Leveraging Changes in Mental Workload during Task Execution to Mitigate Effects of Interruption

BRIAN P. BAILEY AND SHAMSI T. IQBAL
University of Illinois at Urbana-Champaign

Task interruption often has a significant negative impact on a user's productivity and affective state. Cognitive theorists have argued that interrupting the execution of primary tasks at moments of lower mental workload would mitigate effects of interruption, yet knowing just where these moments occur remains elusive. In this article, we present empirical results from three user experiments aimed at further explicating the relationship among mental workload, task execution, and effects of interruption. Principal results include (i) a user's mental workload exhibits momentary decreases at subtask boundaries during task execution; (ii) workload decreases more at boundaries higher in a task model than at boundaries lower in a model; (iii) the decrease in workload is not uniform within the same level of a model; and (iv) interrupting task execution at subtask boundaries with lower workload meaningfully mitigates effects of interruption. These results demonstrate that mental workload is an effective predictor of how opportune different moments in a task are for interruption and contribute further answers to where moments of lower workload exist within the structure of a task. Practical implications of these results are discussed for computational systems that reason about when to interrupt users engaged in tasks.

Categories and Subject Descriptors: H.5.2 [**Information Interfaces and Presentation**]: User Interfaces - *evaluation/methodology*, *user-centered design*, H.1.2 [**Models and Principles**]: User/Machine Systems - *Human Information Processing*
General Terms: Design, Experimentation, Human Factors, Measurement
Additional Key Words and Phrases: Attention, GOMS, Interruption, Mental Workload, Pupil Size, Task Models, User Studies

1. INTRODUCTION

An important and challenging problem in many multi-tasking environments is managing interruption (McFarlane & Latorella, 2002). Proactive systems executing in environments such as aviation cockpits (Dismukes, Young, & Sumwalt, 1998; Latorella, 1996), control rooms (Stanton, 1994), in-vehicle displays (Lee, Hoffman, & Hayes, 2004) and office environments (Bailey & Konstan, 2005; Czerwinski, Cutrell, & Horvitz, 2000b; Jackson, Dawson, & Wilson, 2001) are increasingly interrupting a user's primary tasks. When primary tasks are interrupted at random moments, users take longer to complete the tasks (Bailey & Konstan, 2005; Czerwinski, Cutrell, & Horvitz, 2000a; Rubinstein, Meyer, &

Left blank for Author addresses and copyright notice

Meyer, 2001), commit more errors (Kreifeldt & McCarthy, 1981; Latorella, 1996) and experience increased levels of frustration, annoyance, and anxiety (Adamczyk & Bailey, 2004; Bailey & Konstan, 2005; Zijlstra, Roe, Leonora, & Krediet, 1999).

At the same time, users often desire or need the numerous benefits that proactive systems provide, e.g., supporting near instant communication (Czerwinski et al., 2000a; Dabbish & Kraut, 2004; Latorella, 1996), maintaining awareness of peripheral information (Maglio & Campbell, 2003), being reminded of upcoming activities (Dey & Abowd, 2000), or learning to perform complex tasks (Maes, 1994; Rich & Sidner, 1998).

To maintain benefits of proactive systems while mitigating effects of interruption, systems are being developed that are capable of deferring delivery of information until the cost of interruption is low, e.g., see (Fogarty et al., 2005; Horvitz & Apacible, 2003; Hudson et al., 2003). In these systems, the foremost method used for computing the cost of interruption is to build a probabilistic model based on the use of cues such as desktop activity, visual and acoustical analysis of the task environment, and scheduled activities of the user. However, to compute a more accurate cost of interruption, systems should consider a user' mental workload during task execution (Iqbal & Bailey, 2005).

Researchers have often argued that interruptions would be less disruptive if they occurred during moments of low mental workload (Czerwinski et al., 2000b; Miyata & Norman, 1986; Monk, Boehm-Davis, & Trafton, 2002) and posit that these "opportune" moments occur at subtask boundaries (Miyata & Norman, 1986). However, these theoretical claims have not been empirically validated. Also, since tasks can be hierarchically decomposed, creating many boundaries at many levels in a task model (Card, Moran, & Newell, 1983), it is unclear as to *which* boundaries would be most opportune for interruption, if any.

This article presents empirical results from three user experiments conducted to further explicate the often nebulous relationship among mental workload, task execution, and effects of interruption. The central approach was to align a physiological measure of mental workload (pupil dilation) to hierarchical models of task execution, constructed and validated using established techniques, identify where moments of lower workload occur and any patterns based on the structure of a task that could be reliably generalized, and then test whether these moments are in fact more opportune for interruption.

The principal results of this work include (i) mental workload exhibits momentary decreases at subtask boundaries during task execution; (ii) workload decreases more at boundaries higher in a task model than at boundaries lower in a model; (iii) the decrease

in workload is not uniform within the same level of a task model; and (iv) interrupting task execution at subtask boundaries with lower workload meaningfully mitigates effects of interruption along dimensions of resumption lag, subjective ratings of annoyance, and attribution of respect to the interrupting application.

The latter result (iv) demonstrates that mental workload is an effective predictor of how opportune different moments in a task are for interruption, providing further evidence that systems should integrate consideration of workload into their reasoning framework. A pragmatic approach would be to leverage our former results (i-iii) to encode reasonable approximations of workload within machine-parsable descriptions of tasks. These task descriptions could then be computationally matched to task execution, providing reasonable approximations of workload absent a real-time, external measure.

Though parts of this work have been previously published (Iqbal, Adamczyk, Zheng, & Bailey, 2005; Iqbal & Bailey, 2005; Iqbal, Zheng, & Bailey, 2004), this article provides a coherent synthesis of that work, provides some extended analysis, discusses practical implications of our empirical results for systems that reason about interruption, and situates our results within contemporary cognitive theories of human attention.

2. RELATED WORK

In this section, we discuss effects of interruption, posited moments for interruption that would mitigate those effects, how our work contributes to systems that reason about interruption, and rationale for using pupil size as the measure of workload in this work.

2.1 Effects of Interruption

Experimental studies have convincingly demonstrated that interrupting users engaged in tasks has considerable negative impact on task completion time (Cutrell, Czerwinski, & Horvitz, 2001; Czerwinski et al., 2000a, 2000b; Kreifeldt & McCarthy, 1981; McFarlane, 1999; Monk et al., 2002), error rate (Latorella, 1998), decision making (Speier, Valacich, & Vessey, 1999), and affective state (Adamczyk & Bailey, 2004; Bailey & Konstan, 2005; Zijlstra et al., 1999). For example, when peripheral tasks are delivered at random moments during primary tasks, users can take up to 30% longer to complete them, commit up to twice the errors, and experience up to twice the increase in negative affect than when delivered at opportune moments (Bailey & Konstan, 2005).

In safety critical domains, a small response delay or error committed due to an ill-timed interruption could cost lives or cause catastrophic damage (McFarlane & Latorella,

2002). In office settings, increases in frustration, annoyance, and anxiety due to poorly timed interruption unnecessarily degrades the user experience (Shneiderman, 1997).

Our work advances theoretical understanding of how opportune different moments in a task are for interruption. Systems should be able to leverage our results to better reason about when to deliver information such that effects of interruption would be mitigated.

2.2 Mitigating Effects of Interruption

Researchers have theorized that interruptions would have less negative impact (cost) if they occurred at moments of lower mental workload during task execution, and that these moments occur at (sub)task boundaries (Miyata & Norman, 1986). While studies show that scheduling interruptions at certain boundaries or other moments purported to be of low workload can often mitigate effects of interruption (Adamczyk & Bailey, 2004; Bailey & Konstan, 2005; Cutrell et al., 2001; Czerwinski et al., 2000a, 2000b), researchers can only *assume* why, as ground truth for a user's mental workload is rarely available. This makes it difficult to produce general guidelines about when to interrupt.

In addition, interactive tasks can be hierarchically decomposed into recursive patterns of goal formulation, execution, and evaluation, creating many boundaries at many levels in a task model (Card et al., 1983). It is thus unclear as to *which* of these boundaries would be most opportune for interruption, if any.

By monitoring a user's mental workload during task execution, our work contributes further answers toward knowing just where moments of lower workload occur within the structure of a complex task and provides consistent evidence that interrupting tasks at those moments mitigates effects of interruption.

2.3 Systems that Reason about Interruption

Systems are being developed that can computationally reason about appropriate moments for interrupting users engaged in tasks, e.g., (Bailey, Adamczyk, Chang, & Chilson, 2005; Fogarty et al., 2005; Horvitz & Apacible, 2003; Hudson et al., 2003). The general approach is to weigh the value of delivering information against the cost of interrupting the primary task (Horvitz, Jacobs, & Hovel, 1999), where the focus of current research has been on computing an accurate cost of interruption. Systems typically compute the cost of interruption using non task specific cues such as desktop activity, visual and acoustical analysis of the physical task environment, and scheduled activities of the user.

Our work argues for systems to integrate knowledge of a user's mental workload into their reasoning framework by further demonstrating that the effects of interruption are

strongly tied to workload. We also describe several strategies for integrating workload, including strategies that do not require use of a direct physiological measure. Considering a user's mental workload would allow systems to compute a much more accurate cost of interruption, thus enabling more effective decisions to be made about when to interrupt.

2.4 Use of Pupil Size as our Measure of Workload

Our work required a measure of mental workload and any measure could have been used such that it was continuous, immediate, low latency, and valid. After a thorough review of the literature, combined with availability of needed equipment, we selected pupil dilation as our measure for this work.

Under conditions of controlled illumination, research shows that pupil dilation is a valid and reliable measure of mental workload (Beatty, 1982; Hess & Polt, 1964; Hoecks & Levelt, 1993; Juris & Velden, 1977; Kahneman, 1967; Marshall, 2002; Nakayama & Takahashi, 2002; Takahashi, Nakayama, & Shimizu, 2000). Beatty reviewed a very large corpus of experimental data and concluded that pupil dilation is a reliable indicator of mental workload, that relative increases in pupil size correlate with increases in workload, and that this holds true across tasks and individuals (Beatty, 1982).

Researchers have also investigated many other measures of workload, including event related brain potential (ERP) (Donchin, Kramer, & Wickens, 1986; Kok, 1997; Kramer, Schneider, Fisk, & Donchin, 1986), electroencephalographic activity (Gale & Edwards, 1983; Gevins & Schaffer, 1980; Phelps & Mazziotta, 1985), eye movement and blink rate (Takahashi et al., 2000), heart rate variance (Rowe, Sibert, & Irwin, 1998), performance measures (O'Donnell & Eggemeier, 1986), and subjective ratings (Hart & Staveland, 1988; Yeh & Wickens, 1988).

Relative to these measures, the use of pupil dilation offers many advantages (Kramer, 1991); it is *continuous*, meaning that it provides a steady stream of workload data; it measures allocation of resources in a *holistic* manner rather than specific pools; it has *low latency*, usually responding to a change in workload in about 500 ms; and it is *immediate*, a few recent samples indicates current workload, which simplifies analysis of the data. However, careful experimental control must be maintained with pupil dilation, as it can be considerably affected by factors such as ambient illumination and screen luminance.

A limitation of pupil dilation is that it has rarely been used in interactive computing environments. Thus, the first step of our research was to test whether pupil dilation would correlate with the workload of interactive tasks.

3. EXPERIMENTAL ROADMAP

This article presents results from three user experiments aimed at further explicating the relationship among mental workload, task execution, and effects of interruption. Though it is well established that pupil dilation correlates with mental workload for non-interactive tasks (Beatty, 1982; Hoecks & Levelt, 1993; Juris & Velden, 1977), our first experiment sought to determine whether this correlation would hold for interactive tasks, which also had more complex execution structures than many tasks previously studied. Our results showed that pupil dilation does correlate with the workload induced by interactive tasks, assuming appropriate controls, and suggests that to better understand how workload changes during execution of tasks with complex structures, the response data should be aligned to corresponding models of task execution.

Building on our first study, our second experiment focused on further understanding how mental workload changes during task execution. Users performed tasks while their workload was continuously monitored using pupil dilation. The response data was aligned to corresponding GOMS models of the tasks and statistically analyzed, focusing on subtask boundaries (Miyata & Norman, 1986). Analysis showed that workload exhibits momentary decreases at subtask boundaries and that it decreases more at boundaries higher in a task model than at boundaries lower in the model.

Our third experiment tested whether moments of lower workload, as identified from the workload-aligned models, were more opportune for interruption than other moments. We compared effects of interrupting task execution at boundaries with lower workload, boundaries with higher workload, and random points. Results showed that interrupting at lower workload boundaries allowed users to resume tasks faster, experience less negative affect, and attribute more respect to the interrupting application.

4. EXPLORE USE OF PUPIL DILATION AS A MEASURE OF WORKLOAD IN INTERACTIVE COMPUTING ENVIRONMENTS

Although pupil size has been shown to correlate well with workload induced by stimulus-driven tasks (Beatty, 1982; Hess & Polt, 1964; Hoecks & Levelt, 1993; Hytintk, Tommola, & Alaja, 1995; Juris & Velden, 1977), it is not known whether this correlation holds in interactive environments and for tasks with more complex structures than those previously studied. Our first study was thus designed to answer the following questions:

- How well does pupil dilation correlate with the workload induced by interactive tasks with different levels of difficulty?
- Does this correlation hold across several categories of tasks?

4.1 Users and Equipment

Twelve users (6 female) participated in the study and the average age was 24 years (SD=3.23). As a user performed the tasks, their pupil data was recorded using a head-mounted eye tracking system (Eyelink II). The eye-tracker sampled the pupil at 250 HZ with spatial accuracy to about $1/100^{th}$ of a millimeter (for an average 5 mm pupil) using corneal reflection. Lighting and noise levels of the room were well controlled.

4.2 Tasks

Based on a literature review, an informal questionnaire to eight users, and our own experience, we developed four representative categories of tasks: Email Classification, Reading Comprehension, Mathematical Reasoning, and Search. Each category had two levels of difficulty – Easy and Difficult. Our expectation was that the difference in task workload between the difficulty levels would cause a difference in pupillary response. The four task categories were:

- *Email Classification.* A user had to drag emails and drop them into appropriate folders, based on classification rules; see Figure 1(a). For the easier task, the rules were specific such as using the size of the email, e.g., 1K, 2K, or 3K. For the more difficult task, the rules were less specific, requiring each email to be classified by inferring its topic from the subject header, e.g., travel, course related, fun and humor, announcements, etc.
- *Reading Comprehension.* A user read a given text and answered a few questions about its content; see Figure 1(b). We used the Fry Formula (Fry, 1968) to ensure the texts differed substantially in reading difficulty. The easier task was rated at a grade 9 level while the more difficult task was rated at a grade 17 level.
- *Mathematical Reasoning.* A user performed mathematical calculations; see Figure 1(c). For the easier task, a user mentally added two four digit numbers and then selected the correct answer from a list of three choices. For the more difficult task, a user mentally added 4 five-digit numbers, retained the result in memory, and decided whether the result exceeded a given number.

- *Searching.* A user searched for a specific product from a list of similar products given a set of constraints; see Figure 1(d). For the easier task, a user had to find the product from a list of seven products given just one constraint, e.g., the cheapest camera. For the more difficult task, a user had to identify the correct product using multiple constraints, e.g., the cheapest 3MP camera with 3X digital zoom.

4.3 Procedure

Upon arrival at the lab, the user filled out a background questionnaire and was given general instructions. The eye tracking system was then configured and calibrated. A user performed eight tasks – one easy and one difficult task from each of the four categories. At the beginning of each category, the user received specific instructions and performed a practice. Baseline pupil size was collected by having the user fixate on a blank task screen for a few seconds. The actual tasks were then performed. After completing each task, the user rated its difficulty on a 1-5 scale (1=very easy, 5=very difficult). The presentation order of task category and tasks within each category were randomized. Users were instructed to perform the tasks as quickly and as accurately as possible. The system logged task performance and screen interaction was recorded for later analysis.

4.4 Measurements

A user's subjective rating and task completion time were collected to validate task workload associated with each level of task difficulty. A user's pupil data and on-screen interaction were recorded separately, but were synchronized by correlating timestamps.

For each user, we computed the *percentage change in pupil size* (PCPS), which is the measured pupil size at each sample minus the baseline, divided by the baseline. This calculation is consistent with (Hess, 1972) and normalizes for users having different baselines. Also, the easy and difficult task screens were carefully designed to avoid major differences in overall screen luminance. The average PCPS (APCPS) from the beginning to end of each task was used as the task-evoked pupillary response.

4.5 Results

A 4 Category (Classification, Comprehension, Reasoning, and Search) x 2 Level of Difficulty (Easy and Difficult) repeated measures ANOVA was performed on the data.

4.5.1 Validation of Task Workload

We used task completion time (Figure 2) and subjective ratings of difficulty (Figure 3) to validate task workload. An ANOVA showed that Category had a main effect on task completion time ($F(3,33)=30.07$, $p<0.0005$). Post hoc tests showed that users spent more time on Comprehension ($\mu=66.5$s) than Classification ($\mu=33.3$s, $p<0.001$), Reasoning ($\mu=27.9$s, $p<0.003$) and Search tasks ($\mu=25.2$s, $p<0.001$). Users also spent more time on Classification than Reasoning ($p<0.0005$) and Search tasks ($p<0.001$).

Difficulty also had a main effect on task completion time ($F(1,11)=190.9$, $p<0.0005$). Users spent more time on difficult tasks and post-hoc comparisons showed that this effect existed for all but the Reasoning tasks. The interaction between Category and Difficulty ($F(3,33)=6.75$, $p<0.001$) was significant, mainly due to the Reading category.

For subjective ratings, Difficulty had a main effect ($F(1,11)=34.91$, $p<0.0005$), with higher ratings for the more difficult task in each category. An interaction between Category and Difficulty was detected ($F(3,33)=4.12$, $p<0.014$), mainly due to the easier task in the Classification category. There was no main effect of Category.

These results validate that the more difficult task had higher task workload than the easier task in each category, and thus a similar pattern is expected for pupillary response.

4.5.2 Effects of Task Workload on Pupillary Response

For this analysis, we computed the average percent change in pupillary response (APCPS) for the duration of a task, summarized in Figure 4. An ANOVA showed that Category had a main effect on APCPS ($F(3,33)=4.74$, $p<0.007$). This result was not unexpected, as we did not design each category to have the same average task workload.

Surprisingly, Difficulty did not have a main effect ($F(1,11)=3.12$, $p<0.11$). T-tests between the Easy and Difficult tasks within each category revealed a difference only for Search ($p<0.025$). This was inconsistent with our expectation, especially since the ratings and completion times suggested that a difference should exist between difficulty levels.

Except for the Search task, which induced sustained mental effort, the other tasks could be hierarchically decomposed into multiple subtasks. Since only certain subtasks differed in terms of the cognitive effort required, averaging PCPS over the duration of an entire task may dilute the effects of those subtasks. If pupillary response for only those subtasks with different cognitive demands were considered, then the expected differences between levels of difficulty might be found. This was the focus of our second analysis.

4.5.3 Effects of Cognitive Subtasks on Pupillary Response

We performed a GOMS analysis to decompose the Classification, Comprehension, and Reasoning tasks into their component goals and operators (Card et al., 1983), collectively referred to as *subtasks*. The decomposition continued until there was no observable or meaningful separation between subtasks. The models are shown in Figure 5. Accuracy was measured by how well the models predicted the interaction sequences in the videos. On average, the models were about 95% accurate and there was no pattern to the errors.

Classification was decomposed into three first-level (L1) subtasks, with the third subtask repeated being eight times. The third L1 subtask was further decomposed into two second-level (L2) subtasks. Between easy and difficult Classification tasks, Select Folder was the only subtask expected to induce meaningfully different workload. The Reasoning and Comprehension tasks were decomposed using a similar approach. Between easy and difficult Reasoning tasks, Add Column was the only one expected to induce different workload and, for Comprehension, Read Text was the only one expected to induce different workload. APCPS was then calculated using just these subtasks.

Paired t-tests showed that the cognitive subtasks did induce higher pupil dilation than the other parts of the task ($p<0.01$, $p<0.18$, $p<0.001$ for Classification, Comprehension, and Reasoning, respectively). Although the difference for Comprehension was not significant, the trends were in the right direction. These results indicate that pupil dilation is sensitive to the changing workload demands of subtasks during task execution.

Including only the specified subtasks, we performed ANOVAs similar to our first analysis. Difficulty now had an effect on APCPS ($F(1,11)=4.97$, $p<0.048$). As shown in Figure 6, the more difficult tasks induced higher APCPS ($\mu=8.36$) than the easier tasks ($\mu=6.85$). Paired t-tests showed that a difference existed between the easy ($\mu=7.14$) and difficult ($\mu=10.03$, $p<0.021$) subtasks for Classification. For Comprehension and Reasoning, differences between easy and difficult tasks were not significant ($p<0.14$ and $p<0.79$), but the trends were in the expected direction. While the easy and difficult tasks for these categories differed along completion time and subjective ratings, the differences were apparently not enough to cause detectable changes in pupillary response.

4.6 Discussion

Our results show that pupillary response can be used to reliably measure mental workload for interactive tasks, assuming appropriate environmental controls. For sustained effort tasks, average pupil dilation correlates well with the overall task difficulty. For tasks that

require varying mental effort during execution, pupil dilation demonstrates transient changes that are concomitant with the varying demands of the task, and increased pupil dilation occurs for the more cognitively demanding subtasks. Although measures such as task completion time and user ratings provide overall measures of workload, they do not reflect the changes in workload that a user experiences during task execution (Yeh & Wickens, 1988) and, as our results show, these changes can be meaningfully different.

Our results suggest that to better understand how workload changes during tasks with complex structures, pupillary response data should be aligned to the corresponding models of task execution. When comparing workload between tasks, for example, this would allow subtasks that are non-cognitive or that require similar mental effort to be filtered, allowing for a more effective comparison. Also, this would enable investigation of whether there are detectable decreases in workload at subtask boundaries (Miyata & Norman, 1986), where a user has just completed one subtask and begins activating action schema for the next (Altmann & Trafton, 2002). This was the focus of our next study.

5. UNDERSTAND WORKLOAD CHANGES DURING TASK EXECUTION

Miyata & Norman (1986) have theorized that mental workload should decrease at subtask boundaries during task execution, as there is a momentary shift in cognitive activity, but empirical evidence has never been provided. Also, since tasks can be decomposed into recursive patterns of goal formulation, execution, and evaluation (Card et al., 1983), creating many boundaries at many levels in the task model, it is unclear whether the changes in workload might differ among the boundaries. Building upon lessons from our first study, we wanted to analyze how workload changes during (sub)task execution, focusing on subtask boundaries. Our experiment was designed to answer the following questions:

- How much does a user's mental workload change during subtasks? How much does this change depend on the level in the task model and the type of the subtask?
- How much lower is a user's mental workload at boundaries compared to the average mental workload during execution of subtasks (non-boundary moments)? How much does this change differ for boundaries at different levels in a task model?

5.1 Users and Equipment

Twelve users (1 female) participated in this study and their ages ranged from 23 to 50 (M=27.1, SD=7.45). All users had normal or corrected-to-normal vision. Pupil size was measured using the same eye-tracking system discussed in the first experiment.

5.2 Tasks

Two tasks were developed; Route Planning and Document Editing. For the Route Planning task (Figure 7), a user was given a map with two routes between two cities marked by start/end stars. For each route, there were three segments from the source to the destination. A distance and fare were associated with each segment, and were available through a tooltip that appeared when the user moved the mouse over a segment.

To perform the task, the user moved the mouse over a route segment in the map, committed the distance and fare information that appeared in the tooltip to memory, and entered the data into the corresponding row in the table. When the user moved the mouse away from the segment, the tooltip disappeared. A user completed each row in the table, mentally added the distance and fare columns, and entered the results into the last row. The user repeated this process for the second table and route. After completing the tables, the user selected the shorter and the cheaper of the two routes from drop down lists.

The main cognitive subtasks were storing information from the map to working memory (Store), recalling information for the table (Recall), and adding the numbers (Reasoning). Comparing distance and fare totals and deciding the shorter and cheaper routes also involved reasoning processes.

For Document Editing (Figure 8), a user was given a text document with three annotations. The text was about the social hierarchy of a common pet (cats). This topic was selected because we felt it would be interesting and familiar to most users. A user edited the document according to each comment, which appeared as a tooltip when the mouse was moved over the corresponding highlight. After reading a comment, the user located the text, made the appropriate edit, and repeated two more times. The user saved the document to a specified directory and file name, given a priori. The main cognitive subtasks were understanding comments (Language Comprehension), making edits (Language Generation), and recalling the given directory and file names (Recall).

The tasks were carefully designed to have meaningful subtasks of varying difficulty, a prescribed execution sequence, well defined boundaries between subtasks, and a representative sample of interaction. Although users do not typically follow a prescribed

execution sequence when performing tasks, we had to control the sequence in order to align workload to the models of task execution. The lower-level subtasks were representative of those within many interactive tasks, e.g., selection, memory store and recall, data entry, reasoning, comprehension, and motor control. Each task required about 5 minutes to perform, which is considerably longer than tasks used in many prior studies, e.g., (Bradshaw, 1967; Hytintk et al., 1995; Juris & Velden, 1977; Kahneman, 1967; Takahashi et al., 2000), resulting in several thousand data points being generated per task. These tasks also had more complex execution structures than those used in Experiment 1.

5.3 Procedure

Upon arrival at the lab, a user completed a questionnaire for background information, and received general instructions for the tasks. After questions were answered, we set up the eye-tracker and calibrated the system. At the start of the session, the user was given specific instructions and performed practice tasks. Just before each task, we collected baseline pupil size by having the user fixate on a blank task screen for a few seconds. The user was asked to perform the tasks as quickly and accurately as possible. The ordering of the tasks was counterbalanced. Pupil data was logged to time stamped files while a user's screen interaction was video recorded with eye gaze overlaid. Because the videos and pupil data received time stamps from the same clock, we could precisely align them. The entire experimental session lasted about 30 minutes.

5.4 Task Models and Validation

Figure 9 shows the task model for the Route Planning task, reusing repetitive parts for brevity. The full task model has 4 levels and 81 nodes. The term *subtask* refers to any node in the task model and *subtask boundary* refers to the period between consecutive subtasks. *Level of boundary* between two consecutive subtasks is 1 + the depth of their shared ancestor in the model. For example, in Figure 9, consider the "Locate segment" and "Store data" subtasks at the left of level 4. When a user completes the "Locate segment" subtask and moves to "Store data", this defines a level 4 boundary, since the depth of their shared ancestor "Retrieve segment" is (1 +) 3. When a user completes the "Store data" subtask and then moves to "Recall", this defines a level 3 boundary, since the depth of their shared ancestor "Enter data for segment 1" is (1 +) 2. Finally, *subtask type* refers to whether the subtask represents a store, recall, reasoning, language comprehension/generation, or motor operator.

The task models were developed in an iterative manner. For each task, we developed an initial GOMS model through our own analysis of its execution. We then refined the model based on a set of four sample videos, recorded prior to and independent from the user study. We compared the identified operator sequences to the leaves of our task model and refined them as necessary.

We validated the final task models by matching observable events (keyboard, mouse and gaze) in the videos from the user study to the operators in the models. Gaze events were used to match cognitive operators. An error step was defined to be a deviation from the prescribed sequence. If the user committed an error, each action after that step would count as an error until the user again performed a step in the prescribed sequence, from which point the analysis continued. The average error rate for the Route Planning task was 2.81%, ranging from 0% to 5.66%, consistent with results reported in (Card et al., 1983). We repeated this procedure for Document Editing. Shown partially in Figure 10, the model for this task had 5 levels and 41 nodes. The average error rate was 2.3%.

The GOMS models accurately reflected a user's execution of the tasks and enabled us to precisely align a user's pupillary response. This was challenging since each user performed the tasks at different speeds. We thus aligned the pupil data to specific events in a task model, not to time. A bottom up approach was used for the alignment. For each leaf subtask, we identified the beginning and end time stamp from the screen interaction video and these timestamps were used to index into the pupillary response file. The corresponding data was extracted and associated with the leaf subtask. Data between the end timestamp of one subtask and the begin timestamp of the following subtask was associated with the boundary between them. Data for higher-level subtasks was then calculated from its child subtasks and this process was repeated until the root node was reached.

5.5 Measurements

As in Experiment 1, we calculated the percent change in pupil size (PCPS) at each sample point and the average PCPS for each subtask. The time window of a subtask varied according to its type and level in the task model and ranged from about 24 ms for the lowest-level subtasks to about 63 seconds for the higher level subtasks.

5.6 Results

For both tasks, we discuss how much workload the different types of subtasks induced on a user and how much a user's workload differed between subtask execution and subtask

boundaries. The reader should keep in mind that small changes in pupillary response can represent meaningful changes in workload and that there is an upper bound on how much a user's pupil will increase due solely to workload-based effects.

5.6.1 Route Planning Task

Figure 11 shows the mean APCPS for each subtask in Route Planning. Time moves from left to right and the vertical lines represent first and second level boundaries from the task model in Figure 9. The rise and fall of the curve shows changing mental workload during task execution.

5.6.1.1 Mental workload during subtasks

To validate that cognitive subtasks induced increased workload, we performed a one-sample t-test on the APCPS for the Store, Recall, and Reasoning subtasks. We found that APCPS was greater than 0 across subtasks ($\underline{M}$=12.7, $\underline{SD}$=7.3, $\underline{t}(263)$=28.25, $\underline{p}$<0.001). The standardized effect size index $\underline{d}$ was 1.7, a high value. This represents a 12.7% increase over the baseline and shows that subtasks did impose increased workload on a user. We only used cognitive subtasks in our analysis since the relationship between cognitive effort and pupil size is the one best established by prior work (Beatty, 1982).

A one-way ANOVA with Subtask as the factor showed a main effect on APCPS ($\underline{F}(2,261)$=3.247, $\underline{p}$<0.04). Post hoc tests showed that Reasoning induced more mental workload than Store (difference was 3.4 percentage points, with $\underline{p}$<0.037). This shows that certain subtasks (Reasoning) induce more workload than others (Store) while other subtasks induce similar workload (Store and Recall).

Level also had a main effect on APCPS ($\underline{F}(1,262)$=3.90, $\underline{p}$<0.049). Because the subtasks used in this comparison were the operators in the task model, each subtask existed at either level 3 or 4. Subtasks at level 3 had a higher APCPS than at level 4 (difference was 2.3 percentage points). This difference may be attributed to the cognitive demands of the subtasks rather than their level, since level 3 contained all of the reasoning subtasks, which were shown to be more cognitively demanding. Our results provide evidence showing that execution of subtasks causes increased workload and that different types of subtasks induce different workload on a user.

5.6.1.2 Decrease of mental workload at subtask boundaries

We define a subtask boundary to span the time from the *last* observable operator in a subtask to the *first* observable operator in the subsequent subtask, see Figure 12. There

was a clear boundary between subtasks at each level. For each boundary, we computed *Boundary Decrease* by subtracting the minimum PCPS (taken as the average of 3 values around the min to ensure support) within the boundary from the APCPS of the preceding subtask. Thus, a *positive* change in Boundary Decrease reflects a *decrease* in workload.

A one sample t-test showed that Boundary Decrease was greater than 0 across all subtasks (M=0.28, SD=3.43, $t(611)$=1.995, p<0.047). The standardized effect size d was 0.08. This shows that workload decreases at a subtask boundary, but the decrease is small on average. One reason for the small effect size was that the lowest level boundaries showed little or no decrease in PCPS. PCPS likely did not decrease at these boundaries because the adjacent subtasks were short (about 200 msec), rapid, and closely related. We reran the one sample t-test for Boundary Decrease, excluding level 4 samples. Results showed a stronger effect (M=0.97, SD=3.57, $t(395)$=5.4, p<0.001) with an improved d of 0.3. As level 3 and level 2 samples were removed, results showed increasingly stronger effects. This suggests that changes in workload are meaningful down to the level of boundary just above the elementary operators in a task model.

Level had a main effect on Boundary Decrease ($F(3,608)$=19.68, p<0.001). Post hoc tests showed that decreases at level 1 were marginally greater than at level 2 (about 1.9 percentage points, p<0.056) and were greater than level 3 (2 percentage points, p<0.014) and level 4 (3.9 percentage points, p<0.001). Decreases at level 2 were greater than at level 4 (2 percentage points, p<0.001) and decreases at level 3 were greater than level 4 (2 percentage points, p<0.001). This suggests that workload decreases more at boundaries higher in a model than at boundaries lower in the model.

We not only found that workload changed between levels in the task model, but also that workload changed *within* the same level in the task model. For example, the APCPS over the level 1 boundaries differed ($t(11)$=3.99, p<.002) with a maximum difference of about 4.5 percentage points. The APCPS among level 2 boundaries also differed ($F(3,33)$=3.582, p<0.024) with a maximum difference of about 5 percentage points.

5.6.2 Document Editing Task

Figure 13 shows the mean APCPS for each subtask during Document Editing.

5.6.2.1 Mental workload during subtasks

We performed a one sample t-test for the APCPS across Language Comprehension, Language Generation, and Recall subtasks. These were the observable subtasks and existed only at Levels 2, 3 and 5. APCPS was greater than 0 across subtasks (M=6.72,

SD=6.47, t(299)=17.98, p<0.001) with a standardized effect size d=1.0, a high value. This shows a 6.72% increase over the baseline level, meaning that subtasks did induce mental workload on a user, but not as much as in the route planning task.

An ANOVA with Subtask (Comprehension, Generation, and Recall) as the factor showed a main effect on APCPS (F(2,129)=11.06, p<0.001). Recall induced more workload than Comprehension (difference was 6.1, with p<0.001) and Generation (difference was 3.7, with p<0.036). Generation had a higher APCPS than Comprehension (difference was 2.3), but was not significant. These results are consistent with Route Planning, where different types of subtasks also induced different workload on a user.

An ANOVA with Level as a factor showed a main effect on APCPS (F(2,297)=14.17, p<0.001). Subtasks at levels 2 and 3 induced more workload than at level 5 (difference was 5.4 and 3.7, with p<0.01 and p<0.001, respectively). Subtasks at levels 2 and 3 are Recall, while those at level 5 are Generation and Comprehension, thus this difference may be due to the Type rather than the Level of subtasks.

5.6.2.2 Decrease of mental workload at subtask boundaries

A one-sample t-test showed that Boundary Decrease was greater than 0 across subtasks (M=0.81, t(299)=7.27, p<0.001) with an effect size d=0.42. An ANOVA showed that Level (1-5) had a main effect on Boundary Decrease (F(4, 295)= 8.04, p<0.001). Post hoc tests showed that Boundary Decrease at level 1 was greater than at level 3 (p<0.013) and level 5 (p<0.004). Boundary Decrease at level 2 was greater than level 3 (p<0.001), level 4 (p<0.033) and level 5 (p<0.001). This shows that mental workload decreases more at boundaries higher in a model and less at boundaries lower in the model, consistent with results from the Route Planning task.

5.7 Discussion

Results from our second experiment showed that (i) a user's mental workload exhibits momentary decreases at subtask boundaries; (ii) workload decreases more at boundaries higher in a task model than at boundaries lower in the model, (iii) the decrease in workload at boundaries may differ within the same level of a task model, and (iv) some types of subtasks (e.g., Reasoning) can induce more workload than others (e.g., Store and Recall). We will defer further discussion of these results until the General Discussion.

Let us now turn our attention to a practical application of these results. Much prior work, e.g., (Bailey & Konstan, 2005; Czerwinski et al., 2000b; McFarlane, 1999; Monk et al., 2002) has assumed that interrupting tasks at moments of lower mental workload

would cause less negative impact than interrupting at other moments. The challenge has always been to identify where these moments of lower workload actually occur during task execution. By inspecting the workload-aligned task models, we are able to select precise moments of lower mental workload in the tasks. Comparing the effects of interrupting users at these and other moments was the focus of our third experiment.

6. TEST WHETHER MOMENTS OF LOWER MENTAL WORKLOAD ARE MORE OPPORTUNE FOR INTERRUPTION

Our third study compared effects of interrupting task execution at boundaries with lower workload, boundaries with higher workload, and random moments (simulating today's interface). We focused on boundaries because these are the moments hypothesized to be more opportune for interruption and because it is plausible for systems to automate the process of detecting these moments within many goal-directed tasks (Bailey et al., 2005).

From the workload-aligned models developed in our previous studies, we selected boundaries with the lowest and highest workload. Our expectation was that scheduling interruptions to occur at the Best moments (low workload boundaries) would have less negative impact than interrupting at the Worst (high workload boundaries) and random moments. A no-interruption condition was also included as a control.

6.1 Experimental Design

A repeated measures design was used with Timing (Best, Worst, Random, None) and Task (Route Planning, Document Editing, Email Classification) as factors. Eye tracking equipment was not used in this experiment.

6.2 Users and Tasks

Twelve users (6 female) participated in the study and ages ranged from 21 to 42. The experiment used the Route Planning and Document Editing tasks from our second study and the Email Classification task from our first study. To use the latter task, we performed enough of the workload alignment such that boundaries with the lower and higher workload could be identified. While more tasks could have been used, we felt that using three primary tasks provided a reasonable sample while keeping the length of the experiment practical.

We developed similar tasks for each category, but were careful to design them such that they would not induce workload patterns largely different from the existing models.

For the interrupting task, users read a news article and selected the most appropriate title from three choices. The interrupting task was adapted from (Adamczyk & Bailey, 2004).

6.3 Selected Moments for Interruption

For Route Planning, Best was between completing the second route and selecting the shorter (or cheaper) route. Worst was between recalling and entering information into any cell of the table. For Document Editing, Best was between the completion of the last edit and accessing the Save menu. Worst was between positioning the mouse at the intended location and entering changes to the document text. For Email Classification, Best was between placing an email into a folder and preparing to access the next email. Worst was between selecting an email and starting to drag it towards the destination.

Best and Worst for each task were selected by ordering the boundaries according to increasing APCPS and selecting the boundary with the highest and lowest workload. An ANOVA showed that mental workload was lower at the Best moments (μ=7.85) than the Worst moments (μ=12.46; $F(1,11)=9.31$, $p<0.01$) across tasks. For all tasks, Best existed at a level higher in the task model than Worst.

6.4 Experimental Setup

Delivery of interrupting tasks used a Wizard of Oz model. The experimenter observed a user's task execution using a RealVNC client connected over a high-speed LAN to minimize latency. At pre-defined moments, the experimenter used custom software to send an interrupting task to a user. Best and Worst moments were defined from the task models and Random moments were delivered at times randomly selected from an interval based on average task completion times.

6.5 Procedure

Before each category, specific instructions were given to the user and a practice task was performed. For each category, users performed four task trials, one for each timing condition. Users were instructed to attend to an interrupting task as soon as it appeared and, once complete, resume the primary task. The interrupted task was presented in a modal window and covered the main work area of the primary task. Users were instructed to complete the tasks as quickly and accurately as possible. After each task trial, users completed the NASA TLX and scales for annoyance and respect. The order of the categories, tasks, and timing conditions were randomized. The study lasted an hour.

6.6 Measurements

In the study we measured the following:

- *Subjective workload*. This was measured using the NASA TLX (Hart & Staveland, 1988). Users responded by marking a vertical line along continuous scales from low to high. Scores were weighted equally and combined into a single workload value.
- *Resumption lag*. This is the time needed to meaningfully resume the primary task after interruption. Lag was measured as the time from closing the interrupting task window to the first keyboard or mouse action in the primary task in direction of the task goal.
- *Annoyance*. This was measured on a continuous scale from low to high, similar to the TLX. Annoyance was used as a measure of the user's affective state.
- *Respect*. Users rated how respectful the interrupting system was to the primary task, i.e., social attribution. This measure was included because studies show that users often respond to interactive systems socially (Nass, Steuer, & Tauber, 1994).

Combinations of these measures have been used to measure effects of interruption in many prior studies, e.g., (Adamczyk & Bailey, 2004; Bailey, Konstan, & Carlis, 2001; Trafton, Altmann, Brock, & Mintz, 2003).

7. RESULTS

Two-way ANOVAs (Task x Timing) were used to analyze the dependent measures.

7.1 Subjective Workload

Task had a main effect on subjective workload ($F(2,22)=22.01$, $p<0.001$). Post hoc analysis showed that Route Planning ($\mu=2.74$) induced higher subjective workload than both Document Editing ($\mu=2.25$, $p<0.002$) and Email Classification ($\mu=1.77$, $p<0.001$), while Document Editing induced higher workload than Email Classification ($p<0.003$). Timing did not influence subjective workload and there were no interactions. This shows that the interrupting task did not induce subjective workload beyond that of the primary task – regardless of timing. However, the results do allow the tasks to be rank ordered based on workload (Route Planning > Document Editing > Email Classification), which can be used to help interpret later results.

7.2 Resumption Lag

Figure 14 shows results for resumption lag. Task had a main effect ($F(2,22)=6.27$, $p<0.007$). Post hoc analysis showed that users resumed Email Classification tasks faster ($\mu=2.73$s) than Route Planning ($\mu=5.04$s, $p<0.012$) and Document Editing tasks ($\mu=6.63$s, $p<0.004$) after being interrupted. No other differences were found. This is roughly consistent with the workload ratings, as the lowest workload task had the least resumption lag.

Timing had a main effect ($F(3,33)=6.87$, $p<0.005$). Post hoc tests showed that users resumed the primary task almost 3 times faster after being interrupted at Best ($\mu=2.06$s) than at Worst ($\mu=6.69$s, $p<0.03$) and Random ($\mu=5.66$s, $p<0.007$) moments. This result may be explained by the executive system needing to acquire fewer attentional resources to resume a primary task interrupted at a moment of lower workload (Wickens, 2002). No other differences were found and there were no interactions.

7.3 Annoyance

Figure 15 shows ratings of annoyance. Task had a main effect ($F(2,22)=7.06$, $p<0.004$). Post hoc tests showed that Route Planning ($\mu=2.43$) caused users to experience more Annoyance than Document Editing ($\mu=1.6$, $p<0.025$) and Email Classification ($\mu=1.5$, $p<0.004$). No other differences were found. The results are mostly consistent with the ratings of subjective workload, as the highest workload task caused the most annoyance and the lowest workload task caused the least annoyance.

Timing had a main effect on ratings of annoyance ($F(3,33)=7.41$, $p<0.001$). Post hoc tests showed that interruptions at Best moments ($\mu=1.69$) caused 18% less annoyance than at Worst moments ($\mu=2.07$, $p<0.079$) and 28% less annoyance than at Random moments ($\mu=2.35$, $p<0.052$). Not surprisingly, users experienced the least annoyance when not interrupted ($\mu=1.26$) than when interrupted ($p<0.043$ in all cases). No other differences were found and there were no interactions.

7.4 Respect

Figure 16 shows the ratings of respect attributed to the interrupting system. Task had no main effect ($F(2,22)=0.70$, $p<0.51$), while Timing did have a main effect on user ratings ($F(3,33)=12.11$, $p<0.001$). Post hoc tests showed that when interrupted at Best moments ($\mu=3.02$), users rated the system to be 63% more respectful to their primary task than when interrupted at Worst moments ($\mu=1.85$, $p<0.015$) and 39% more respectful than when interrupted at Random moments ($\mu=2.17$, $p<0.086$). Users rated the system most

respectful to their primary task when not interrupted (μ=4.02) than when it was interrupted ($p<0.001$). No interactions were detected in the data.

8. DISCUSSION

The moment at which an ongoing primary task is interrupted influences the disruptive effects caused by the interruption. Results showed that interrupting at the Best moments was less disruptive across tasks. Users resumed primary tasks 69% faster, experienced 18% less annoyance, and attributed 63% more respect to the interrupting system compared to being interrupted at the Worst moments. Similar differences were found between Best and Random. However, interrupting at the Worst moments demonstrated no measurable improvement over Random moments, which shows that not all boundaries are opportune for interruption, especially those that are lower in a task model.

For each task, Best and Worst moments were selected by identifying boundaries with the lowest and highest workload in the workload-aligned models. Since interrupting primary tasks at Best moments consistently caused less negative impact, mental workload can and should be leveraged to help predict how opportune various moments in a task are for interruption. Also, the average difference in time between the Best and other moments across tasks was relatively short, less than a minute on average (μ_{RP}=57.3s, μ_{DE}=41.1s, μ_{EC}=4.5s), but the corresponding mitigation of negative impact was meaningfully large.

Apart from boundaries, there may have been other moments of low mental workload during the tasks, which may have been opportune for interruption, but were not considered. We focused on selecting opportune moments from boundaries since systems could feasibly automate their detection and this process could generalize to many goal-directed tasks. While considering only boundaries may not always produce the optimal solution, it can still meaningfully mitigate effects of interruption, as shown by our results.

9. GENERAL DISCUSSION

This work investigated the relationship among mental workload, task execution, and effects of interruption. The overarching approach was to leverage a physiological measure to monitor a user's mental workload during task execution, identify where moments of lower workload occur and any patterns based on the structure of a task that could be generalized, and test whether selected moments of lower workload were in fact more opportune for interruption.

In our first experiment, we showed that pupil dilation can provide a reliable measure of mental workload for interactive computing tasks, assuming appropriate environmental controls. We also learned that pupillary response is sensitive to the varying mental demands of a task and thus that the response data is best understood by aligning it to the corresponding model of task execution. Aligning pupil dilation to the execution structure of a task has been previously used, for example, to understand how mental effort relates to syntactic ambiguity during sentence comprehension (Schluroff et al., 1986). Our work has extended the use of workload alignment to hierarchical models of goal-directed tasks. For example, interface designers could leverage our alignment method to identify areas of an interface that induce unacceptably high workload and target those areas for re-design.

Building on these lessons, in our second experiment, we developed tasks with more complex structures and monitored workload (pupil dilation) while users performed them. The workload response data was then aligned to corresponding GOMS models of the tasks. Analysis of these workload-aligned task models produced three main results.

First, our results showed that a user's mental workload exhibits momentary decreases at subtask boundaries. Though predicted by Miyata & Norman (1986), our work has provided the first physiological evidence showing this effect. From the perspective of resource theories of attention (Kahneman, 1973; Wickens, 1980, 1991, 2002), this result suggests that attentional resources are not statically allocated at the onset of a task, but are dynamically allocated and released throughout task execution. When a user reaches a boundary (i.e., breakpoint) in a task, some resources are released, making more resources available for the interrupting task. Also, the executive system may require fewer resources to resume the suspended task, as fewer goals and action schema may need to be (re)activated at boundary points (Rubinstein et al., 2001). A practical implication of this result is that systems should consider a user's position within the structure of a task when making finer-grained temporal decisions about when to interrupt.

Second, our results showed that the decrease in workload at subtask boundaries is not uniform, as workload decreased more at boundaries higher in a task model than at boundaries lower in the model. In general, it appears that more attentional resources are released when more significant cognitive breakpoints are reached in the task. In contrast, the amount of resources released when lower-level boundaries are reached is likely small, possibly due to cognitive chunking of repetitive or skilled action sequences (Newell &

Rosenbloom, 1981). When making fine-grained decisions, our results show that systems would only need to consider boundaries at the first few levels of a task model.

Third, the decrease in workload at boundaries within the same level of a task model was also not uniform. This indicates that the level of a boundary itself cannot always be used to determine whether a particular boundary is more opportune for interruption than other points in a task. An accurate determination of which boundaries are most opportune currently requires aligning a continuous measure of workload to the corresponding model of task execution. However, other modeling techniques such as those employed to study event perception (Zacks, Tversky, & Iyer, 2001) have been successfully used to identify "natural" breakpoints in an interactive task without the use of a workload measure (Adamczyk & Bailey, 2004). Understanding how well these natural breakpoints align with the low workload boundaries in a task offers a promising direction for future work.

Our final experiment compared effects of interrupting a task at boundaries with low and high workload and random moments. Results showed that interrupting at boundaries with lower workload resulted in less resumption lag, less subjective annoyance, and greater attribution of respect than interrupting at the other moments. This indicates that workload is an effective predictor of how opportune various moments in a task are for interruption. Further, the time at which the interrupting task was delivered in the various conditions differed by less than one minute on average. Thus, systems need not defer delivery of secondary information for a long period of time – a short deferral can result in meaningful mitigation of negative impact due to interruption.

9.1 Integrating Workload into Systems that Reason about Interruption

Our work has shown that the mental workload being induced at different moments in a task is an effective predictor of how opportune those moments are for interruption. This provides further evidence that systems should integrate consideration of workload into their reasoning framework. A more accurate cost of interruption could then be computed, which would enable more effective decisions to be made about when to interrupt.

To integrate workload, systems could link a real-time measure of workload directly into their reasoning framework, e.g., a Bayesian model could include an evidence variable for workload. Rather than use raw pupillary response or other raw physiological data, the incoming source data could be computationally mapped onto a discrete scale that maintains sufficient sensitivity, but is much easier to interpret, e.g., see the IOP scale described in (Iqbal et al., 2005). The stream of mapped values could also be analyzed for

temporal patterns to provide a more robust sense of a user's cognitive engagement in a task than what a single value could otherwise indicate. The input source for workload could be provided, for example, by eye tracking systems embedded within standard desktop monitors (Tobii-Systems) or by inexpensive heart rate sensors built into office chairs (Anttonen & Surakka, 2005).

However, integrating a real-time measure of workload may often not be possible or desirable, due to the expense, intrusiveness of required equipment, or lack of necessary environmental controls. In these cases, the cost of interrupting (the workload of) a task at different moments could be encoded within a machine-parsable description of the task (Bailey et al., 2005). To assign cost values, one approach is to develop workload-aligned models for the necessary tasks in controlled settings, map workload at salient points onto a discrete cost scale, and then encode the mapped values within the corresponding parts of the task descriptions. When a user later executes the tasks, a system would monitor their task execution, match it to the task descriptions, and pass the encoded costs at each step on to a broader reasoning framework.

Building workload-aligned task models to acquire precise cost values would be most appropriate in safety critical or other domains where the set of possible tasks is small, but the cost of poorly timed interruption could be large. An alternative is to assign cost values by applying workload heuristics to the hierarchical structure of a task. For example, one could assign higher costs to lower-level boundaries and successively lower costs to higher-level boundaries in a task description. Similar heuristics could be developed for different types of subtasks such as memory store/recall, language comprehension/ generation, mental reasoning, etc. Though applying heuristics could only offer reasonable approximations, these estimates may be sufficient for a broad range of tasks and could be expediently applied.

9.2 Limitations

Aligning workload to a task model requires a prescribed execution sequence. As a result, cost values computed from the aligned models may only apply when a user executes that task in the same or similar sequence. Developing workload-aligned models is thus most appropriate for safety critical or high frequency tasks, which are often performed in a repetitive sequence (Degani & Wiener, 1993). Though building workload-aligned models currently requires a large effort, the aligned model for each desired task only needs to be developed once in a controlled setting, and the results could then be broadly applied in

uncontrolled settings. Automated tools may also help lower the effort needed to build the models in the future. For other situations, as mentioned previously, we believe that effective heuristics could be developed that would allow reasonable approximations of interruption cost to be quickly assigned to various moments within a task, completely eliminating the need to create workload-aligned models.

Another concern is whether the observed patterns in workload changes would remain if our experimental tasks were embedded within broader interactive activities, e.g., when additional task goals or data must be carried through part or all of the tasks. This would almost certainly have an effect, but we suspect that this effect would be manifested as a shift in *absolute* workload, whereas the *relative* changes in workload would remain similar, e.g., workload would still decrease at boundaries even though the absolute values at those points and surrounding subtasks might be different. However, additional empirical studies are needed to verify these claims.

Similarly, the presence or location of boundaries may change as a user's knowledge of performing a task transitions from novel to skilled behavior. As a task becomes skilled, mental representations of the task may become coarser (Newell & Rosenbloom, 1981), eliminating some of the perceived boundaries. However, recent experimental studies have shown that familiarity with a task seems to have little effect on how users perceive its hierarchical structure (Zacks et al., 2001), suggesting that the mental representations for tasks remain fairly stable. Even so, skilled tasks are typically performed in larger chunks. Thus, while interrupting at boundaries between chunks may still show a mitigating effect similar to our results, interrupting at boundaries within chunks may have less of an effect.

10. CONCLUSION AND FUTURE WORK

A recent thrust in the HCI community has been to understand when to deliver peripheral information such that it would have the least negative impact on a user's productivity and affective state. Many have argued that interrupting tasks during periods of low workload would have the least negative impact, but knowing just where these opportune moments occur has been elusive. Our work has made several contributions toward understanding the relationship among mental workload, task execution, and effects of interruption.

First, our work has demonstrated that pupil dilation can be used as a reliable measure of mental workload for interactive tasks, which may encourage more researchers to consider this particular measure. Second, our empirical results have advanced theoretical understanding of how a user's mental workload changes during task execution. Our

results showed that there are momentary decreases in workload at subtask boundaries, workload decreases more at boundaries higher in a task model and less at boundaries lower in the model, and the decrease in workload differs within the same level of a model. Third, we validated that interrupting users at boundaries with lower workload mitigates effects of interruption relative to interrupting at other moments. Finally, we situated our results within contemporary theories of attention as well as described several methods for how our results could be integrated within interruption reasoning systems.

For future work, we plan to analyze workload-aligned models for more tasks in order to produce additional heuristics for assigning costs of interruption. We plan to implement these heuristics within an existing task monitoring framework. The framework would provide the cost of interrupting a task's moment-by-moment execution to broader reasoning systems, enabling them to make finer-grained and more effective decisions about when to interrupt.

11. REFERENCES

Adamczyk, P. D., & Bailey, B. P. (2004). If Not Now When? The Effects of Interruptions at Different Moments Within Task Execution. *Proceedings of the ACM Conference on Human Factors in Computing Systems*, 271-278.

Altmann, E. M., & Trafton, J. G. (2002). Memory for Goals: An Activation-based Model. *Cognitive Science, 26*(1), 39-83.

Anttonen, J., & Surakka, V. (2005). Emotions and heart rate while sitting on a chair. *Proceedings of the ACM conference on Human Factors in Computing Systems*, 491-499.

Bailey, B. P., Adamczyk, P. D., Chang, T. Y., & Chilson, N. A. (2005). A Framework for Specifying and Monitoring User Tasks. *Journal of Computers in Human Behavior, special issue on attention aware systems*, (July/August).

Bailey, B. P., & Konstan, J. A. (2005). On the Need for Attention Aware Systems: Measuring Effects of Interruption on Task Performance, Error Rate, and Affective State. *Journal of Computers in Human Behavior, special issue on attention aware systems, (July/August)*.

Bailey, B. P., Konstan, J. A., & Carlis, J. V. (2001). The Effects of Interruptions on Task Performance, Annoyance, and Anxiety in the User Interface. *Proceedings of the IFIP TC.13 International Conference on Human-Computer Interaction*, 593-601.

Beatty, J. (1982). Task-evoked Pupillary Responses, Processing Load, and the Structure of Processing Resources. *Psychological Bulletin, 91*(2), 276-292.

Bradshaw, J. L. (1967). Pupil Size as a Measure of Arousal during Information Processing. *Nature, 216*, 515-516.

Card, S., Moran, T., & Newell, A. (1983). *The Psychology of Human-computer Interaction*. Hillsdale: Lawrence Erlbaum Associates.

Cutrell, E., Czerwinski, M., & Horvitz, E. (2001). Notification, Disruption and Memory: Effects of Messaging Interruptions on Memory and Performance. *Proceedings of the IFIP TC.13 International Conference on Human-Computer Interaction*, 263-269.

Czerwinski, M., Cutrell, E., & Horvitz, E. (2000a). Instant Messaging and Interruption: Influence of Task Type on Performance. *Annual Conference of the Human Factors and Ergonomics Society of Australia (OZCHI)*, 356-361.

Czerwinski, M., Cutrell, E., & Horvitz, E. (2000b). Instant Messaging: Effects of Relevance and Timing. *People and Computers XIV: Proceedings of HCI*, 71-76.

Dabbish, L., & Kraut, R. E. (2004). Controlling interruptions: awareness displays and social motivation for coordination. *Proceedings of the ACM Conference on Computer Supported Cooperative Work*, 182-191.

Degani, A., & Wiener, E. (1993). Cockpit checklists: Concepts, design, and use. *Human Factors, 35*(2), 345-359.

Dey, A. K., & Abowd, G. D. (2000). CybreMinder: A Context-Aware System for Supporting Reminders. *Proceedings of 2nd International Symposium on Handheld and Ubiquitous Computing*, 172-186.

Dismukes, K., Young, G., & Sumwalt, R. (1998). Cockpit Interruptions and Distractions. *ASRS Directline, 10*.

Donchin, E., Kramer, A. F., & Wickens, C. D. (1986). *Applications of brain event related potentials to problems in engineering psychology*. New York: Guildford.

Fogarty, J., Ko, A. J., Aung, H. H., Golden, E., Tang, K. P., & Hudson, S. E. (2005). Examining task engagement in sensor-based statistical models of human interruptibility. *Proceeding of the ACM Conference on Human Factors in Computing Systems*, 331-340.

Fry, E. (1968). A readability formula that saves time. *Journal of Reading, 11*(7), 265-271.

Gale, A., & Edwards, J. (1983). *The EEG and human behavior*. New York: Academic Press.

Gevins, A., & Schaffer, R. (1980). A critical review of electroencephalographic (eeg) correlates of higher cortical functions. *CRT Critical Reviews in Bioengineering, 4*, 113-164.

Hart, S. G., & Staveland, L. E. (1988). Development of a Muli-dimensional Workload Rating Scale: Results of Empirical and Theoretical Research. In P. A. Hancock & N. Meshkati (Eds.), *Human Mental Workload* (pp. 138-183). Amsterdam, The Netherlands: Elsevier.

Hess, E. H. (1972). Pupillometrics: A method of studying mental, emotional and sensory processes. In N. S. Greenfield & R. A. Sternbach (Eds.), *Handbook of Psychophysiology* (pp. 491-531). New York: Holt, Rinehart & Winston.

Hess, E. H., & Polt, J. M. (1964). Pupil Size in Relation to Mental Activity during Simple Problem Solving. *Science, 132*, 1190-1192.

Hoecks, B., & Levelt, W. (1993). Pupillary Dilation as a Measure of Attention: A Quantitative System Analysis. *Behavior Research Methods, Instruments, & Computers, 25*, 16-26.

Horvitz, E., & Apacible, J. (2003). Learning and Reasoning about Interruption. *Proceedings of the Fifth ACM International Conference on Multimodal Interfaces*, 20-27.

Horvitz, E., Jacobs, A., & Hovel, D. (1999). Attention-Sensitive Alerting. *Conference Proceedings on Uncertainty in Artificial Intelligence*, 305-313.

Hudson, S. E., Fogarty, J., Atkeson, C. G., Avrahami, D., Forlizzi, J., Kiesler, S., et al. (2003). Predicting Human Interruptibility with Sensors: A Wizard of Oz Feasibility Study. *Proceedings of the ACM Conference on Human Factors in Computing Systems*, 257-264.

Hytintk, J., Tommola, J., & Alaja, A. (1995). Pupil Dilation as a Measure of Processing Load in Simultaneous Interpretation and Other Language Tasks. *The Quarterly Journal of Experimental Psychology, 48A*(3), 598-612.

Iqbal, S. T., Adamczyk, P. D., Zheng, S., & Bailey, B. P. (2005). Towards an Index of Opportunity: Understanding Changes in Mental Workload during Task Execution. *Proceedings of the ACM Conference on Human Factors in Computing Systems*, 311-320.

Iqbal, S. T., & Bailey, B. P. (2005). Investigating the Effectiveness of Mental Workload as a Predictor of Opportune Moments for Interruption. *Proceedings of the ACM Conference on Human Factors in Computing Systems*, 1489-1492.

Iqbal, S. T., Zheng, X. S., & Bailey, B. P. (2004). Task Evoked Pupillary Response to Mental Workload in Human-Computer Interaction. *Proceedings of the ACM Conference on Human Factors in Computing Systems*, 1477-1480.

Jackson, T. W., Dawson, R. J., & Wilson, D. (2001). The cost of email interruption. *Journal of Systems and Information Technology, 5*(1), 81-92.

Juris, M., & Velden, M. (1977). The Pupillary Response to Mental Overload. *Physiological Psychology, 5*(4), 421-424.

Kahneman, D. (1967). Pupillary Responses in a Pitch-discrimination Task. *Perception & Psychophysics, 2*, 101-105.

Kahneman, D. (1973). *Attention and Effort*. Englewood Cliffs, N.J: Prentice-Hall.

Kok, A. (1997). Event-related-potential (ERP) reflections of mental resources: a review and synthesis. *Biological Psychology, 45*, 19-56.

Kramer, A. F. (1991). Physiological Metrics of Mental Workload: A Review of Recent Progress. In D. L. Damos (Ed.), *Multiple-Task Performance* (pp. 279 - 328). London: Taylor and Francis.

Kramer, A. F., Schneider, W., Fisk, A. D., & Donchin, E. (1986). The effects of practice and task structure on components of event related brain potential. *Psychophysiology, 23*, 33-47.

Kreifeldt, J. G., & McCarthy, M. E. (1981). Interruption as a Test of the User-computer Interface. *Proceedings of the 17th Annual Conference on Manual Control*, 655-667.

Latorella, K. A. (1996). Investigating Interruptions: An Example From the Flight Deck. *Proceedings of the Human Factors and Ergonomics Society 40th Annual Meeting*, 249-253.

Latorella, K. A. (1998). Effects of modality on interrupted flight deck performance: Implications for data link. *42nd Annual Meeting of the Human Factors and Ergonomics Society*, 87-91.

Lee, J. D., Hoffman, J. D., & Hayes, E. (2004). Collision warning design to mitigate driver distraction. *Proceedings of the ACM Conference on Human factors in Computing Systems*, 65-72.

Maes, P. (1994). Agents that Reduce Work and Information Overload. *Communications of the ACM, 37*(7), 30-40.

Maglio, P., & Campbell, C. S. (2003). Attentive agents. *Communications of ACM, 46*(3), 47-51.

Marshall, S. P. (2002). The Index of Cognitive Activity: Measuring cognitive workload. *Proceedings of the 2002 IEEE 7th Conference on Human Factors and Power Plants*, 7.5-7.9.

McFarlane, D. C. (1999). Coordinating the Interruption of People in Human-computer Interaction. *Proceedings of the IFIP TC.13 International Conference on Human-Computer Interaction*, 295-303.

McFarlane, D. C., & Latorella, K. A. (2002). The Scope and Importance of Human Interruption in HCI Design. *Human-Computer Interaction, 17*(1), 1-61.

Miyata, Y., & Norman, D. A. (1986). Psychological Issues in Support of Multiple Activities. In D. A. Norman & S. W. Draper (Eds.), *User Centered System Design: New Perspectives on Human-Computer Interaction* (pp. 265-284). Hillsdale, NJ: Lawrence Erlbaum Associates.

Monk, C. A., Boehm-Davis, D. A., & Trafton, J. G. (2002). The Attentional Costs of Interrupting Task Performance at Various Stages. *Proceedings of the Human Factors and Ergonomics Society 46th Annual Meeting*

Nakayama, M., & Takahashi, K. (2002). The Act of Task Difficulty and Eye-movement Frequency for the Ocul-motor Indices. *Proceedings of Eye Tracking Research and Applications*, 37-42.

Nass, C., Steuer, J., & Tauber, E. R. (1994). Computers Are Social Actors. *Proceedings of the ACM Conference on Human Factors in Computing Systems*, 72-78.

Newell, A., & Rosenbloom, P. S. (1981). Mechanisms of skill acquisition and the law of practice. In J. R. Anderson (Ed.), *Cognitive Skills and their Acquisition* (pp. 1-55). Hillsdale, NJ: Erlbaum.

O'Donnell, R. D., & Eggemeier, F. T. (1986). Workload assessment methodology. In K. R. Boff, L. Kaufman & J. P. Thomas (Eds.), *Handbook of Perception and Human Performance. Volume II, Cognitive Processes and Performance* (pp. 42/41-42/49). New York: Wiley.

Phelps, M. P., & Mazziotta, J. (1985). Positron emission tomography: Human brain function and biochemistry. *Science, 228*, 799-809.

Rich, C., & Sidner, C. L. (1998). COLLAGEN: A Collaboration Manager for Software Interface Agents. *User Modeling and User-Adapted Interaction, 8*(3/4), 315-350.

Rowe, D. W., Sibert, J., & Irwin, D. (1998). Heart Rate Variability: Indicator of User State as an Aid to Human-computer interaction. *Proceedings of the ACM Conference on Human Factors in Computing Systems*, 480-487.

Rubinstein, J. S., Meyer, D. E., & Meyer, D. E. (2001). Executive Control of Cognitive Processes in Task Switching. *Journal of Experimental Psychology: Human Perception and Performance, 27*(4), 763-797.

Schluroff, M., Zimmermann, T. E., Freeman, R. B., Hofmeister, K., Lorscheid, T., & Weber, A. (1986). Pupillary Responses to Syntactic Ambiguity of Sentences. *Brain and Language, 27*, 322-344.

Shneiderman, B. (1997). *Designing the User Interface* (Third ed.): Pearson Addison Wesley, Third Edition.

Speier, C., Valacich, J. S., & Vessey, I. (1999). The influence of task interruption on individual decision making: An information overload perspective. *Decision Sciences, 30*(2), 337-360.

Stanton, N. (Ed.). (1994). *Human Factors in Alarm Design*. London: Taylor and Francis.

Takahashi, K., Nakayama, M., & Shimizu, Y. (2000). The Response of Eye-movement and Pupil Size to Audio Instruction while Viewing a Moving Target. *Proceedings of the ACM Conference on Eye Tracking Research & Applications*, 131-138.

Tobii-Systems. *http://www.tobii.se/*.

Trafton, J. G., Altmann, E. M., Brock, D. P., & Mintz, F. E. (2003). Preparing to resume an interrupted task: effects of prospective goal encoding and retrospective rehearsal. *International Journal of Human-Computer Studies, 58*, 583-603.

Wickens, C. D. (1980). The structure of attentional resources. In R. Nickerson (Ed.), *Attention and Performance VIII* (pp. 239-257). Hillsdale, NJ: Lawrence Erlbaum.

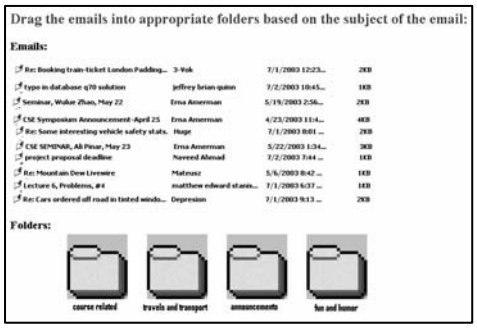
Wickens, C. D. (1991). Processing resources and attention. In D. L. Damos (Ed.), *Multiple-task performance* (pp. 3-34). London: Taylor & Francis.

Wickens, C. D. (2002). Multiple Resources and Performance Prediction. *Theoretical Issues in Ergonomic Science, 3*(2), 159-177.

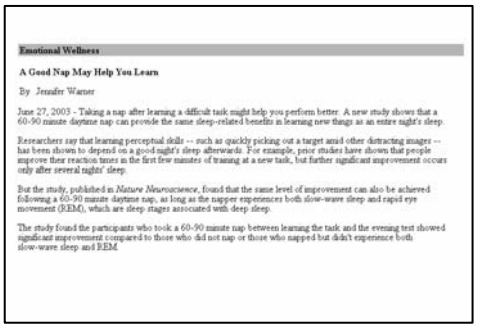
Yeh, M., & Wickens, C. D. (1988). Dissociation of performance and subjective measures of workload. *Human Factors, 30*, 111-120.

Zacks, J., Tversky, B., & Iyer, G. (2001). Perceiving, remembering, and communicating structure in events. *Journal of Experimental Psychology: General, 130*(1), 29-58.

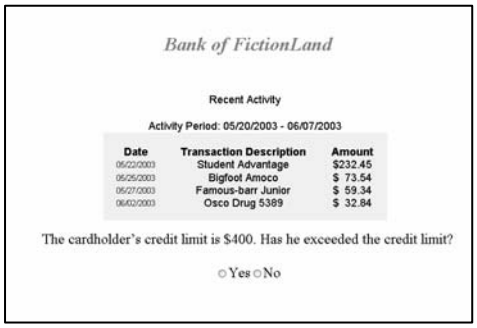
Zijlstra, F. R. H., Roe, R. A., Leonora, A. B., & Krediet, I. (1999). Temporal Factors in Mental Work: Effects of Interrupted Activities. *Journal of Occupational and Organizational Psychology, 72*, 163-185.



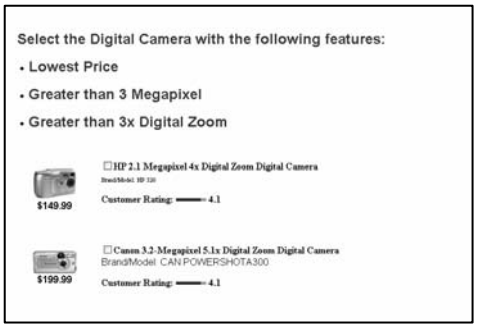
a) Email Classification Task



b) Reading Comprehension Task



c) Mathematical Reasoning Task



d) Search Task

Figure 1: Screenshots of the four task categories used in Experiment 1.

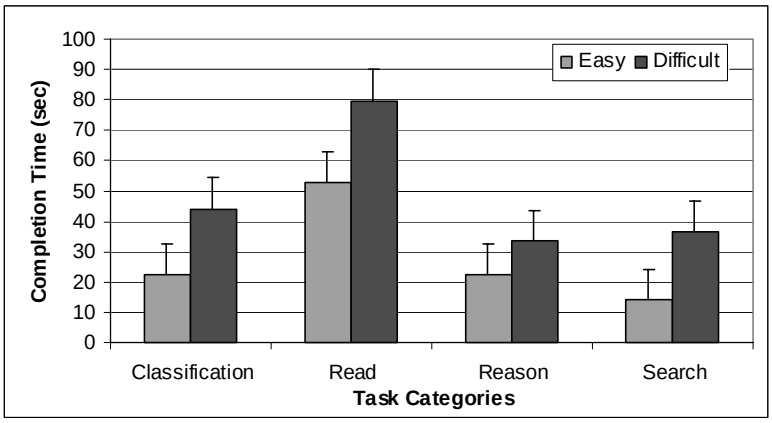


Figure 2: Average completion time for each task. Error bars show 95% CI of mean.

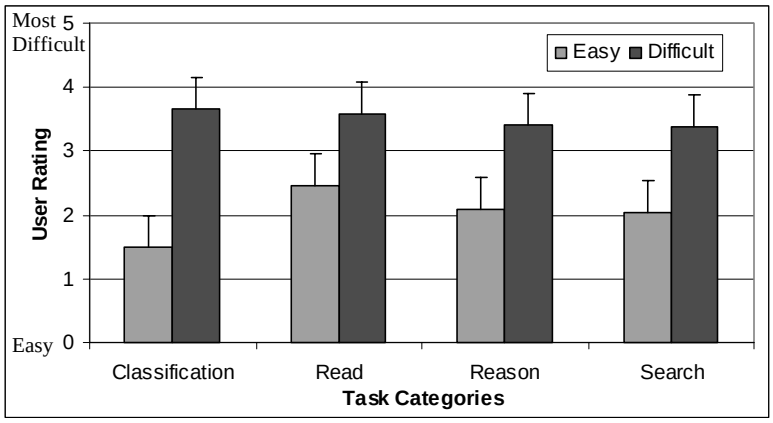


Figure 3: Average user rating for each task.

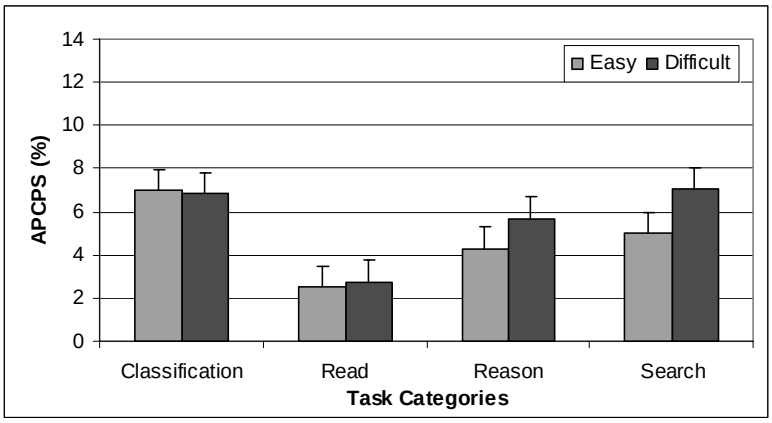
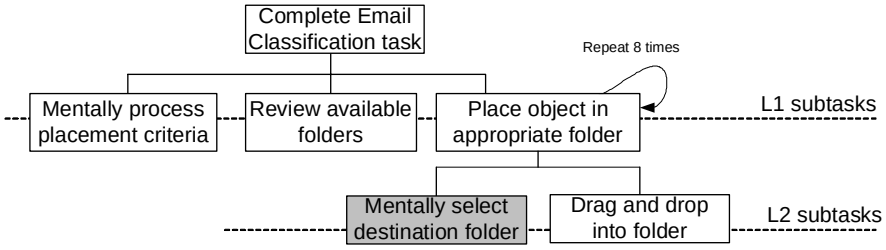
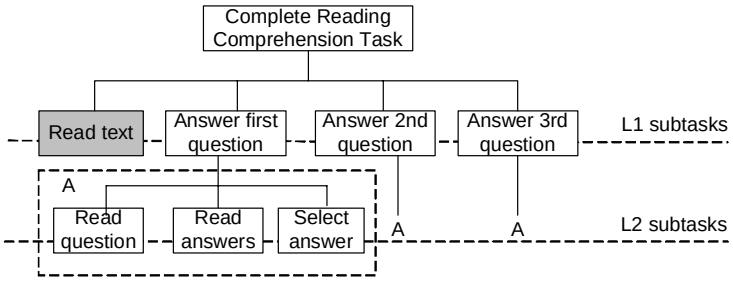


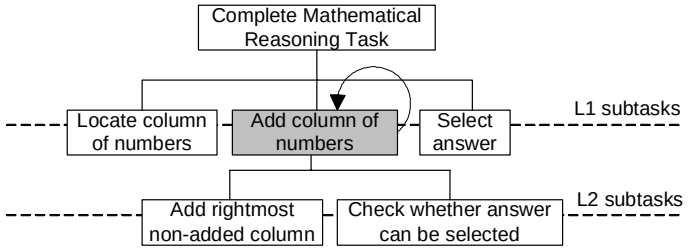
Figure 4: Average PCPS for each task.



(a) Email classification



(b) Reading comprehension



(c) Mathematical reasoning

Figure 5: GOMS models for the classification, comprehension, and reasoning categories. Tasks within each level of difficulty had a similar execution structure. Shaded areas represent those parts of the models that required different mental effort between the difficulty levels.

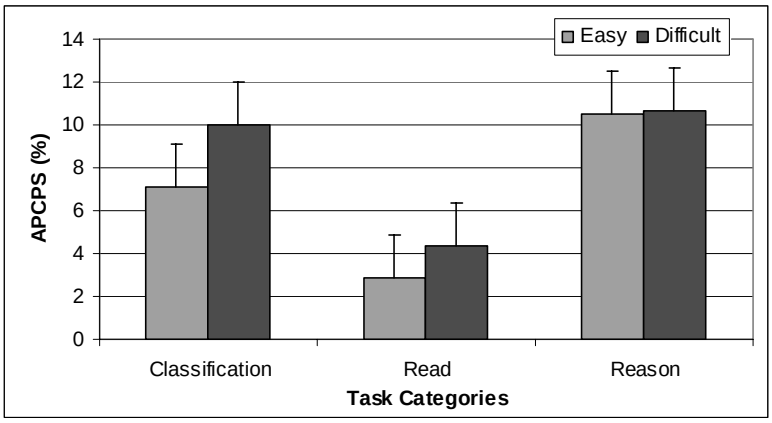


Figure 6: Average PCPS for cognitive subtasks.

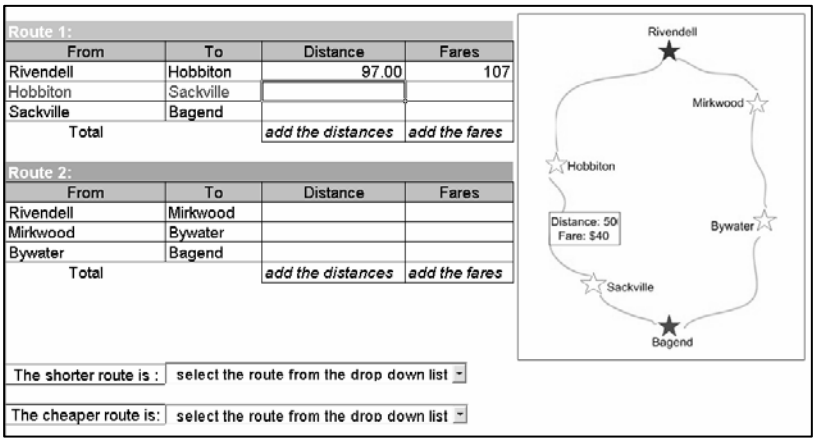


Figure 7: The interactive route planning task. A user retrieves distance and fare information from the map, enters the data into the tables, adds the distances and fares, and selects the cheaper and the shorter of the two routes.

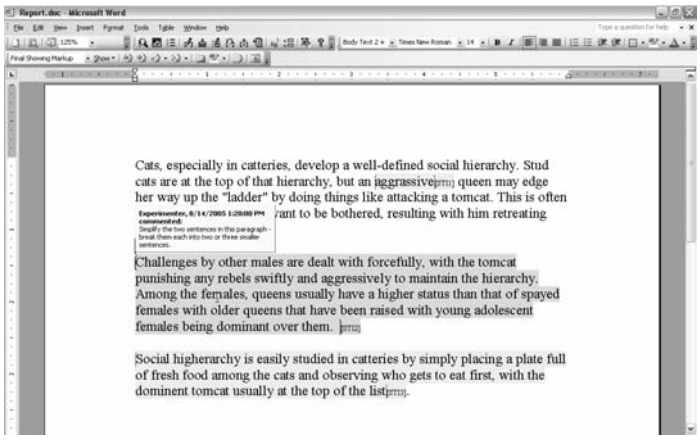


Figure 8: The document editing task. A user edits the document based on the given comments. Once edited, the document is saved to a specified directory and file name.

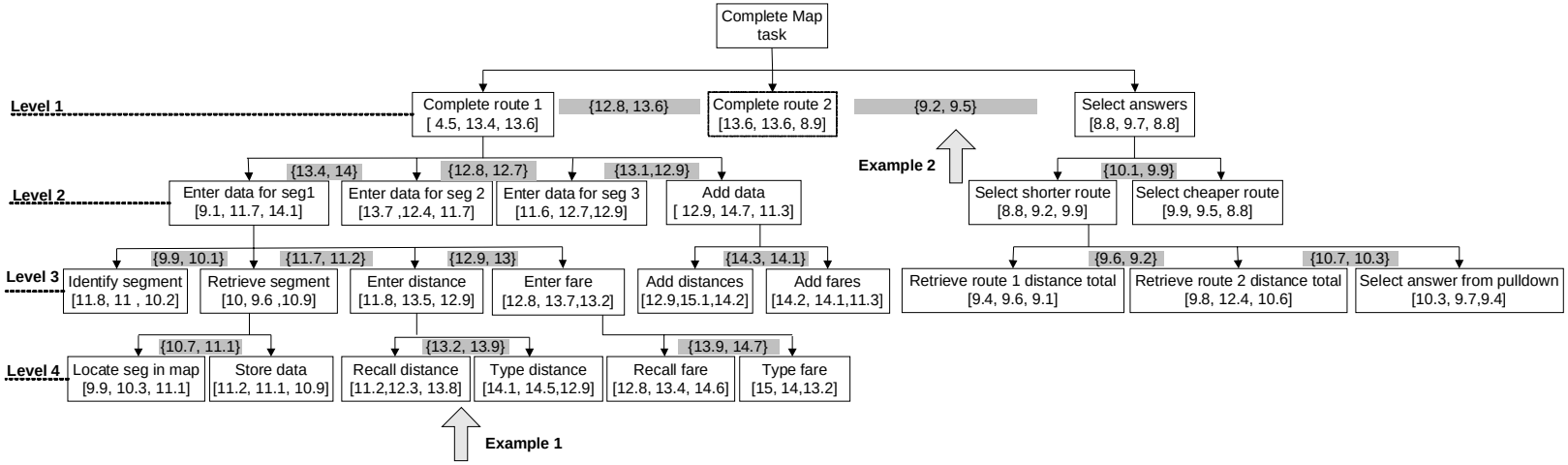


Figure 9: A workload aligned task model for Route Planning. The interior nodes represent goal nodes, the leaf nodes represent operators, and time moves from left to right. Regions A, B and C show parts of the task repeated elsewhere in the model. Within each subtask, we provide the [*beginning PCPS, average PCPS, last PCPS*] for that subtask. Each shaded area indicates a boundary and contains the [*minimum PCPS, average PCPS*] across it.

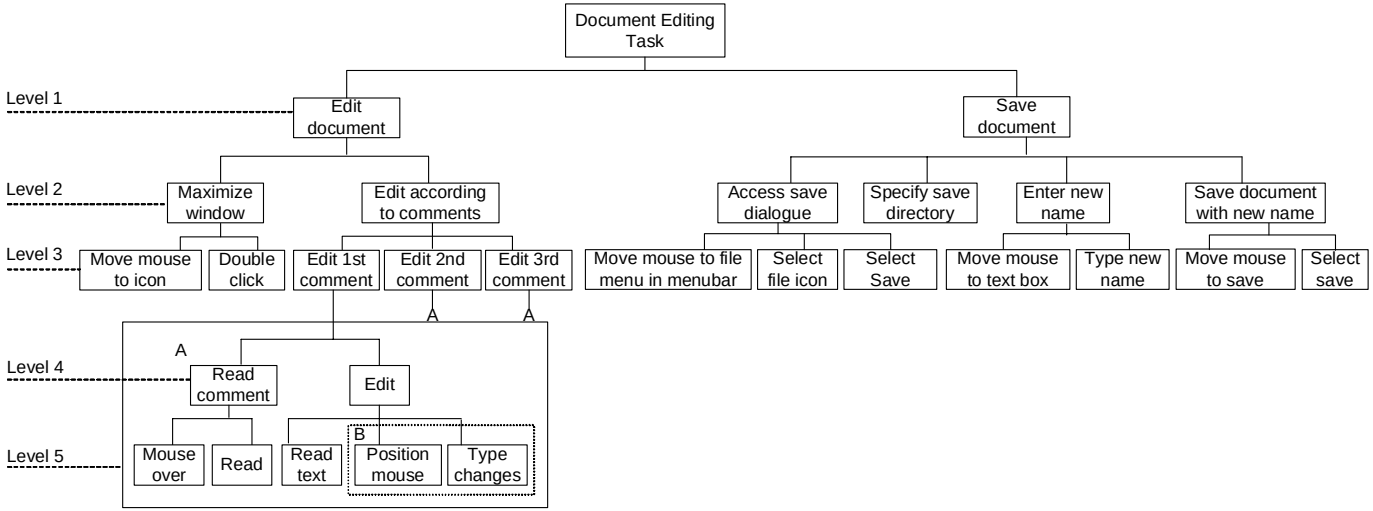


Figure 10: Validated GOMS model of the document editing task. The interior nodes represent goal nodes, the leaf nodes represent operators, and time moves from left to right. Regions A and B show parts of the task repeated elsewhere in the model.

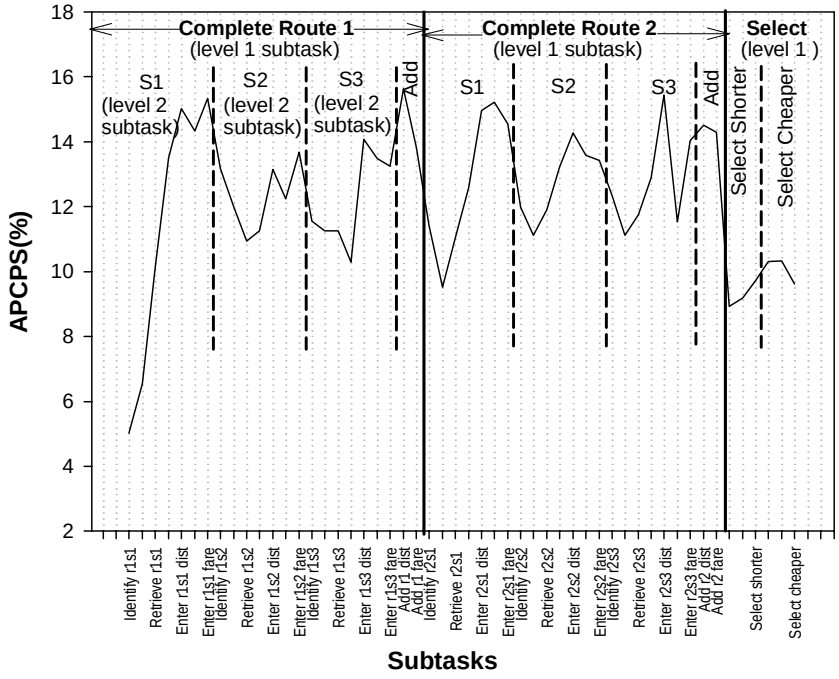


Figure 11: APCPS for the subtasks in the route planning task. Solid lines indicate level 1 boundaries and dashed lines indicate level 2 boundaries. The x-axis enumerates level 3 subtasks. Notice how the graph dips lower at level 1 boundaries than at level 2 boundaries – showing how mental workload decreases more at boundaries higher in the model.

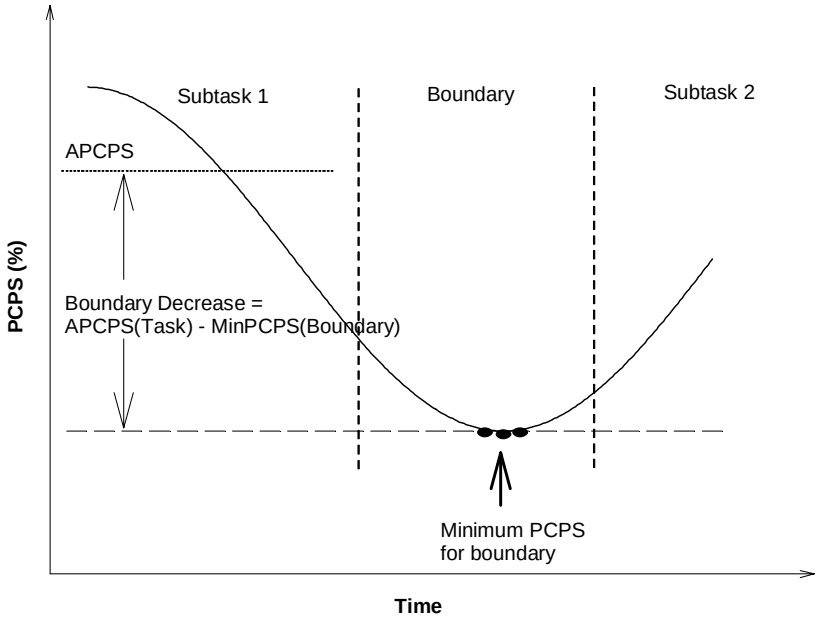


Figure 12: Illustration of metrics used in the analysis. The vertical dashed lines mark the last observable operator in subtask 1 and the first observable operator in subtask 2 (taken as the average of the three surrounding values) and define the boundary between the two subtasks. Our analysis compared the differences between the minimum PCPS at a boundary and the APCPS of its preceding subtask.

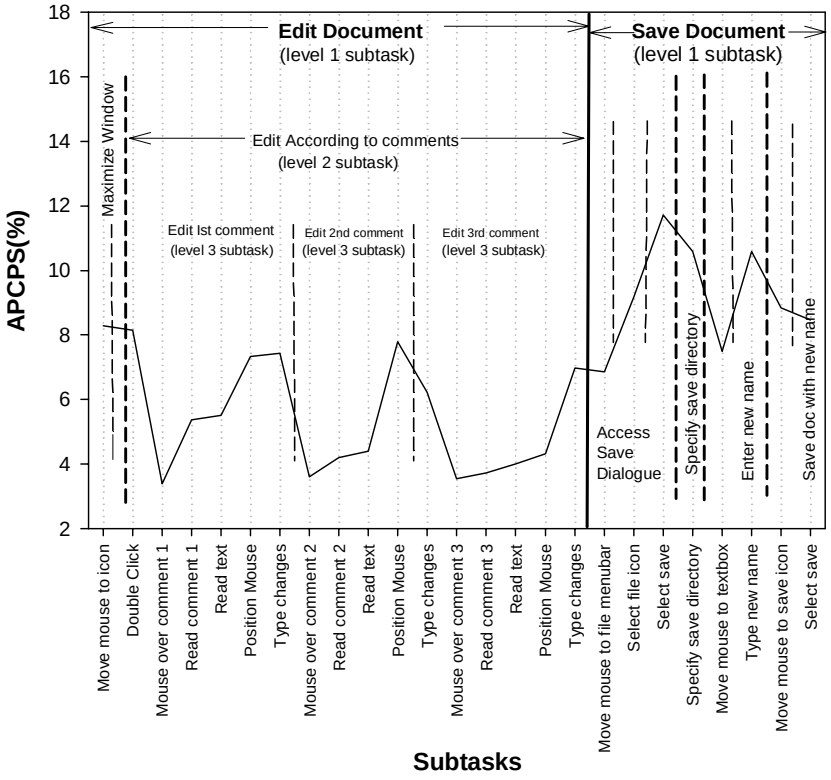


Figure 13: APCPS for subtasks in the document editing task. Solid lines indicate level 1, heavier dashed lines indicate level 2, and lighter dashed lines indicate level 3 boundaries. The x-axis enumerates the observable operators.

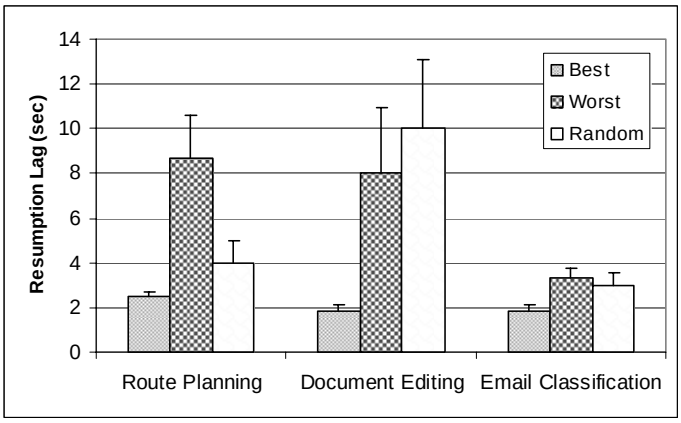


Figure 14: Time to resume a primary task after being interrupted in each timing condition.

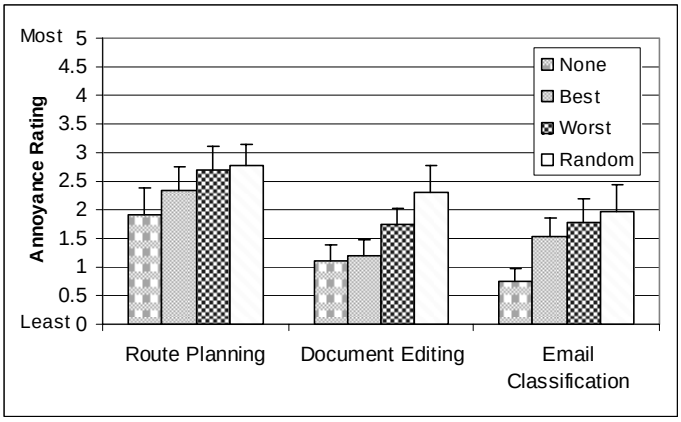


Figure 15: Ratings of annoyance when interrupted in each timing condition.

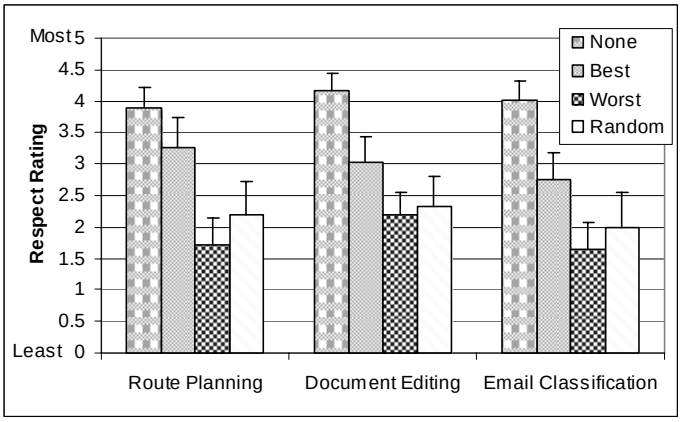


Figure 16: Ratings of respect for the interrupting application.