow your mouse cursor around the screen.

Do you have any questions about the notifications?

Once you notice a notification, you may click on the icon to see the hint. This is the second stage of the hint. There are three kinds of hints (demonstrated on screen):

1. A popup box with a text message.

2. A popup box combined with 1 highlighted card. This kind of hint tries to show you the location of 1 match. Unfortunately, sometimes it makes mistakes. So, sometimes the highlighted card will be the match for the selected card, and sometimes it will not be.

   If you have a card selected at the time that you click on the notification, the hint will tell you about the match for that card. If you do not have a card selected at the time you click on the notification, a card will be randomly selected for you, and the hint will tell you about the match for that card.

   You dismiss the hint popup box by clicking on OK. Note that the card will continue to be highlighted after you dismiss the popup box. It will stop being highlighted as soon as you click on the card.

3. A popup box combined with 10 highlighted cards. This hint shows you the location of 5 matches, using different colours for the different pairs. You will see 2 cards highlighted in yellow, 2 in pink, 2 in blue, 2 in orange, and 2 in green. The 2 pink cards are a match; the 2 yellow cards are a match, and so on. If you have a card selected at the time that you click on the notification, that card will become one of the highlighted cards.

   You dismiss the hint popup box by clicking on OK. Note that the cards will continue to be highlighted after you dismiss the popup box. Each card will continue to be highlighted until you click on that card. If you have fewer than 10 cards left on the board during this type of hint, the highlighted matches will carry over to the next board, i.e. if you have 4 cards left on the board, you will see hints for the last 2 pairs on the current board, and there will be hints for 3 more matches on the new board after it resets.

Do you have any questions about the hints?

- Once you click on the icon during a notification, the notification will stop being displayed (i.e. the icon will return to its normal size and colour).

- If you do not click on the notification within a certain amount of time, then the notification will stop and you will no longer be able to access that particular hint.

Do you have any questions?

Your performance during the game play will be recorded. Your score will be calculated as the total number of matches you find during all 3 sessions. If your score falls within the top third of participants, you will be paid an additional $10, so you should try to perform to the best of your ability.

Do you have any questions?

During each of the three sessions, you will see a slightly different version of the game with respect to the notifications and hints available, but I will let you know at the beginning of each session what exactly you will see.

Again, there are three sessions and they are each 17 minutes. That is a pretty long time to be playing the Memory game, so you might want to think about maybe pacing yourself so that you're not exhausted before you finish the last session. But of course, it's up to you to decide.

At the start of the Control condition: You will not see any hints during this session.

At the start of the Match condition: During this session you will see all three of the notifications and all three types of hints.

At the start of the Static condition: During this session you will see all three types of hints but you will only see one type of notification: the signal that grows and shrinks.

D.2 Study 3 Interview

Part 1: Preference

1. If you had the option to play one more round of the game (after enough rest, and the $10 bonus was still offered), which version of the game would you prefer?

    - The session that used 3 different types of notification to indicate that a hint was available
    - The session that used only 1 type of notification to indicate that a hint was available
    - The session that did not offer any hints

    Which version would you prefer the least?

Part 2: Effect of hints

In the next few questions I will be asking about the two sessions that had hints.

1. First of all, how do you feel the hints affected your performance?

    Was this effect different for the two sessions that had hints?

2. (If not already answered) Were any of the hints helpful? Which ones?

    Was the overall amount of help different for the two sessions that had hints?

3. Did any hints hinder your performance? Which ones? How?

    Did these hints hinder (or annoy) more in one session than another?

4. Do you feel that you saw the same number of hints in both sessions?

Part 3: Notifications

1. Do you have any thoughts you'd like to share about the different kinds of notifications? As a reminder, the three different kinds of notifications were:

   Exclamation mark:

   Grow and shrink:

   Follow:

2. Did you notice any relationship between the notification types and the helpfulness of the hints?

   [if yes] Did this relationship affect how you perceived the notifications?

Part 4: Strategy

The next few questions have to do with strategies for using the notifications and hints

1. What strategy/approach did you use to detect the notifications?

   - No polling, all peripheral.
   - Polled a little. Frequency? _________
   - Polled when I was hoping for a hint. How often? _________
   - Polled a lot. Frequency? _________

   Was your strategy different for the two sessions that had hints?

2. Once you noticed a notification, did you use a strategy for deciding when to click to see the hint?

   - Clicked as soon as I noticed the notification
   - If I was about to make a match, I would do that first
   - If there were only a few cards left, I finished the board first
   - Selected the card I wanted to match, then clicked

- Clicked right away only if I had no idea for a match

Was your strategy different for the two sessions that had hints?

Part 5: Ignoring notifications

1. Did you purposely ignore any of the notifications? If yes, which ones did you ignore, in which session, and why?

   - No
   - Ignored the Utility 1/Signal 1 match
   - When I was about to make a match
   - When I didn't want my concentration to be broken
   - When I was almost finished the board

Part 6: Wrap-up

1. (If reason for preference ranking is not obvious)

   (a) At the beginning of this interview, you said that you would prefer to play _________ Why was this your favourite session?

   (b) You also said that you preferred _________ the least. Why was that?

2. Do you have any other comments about this experiment that you would like to share?

D.3 Study 3 Questionnaires

The first questionnaire presented was administered after the Match and Static conditions. The questionnaire that follows was administered after the Control condition

Subject #: ____
Condition: ____
Session: ____

Post-Session Questionnaire

With respect to **this session** of game play, please answer the following questions by marking an 'X' along the scale beside the corresponding question.

How much mental and perceptual activity was required to play the game and attend to the hints? (e.g., thinking, remembering, looking, searching, deciding, etc.)?	**MENTAL DEMAND** **Low** **High**
How much physical activity was required to play the game? (e.g. moving the mouse, clicking the mouse button, etc.)	**PHYSICAL DEMAND** **Low** **High**
How much time pressure did you feel due to the rate or pace at which the tasks or task elements occurred?	**TEMPORAL DEMAND** **Low** **High**
How hard did you have to work (mentally and physically) to accomplish your level of performance?	**EFFORT** **Low** **High**
How successful do you think you were in accomplishing the goals of the task set by the experimenter (or yourself)?	**PERFORMANCE** **Poor** **Good**
How insecure, discouraged, irritated, stressed and annoyed versus secure, gratified, content, relaxed and complacent did you feel during the task?	**FRUSTRATION** **Low** **High**

Subject #: ____
Condition: ____
Session: ____

To what extent did your performance benefit from the hints?	**BENEFIT** **Low** **High**
How annoyed (i.e. pestered, harassed, disturbed or irritated) did you feel **during the task in general**?	**GENERAL ANNOYANCE** **Low** **High**
How annoyed (i.e. pestered, harassed, disturbed or irritated) were you **by the notifications and hints in particular**?	**NOTIFICATION/HINT ANNOYANCE** **Low** **High**

For "General Annoyance," please describe what it was that annoyed you:

For "Notification/Hint Annoyance," please describe what it was that annoyed you:

Please indicate the extent to which you agree or disagree with the following statements (circle one):

I was motivated to look at the hints.	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
I felt fatigued during this session.	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree

Subject #: ____
Condition: ____
Session: ____

Post-Session Questionnaire

With respect to **this session** of game play, please answer the following questions by marking an 'X' along the scale beside the corresponding question.

How much mental and perceptual activity was required to play the game? (e.g., thinking, remembering, looking, searching, deciding, etc.)?	**MENTAL DEMAND** **Low** **High**
How much physical activity was required to play the game? (e.g. moving the mouse, clicking the mouse button, etc.)	**PHYSICAL DEMAND** **Low** **High**
How much time pressure did you feel due to the rate or pace at which the tasks or task elements occurred?	**TEMPORAL DEMAND** **Low** **High**
How hard did you have to work (mentally and physically) to accomplish your level of performance?	**EFFORT** **Low** **High**
How successful do you think you were in accomplishing the goals of the task set by the experimenter (or yourself)?	**PERFORMANCE** **Poor** **Good**
How insecure, discouraged, irritated, stressed and annoyed versus secure, gratified, content, relaxed and complacent did you feel during the task?	**FRUSTRATION** **Low** **High**

Subject #: ____
Condition: ____
Session: ____

How annoyed (i.e. pestered, harassed, disturbed or irritated) did you feel **during the task in general**?	**GENERAL ANNOYANCE** **Low** **High**

For "General Annoyance," please describe what it was that annoyed you:

Please indicate the extent to which you agree or disagree with the following statement (circle one):

I felt fatigued during this session.	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree

Appendix E

UBC Research Ethics Board Certificates

This section contains all Certificates of Approval administered by the UBC Research Ethics Board in relation to the research reported in this thesis.