# Understanding Changes in Mental Workload during Execution of Goal-Directed Tasks and Its Application for Interruption Management

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

Notifications can have reduced interruption cost if delivered at moments of lower mental workload during task execution. Cognitive theorists have speculated that these moments occur at subtask boundaries. In this article, we empirically test this speculation by examining how workload changes during execution of goal-directed tasks, focusing on regions between adjacent chunks within the tasks, that is, the subtask boundaries. In a controlled experiment, users performed several interactive tasks while their pupil dilation, a reliable measure of workload, was continuously measured using an eye tracking system. The workload data was extracted from the pupil data, precisely aligned to the corresponding task models, and analyzed. Our principal findings include (i) workload changes throughout the execution of goal-directed tasks; (ii) workload exhibits transient decreases at subtask boundaries relative to the preceding subtasks; (iii) the amount of decrease tends to be greater at boundaries corresponding to the completion of larger chunks of the task; and (iv) different types of subtasks induce different amounts of workload. We situate these findings within resource theories of attention and discuss important implications for interruption management systems.

Categories and Subject Descriptors: H.1.2 [**Models and Principles**]: User/Machine Systems—*Human information processing*; H.5.2 [**Information Interfaces and Presentation**]: User Interfaces—*Evaluation / methodology, user-centered design*

General Terms: Design, Experimentation, Human Factors, Measurement

Additional Key Words and Phrases: Attention, interruption, workload, pupil size, task models, user studies

**ACM Reference Format:**

Bailey, B. P. and Iqbal, S. T. 2008. Understanding changes in mental workload during execution of goal-directed tasks and its application for interruption management. ACM Trans. Comput.-Hum. Interact. 14, 4, Article 21 (January 2008), 28 pages. DOI = 10.1145/1314683.1314689 http://doi.acm.org/10.1145/1314683.1314689

This research was supported in part by the National Science Foundation (NSF) under award no. IIS-0534462.

Authors' address: Department of Computer Science, University of Illinois, Urbana, IL 61801; email: {bpbailey;siqbal}@uiuc.edu.

## 1. INTRODUCTION

An increasing number of systems are seeking to proactively deliver information to end users through the use of notifications [McCrickard et al. 2003]. This trend is occurring in many multi-tasking environments [McFarlane and Latorella 2002] such as the desktop computing environment [Bailey and Konstan 2006; Czerwinski et al. 2000b; Jackson et al. 2001], command and control rooms [Stanton 1994], and aviation and automobile cockpits [Dismukes et al. 1998; Latorella 1996; Lee et al. 2004].

On the one hand, users often desire or need the information delivered through the use of notifications, for example, to facilitate near instant communication [Czerwinski et al. 2000a; Dabbish and Kraut 2004; Latorella 1996], to maintain awareness of peripheral information [Maglio and Campbell 2003], or to be reminded of upcoming activities [Dey and Abowd 2000]. On the other hand, delivering notifications in a proactive manner runs the serious risk of interrupting the user's ongoing task. For example, studies have shown that interrupting a user's task at random moments can cause decreased performance on the main task [Bailey and Konstan 2006; Czerwinski et al. 2000a; Kreifeldt and McCarthy 1981; Latorella 1996; Rubinstein et al. 2001] as well as increased feelings of frustration and anxiety [Adamczyk and Bailey 2004; Bailey and Konstan 2006; Zijlstra et al. 1999].

An emerging body of empirical research is now beginning to probe how manipulating the *time* at which a notification is delivered relative to the execution of the ongoing task impacts costs of interruption, for example, see Bailey and Konstan [2006], Czerwinski et al. [2000b], and Monk et al. [2002]. A central goal of this corpus of empirical work is to identify effective strategies that can meaningfully reduce costs of interruption and that can be practically implemented within interruption management systems.

A common theoretical foundation for much of this empirical research has been Miyata and Norman's [1986] influential article arguing that notifications would have lower interruption cost if delivered at moments of lower mental workload. It was further argued that these moments occur at the regions between adjacent chunks within the structure of the task's execution, that is, at *subtask boundaries*. At these moments, for example, the executive system may have just released attentional resources allocated for the previous subtask, but not yet acquired resources for the next [Wickens 2002]. This presumed lull in resource allocation may be commensurate with a temporary decrease in mental effort.

Though the arguments are certainly reasonable, whether a user's mental workload is indeed lower at subtask boundaries has never been explicitly tested, and assuming the veracity of these arguments for interruption management could thus be premature. Furthermore, since interactive tasks can typically be decomposed into recursive patterns of goal formulation and execution, there are many boundaries at many levels within a task's hierarchical structure [Card et al. 1983]. In this article, we refer to this hierarchical structure as the *task model*. It is thus unclear whether workload would decrease at all, some, or none of the boundaries within a task model, or whether the workload patterns exhibited within the structure of one task would hold across others.

In this work, we seek to empirically examine how a user's mental workload changes during the execution of goal-directed tasks, focusing on subtask boundaries and their level (depth) within the corresponding task model. In a carefully controlled experiment, users performed several interactive tasks; route planning, document editing, and e-mail classification. The tasks were necessarily simple, but exhibited a relatively complex execution structure given the type of analysis to be performed. While the tasks were performed, users' pupil dilation was continuously measured using a head-mounted eye tracking system. Research has shown that pupil dilation is a reliable and valid indicator of mental workload [Beatty 1982; Just et al. 2003; Pomplun and Sunkara 2003; Verney et al. 2004]. The workload data was extracted from the raw pupil data using known techniques (see Section 3.5). To analyze the resulting workload data, we first developed and validated models describing the execution structure of each task. We then precisely aligned each user's workload data to the corresponding task model, aggregated the aligned models across users for each task, and analyzed the workload at various regions of the model.

Our principal results from this analysis include (i) workload changes throughout the execution of goal-directed tasks; (ii) workload exhibits transitory decreases at subtask boundaries relative to the preceding subtask; (iii) the amount of this decrease is greater for boundaries higher in the task model, that is, those that correspond to the completion of larger chunks of the task; and (iv) different types of subtasks induce different amounts of workload. We discuss these results from the perspective of resource theories of attention and discuss important implications of these results for interruption management systems.

Though parts of this work have been previously reported in Iqbal et al. [2005] and Iqbal and Bailey [2005], this article provides results from a revised analysis of the data (explained in Section 3.6); results from the analysis of an additional experimental task not previously reported, giving further confidence in our findings; and a much more thorough discussion of the implications of our findings for interruption management.

## 2. RELATED WORK

In this section, we discuss costs of interruption, how our work contributes to strategies for mitigating those costs as well as systems that reason about when to interrupt, and our rationale for using pupil dilation as the measure of mental workload in this research.

### 2.1 Costs of Interruption

Controlled studies have clearly demonstrated that interrupting users engaged in tasks has considerable negative impact on task completion time [Cutrell et al. 2001; Czerwinski et al. 2000a, 2000b; Kreifeldt and McCarthy 1981; McFarlane 1999; Monk et al. 2002], error rate [Latorella 1998], decision making [Speier et al. 1999], and affective state [Adamczyk and Bailey 2004; Bailey and Konstan 2006; 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 the tasks, commit up to twice the errors, and experience up to twice the increase in negative affect compared to when those same peripheral tasks are scheduled at task or subtask boundaries [Adamczyk and Bailey 2004; Bailey and Konstan 2006]. A leading explanation for why this particular strategy has shown positive effects is because users are experiencing transient reductions in workload as they cross through boundaries during the task sequence. The experiment reported in this article seeks to test the veracity of this argument and produce further guidelines for understanding when notifications can be delivered such that the costs of the ensuing interruption are mitigated.

It is also important to note that interruption costs can have significant consequences. For example, in safety-critical domains, a short response delay or error committed due to an ill-timed notification could cause loss of life or catastrophic damage [McFarlane and Latorella 2002]. In office settings, unnecessary increases in frustration caused by poorly timed notifications could seriously degrade the user experience [Shneiderman 1997].

### 2.2 Leveraging Workload as a Means for Reducing Costs of Interruption

Researchers have theorized that notifications would have reduced interruption cost if they were delivered at moments of lower workload during execution of the ongoing task, and that these moments occur at (sub)task boundaries [Miyata and Norman 1986]. While empirical studies show that scheduling interruptions at certain boundaries or other moments can mitigate interruption cost [Adamczyk and Bailey 2004; Bailey and Konstan 2006; Cutrell et al. 2001; Czerwinski et al. 2000a, 2000b], researchers can only *assume* why as the workload experienced at those moments was never empirically measured. Without a tested theoretical basis for understanding why certain moments exhibit lower cost, it is difficult to distill these empirical findings into more general principles for identifying lower cost moments within a task for delivering notifications.

In addition, interactive tasks can be decomposed into recursive patterns of goal formulation and execution, creating many boundaries at many levels within the task's hierarchical model [Card et al. 1983]. It is thus unclear as to *which* of these boundaries would have the lowest workload (and thus cost) for interruption. By analyzing workload during task execution, our work seeks to contribute further understanding of just where moments of lower workload occur within the structure of goal-directed tasks.

### 2.3 Systems that Reason about Interruption

Systems are being developed that computationally reason about appropriate moments for interrupting users engaged in tasks (e.g., Bailey et al. [2006], Fogarty et al. [2005], Horvitz and Apacible [2003], and Hudson et al. [2003]). The general approach is to weigh the value of delivering information against the cost of interrupting the user's ongoing task within a decision-theoretic framework [Horvitz et al. 1999]. Given the complexity of creating and applying such frameworks, the focus of research to date has been on understanding how to compute accurate costs of interruption. Systems typically compute cost values using non task-specific cues such as mouse and keyboard activity, visual and acoustical analysis of the task environment, and scheduled activities of the user.

An important source of information that is missing from current systems is knowledge about the current workload of a user, as workload is directly related to interruption cost [Kreifeldt and McCarthy 1981]. One method for acquiring this information is to connect a physiological measure of workload directly to a reasoning engine [Chen and Vertegaal 2004]. However, in situations where using such a measure is not practical or desirable, which is arguably the more common case, it would be useful to be able to approximate workload given knowledge about the ongoing task such as it its hierarchical structure.

Results from our work provide useful heuristics for assigning costs of interruption at subtask boundaries and other moments within the structure of goal-directed tasks based on understanding patterns of workload exhibited during their execution. Considering this information will allow systems to make finer-grained decisions about when to interrupt.

### 2.4 Use of Pupil Size as a Measure of Workload

The methodology used in our experiment required the use of a measure of mental workload. Any measure could have been used as long as it was continuous, immediate, low latency, and valid. After a review of the literature, combined with local 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 indicator of mental workload [Beatty 1982; Hess and Polt 1964; Hoecks and Levelt 1993; Juris and Velden 1977; Kahneman 1967; Marshall 2002; Nakayama and Takahashi 2002; Takahashi et al. 2000]. Though some experiments have not detected a relationship between task difficulty and pupil dilation (e.g., see Lin et al. [2003]), the relationship does seem to hold in the general case. For example, Beatty [1982] reviewed a 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.

Researchers have also pursued many other measures of workload, including event-related brain potential [Donchin et al. 1986; Kok 1997; Kramer et al. 1986], electro-encephalographic activity [Gale and Edwards 1983; Gevins and Schaffer 1980; Phelps and Mazziotta 1985], eye movement and blink rate [Takahashi et al. 2000], heart rate variance [Rowe et al. 1998], performance measures [O'Donnell and Eggemeier 1986], and subjective ratings [Hart and Staveland 1988].

Relative to these measures, the use of pupil dilation offers many advantages [Kramer 1991]. For example, this measure is *continuous* meaning that it provides a steady stream of workload data; it measures allocation of attentional resources in a *holistic* manner rather than specific pools; it has *low latency*, usually responding to a change in workload in 300–500 ms; and it is *immediate*, a few recent data samples indicates workload, which simplifies analysis of the data. However, careful experimental control must be maintained with pupil dilation, as it can be considerably affected by environmental factors such as changes in ambient illumination or screen luminance.

A caveat of using pupil dilation as an indicator of mental workload is that it has rarely been used in interactive computing environments. We explored this issue in prior work [Iqbal et al. 2004] and our results showed that pupil dilation does correlate with the workload induced by interactive tasks, assuming appropriate environmental controls.

Building on this prior work, our current experiment seeks to better understand how workload changes within the structure of goal-directed tasks, paying particular attention to subtask boundaries. Note that we are using pupil dilation as a means for studying how workload changes in relation to a task's hierarchical structure, and distilling the empirical results into guidelines for interruption management. We are not necessarily advocating the use of pupil dilation or any other physiological measure as a real-time component of an interruption management system.

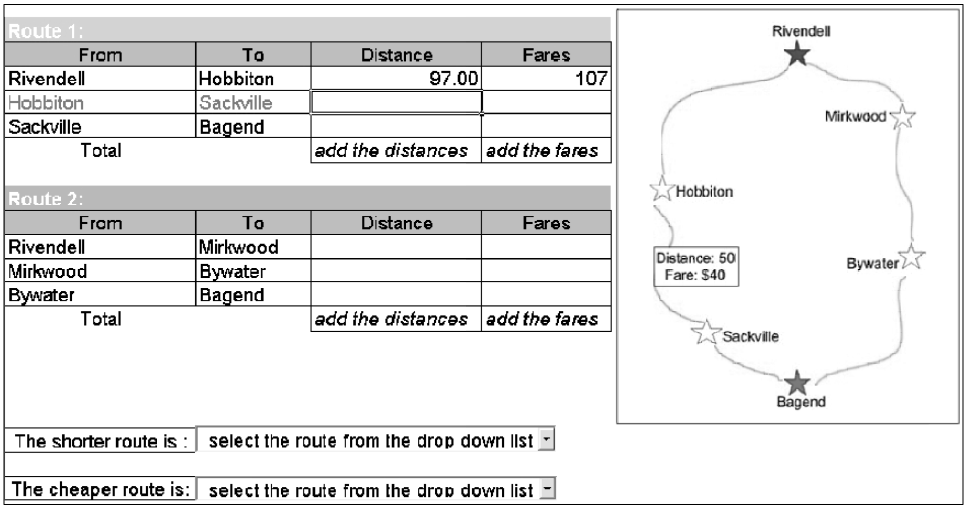
## 3. UNDERSTANDING WORKLOAD CHANGES DURING TASK EXECUTION

The purpose of our experiment is to develop further understanding of the relationship between mental processing effort (i.e., *workload*) and the structural characteristics of goal-directed tasks. As a starting point, our focus is on examining how workload changes at subtask boundaries within the hierarchical structure of a task's execution, how much this change differs at different levels within the task hierarchy, and how much workload changes among different types of subtasks. Answers to these questions will advance understanding of how to compute accurate costs of interruption during interactive tasks.

### 3.1 Experimental Tasks

For the experiment, three interactive tasks were developed:

—*Route Planning*. An interactive map was provided that showed two separate routes between two cities marked with start and end symbols (see Figure 1). 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 cursor over a segment. To perform the task, the user moved the cursor over the first segment in the map corresponding to the first route, committed the distance and fare information shown in the tooltip to memory (the tooltip disappeared when the cursor was moved away), and entered the data into the corresponding row in the table. A user completed each row in the table for the first route, mentally added the distance and fare columns, and entered the results into the last row. The user then repeated this process for the second table and route. Distance and fare values were manipulated (number of digits) to affect the difficulty of storing and recalling their values from memory as well



Route 1:

| From | To | Distance | Fares |
|---|---|---|---|
| Rivendell | Hobbiton | 97.00 | 107 |
| Hobbiton | Sackville | | |
| Sackville | Bagend | | |
| Total | | *add the distances* | *add the fares* |

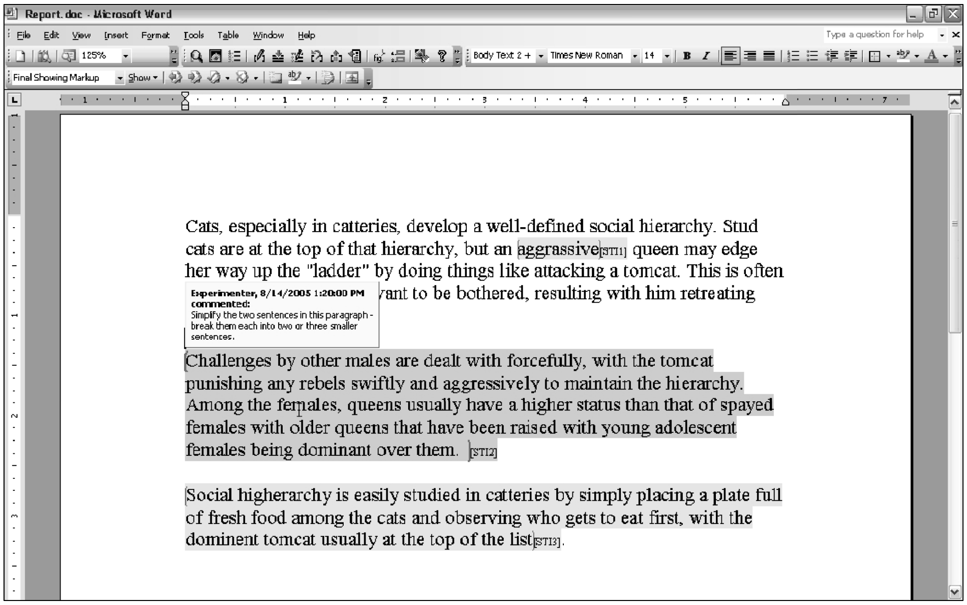
Route 2:

| From | To | Distance | Fares |
|---|---|---|---|
| Rivendell | Mirkwood | | |
| Mirkwood | Bywater | | |
| Bywater | Bagend | | |
| Total | | *add the distances* | *add the fares* |

The shorter route is : select the route from the drop down list

The cheaper route is: select the route from the drop down list

Fig. 1. 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 shorter and the cheaper of the two routes.



Cats, especially in catteries, develop a well-defined social hierarchy. Stud cats are at the top of that hierarchy, but an aggrassive[STI1] queen may edge her way up the "ladder" by doing things like attacking a tomcat. This is often

Experimenter, 8/14/2005 1:20:00 PM commented:
Simplify the two sentences in this paragraph - break them each into two or three smaller sentences.

want to be bothered, resulting with him retreating

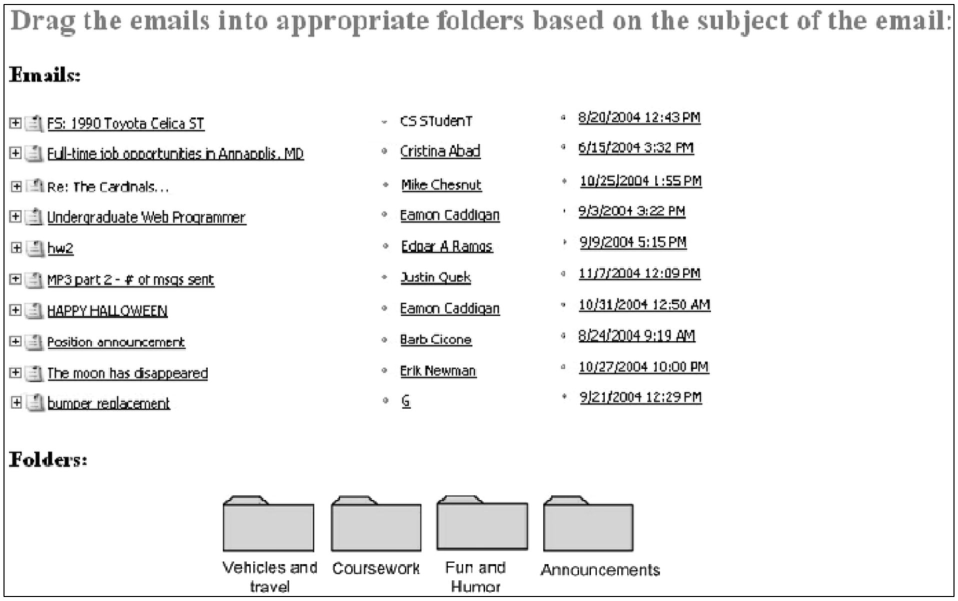
Challenges by other males are dealt with forcefully, with the tomcat punishing any rebels swiftly and aggressively to maintain the hierarchy. Among the females, queens usually have a higher status than that of spayed females with older queens that have been raised with young adolescent females being dominant over them. [STI2]

Social higherarchy is easily studied in catteries by simply placing a plate full of fresh food among the cats and observing who gets to eat first, with the dominent tomcat usually at the top of the list[STI3].

Fig. 2. The document editing task. A user edited the document based on each of three annotations. Once edited, the document was saved to a specified directory and file name.

as computing their sum. After completing both tables, the user selected the shorter and the cheaper of the two routes from drop down lists.

—*Document Editing*. A user was given a document with three annotations (see Figure 2). The content of the document was about the social hierarchy of a common pet (cats), selected because we felt it would be familiar



Drag the emails into appropriate folders based on the subject of the email:

**Emails:**

| | | |
|---|---|---|
| FS: 1990 Toyota Celica ST | CS STudenT | 8/20/2004 12:43 PM |
| Full-time job opportunities in Annapolis, MD | Cristina Abad | 6/15/2004 3:32 PM |
| Re: The Cardinals... | Mike Chesnut | 10/25/2004 1:55 PM |
| Undergraduate Web Programmer | Eamon Caddigan | 9/3/2004 3:22 PM |
| hw2 | Edgar A Ramos | 9/9/2004 5:15 PM |
| MP3 part 2 - # of msgs sent | Justin Quek | 11/7/2004 12:09 PM |
| HAPPY HALLOWEEN | Eamon Caddigan | 10/31/2004 12:50 AM |
| Position announcement | Barb Cicone | 8/24/2004 9:19 AM |
| The moon has disappeared | Erik Newman | 10/27/2004 10:00 PM |
| bumper replacement | G | 9/21/2004 12:29 PM |

**Folders:**

Fig. 3. The email classification task. Users reasoned about the classification of each email (starting from the top) using its subject descriptor, and then dragged the e-mail into the corresponding folder below. These actions were repeated for each of the emails in the list.

and understandable to most users. A user edited the document according to each annotation, which appeared as a tooltip when the cursor was moved over the corresponding highlight. After reading an annotation, the user located the corresponding text, made the desired edit, and repeated two more times. The document was saved to a specified directory and file, given a priori. The edits were manipulated to have varying difficulty, for example, the easiest edit was to correct one misspelled word, the medium edit was to locate and correct two misspelled words, and the most difficult was to rephrase a sentence so that it was grammatically correct.

—*E-mail Classification*. For this task (see Figure 3), a user was asked to classify a set of ten email messages into a set of supplied categories; for example, coursework, vehicles and travel, announcements, and fun and humor. The user would review the subject descriptor of a message, reason about which category it belonged to, and drag the message into the corresponding folder. The user then repeated this sequence for the remaining messages. The content of the subject descriptors was manipulated to affect the difficulty of the classification subtask, for example, some descriptors had the name of the destination category within it while others were more ambiguous.

These tasks were carefully designed to have meaningful subtasks of varying difficulty, well-defined boundaries between subtasks, a representative sample of interaction, and a prescribed execution sequence. A prescribed sequence was necessary to be able to align each user's workload data to the corresponding model of task execution. The lower-level cognitive subtasks, for example, memory store and recall, comprehension, and reasoning are representative of those within many other tasks. Though the tasks are relatively simple, it is important to note that they are more complex and of longer duration than tasks used in many prior experiments involving pupillary response, for example, see tasks used in Bradshaw [1967], Hytintk et al. [1995], Juris and Velden [1977], Kahneman [1967], and Takahashi et al. [2000].

### 3.2 Users and Equipment

A total of 24 users (7 female) participated in the experiment, with ages ranging from 19 to 50 ($M = 25.4$). All users had normal or corrected-to-normal vision. As users performed tasks, their pupil data was recorded using a head-mounted eye tracking system (Eyelink II). The eye-tracker sampled the pupil at a high temporal frequency of 250 Hz with spatial accuracy to about 1/100th of a millimeter using corneal reflection. Lighting and noise levels of the task environment were well controlled. Twelve users performed both the route planning and document editing tasks while the remaining twelve performed the email classification task. This reduced the time that any one user had to wear the eye tracking equipment, but did not impact the results as each task was analyzed separately.

### 3.3 Procedure

Upon arrival at the lab, we went through an informed consent process with the user and provided 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 experimental 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. Time-stamped samples of pupil data were logged to a file while the user's screen interaction was recorded with eye gaze overlaid. Because the videos and pupil data received time stamps from the same clock, we could precisely align the two data sets. The entire experimental session lasted about 30 minutes.

### 3.4 Task Models and Validation

Figures 4, 5, and 6 show the task models for the Route Planning, Document Editing, and Email Classification tasks, respectively, reusing repetitive parts for brevity. *Subtask* refers to any node in the model, and *subtask boundary* refers to the period between adjacent subtasks. *Level of boundary* between two adjacent subtasks is 1 + the depth of their shared ancestor in the model. For example, in Figure 4, 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 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 memory store,

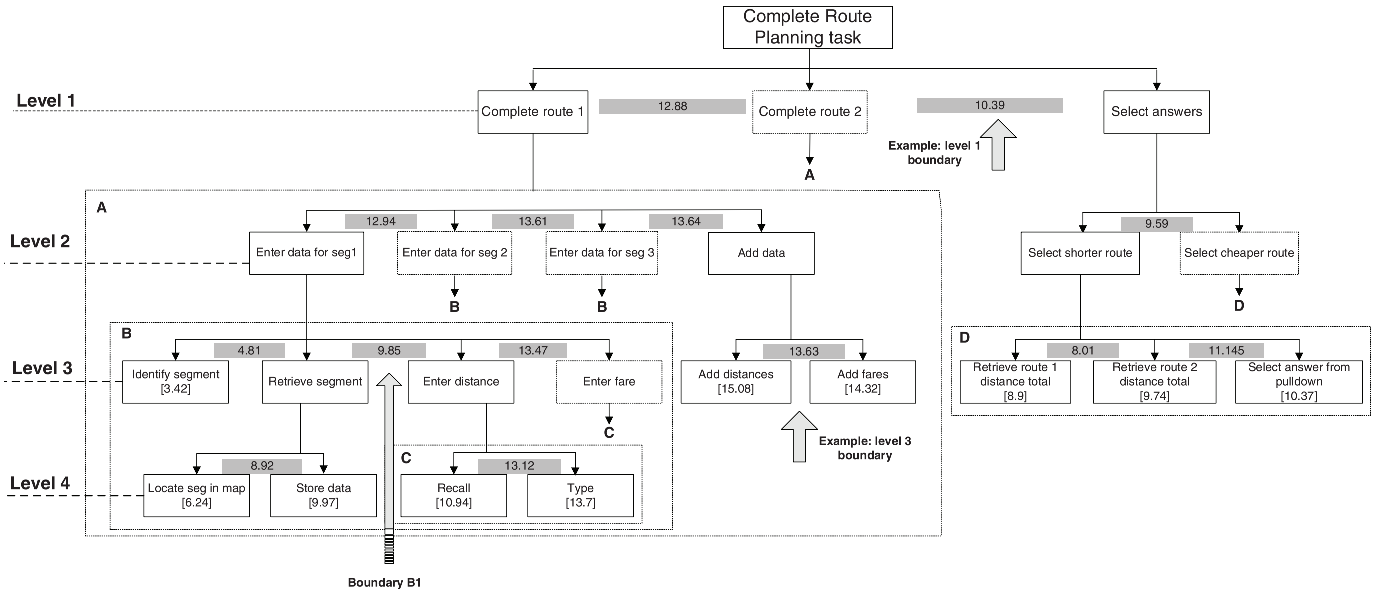


Fig. 4. A workload aligned task model for Route Planning. The interior nodes represent goal nodes, the leaf nodes represent operators, and time moves from left to right. Regions A, B and C show parts of the task repeated elsewhere in the model. Within each subtask, we provide the [APCPS] for that subtask. Each shaded area indicates a boundary and contains the [*APCPS*] across it.

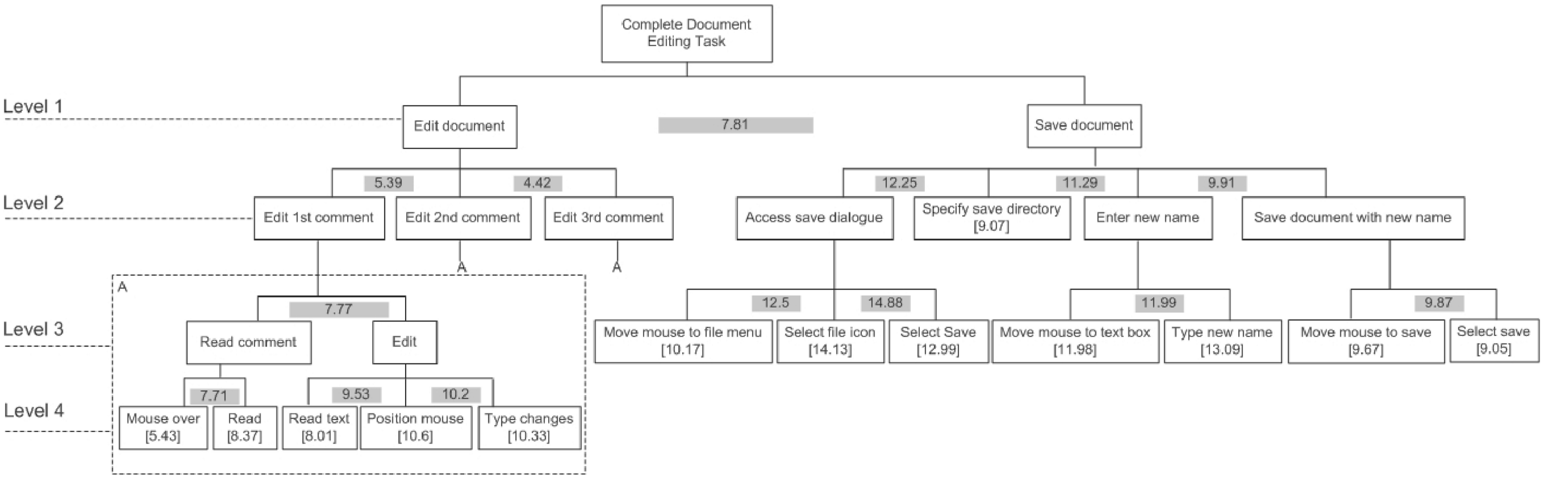


Fig. 5. A workload aligned model for Document Editing. The interior nodes represent goal nodes, the leaf nodes represent operators, and time moves from left to right. Regions A shows parts of the task repeated elsewhere in the model.

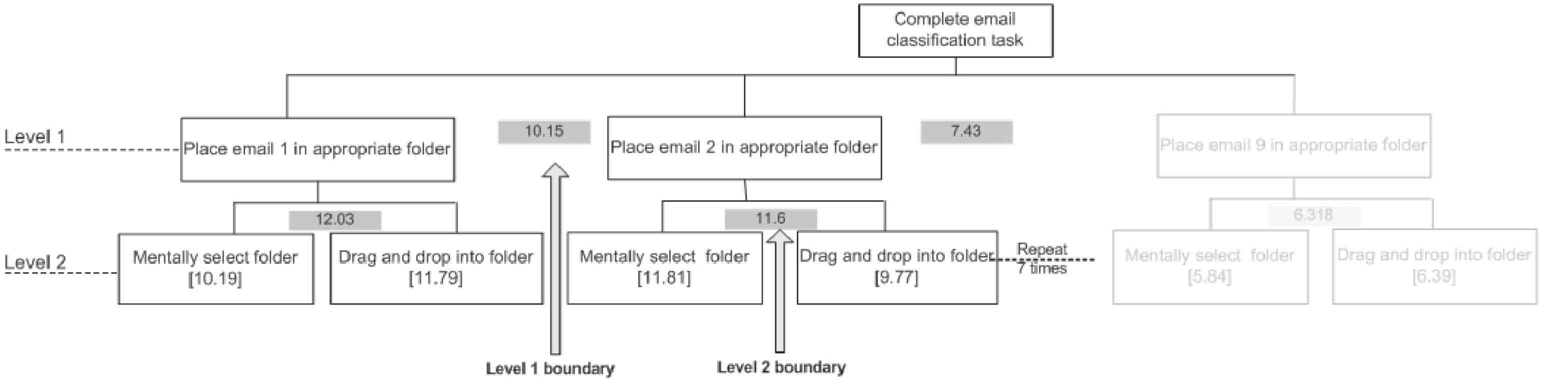

Complete email classification task
Level 1
Level 2
Place email 1 in appropriate folder
10.15
Place email 2 in appropriate folder
7.43
Place email 9 in appropriate folder
12.03
Mentally select folder [10.19]
Drag and drop into folder [11.79]
11.6
Mentally select folder [11.81]
Drag and drop into folder [9.77]
Repeat 7 times
6.318
Mentally select folder [5.84]
Drag and drop into folder [6.39]
Level 1 boundary
Level 2 boundary

Fig. 6. A workload aligned task model for Email Classification. The interior nodes represent goal nodes, the leaf nodes represent operators, and time moves from left to right. The level 1 subtask is repeated 9 times. Within each subtask, we provide the [APCPS] for that subtask. Each shaded area indicates a boundary and contains the [*APCPS*] across it.

memory recall, reasoning, language comprehension, language generation, or motor operator.

The task models were developed in an iterative manner. For each task, we developed an initial model through our own analysis of the task's execution. The initial models were refined based on screen interaction videos of four users performing the tasks prior to and independent from the reported experiment. The interaction sequences predicted by the leaves in our task models were compared to the sequences observed in the interaction videos, and our models were refined until a high degree of agreement was reached.

We measured the accuracy of our final task models by comparing the operators in the models to the observable events (keyboard, mouse, and eye gaze) in the interaction videos recorded during the experiment. 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 as discussed in Card et al. [1983].

The final task model for Route Planning has 4 levels and 81 nodes. The average error rate was 2.81% with no detectable pattern to the errors. Repeating this same process for Document Editing, the resulting model had 4 levels and 38 nodes with an average error rate of 2.3%. The model for E-mail Classification had 2 levels and 25 nodes and matched users' execution of the task without error.

### 3.5 Measurements

Following prior work [Hess 1972], workload was calculated as the percent change in pupil size (PCPS) for each sample of data relative to the baseline. Eye blinks, which were identified by the eye tracking system, were accounted for by linearly interpolating the missing values [Verney et al. 2001]. For each subtask and boundary, we also computed the average PCPS (APCPS) for that region of data. The duration of subtasks ranged from about 25 ms for the lowest-level subtasks, to about 1 min for the higher level subtasks, to about 5 min for the root node (the entire task). The duration of boundaries ranged from about 8 ms to 6 seconds ($M = 487$ ms, $SD = 574$ ms), with higher level boundaries generally being of longer duration. A more detailed analysis of the durations of the boundaries within each task will be provided within the Results section.

### 3.6 Alignment and Revised Analysis

As the models accurately reflected a user's execution sequences in the tasks, we were able to precisely align a user's pupillary response to the task models. Since each user performed the tasks at different speeds, our approach was to align the pupil data to the subtasks in the model, not to time, starting from the leaf operators and working upwards.

For each leaf subtask, we identified the beginning and end time stamp from the screen interaction video and used these timestamps to index the pupillary response file. The corresponding PCPS data was then extracted and associated with that subtask. APCPS for higher-level subtasks was calculated by averaging the PCPS values associated with their child subtasks, and this process was repeated until the root node was reached. For boundaries, we extracted the PCPS data from the end timestamp of the preceding subtask to the begin timestamp of the subsequent subtask and computed APCPS values as before.

Overall, analysis of the data was consistent with our prior work [Iqbal et al. 2005], but several improvements were included. First, to account for latency in the response of the pupil to the onset of a stimulus, we temporally shifted the pupil data by a small amount (500 ms) prior to aligning it with the task model [Kramer 1991].

Second, to compute decrease in workload at a boundary, we now compare the APCPS of the boundary to the APCPS of its preceding subtask. In prior work [Iqbal et al. 2005], the decrease was computed as the difference between the minimum value within the boundary (taken as the average of multiple surrounding points) and the APCPS of the preceding subtask. Our current analysis was revised to ensure that we are offering a fair comparison between these two regions of data.

Finally, when a non-leaf subtask precedes a boundary, we compare the APCPS of the boundary to the APCPS of the *last leaf operator* of that subtask. For example, in Figure 4, the decrease at boundary B1 is computed as the difference in APCPS between B1 and the operator *Store Data* (the last operator of *Retrieve segment*). This was done to ensure that earlier parts of longer subtasks were not unfairly affecting the comparison.

## 4. RESULTS

To provide further confidence in our measure of workload, we first check that regions within the execution structure of the tasks that were expected to induce lower/higher workload did indeed correspond to lower/higher values of PCPS. Then, for each task, we report results of how the different types of subtasks affected workload, how the level within a task model affected workload at the subtask boundaries, and how much workload differed between a boundary and its preceding subtask.

The reader should keep in mind that small changes in pupillary response can represent meaningful changes in workload, but that there is also an upper bound on how much a user's pupil size will increase due solely to increases in mental processing effort.

### 4.1 Validation of Workload Measure

To validate our workload measure, we compared pupillary response between different regions of the tasks that would presumably require different amounts of mental processing effort. For Route Planning, we performed an ANOVA with Load (fewest, middle, and most digits) as the factor on the APCPS of Recall subtasks.

Results showed that Load had a main effect on APCPS ($F(2, 46) = 6.24$, $p < 0.01$), where Recall subtasks that required more digits to be retrieved from memory had higher APCPS. For Email Classification, an ANOVA with Classification Difficulty (easier, more difficult) as the factor showed that the

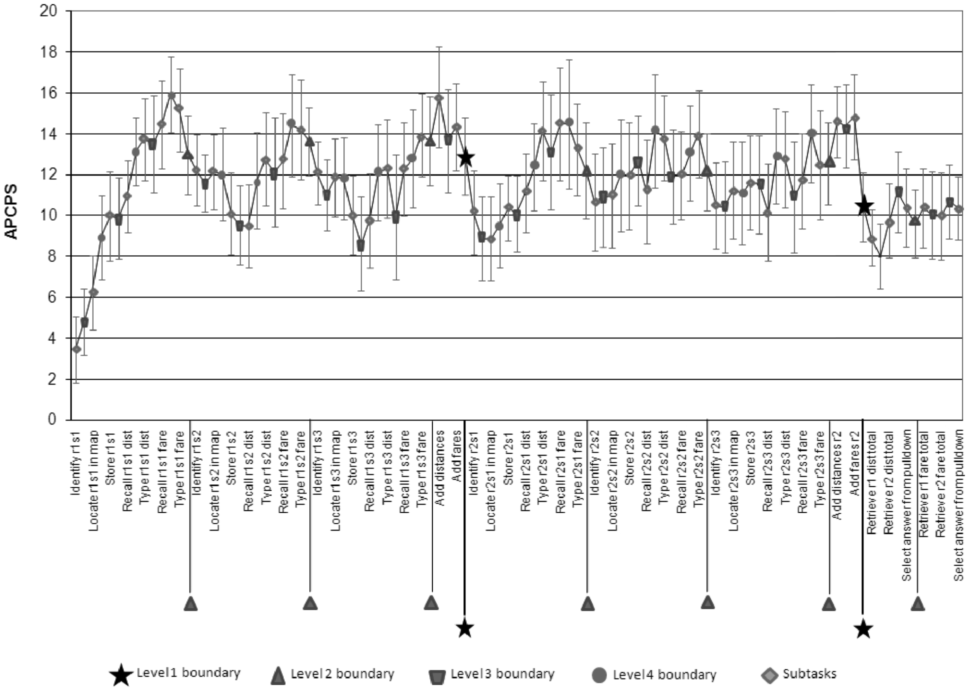


Fig. 7. APCPS of the leaf subtasks and all boundaries within the model for Route Planning. Vertical lines within the subtask labels demarcate Level 1 and 2 boundaries.

more difficult classification (where more mental effort was required to select the target folder for the e-mail message) had higher APCPS than for the easier classification (F(1, 7) = 9.81, $p < 0.05$). For Document Editing, Difficulty (simple, medium, and difficult edit) did not have a main effect, though the trends were in the expected direction. We attribute this lack of significance to the three types of edits being closer in terms of the mental effort required relative to the subtasks being compared for the other two tasks. Overall, these results confirm that users' pupil size was changing in response to the changing difficulty of the subtasks.

## 4.2 Route Planning

Figure 7 shows the average (across users) APCPS for each leaf (operator) subtask and all boundaries within the model for Route Planning. In the graph, note how workload rises quickly at the onset of the task and then rises and falls throughout execution. Inspection of the graph clearly shows transitory decreases in workload at the two level 1 boundaries.

4.2.1 *Workload During Subtasks.* To test whether performing subtasks induced workload over the baseline value, we performed a t-test on the APCPS values of the subtasks. Our analysis included only those subtasks that required cognitive effort such as storing, recalling, or reasoning about distance and fare information, rather than motor subtasks, as the relationship between cognitive effort and pupillary response is the one best established [Beatty 1982]. Results showed that APCPS was greater than 0 across subtasks ($M = 11.98$, $SD = 7.6$, $t(263) = 25.75$, $p < 0.001$). This represents about a 12% increase over the baseline value and shows that the subtasks did impose increased workload on a user.

An ANOVA with Subtask (Store, Recall, and Reasoning) as the factor showed a main effect on APCPS ($F(2,261) = 4.87$, $p < 0.01$). Post hoc tests showed that Reasoning induced more workload than Store (difference was 4.3 percentage points, $p < 0.01$) and Recall (difference was 3.2 percentage points, $p < 0.05$), while there was no difference found between Store and Recall subtasks.

4.2.2 *Workload at Subtask Boundaries.* A t-test showed that APCPS at boundaries was greater than 0 ($M = 11.68$, $SD = 7.62$, $t(611) = 37.91$, $p < 0.001$) and that Level had a main effect on the APCPS of boundaries ($F(3,608) = 2.61$, $p < 0.05$). Post hoc tests showed that the APCPS of boundaries at Level 3 ($M = 10.78$) was less than the APCPS of boundaries at Level 4 ($M = 12.59$, $p < 0.05$). Other pairs were not significant, though the means were in the expected direction ($M = 11.64$ for Level 1 and $M = 12.37$ for Level 2). Among all boundaries in the task model, the Level 2 boundary between *Retrieve route* 1 *total* and *Retrieve route* 2 *total* had the lowest APCPS (9.57) while the Level 3 boundary between *Add distances* and *Add fares* had the highest (13.99). This indicates that the workload carried through a boundary depends not just on the level in a model, but also on the mental demands of the surrounding subtasks.

The overall average duration of the boundaries was 590 ms. Level had a main effect on boundary duration ($F(3,593) = 11.26$, $p < 0.001$). Post hoc tests showed that level 1 boundaries ($M = 887$ ms) were of longer duration than level 3 ($M = 456$ ms, $p < 0.001$) and level 4 ($M = 483$ ms, $p < 0.01$) boundaries, and that level 2 boundaries ($M = 691$ ms) were of longer duration than level 3 ($p < 0.001$) and level 4 ($p < 0.001$) boundaries.

4.2.3 *Decrease of Workload at Subtask Boundaries.* Boundary Decrease is computed as the difference between APCPS at the boundary and the preceding subtask, as discussed in Section 3.6. With values for all boundaries included, a t-test showed that Boundary Decrease was slightly greater than 0, but did not reach a level of significance ($M = 0.029$, $SD = 3.91$, $t(612) = 0.18$, $p = 0.85$). This indicates that not all boundaries exhibit a detectable decrease in workload, likely because the numerous lower level boundaries were closely related and had a high degree of mental carryover.

However, when the lowest level boundaries (Level 4) are excluded, the same analysis now shows Boundary Decrease to be greater than 0 ($M = 0.7905$, $SD = 3.8$, $t(397) = 4.15$, $p < 0.001$). This effect continues to become stronger as lower-level boundaries are successively excluded from the analysis. This result indicates that workload temporarily decreases as a user crosses through a boundary during execution of a task, but the result only holds for boundaries that are above a certain level (depth) within the task model.

Exploring this pattern further, we found that Level had a main effect on Boundary Decrease ($F(3,608) = 18.42$, $p < 0.001$). Post hoc tests showed that

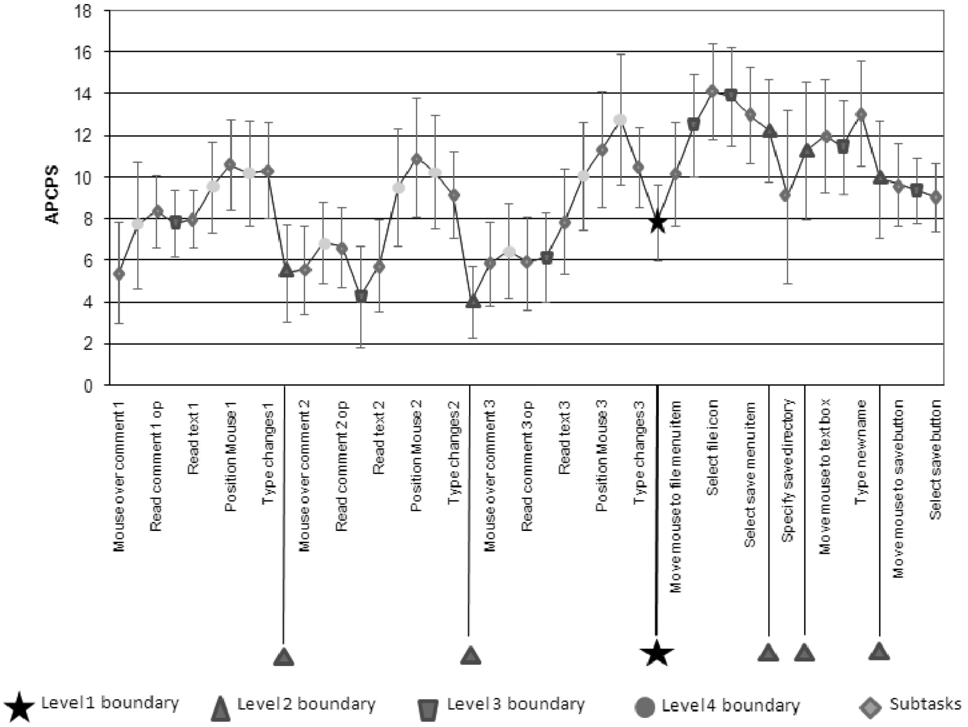


Fig. 8. APCPS of leaf subtasks and all boundaries within the model for Document Editing. Vertical lines within the subtask labels differentiate Level 1 and 2 boundaries.

decreases at level 1 were greater than at level 2 (1.95 percentage points), level 3 (2.38 percentage points, $p < 0.05$) and level 4 (4.3 percentage points, $p < 0.001$). Level 2 decreases were greater than level 3 (0.43 percentage points) and level 4 (2.34 percentage points, $p < 0.001$), and Level 3 decreases were greater than level 4 (1.91 percentage points, $p < 0.001$).

Overall, this pattern shows that workload tends to decrease more at boundaries higher in the task model than at boundaries lower in the model. We also found that workload changed *within* the same level in a task model. For example, APCPS between the two level 1 boundaries was different ($F(1,22) = 5.31$, $p < 0.05$) with an absolute difference of 2.96 percentage points.

### 4.3 Document Editing

Figure 8 shows the average APCPS for each leaf subtask and all boundaries within the model for the Document Editing task. As in the Route Planning task, workload rises at the onset of the task, rises and falls throughout task execution, and temporarily decreases at salient boundaries within the task, that is, after completing each of the three edits.

4.3.1 *Workload During Subtasks.* Including only cognitive subtasks (language comprehension and generation, and recall), a t-test showed that APCPS for subtasks was greater than 0 ($M = 8.62$, $SD = 7.35$, $t(111) = 12.41$, $p <$

0.001). This shows an 8.62% increase over the baseline, meaning that the subtasks did induce increased workload, 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(2,109) = 4.19, $p < 0.05$). Recall induced more workload than Comprehension (difference was 5.9 percentage points, $p < 0.05$) and Generation (difference was 3.04), and Generation induced higher workload than Comprehension (difference was 2.86). These results are consistent with Route Planning where different types of subtasks also induced different amounts of workload.

4.3.2 *Workload at Subtask Boundaries.* A t-test showed that the APCPS of boundaries was greater than 0 ($M = 9.02$, $SD = 8.4$, t(233) = 16.43, $p < 0.001$). Level did not have a main effect, but the trends were in the expected direction ($M = 7.82$ for Level 1, $M = 8.22$ for Level 2, $M = 9.40$ for Level 3, and $M = 9.26$ for Level 4). Among all boundaries, the Level 2 boundary between *Edit second comment* and *Edit third comment* had the lowest APCPS (4.42), while the highest APCPS was at the Level 3 boundary between *Select file menu* and *Select save* (14.99).

The overall average duration of boundaries was 528 ms. Level had a main effect on the duration of a boundary (F(3,241) = 5.31, $p < 0.001$). Post hoc tests showed that boundaries at level 1 ($M = 1.1$s) were of longer duration than boundaries at levels 3 ($M = 461$ ms, $p < 0.05$) and 4 ($M = 401$ ms, $p < 0.01$), and boundaries at level 2 ($M = 746$ ms) were of longer duration than those at level 4 ($p < 0.05$).

4.3.3 *Decrease of Workload at Subtask Boundaries.* With all boundaries included, a t-test did not show Boundary Decrease to be greater than 0. As with the Route Planning task, excluding the lowest level boundaries (Level 4) and re-running the t-test now showed Boundary Decrease to be greater than 0 ($M = 1.34$, $SD = 4.01$, t(133) = 3.87, $p < 0.001$). An ANOVA with Level as the factor showed a main effect on the APCPS of boundaries (F(3, 234) = 16.81, $p < 0.001$). Boundary Decrease at Level 1 was similar to Level 2, but larger than level 4 (3.99 percentage points, $p < 0.01$). Level 1 also had quantitatively higher decrease than level 3 (2.61 percentage points), but the difference did not reach a level of significance. Boundaries at level 2 were found to have a larger decrease than at level 3 (2.94 percentage points, $p < 0.001$) and level 4 (4.32 percentage points, $p < 0.001$). Boundaries at level 3 tended to have a larger decrease than at level 4 (1.39 percentage points), but did not reach significance. Overall, this pattern of results shows that workload tends to decrease more when a boundary higher in the model is reached during an interaction sequence, consistent with results from Route Planning.

### 4.4 Email Classification

Figure 9 shows the average APCPS for subtasks and all boundaries within the model for Email Classification. Analogous with the other tasks, the graph shows a temporary decrease in workload at the top-level boundary corresponding to

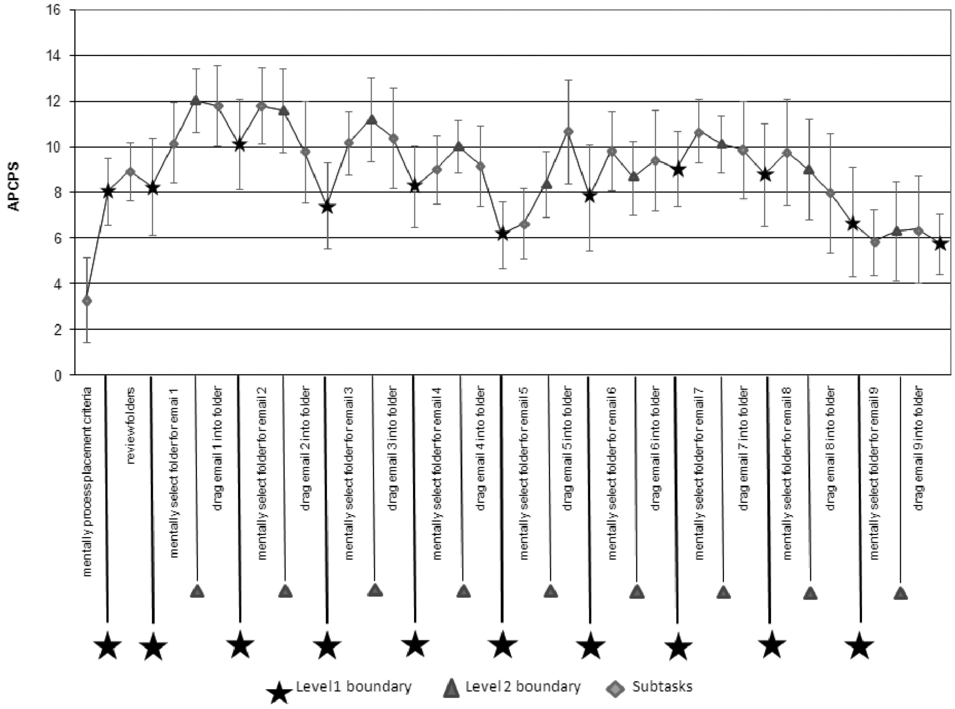


Fig. 9. APCPS of leaf subtasks and all boundaries within the model for Email Classification. Vertical lines within the labels demarcate Level 1 and 2 boundaries.

the completion of the classification of each mail message. Also, note that the structure of this task is simpler than the previous tasks, as there were only two levels in the task model and only one type of cognitive subtask—reasoning about the destination folder.

4.4.1 *Workload During Subtasks and at Boundaries.* For the cognitive subtasks (reasoning), a t-test showed that APCPS was greater than 0 ($M = 9.48$, $SD = 5.6$, $t(71) = 14.45$, $p < 0.001$). This represents a 9.48% increase over the baseline and shows that the subtasks did induce increased workload, as with the previous two tasks. In addition, a t-test showed that the APCPS of boundaries was greater than 0 ($M = 8.93$, $SD = 6.05$, $t(142) = 17.65$, $p < 0.001$), with Level 1 boundaries having lower APCPS ($M = 7.93$) than the Level 2 boundaries ($M = 9.92$). Boundaries at level 1 ($M = 484$ ms) were of longer duration than those at level 2 ($M = 320$ ms; $F(1, 167) = 7.41$, $p < 0.01$), and the overall average duration for a boundary was 405 ms.

4.4.2 *Decrease of Workload at Subtask Boundaries.* A t-test showed that Boundary Decrease was greater than 0 ($M = 0.6816, SD = 3.73$, $t(142) = 2.19$, $p < 0.05$). An ANOVA showed that Level had a main effect on Boundary Decrease ($F(1,141) = 14.41$, $p < 0.001$), with the decrease at Level 1 boundaries ($M = 1.82$, S.D. $= 2.94$) being larger than at Level 2 boundaries ($M = -0.44$, S.D. $= 4.08$). These results show that workload decreased at boundaries and that

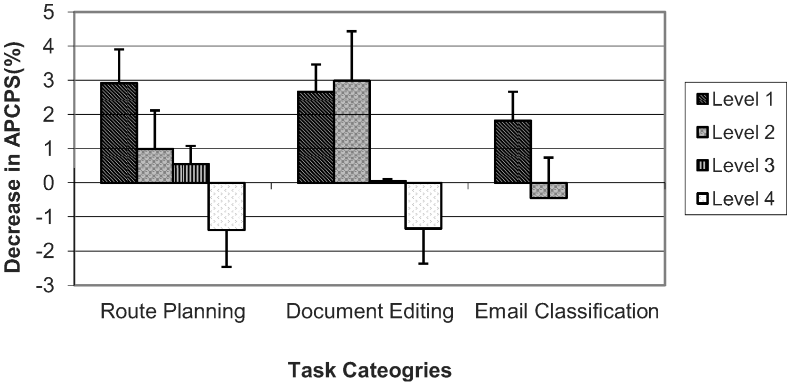


Fig. 10. Decrease in the APCPS of boundaries, shown as a function of level and task. Higher level boundaries (1 and 2) show larger decreases in workload than lower level boundaries (3 and 4), which exhibited little or no decrease.

it decreased more at boundaries higher in the task model, which is consistent with results from the other two tasks.

## 5. DISCUSSION

The purpose of our experiment was to better understand how workload changes during execution of goal-directed tasks. Here, we summarize our primary findings, situate these findings within resource theories of human attention, and discuss implications of these findings for systems that reason about when to interrupt users engaged in tasks.

First, our results provide further evidence showing that a user's mental workload changes throughout execution of goal-directed tasks. From the perspective of resource theories of attention [Kahneman 1973; Wickens 1980, 1991, 2002], this result indicates that the executive system does not statically allocate attentional resources at the onset of a task stimulus, but dynamically allocates and releases resources throughout its execution.

For interruption management, this implies that the moment at which a notification is delivered relative to a user's ongoing task will affect interruption cost. Indeed, several studies have shown that the moment that a task is interrupted does affect interruption cost (e.g., see Bailey and Konstan [2006], Czerwinski et al. [2000b], Iqbal and Bailey [2006], and Monk et al. [2002]). In these studies, cost was measured in terms of the time needed to resume the primary task. Our results strongly suggest that the interruptions resulting in lower cost likely occurred at moments of lower workload, when more attentional resources were available for the interrupting task and fewer resources were needed to resume the previously suspended task [Rubinstein et al. 2001].

Second, we found that transitory decreases in workload are experienced as subtask boundaries are reached during task execution. This result is summarized in Figure 10. A plausible explanation is that the executive system releases attentional resources allocated for the just completed subtask, but has not yet acquired resources for the subsequent subtask. An important implication of this result is that it establishes the principle of using *defer-to-boundary* policies for reducing costs of interruption caused by notifications. For example, if notifications could be deferred until a boundary is reached, the executive system would have more resources available for performing the interrupting task [Rubinstein et al. 2001]. Such policies would also be beneficial because boundaries typically represent moments between explicit interactions with a system. For example, this would prevent notifications from being delivered during text entry or other motor subtasks.

Third, our results showed that the transitory decrease in workload tended to be larger when boundaries higher in a model were reached during task execution. In addition, higher-level boundaries were found to be of longer duration than lower-level boundaries. These results indicate that more resources are released when more salient breakpoints are reached during a task. Whereas, when lower-level boundaries are reached, the amount of resources released is apparently small, possibly due to cognitive chunking of repetitive or skilled actions [Newell and Rosenbloom 1981] or to large carryover of information being actively maintained in short-term memory. The implication of this finding is that interruption management systems should favor boundaries that represent more salient breakpoints during a user's ongoing interaction, as these should result in lower costs of interruption. In addition, systems need not consider boundaries that are lower in the task model (roughly beyond the third level) as these appear to provide little benefit over non-boundary moments.

Finally, the level of a boundary in a task model cannot always predict whether there would be a larger decrease or lower absolute value of workload at a boundary. For example, in the document editing task, the boundary with the lowest workload was between the second and the third edits, which was not a top level boundary. Similarly, for route planning, the lowest workload boundary was between selecting the shorter and the cheaper of the routes, which also was not a top-level boundary. A plausible explanation is that the executive system may be maintaining information in short-term memory or prospectively allocating resources in anticipation of a subsequent subtask across some boundaries, but not others [Trafton et al. 2003]. The implication is that using knowledge related to the hierarchical decomposition of a task within an interruption management system can offer only a rough approximation of interruption cost at various boundaries. A more precise determination of cost would require aligning a measure of workload to the model of task execution and rank ordering the boundaries based on their workload, similar to the methodology demonstrated in this work.

### 5.1 Linking Workload and Interruption Cost

Results from our experiment suggest that interrupting ongoing tasks at moments of lower workload should result in lower costs of interruption. As a first step towards testing this claim, we conducted a follow-up experiment where users were interrupted at different moments—*better*, *worse*, and *random*—while performing the same tasks presented in this work. As details of that experiment can be found in Iqbal and Bailey [2005], here we briefly summarize our methodology and findings.

Given the workload-aligned task models (Figures 4–6), *better* moments were selected as boundaries with lower workload, *worse* moments were selected as boundaries with higher workload, and *random* moments were any moments during execution of the tasks. As users (N = 12) performed the primary tasks, they were interrupted with a peripheral task at each of these moments, chosen randomly. Measures included time to resume the primary task, annoyance due to interruption, and level of respect attributed to the interrupting system.

Results showed that interrupting at the better moments had lower cost of interruption across tasks. Users resumed primary tasks 69% faster, experienced 18% less annoyance, and attributed 63% more respect to the interrupting system relative to being interrupted at worse moments. Similar differences were found between the better and random moments. However, interrupting at worse vs. random moments showed no difference, indicating that not all boundaries result in lower interruption cost relative to non-boundary moments. This may be particularly true for finer-grained boundaries that are deeper in a task model where there is little to no measurable decrease in workload. Overall, our results provided an important first step towards showing that workload can be used to predict which boundaries within a task model have higher/lower costs of interruption. As only boundaries with larger differences in workload were tested, future work is needed to assess the amount of change in workload that is required before a meaningful change in interruption cost could be detected at a boundary or other moment in a task.

### 5.2 Enabling Systems to Consider Workload for Reasoning about Interruption

Though our follow-up experiment showed that the cost of interruption could be reduced by deferring delivery of notifications until lower workload boundaries, methods are needed that would allow systems to consider similar information in practice. Here we discuss three such methods that would allow interruption management systems to directly or indirectly consider workload when reasoning about when to interrupt.

One approach follows directly from the methodology used in this paper. For example, workload-aligned task models would be developed by aligning a continuous measure of workload to the corresponding models of task execution. From these workload-aligned task models, the boundaries and subtasks would be rank ordered based on their workload and then mapped to a cost value. The model of the task and associated cost information would then be formally described using a task specification language such as that presented in Bailey et al. [2006]. As a user performs tasks, a monitoring system would match the ongoing interaction to the specifications, allowing the monitor to identify when a specific boundary or other moment was reached. The associated cost value could then be extracted from the specification and directly used to determine whether to interrupt, or passed to a broader reasoning framework [Horvitz et al. 2004]. This approach would be most appropriate for safety critical or other domains where tasks have fairly prescribed sequences and the range of possible tasks is somewhat constrained, but the cost of poorly timed interruption could cause loss of life or catastrophic damage. Example situations might include working through aviation checklists [Degani and Wiener 1993] or entering target information in a command/control interface [Guerlain and Willis 2001].

In other situations where high precision is not necessary, an alternative is to utilize a set of workload-based heuristics to assigns costs within static specifications of tasks. For example, lower costs could be assigned to higher-level boundaries and successively higher costs could be assigned to lower-level boundaries. Similar heuristics could be developed for different types of subtasks such as memory store and recall, language comprehension and generation, and reasoning. Though applying heuristics could only offer approximations, they could be expediently applied to many tasks, the values would still allow systems to better reason about when to interrupt, and the use of heuristics would eliminate the need to develop workload-aligned task models (a very large effort).

A third approach would be to detect boundaries directly from a user's task execution data, bypassing the need to construct any specifications of the tasks. Newly generated notifications would then be deferred, for example, until the next boundary was detected. This approach is feasible, as researchers have recently demonstrated the feasibility of building statistical models that are able to detect and differentiate boundaries within task execution data [Iqbal and Bailey 2007; Nair et al. 2005]. The main advantage of this approach is that formal specifications of user tasks are not needed to detect perceptually meaningful boundaries. The disadvantages are that only a subset of the boundaries that would otherwise be available from the hierarchical model of a task could be detected and different types of subtasks could not be easily considered. This approach may be most appropriate in situations where users perform a range of diverse tasks that have highly variable execution sequences, for example, as exemplified in office computing environments.

The last approach would be to link a real-time measure of workload directly into a system's reasoning framework [Chen and Vertegaal 2004]. Input could be provided, for example, by eye tracking systems embedded within a desktop monitor [Tobii-Systems] or by inexpensive heart rate sensors embedded within office chairs [Anttonen and Surakka 2005]. Though immediate knowledge of a user's changing workload could help systems better manage delivery of notifications, using such a measure may not be possible or desirable in many situations due to the expense or intrusiveness of the hardware or the lack of necessary controls within the task environment. In addition, systems that consider only the current workload being induced would still run the risk of interrupting a user's ongoing action, for example, entering text into a control, as opposed to interrupting at a moment *between* such actions.

### 5.3 Limitations

One limitation of our work involves the accuracy of the subgoal structure of the models of task execution. When developing the models, the sequential ordering of the leaf (operator) subtasks could be objectively compared to the actual execution sequences, and the models could be revised until high agreement was reached. Unfortunately, there is no known technique for measuring the accuracy of the subgoal structure of a model. Though we followed best practices, it is possible to have models that express the same operator sequence, but that have different subgoal structures. For example, in the model for document editing, if the level 1 subtask "Edit document" was collapsed, then each of the three edit subtasks and their corresponding boundaries would be shifted up to level 1. As a result, the findings in this paper on how workload changes in relation to the subgoal structure of a model should be considered only as general guidelines.

A second limitation is whether the observed changes in workload would remain if our experimental tasks were embedded within broader interactive activities, for example, 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. For example, workload would still decrease at boundaries even though the absolute values at those points and the surrounding subtasks might be different. Further empirical studies are needed to verify these claims.

Third, 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 and Rosenbloom 1981], eliminating some of the perceived boundaries. Indeed, studies of event perception have shown that increased familiarity with a task causes users to generate a similar, but less detailed description of its hierarchical structure [Zacks et al. 2001]. This suggests that mental representations of tasks remain fairly stable, but are performed in larger chunks as skill level increases. In these cases, it is plausible that larger transitory decreases in workload would occur at the boundaries separating these larger chunks of the task.

## 6. CONCLUSION AND FUTURE WORK

A recent thrust in the HCI research community has been to understand when notifications could be delivered to users such that costs of the ensuing interruption would be mitigated. Researchers have argued that interrupting tasks during moments of lower mental workload would have lower interruption cost, but it has been unclear as to just where these moments occur during task execution. Theorists have speculated that workload should be lower at subtask boundaries, but this speculation has never been empirically tested. Our work has made several contributions towards understanding how workload changes in relation to the structure of goal-directed tasks and how this knowledge can be leveraged to improve the design of systems that manage delivery of notifications.

First, we showed that a user's mental workload changes throughout execution of a goal-directed task. This indicates that the moment at which notifications are delivered relative to the ongoing task will impact the cost of interruption. Second, we showed that transitory decreases in workload are experienced as a user moves through a boundary during task execution. We further showed that the decreases in workload tend to be larger at boundaries higher in the task model, as these boundaries correspond to the completion of larger chunks of the task. This indicates that interruption management systems should differentiate among the many boundaries within a task model and favor those representing more salient breaks in the task. Finally, we situated our empirical results within resource theories of attention and described several methods that would enable systems to consider workload as a central part of a broader reasoning framework.

For future work, we plan to analyze workload patterns within additional tasks in order to develop further theoretical understanding of workload changes and produce additional heuristics for assigning costs of interruption to various moments within a task. Also, we will be continuing our implementation of a system that reasons about when to deliver notifications given an explicit specification of the hierarchical structure of a task, the cost of interrupting at various boundaries and other moments within it, and the urgency and relevance of a notification.

## REFERENCES

Adamczyk, P. D. and Bailey, B. P. 2004. If not now when? The effects of interruptions at different moments within task execution. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*. ACM, New York, 271–278.

Anttonen, J. and Surakka, V. 2005. Emotions and heart rate while sitting on a chair. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*. ACM, New York, 491–499.

Bailey, B. P., Adamczyk, P. D., Chang, T. Y., and Chilson, N. A. 2006. A framework for specifying and monitoring user tasks. *J. Comput. Human Behav. 22*, 4, 685–708.

Bailey, B. P. and Konstan, J. A. 2006. On the need for attention aware systems: Measuring effects of interruption on task performance, error rate, and affective state. *J. Comput. Human Behav. 22*, 4, 709–732.

Beatty, J. 1982. Task-evoked pupillary responses, processing load, and the structure of processing resources. *Psych. Bull. 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., and Newell, A. 1983. The psychology of human-computer interaction. Lawrence Erlbaum Associates, Hillsdale.

Chen, D. and Vertegaal, R. 2004. Using mental load for managing interruptions in a physiologically attentive user interface. In *Proceedings of ACM Conference on Human Factors in Computing Systems*. ACM, New York, 1513–1516.

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

Czerwinski, M., Cutrell, E., and Horvitz, E. 2000a. Instant messaging and interruption: Influence of task type on performance. In *Proceedings of the Annual Conference of the Human Factors and Ergonomics Society of Australia (OZCHI)* (Sydney, Australia). C. Paris, N. Ozkan, S. Howard and S. Lu, Eds, 356–361.

Czerwinski, M., Cutrell, E., and Horvitz, E. 2000b. Instant messaging: Effects of relevance and timing. In *People and Computers XIV: Proceedings of HCI*, S. Turner and P. Turner, Eds, 71–76.

Dabbish, L. and Kraut, R. E. 2004. Controlling interruptions: Awareness displays and social motivation for coordination. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work*. ACM, New York, 182–191.

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

Dey, A. K. and Abowd, G. D. 2000. CybreMinder: A context-aware system for supporting reminders. In *Proceedings of 2nd International Symposium on Handheld and Ubiquitous Computing*, 172–186.

Dismukes, K., Young, G., and Sumwalt, R. 1998. Cockpit interruptions and distractions. *ASRS Direct. 10*.

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

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

Gale, A. and Edwards, J. 1983. The EEG and human behavior. Academic Press, New York.

Gevins, A. and Schaffer, R. 1980. A critical review of electroencephalographic (EEG) correlates of higher cortical functions. *CRT Crit. Rev. Bioeng. 4*, 113–164.

Guerlain, S. and Willis, R. 2001. The tactical tomahawk weapons control system: Operator interface design project. In *Human Systems Integration Symposium*.

Hart, S. G. and Staveland, L. E. 1988. Development of a multi-dimensional workload rating scale: Results of empirical and theoretical research. In *Human Mental Workload* P. A. Hancock and N. Meshkati (Eds.), Elsevier, Amsterdam, The Netherlands, pp. 138–183.

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

Hess, E. H. and Polt, J. M. 1964. Pupil size in relation to mental activity during simple problem solving. *Science 132*, 1190–1192.

Hoecks, B. and Levelt, W. 1993. Pupillary dilation as a measure of attention: A quantitative system analysis. *Behav. Res. Meth. Instrum. Comput. 25*, 16–26.

Horvitz, E. and Apacible, J. 2003. Learning and reasoning about interruption. In *Proceedings of the 5th ACM International Conference on Multimodal Interfaces*. ACM, New York, 20–27.

Horvitz, E., Jacobs, A., and Hovel, D. 1999. Attention-sensitive alerting. In *Conference Proceedings on Uncertainty in Artificial Intelligence*, 305-313.

Horvitz, E., Koch, P., and Apacible, J. 2004. BusyBody: Creating and fielding personalized models of the cost of interruption. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work* (Chicago, IL). ACM, New York, 507–510.

Hudson, S. E., Fogarty, J., Atkeson, C. G., Avrahami, D., Forlizzi, J., Kiesler, S., Lee, J., and Yang, J. 2003. Predicting human interruptibility with sensors: A wizard of Oz feasibility study. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*. ACM, New York, 257–264.

Hytintk, J., Tommola, J., and Alaja, A. 1995. Pupil dilation as a measure of processing load in simultaneous interpretation and other language tasks. *Quart. J. Exper. Psych. 48A*, 3, 598–612.

Iqbal, S. T., Adamczyk, P. D., Zheng, S., and Bailey, B. P. 2005. Towards an index of opportunity: Understanding changes in mental workload during task execution. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*. ACM, New York, 311–320.

Iqbal, S. T. and Bailey, B. P. 2005. Investigating the effectiveness of mental workload as a predictor of opportune moments for interruption. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*. ACM, New York, 1489–1492.

Iqbal, S. T. and Bailey, B. P. 2006. Leveraging characteristics of task structure to predict costs of interruption. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*. ACM, New York, 741-750.

Iqbal, S. T. and Bailey, B. P. 2007. Understanding and developing models for detecting and differentiating breakpoints during interactive tasks. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*. ACM, New York, 697–706.

Iqbal, S. T., Zheng, X. S., and Bailey, B. P. 2004. Task evoked pupillary response to mental workload in human-computer interaction. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*. ACM, New York, 1477–1480.

Jackson, T. W., Dawson, R. J., and Wilson, D. 2001. The cost of email interruption. *J. Syst. Inf. Tech. 5*, 1, 81–92.

Juris, M. and Velden, M. 1977. The pupillary response to mental overload. *Phys. Psych. 5*, 4, 421–424.

Just, M. A., Carpenter, P. A., and Miyake, A. 2003. Neuroindices of cognitive workload: Neuroimaging, pupillometric, and event-related potential studies of brain work. *Theoret. Issues Ergonom. 4*, 56–88.

Kahneman, D. 1967. Pupillary responses in a pitch-discrimination task. *Percept. Psych. 2*, 101–105.

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

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 *Multiple-Task Performance*, D. L. Damos, Ed. Taylor and Francis, London, England, pp. 279–328.

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

Kreifeldt, J. G. and McCarthy, M. E. 1981. Interruption as a test of the user-computer interface. In *Proceedings of the 17th Annual Conference on Manual Control*, Jet Propulsion Laboratory, California Institute of Technology, JPL Publication 81–95, 655–667.

Latorella, K. A. 1996. Investigating interruptions: An example from the flight deck. In *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. In *42nd Annual Meeting of the Human Factors and Ergonomics Society*. 87–91.

Lee, J. D., Hoffman, J. D., and Hayes, E. 2004. Collision warning design to mitigate driver distraction. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*. ACM, New York, 65–72.

Lin, Y., Zhang, W. J., and Watson, L. G. 2003. Using eye movement parameters for evaluating human–machine interface frameworks under normal control operation and fault detection situations. *Int. J. Human-Comput. Stud. 59*, 6, 837–873.

Maglio, P. and Campbell, C. S. 2003. Attentive agents. *Commun. ACM 46*, 3, 47–51.

Marshall, S. P. 2002. The index ofs cognitive activity: Measuring cognitive workload. In *Proceedings of the 2002 IEEE 7th Conference on Human Factors and Power Plants*. IEEE Computer Society Press, Los Alamitos, CA, 7.5–7.9.

McCrickard, S., Chewar, C. M., Somervell, J. P., and Ndiwalana, A. 2003. A model for notification systems evaluation–assessing user goals for multitasking activity. *ACM Trans. Computer-Hum. Interact. 10*, 4, 312–338.

McFarlane, D. C. 1999. Coordinating the interruption of people in human-computer interaction. In *Proceedings of the IFIP TC.13 International Conference on Human-Computer Interaction*. 295–303.

McFarlane, D. C. and Latorella, K. A. 2002. The scope and importance of human interruption in HCI design. *Human-Computer Interaction 17*, 1, 1–61.

Miyata, Y. and Norman, D. A. 1986. Psychological issues in support of multiple activities. In *User Centered System Design: New Perspectives on Human-Computer Interaction*. D. A. Norman and S. W. Draper, Eds. Lawrence Erlbaum Associates, Hillsdale, N.J., pp. 265–284.

Monk, C. A., Boehm-Davis, D. A., and Trafton, J. G. 2002. The attentional costs of interrupting task performance at various stages. In *Proceedings of the Human Factors and Ergonomics Society 46th Annual Meeting*.

Nair, R., Voida, S., and Mynatt, E. 2005. Frequency-based detection of task switches. In *Proceedings of the 19th British HCI Group Annual Conference* (Edinburgh, Scotland). 94–99.

Nakayama, M. and Takahashi, K. 2002. The act of task difficulty and eye-movement frequency for the oculo-motor indices. In *Proceedings of Eye Tracking Research and Applications*. 37–42.

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

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

PHELPS, M. P. AND MAZZIOTTA, J. 1985. Positron emission tomography: Human brain function and biochemistry. *Science 228*, 799–809.

POMPLUN, M. AND SUNKARA, S. 2003. Pupil dilation as an indicator of cognitive workload in human-computer interaction. In *Proceedings of the 10th International Conference on Human-Computer Interaction*, V. D. D. Harris, M. Smith and C. Stephanidis, Eds.

ROWE, D. W., SIBERT, J., AND IRWIN, D. 1998. Heart rate variability: indicator of user state as an aid to human-computer interaction. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*. ACM, New York, 480–487.

RUBINSTEIN, J. S., MEYER, D. E., AND EVANS, J. E. 2001. Executive control of cognitive processes in task switching. *J. Exper. Psych.: Human Percept. Perform. 27*, 4, 763–797.

SHNEIDERMAN, B. 1997. Designing the user interface. Pearson Addison Wesley, Reading, MA, Third Edition.

SPEIER, C., VALACICH, J. S., AND VESSEY, I. 1999. The influence of task interruption on individual decision making: An information overload perspective. *Decis. Sci. 30*, 2, 337–360.

STANTON, N., ED. 1994. *Human Factors in Alarm Design*. Taylor and Francis, London, UK.

TAKAHASHI, K., NAKAYAMA, M., AND SHIMIZU, Y. 2000. The response of eye-movement and pupil size to audio instruction while viewing a moving target. In *Proceedings of the ACM Conference on Eye Tracking Research and Applications*. ACM, New York, 131-138.

TOBII-SYSTEMS. *http://www.tobii.se/*.

TRAFTON, J. G., ALTMANN, E. M., BROCK, D. P., AND MINTZ, F. E. 2003. Preparing to resume an interrupted task: effects of prospective goal encoding and retrospective rehearsal. *Int. J. Human-Comput. Stud. 58*, 583–603.

VERNEY, S. P., GRANHOLM, E., AND DIONISIO, D. P. 2001. Pupillary responses and processing resources on the visual backward masking task. *Psychophysiology 38*, 1, 76–83.

VERNEY, S. P., GRANOLM, E., AND MARSHALL, S. 2004. Pupillary responses during the visual backward masking task predict cognitive ability. *Int. J. Psychophys. 52*, 23–36.

WICKENS, C. D. 1980. The structure of attentional resources. In *Attention and Performance VIII*, R. Nickerson, Ed. Lawrence Erlbaum, Hillsdale, NJ, pp. 239–257.

WICKENS, C. D. 1991. Processing resources and attention. In *Multiple-Task Performance*, D. L. Damos, Ed. Taylor & Francis, London, UK, pp. 3–34.

WICKENS, C. D. 2002. Multiple resources and performance prediction. *Theoret. Iss. Ergon. Sci. 3*, 2, 159–177.

ZACKS, J., TVERSKY, B., AND IYER, G. 2001. Perceiving, remembering, and communicating structure in events. *J. Exper. Psych. General 130*, 1, 29–58.

ZIJLSTRA, F. R. H., ROE, R. A., LEONORA, A. B., AND KREDIET, I. 1999. Temporal factors in mental work: Effects of interrupted activities. *J. Occupat. Org. Psych. 72*, 163–185.

Received August 2005; revised October 2006; accepted May 2007 by Alistair Sutcliffe