Matching Attentional Draw with Utility in Interruption

Jennifer Gluck, Andrea Bunt, and Joanna McGrenere
Department of Computer Science, University of British Columbia
Vancouver, BC, Canada
jengluck@gmail.com, {bunt, joanna}@cs.ubc.ca

ABSTRACT

This research examines a design guideline that aims to increase the positive perception of interruptions. The guideline advocates matching the amount of attention attracted by an interruption's notification method (attentional draw) to the utility of the interruption content. Our first experiment examined a set of 10 visual notification signals in terms of their detection times and established a set of three significantly different signals along the spectrum of attentional draw. Our second experiment investigated matching these different signals to interruption content with different levels of utility. Results indicate that the matching strategy decreases annoyance and increases perception of benefit compared to a strategy that uses the same signal regardless of interruption utility, with no significant impact on workload or performance. Design implications arising from the second experiment as well as recommendations for future work are discussed.

Author Keywords

Interruption, attention, utility, detection, annoyance, benefit, mental workload

ACM Classification Keywords

H5.2. User Interfaces, *evaluation/methodology*. H1.2. User/Machine Systems, *human factors*

INTRODUCTION

Interrupting technologies such as telephones, email, instant messaging (IM), and calendar systems pervade our everyday lives. The ubiquity of interruption can be overwhelming, and studies commonly fixate on disruptive effects of interruption on task performance [8, 13] and emotional state of users [3]. This focus on negative effects ignores the potential value of interruption. Studies that are foundational to the research literature examine interruptions that are often relevant to neither the primary task nor the user (e.g., [22]); however, in practice, interruption content is often relevant. Email-alerting and IM software are often

implicated as disruptive interruption offenders [2, 19], yet their rampant popularity testifies to their usefulness. The fact that we continue to propagate and to tolerate such computer-based interruption suggests that it has some value. Other potentially valuable interruption-based technologies are recommender [5] and mixed-initiative [10, 15] systems, which strive to improve the user experience by making real-time, context-sensitive suggestions aimed to assist users in performing a task. A key component to the success of such systems, however, is that users perceive the interruptions positively. Interruptions must be presented tactfully so that users neither ignore suggestions, nor are driven by annoyance to abandon use of the system, as in the case of the Microsoft Office Assistant [20].

A design guideline proposed by Obermayer and Nugent [25] may help to promote the positive perception of interruption. The guideline recommends setting the level of attention attracted by an interruption's notification signal relative to the utility of the interruption content. Using this strategy, systems present interruptions that are highly important using notification signals with high *attentional draw* (AD) so that they are noticed immediately, while presenting less important interruptions more subtly so that they will be noticed only during a natural break. AD for interruptions with utilities between these endpoints is scaled accordingly, and so users are only truly interrupted from a task when it is important to do so. Some [23] have argued that this design guidance is simplistic: alone, it cannot solve the disruptive aspects of interruption. Meanwhile, few commercially available interruption systems have adopted the strategy. We suspect that the value of the guideline has been underestimated, however, empirical investigation of the design approach is absent in the literature. Thus, we studied the effects of matching AD and utility to determine if this strategy alone can in fact help to ease the disruptive effects and facilitate positive perception of interruption.

In our research we define the attentional draw (AD) of a notification signal as the time elapsed between when the signal is presented and when the user notices its presence. Interruption content may be examined in terms of both relevance (i.e., how pertinent the content is to the recipient) and utility (i.e., how useful, important, or urgent the content is to the recipient). Relevance is a component of utility but does not define it. Interruption content that is highly utilitarian must be relevant to the user in some way; however, it is possible for content to be relevant but

unimportant. In this work, we define utility in terms of relevance to a specific primary task.

For our first experiment, we created 10 prospective notification signals and then empirically distinguished a subset of three signals, where each signal in the set was significantly different than the other two in terms of AD. Our second experiment matched these signals with interruptions with varying levels of utility, and investigated effects in terms of annoyance, perceived benefit, workload, and performance. The interrupting task comprised context-sensitive hints designed to help subjects perform a primary task. By allowing subjects to decide if and when to utilize each hint, we effectively emulated a mixed-initiative system. Results from our second experiment indicate that, when a measure of utility is available, the matching strategy does result in decreased annoyance and increased perception of benefit compared to a strategy that employs the same level of AD regardless of interruption utility, with no significant impact on workload or performance.

RELATED WORK
A substantial amount of research [e.g., 1, 11, 18] has examined the effect of timing of interruption onset to determine if negative effects can be reduced by presenting an interruption at an "ideal moment" and postponing the interruption if the moment is inopportune. Our work takes a complimentary approach by investigating how to present interruptions to users without delaying delivery of important messages.

Few evaluations have examined the AD associated with how an interruption is presented. Bartram et al. [4] studied the perceptual properties of visual motion applied to notification in terms of detection and distraction. Robertson et al. [26] examined notification in terms of intensity, comparing "high-intensity" and "low-intensity" interruptions, where the notion of intensity is similar in spirit to AD. Robertson et al. did not quantify or measure intensity nor did they link intensity to the utility of interruption content. Finally, McCrickard et al. [21] explored notifications that trade off utility and attention in peripheral awareness systems, which provide a stream of continuous content in a divided-attention situation. Our work, on the other hand, focuses on interruptions that occur at discrete moments and contain explicit content. The difference between these two contexts is further evidenced by diverging definitions of utility. McCrickard et al. discussed utility as the value provided by the peripheral system as a whole and did not directly manipulate utility within their experiment, while we consider utility as the importance of the content of a particular interruption.

A number of systems have matched notification signals to utility; however, these matching strategies have not been explicitly or systematically compared to a static one. The Notification Platform (e.g., [16]) took into account human attentional state, expected value of interruption content, and cost of disruption to select an appropriate device, modality, and intensity with which to present notification signals. Intensity was based on direct subjective assessment of the different signals. The Priorities system (e.g., [17]) was designed to appraise the criticality of email and sort incoming messages accordingly. Users could also configure the multimodal properties of email notifications according to message criticality. Building on the Notification Platform, the Scope [28] project explored automatic mapping of utility and AD. Both the QnA IM Client [2] and the FXPAL Bar [5] also matched AD to utility automatically, using two levels of AD and two levels of utility. Finally, Oberg and Notkin [24] conveyed a gradient of AD and utility by using colour saturation and intensity to communicate the age and importance of code errors. In addition to a controlled evaluation that investigates the benefits of automatically mapping AD to utility, our research extends this prior work by exploring AD in detail.

Our research has not included how to appraise the utility of interruption content computationally. Instead, we selected a primary task for which we could generate interruption content with objective levels of utility. Appraisal of utility of interruption content has been investigated by a number of researchers (e.g., [2, 16, 28]).

EXPERIMENT 1: SELECTING NOTIFICATION SIGNALS
The goal of Experiment 1 was to identify three to five signals whose mean detection times (i.e., AD) were significantly different from one another. Since most existing interruption research and interrupting applications utilize the visual field, we focused on visual notification rather than using another modality. We based our experimental design on research by Bartram et al. [4]. Taking into account well-understood properties in the psychology and information visualization literature (i.e., 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.

Primary Tasks
We used two primary tasks with different workloads. The high-workload task, which was also needed for our planned second experiment, had to meet two key requirements: (1) the ability to generate interruptions with an objective measure of utility; and (2) the need to involve concentration such that a cognitive context switch is required to go from primary to interrupting task.

A computer-based version of the game Memory satisfied these requirements. This traditional game involves a set of picture cards consisting 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 session, the board was reset with a different deck of cards and the matching task continued. The deck size was 64

cards (32 pairs). This large number of cards ensured that the task required a significant amount of concentration and thus provided a high workload.

The low-workload task allowed us to gauge reaction times when subjects did not need to be pulled out of heavy concentration. It was based on the simple editing task in Bartram et al.'s Moticon research [4]. A large non-scrollable editing window contained a 20x20 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 a subject completed all necessary edits before the end of a session, 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.

Interruption Detection Task
We designed and studied a base set of 10 different signals. Signals were presented sequentially while subjects performed one of the two primary tasks. Subjects were asked to respond by pressing the space bar with their non-dominant hand whenever they noticed a signal. Signals were comprised of transformations applied to an icon that was present on the screen at all times. The base icon was a blue circle with a diameter of 21 pixels (0.62cm). We placed this icon in the bottom right-hand corner of the screen in order to emulate the Windows OS system tray. The notification signals were designed across four categories that we hypothesized would span the spectrum of AD. Parameters such as colour change rates and movement velocities were based on informal piloting. Table 1 provides a description of the signals, by category, and the accompanying video demonstrates the signals.

Our experiment was designed so that subjects would *be interrupted* rather than *wait for an interruption*. Similarly to Bartram et al. [4], 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, at which point the trial ended and a new trial began. We also 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, no signal was presented. 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.

Design
The experiment used a within-subjects 2 x 10 x 3 (primary task x notification signal x block) design. There were also two orders of presentation of the primary task, a between-

Category A: Single State Change	
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, centered on its origin, over a period of 500ms.
Category B: Continuous Slow State Change	
OSCILLATE (OS)	The icon moved slowly up and down a path of 17 pixels (0.5cm) with sinusoidal motion. It took 1700ms to complete one 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. It took 1500ms for the icon to complete one full grow/shrink cycle.
SLOW BLINK (SB)	The icon continuously flashed back and forth from blue to yellow every 1000ms.
Category C: Continuous Fast State Change	
BOUNCE (BC)	The icon moved up and down with a bouncing motion. Each bounce took 800ms to complete.
FAST ZOOM (FZ)	The icon smoothly and continuously grew and shrank between 100% and 200% of its original size, centered on its origin. It took 780ms for the icon to complete one full grow/shrink cycle.
FAST BLINK (FB)	The icon continuously flashed back and forth from blue to yellow every 300ms.
Category D: Continuous Location Change	
FOLLOW (FL)	A copy of the base icon appeared directly beside the mouse cursor and continued to follow the cursor until detection occurred or the trial timed out.

Table 1. Notification signals used in Experiment 1.

subjects control variable introduced to account for order effects.

Participants
Twelve subjects (1 female) between 18 and 39 years of age participated in the experiment 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 and staff at the University of British Columbia.

Motivation
To motivate subjects to focus on the primary task but not entirely ignore 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. The explanation of scoring was deliberately vague so that participants would not try to fit their performance to the specifics of the scoring system.
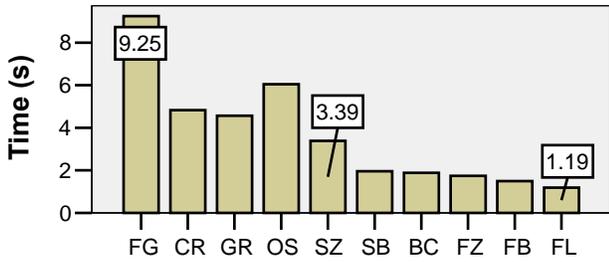
Figure 1. Mean detection times by signal (N=12).

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 1280x1024. The experimental software, including all notifications signals, was fully automated and was coded in Java 1.5.0.

Procedure

The experiment was designed to fit in a single 90-minute session. The procedure was as follows. (1) A questionnaire obtained 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. 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 when it was necessary to obtain clarification on questionnaire responses.

Measures

Our main dependent variable was detection time. This measure was capped at 30 seconds if the notification signal timed out. Annoyance measures were collected via a questionnaire at the end of the experiment; subjects were asked 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." The timeout and false detection rates were also measured.

Results

A 2 (task) by 10 (signal) by 3 (block) by 2 (presentation order) ANOVA showed no significant main or interaction effects of presentation order, so in all subsequent analysis we examine only the effects of task, notification signal, and block. A series of 2 (task) by 10 (notification signal) by 3 (block) ANOVAs were performed on the data. We applied the Greenhouse-Geisser adjustment 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 [7].

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

Table 2. Pairwise comparisons of detection times: a square indicates that the row signal had a significantly faster detection time than the column signal. (N=12).

Detection times for each signal are summarized in Figure 1. As expected, the signal design had a significant impact on detection times (a large main effect of signal on detection time, $F(2.632, 28.954) = 14.204$, $p < .001$, $\eta^2 = 0.564$). Pairwise comparisons, summarized in Table 2, show one subset of three signals with mean detection times significantly different from one another: 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 FOLLOW ($p = .044$). There were no significant comparisons of four or more signals.

Counter to our expectations, the low- and high-workload tasks did not impact detection times differently ($F(1,11) = .781$, $p = .396$, $\eta^2 = .066$). 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 explainable pattern. Looking only at the three signals identified above, the interaction effect becomes even less of a 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.

The self-reported measures show that FOLLOW was ranked most annoying by 82% of subjects. FLAG was ranked least annoying by 55% of subjects. Data for one subject was excluded from the qualitative results because of a misunderstanding of the questionnaire.

The timeout and false detection rates were in line with the detection rates and are discussed more fully in [12].

Conclusions

The goal of this experiment was to identify a set of significantly different signals in terms of detection times. Results revealed one subset of the signals that fit our selection criteria: FLAG, SLOW ZOOM, and FOLLOW. In

Experiment 2, 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.

An insignificant effect of task on the three signals suggests that the mean detection times for these signals generalized across primary task workload, pointing to their potential usefulness in future studies.

In addition to investigating signals under different workload, Experiment 1 adds to existing research on notification signals in three ways. First, by identifying four categories during our initial signal-selection process, we provide a characterization of the space of visual notification signals. Second, we undertook to motivate a division of focus between primary and secondary detection task that mimicked a realistic interruption scenario. Consequently, we hope our results generalize more readily to interruption contexts. Finally, we established baseline annoyance ratings for a variety of notification signals.

EXPERIMENT 2: MATCHING NOTIFICATION TO UTILITY

In this experiment, we examined the effects of matching the degree of attentional draw (AD) associated with an interruption signal to the utility of its content, in terms of annoyance, perceived benefit, workload, and performance. We compared three conditions: *Match*, *Static*, and *Control*.

Match: Low-utility hints were presented 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). Subjects were not informed of the relationship between notification type and interruption utility, allowing us to probe whether benefits could be perceived on an unconscious level by subjects who did not consciously decipher the relationship.

Static: All hints were presented using the medium-AD signal (SLOW ZOOM). This condition was designed to emulate current practices, where all notification takes the same form. Billsus et al. [5] state that a key problem with interrupting interfaces is that notification is either too subtle or too obtrusive. We used the medium-AD signal to avoid this problem.

Control: An interruption-free condition was included establish baseline workload and performance measures.

Primary and Interrupting Tasks

The primary task was the Memory Game used in Experiment 1. The interrupting task comprised context-sensitive hints and comments from the system, many of which aimed to aid the subject in playing the game. A 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. Experiment 1 provided us with a set of three notification signals and we defined three corresponding 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 both interruption conditions (Match and Static).

3 Different Hint Utilities

A primary goal of this experiment was to investigate the perceived benefit of matching utility with the type of notification signal. To achieve *perceived* benefit, we felt that it was necessary for the interruption system to *actually* improve primary task performance. Based on results from a preliminary study [12], we expected an average performance boost of 15% if subjects looked at all hints

A *high-utility hint* showed the location of five matches by highlighting 10 cards, using different colours to indicate which of the cards matched.

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. This type of hint was designed to be "somewhat helpful" and needed to be appreciably different from the high-utility hint. 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 hint replications did not allow this.

A *low-utility hint* did not provide any assistance in finding a match. Instead, a text message unrelated to the game was displayed (e.g., "Nice weather we're having.").

Frequency and Structure of Interruption

Our design used five replications of each type of hint 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. The Control condition also lasted 17 minutes but contained no interruptions. As in Experiment 1, an interruption timed out after 30 seconds. If a subject did not respond to the notification signal within that time, the notification stopped and the subject missed that particular hint. 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. (Piloting revealed an upper bound of 6 seconds on the amount of time required to attend to a hint. Thus, even in the tightest scenario when two interruptions were 45 seconds apart, the first interruption signal played for 29 seconds before it was accessed, and the subject took the full 6 seconds to attend to the hint, there would still be an additional 10 seconds before the second interruption occurred.)

Our interruption frequency of 65 seconds falls within the range used in previous work [1, 3, 8, 9, 11, 18, 22], in which frequency varied from 3 seconds to 5 minutes with an average of 2 minutes. We balanced maximizing the

number of replications of each hint utility (i.e., to give subjects the opportunity to comprehend the relationship between notification signal and hint utility in the Match condition) and minimizing the duration of play to avoid excessive subject fatigue.

Expected Causes 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. To elicit the latter type of annoyance, following each of the Match and Static conditions, subjects were informed of the number and types of hints that were missed during that condition.

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. Level 1 was also nested with three notification signals. To minimize order effects, we fully counterbalanced the order of presentation of the conditions, producing six configurations.

Participants

Twenty-four subjects (15 female) between 18 and 39 years of age participated in the experiment 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 Experiment 1, as well as through advertisements posted throughout the university campus. None of the subjects participated in Experiment 1.

Motivation

Subjects were told that an extra $10 would be provided to the 1/3 of the subjects who made the most number of matches over all three conditions. 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.

Apparatus

The apparatus was identical to Experiment 1.

Procedure

The experiment was designed to fit in a single 2-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. (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 dialog 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. (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.

Measures

Our main dependant measures were perceived benefit, annoyance, workload, and performance. Performance was measured as the number of matches made in each condition. The remaining three measures were self-reported through questionnaires, which also elicited fatigue ratings on a 5-point Likert scale.

To assess workload, we used the NASA-TLX scales [14], a standardized instrument for assessing various dimensions of workload. Perceived benefit and annoyance were assessed through additional questions we added to the TLX in a manner similar to [1], where subjects rated statements on a 20-point scale. The statements rated were as follows:

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

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

Because piloting indicated that good performance tended to mitigate annoyance specific to the interruptions, we defined two measures of annoyance: one related to the task in general, and one specific to the interruptions. We report on only the latter measure in this paper.

Secondary measures were gathered in 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.

Detection times for the notification signals and the number of missed hints were also measured.

Hypotheses

Our hypotheses were as follows:

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, 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 and Static conditions. H3 and H4 concern all three conditions.
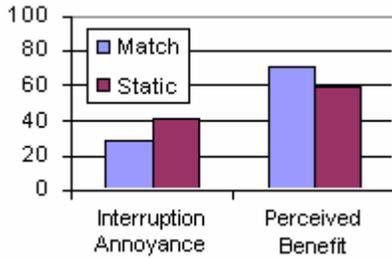
Figure 2. Mean ratings for interruption annoyance and benefit (scale: 5-100) (N=20).

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 3. Results of ANOVA on NASA-TLX workload measures (N=20).

RESULTS

Data for four outlier subjects were removed from the analysis.[1] Statistical adjustment strategies were identical to those employed in Experiment 1, and we again report effect sizes.

To test H1 and H2, a 2 (condition: Match, Static) by 2 (presentation order) ANOVA was performed on the annoyance and benefit ratings. Results for these ratings are illustrated in Figure 2. 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.

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. Results for the NASA-TLX workload measures in Table 3 show no significant differences among the three conditions for any of the workload measures. Furthermore, no effect of presentation order was present. This was consistent with our hypothesis

---

[1] In order to ensure that subjects saw enough interruptions to perceive the difference between the Match and Static conditions, outliers were defined as subjects whose number of missed hints was more than two standard deviations from the mean in either condition. In the Static condition we counted the total number of hints missed. In the Match condition we considered only high-utility hints, since subjects who deciphered the signal-utility relationship could ignore low and medium-utility hints.
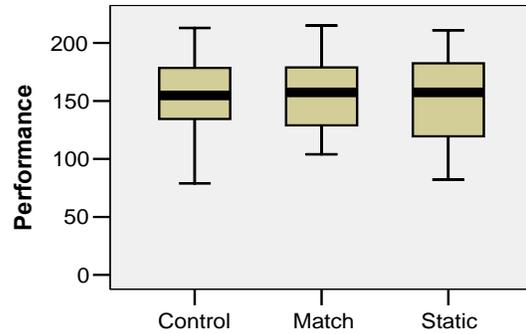

Figure 3. Boxplot of condition versus performance (N=20); the line through each box represents the median. Performance is measured as the number of matches made.

Dependent Variable	M	S	C	df	Chi-square	p
Preferred overall	15	3	2	2	15.70*	<.001
	M	**S**	**Same**			
More helpful	12	3	5	2	6.70*	.035
More hindering	2	11	7	2	6.10*	.047

Table 4. Chi-square statistic for qualitative results (M=match; S=static; C=control; Same=same amount of help/hindrance in both match and static) (N=20).

H3 in which we speculated that workload would be no worse in the Match condition than in the other conditions.

Performance results are presented in Figure 3. Counter to our hypothesis H4, condition did not impact performance ($F(2, 28) = .812$, $p = .454$, $\eta^2 = .055$). However, an interaction effect of condition and presentation order ($F(10, 28) = 2.035$, $p = .068$, $\eta^2 = .421$) approached significance with a large effect size, but large individual differences and sparse data per cell revealed no clear trends. Not unexpectedly, subjects performed (borderline) significantly better as the experiment progressed, i.e., a learning effect on performance ($F(2,38) = 3.171$, $p = .053$, $\eta^2 = .143$). They also reported significantly greater fatigue over time ($F(2,38) = 5.327$, $p = .009$, $\eta^2 = .219$), where they were more fatigued in the third block than the first ($p = .017$).

Detection times for the notification signals were comparable to Experiment 1, and the number of missed hints was not a concern. Full details can be found in [12].

Interview Results

We calculated the Chi-square statistic for preference, helpfulness, and hindrance responses. A summary of the results in Table 4 shows that 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.

The interviews revealed that 25% of subjects made no comprehension of the relationship between the hints and the degree of AD in the Match condition. 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.

The interruption conditions also shaped subject perception of the different types of hints. In surveys distributed following each condition, we asked what aspects of the interruptions annoyed subjects during that condition. In the Static condition, 85% of subjects indicated being annoyed by the low-utility hints. In the Match condition, only 60% of subjects admitted to being annoyed by the low-utility hints, including 20% who stated that annoyance associated with low-utility hints lowered significantly – if not ceased – once subjects purposely began 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-AD 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-AD 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 15% of subjects – those who comprehended and trusted the relationship – appreciated the subtlety of the low-AD signal because it was easy to ignore. The remaining 55% gave no opinion about the low-AD signal.

The advantages of matching AD and utility were best summarized by two subjects. One said of the high-AD 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."

Summary of Results

We summarize our results according to our hypotheses:

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.

DISCUSSION

Perception of Notification Signals

The differences in qualitative feedback on the notification signals between Experiments 1 and 2 highlight the importance of context in interruption systems. In Experiment 1, where notifications had no content and were irrelevant to the task, the signal with highest attentional draw (AD) was perceived by subjects to be the most annoying (82%), while the signal with lowest AD was ranked as least annoying (55%). When content and utility became a factor in Experiment 2, perceptions reversed. The signal with high AD fell into favour with subjects (35%) who realized that its content improved their primary task performance. Conversely, the low-AD signal drew mixed reviews: subjects who either did not comprehend the relationship between utility and AD, or who comprehended but did not trust it, complained that the low-AD 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 also highlight the significance of Billsus et al.'s [5] 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 experiment shows, when the utility of an interruption is known, 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.

Performance and Workload

Hart and Staveland [14] define workload as the cost incurred by a user to achieve a particular level of performance; thus, 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. If the interruption content increases performance on the primary task, however, there is a potential to actually reduce workload. Our results showed that interruption boosted task performance enough to mitigate the increase in cost such that workload under interruption was no worse than workload in the no-interruption condition. Although studying the performance impacts of helpful interruptions was not the primary goal of our experiment, we had hoped that our matched interruption presentation strategy would yield performance benefits in addition to improving annoyance and benefit as well as balancing workload. Unfortunately, fatigue, learning, and interaction effects made it impossible to interpret the performance results. Future work is required to determine if performance gains can be expected when interruptions are specific to the primary task.

Although neither performance nor workload varied significantly across the conditions, annoyance and benefit responses were significantly better in the Match condition than in the Static 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 effects would have been even stronger.

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 Experiment 1, and also distinguishes our work from previous research investigating interruption content with varying levels of relevance [9]. 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 more than three levels of utility.

Further study is necessary to understand the tradeoffs between increasing the set of notification signals to permit a wider range utilities to be conveyed and the potential cognitive overhead associated with having to interpret the meaning behind this increased set. Going beyond three levels of AD likely requires notification signals with some continuous property (e.g., for a signal that uses motion, velocity) that can be manipulated to convey multiple levels of AD using the same signal. In the motion example, users would not be expected to recognize differences in velocity; rather, faster velocities would simply grab user attention more quickly.

Finally, there is the question of generalizability of scope and context. We examined utility in the scope of a primary task. Another option would be to define utility in terms of personal relevance to the user, using content typically delivered via personal systems such as IM, email, or calendar software. We hypothesize that our results could generalize to these contexts, but further research is needed

and determining objective levels of utility in such contexts may be very difficult.

Design Implications
The goal of our work was to explore the validity of Obermayer and Nugent's design guideline to match the amount of attention attracted by a notification signal to the utility of interruption content. In our research, identical interruptions were presented to subjects, and 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 design solution can in fact provide significant improvement over current methods of interruption with static notification signals, suggesting that the value of Obermayer and Nugent's design guidance has been underestimated in past research [23]. 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.

Mixed-initiative and recommender systems capable of assessing utility do currently exist [5, 6]; auspiciously, these are the types of systems for which a positive perception of interruption is most crucial. Alternatively, when interruptions are human-generated, senders can designate utility [27]. 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 our preliminary investigation [12] 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 on 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., [1, 11, 18]): 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 interruption.

CONCLUSIONS
We conducted an empirical investigation to examine the effects of matching attentional draw (AD) of notification to interruption utility in terms of annoyance, perceived benefit, workload, and performance. Our results indicate that when interruption utility is known, interfaces that vary

AD 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. Because we emulated a mixed-initiative context, we expect our findings to apply most readily to mixed-initiative and recommender systems that are able to appraise utility. The generalizability to other interruption systems will depend on the eventual availability of reasonable utility estimates.

Future study is recommended to define notification signals that can maximize the number of signal-utility pairs without cognitively overloading users. Further work is also motivated in computational assessment of utility so that the multi-level AD 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. Future study may also investigate whether our results generalize to utility in the context of personally relevant interruptions (e.g., IM or email systems). This research should examine the matching strategy in more realistic contexts in order to develop specific guidelines that will be appropriate for a wide range of applications.

ACKNOWLEDGEMENTS

REFERENCES

1. Adamczyk, P. D., and Bailey, B. P. (2004). If not now, when?: the effects of interruption at different moments within task execution. *CHI'04*, 271–278.

2. Avrahami, D., and Hudson, S. E. (2004). QnA: augmenting an instant messaging client to balance user responsiveness and performance. *CSCW 2004*, 515–518.

3. Bailey, B. P., Konstan, J. A, and Carlis, J. V. (2001). The effects of interruptions on task performance, annoyance, and anxiety in the user interface. *INTERACT 2001*, 593–601.

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

5. Billsus, D., Hilbert, D. M., and Maynes-Aminzade, D. (2005). Improving proactive information systems. *IUI 2005,* 159–166.

6. Bunt, A. (2005). User modelling to support user customization. *UM 2005*, 499–501.

7. Cohen, J. (1973). Eta-squared and partial eta-squared in communication science. *Human Communication Research*, 28:473–490.

8. Cutrell, E., Czerwinski, M., and Horvitz, E. (2001). Notification, disruption, and memory: effects of messaging interruptions on memory and performance. *INTERACT 2001*, 263–269.

9. Czerwinski, M., Cutrell, E., and Horvitz, E.. (2000). Instant messaging and interruption: influence of task type on performance. *OZCHI 2000*, 356–361.

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

11. Fogarty, J., Hudson, S. E., and Lai., J. (2004). Examining the robustness of sensor-based statistical models of human interruptibility. *CHI'04*, 207–214.

12. Gluck, J. (2006). An investigation of the effects of matching attentional draw with utility in computer-based interruption. M.Sc Thesis, University of British Columbia, Canada.

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

14. Hart, S. G., and Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): results of empirical and theoretical research. In P. Hancock & N. Meshkati (Eds.), *Human Mental Workload* (pp. 139-183). Amsterdam: Elsevier Science.

15. Horvitz, E. (1999). Principles of mixed-initiative user interfaces. *CH'99*, 159–166.

16. Horvitz, E., Apacible, J. Learning and reasoning about interruption. *ICMI'03*, 20–27.

17. Horvitz, E., Jacobs, A., and Subramani, M. (2005). Balancing Awareness and Interruption: Investigation of Notification Deferral Policies.. *UM 2005*, 433-437.

18. Iqbal, S. T., and Bailey, B. P. (2006). Leveraging characteristics of task structure to predict the cost of interruption. *CHI'06*, 741–750.

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

20. Levitt, J. (2001). Internet Zone: Good help is hard to find. Information Week: Listening Post. http://www.informationweek.com/835/35uwjl.htm

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

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

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

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

25. Obermayer, R. W, and Nugent, W. A. (2000). Human-computer interaction for alert warning and attention allocation systems of the multi-modal watchstation. *SPIE, Vol. 4126*, 14–22.

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

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

28. van Dantzich, M., Robbins, D., Horvitz, E., and Czerwinski, M. (2002). Scope: Providing awareness of multiple notifications at a glance. *AVI 2002*.