Towards an Index of Opportunity: Understanding Changes in Mental Workload during Task Execution

Shamsi T. Iqbal, Piotr D. Adamczyk[♦], Xianjun Sam Zheng[†], and Brian P. Bailey

Department of Computer Science, School of Library and Information Science[♦], Beckman Institute[†]
University of Illinois, Urbana-Champaign 61801
{siqbal, pdadamcz, xzheng, bpbailey}@uiuc.edu

ABSTRACT

To contribute to systems that reason about human attention, our work empirically demonstrates how a user's mental workload changes during task execution. We conducted a study where users performed an interactive hierarchical task and measured mental workload through the use of pupil size. Results show that (i) different types of subtasks impose different mental workload, (ii) workload decreases at subtask boundaries, (iii) workload decreases more at higher level boundaries and less at lower level boundaries, (iv) workload changes among subtask boundaries within the same level of a task model, and (v) effective understanding of why changes in workload occur requires that the measure be tightly coupled to a validated task model. We offer an Index of Opportunity that maps a user's mental workload to the disruptive impact of an interruption. This mechanism can help systems better compute the cost of interruption.

CATEGORIES AND SUBJECT DESCRIPTORS

H.5.2 [Information Interfaces and Presentation]: User Interfaces — evaluation/methodology, user-centered design

GENERAL TERMS

Human Factors, Design, Experimentation, Measurement

KEYWORDS

Interruption, task models, attention, affective state

INTRODUCTION

When interacting with applications, users often suffer *interruption overload*. E-mail notifications [20], instant messages [7], agent requests [25], and system alerts all contribute to this burgeoning epidemic of interruption that affects almost every user. When a background application interrupts a user at an inopportune moment during task execution, the user performs tasks slower [2, 29], commits more errors [24], makes worse decisions [35], and experiences more frustration, annoyance, and anxiety [1, 2, 5] than if it had interrupted at a more opportune moment. To mitigate the disruptive effects of interruption, researchers are investigating systems that reason about when to interrupt users [15, 17, 18]. These systems compute the cost of interruption using external cues such as desktop activity, visual, and acoustical analyses of the physical environment, and scheduled activities of the user.

To compute a more accurate cost of interruption, however, these systems need a direct measure of a user's mental workload. Researchers have long argued that opportune moments for interruption occur at periods of low mental workload during task execution [2, 8, 11]. Miyata and Norman posit that moments of low mental workload occur at subtask boundaries during task execution [30]. However, interactive tasks are composed of hierarchical patterns of goal formulation, execution, and evaluation, creating many levels of boundaries in the task model [6]. Our work seeks to empirically demonstrate how a user's mental workload changes during execution of an interactive task, focusing on subtask boundaries in the task model.

We conducted a user study where we asked 12 users to perform an interactive route planning task. While a user performed the task, we measured mental workload by measuring relative changes in the user's pupil size using an eye tracking system. Research shows that pupil size is a reliable measure of mental workload [3, 22]. We developed and validated a GOMS model for the task and precisely aligned it with the pupillary response data.

Our results show that (i) different types of subtasks within a task model impose different mental workload on a user, (ii) workload decreases at subtask boundaries, (iii) workload decreases more at boundaries higher in the task model and less at boundaries lower in the task model, (iv) workload changes among subtask boundaries within the same level of a task model, and (v) effective understanding of why changes in mental workload occur requires that the measure be tightly coupled to a validated task model.

Our work contributes the first empirical evidence showing how much mental workload changes at different levels of subtask boundaries in a hierarchical task. Informed by our results, we develop an *Index of Opportunity* that maps a user's mental workload - as measured by pupillary response - to the disruptive impact of an interruption. This mechanism would be useful in any system that tries to manage human attention – not only on the desktop but also in control rooms, aviation cockpits, and in-vehicle displays.

In her UIST 2003 closing keynote [27], Marshall argues that mental workload as measured by pupil size offers a new metric by which to evaluate user interfaces. Our research demonstrates further methodology for how to analyze pupillary response for interactive computing tasks. By leveraging our methodology of aligning task models with pupillary response data, interface designers could better associate periods of high mental workload with specific elements of an interface, targeted for re-design.

RELATED WORK

We discuss posited moments for interruption, discuss systems that reason about when to interrupt users, and justify our use of pupil size to measure mental workload.

Posited Moments for Interruption

During task execution, an opportune moment to interrupt a user is during a period of low mental workload. This finding has been supported empirically [7-11, 26, 28]. When a user is interrupted during a period of low mental workload, the interruption disrupts the user less than if it had occurred during a period of high mental workload. The challenge is to understand when a user's mental workload changes during task execution.

In [30], Miyata and Norman theorize that moments of lower mental workload occur between the completion (evaluation) of one subtask and the beginning (goal formulation) of the next subtask, i.e. at a subtask boundary. Interactive tasks, however, are typically performed through recursive patterns of goal formulation, execution and evaluation [6], where higher-level goals are recursively decomposed into lower-level subgoals and operators.

Our work provides the first empirical findings of how much a user's mental workload changes at subtask boundaries and how much that change differs at different levels of boundaries in a hierarchical task. Results of this work contribute to a further theoretical understanding of how mental workload changes during task execution and contributes to systems that reason about when to interrupt users.

Reasoning About When to Interrupt Users

In [15, 17, 18], researchers are constructing computational systems that reason about when to interrupt a user by weighing the value of information against the cost of interruption. The underlying models use external cues such as desktop activity, visual and acoustical analyses of the physical environment, and scheduled activities to compute the cost of interruption.

Although researchers recognize the importance of including a measure of mental workload in an interruption reasoning system, there is no such computational measure available. Without a more accurate assessment of a user's mental workload, systems can make poor decisions about when to interrupt the user. For example, in [15], researchers model desktop inactivity as better moments for interruption than periods of activity. Miyata and Norman [30] argues however, that inactivity is generally *worse* for an interruption because those periods represent moments of planning or evaluation, which can require more mental workload than subtask execution.

Informed by our empirical findings, we have developed a computational Index of Opportunity that provides a system with a real-time measure of how disruptive an interruption would be if it was delivered at the present moment during task execution. The Index of Opportunity enables a reasoning system to make a more accurate assessment of the cost of interruption, leading to more effective decisions about when to interrupt the user.

Use of Pupil Size to Measure Mental Workload

Under conditions of controlled illumination, research shows that pupil size is a reliable measure of mental workload [12, 14, 22, 31, 36], where the increase in pupil size correlates with the increase in workload. In [3], Beatty reviewed a large corpus of experimental data and concluded that pupillary response is a reliable indicator of mental workload for a task, that the degree of pupillary response correlates with the workload of the task, and that this phenomenon holds true between tasks and individuals. UI researchers are already using pupil size to evaluate the mental workload imposed by user interface designs [27].

In [19], we showed that pupillary response correlates with the mental workload of *interactive* tasks and discovered that changes in mental workload seem to align well with the hierarchical model of the task being performed. Our current study seeks to better understand this relationship.

Although researchers have also investigated the use of eye movement [36], blink rate [23], and heart rate variance [33] to approximate mental workload, pupil size offers a *direct* quantitative measure, which simplifies the statistical and computational analysis of the response data. Of course, we do not expect users to wear eye-tracking equipment while performing daily computing tasks. We believe that future technology will provide cost-effective, non-intrusive means to effectively measure pupil size, e.g., remote eye trackers built into LCD monitors [37] or even eye glasses [34], thus justifying our use of pupil size in the present work.

Because pupil size has been repeatedly shown to correlate well with changes in mental workload, we believe our use of pupil size alone provides a sufficient measure of mental workload for this work. There is a need to cross-validate mental workload across multiple measures, but how to effectively align different physiological measures of workload is not well developed. Our Index of Opportunity, however, may contribute to a common scale appropriate for aligning multiple measures in the future.

USER STUDY

The purpose of our study is to better understand how much a user's mental workload decreases at subtask boundaries and whether that decrease changes for boundaries that are at different levels in a task model. Also investigated is whether different types of cognitive subtasks induce different mental workloads. Specifically, we designed a user study 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 does a user's mental workload change at subtask boundaries? How much does this change differ for boundaries at different levels in a task model?
- How much lower is a user's mental workload at boundaries compared to the average mental workload during subtask execution (non-boundary) moments?
- How can we use our findings to develop a computational mechanism that measures how opportune different moments are for an interruption during task execution?

Users and Equipment

Twelve users (1 female) participated in the study. Users ranged from 23 to 50 years of age, with distribution (M=27.1, SD=7.45). All had normal or corrected-to-normal vision. Though we did not balance for gender, previous research has not shown a gender effect of pupillary response to mental workload [3]. Users were not compensated for their participation.

We recorded pupillary response using a head-mounted SR Inc. Eyelink II eye-tracking system with a sampling rate of 250 HZ and high spatial resolution of 0.005 degrees. The study was conducted in a room where light and noise levels were well controlled.

Task and Subtasks

For the study, we developed an interactive route planning task, shown in Figure 1. We designed the task to be comprised of meaningful subtasks of varying difficulty, to have a prescribed execution sequence, to have well defined boundaries among subtasks, and to provide a representative sample of user interaction.

Although a user does not typically follow a prescribed execution sequence when performing a task, the sequence had to be controlled to reliably link changes in mental workload to task execution. The subtasks that comprised the higher level task tap the same cognitive mechanisms, such as memory and reasoning, often needed to perform interactive computing tasks. The task required a user to perform mouse movements, item selections, data entry, and calculations for about five minutes.

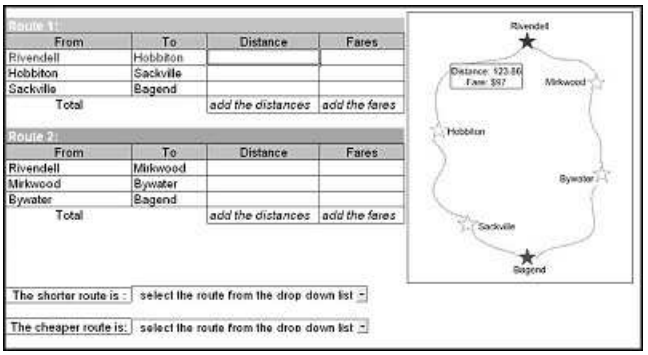

Figure 1: The interactive route planning task. A user retrieves distance and fare information from the map and enters it into the tables. Each route had three segments and there were two routes. To retrieve the data, the user moves the mouse over a segment, notes the distance and fare information shown in the tooltip, and enters the data into the corresponding row in the table. Once the three rows have been filled in, the user adds the distance and fare information and enters it into the fourth row. The user repeats the process for the second table. The user then selects the shorter and the cheaper of the two routes from the drop down lists. The map shows the data for the first segment of the first route. The information disappears when the user moves the mouse away from the segment.

In the task, a user was shown a map with two routes between two cities marked by differently colored 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 then entered the data into the corresponding row in the table. When the user moved the mouse away from the segment, the tooltip disappeared. The user completed each of the three rows in the table and then mentally added the distance and fare columns and entered the results into the fourth row. The user then repeated the process for the second table and route. After completing the tables, the user selected the shorter and the cheaper of the two routes from the drop down lists near the bottom of Figure 1.

The main cognitive subtasks were storing information from the map in working memory (Store), recalling information in order to fill in the table (Recall), and adding numbers mentally (Computation). Comparing the distance and fare totals and deciding the shorter and cheaper routes also involved cognitive skills.

To vary the cognitive load of the subtasks, we varied the complexity of the distance and fare information. For example, for the easier subtasks, whole numbers were used for the distance and fare information (e.g. 80). For the more difficult subtasks, we used numbers with more digits (e.g. 147.53) and that required carries in the add computations.

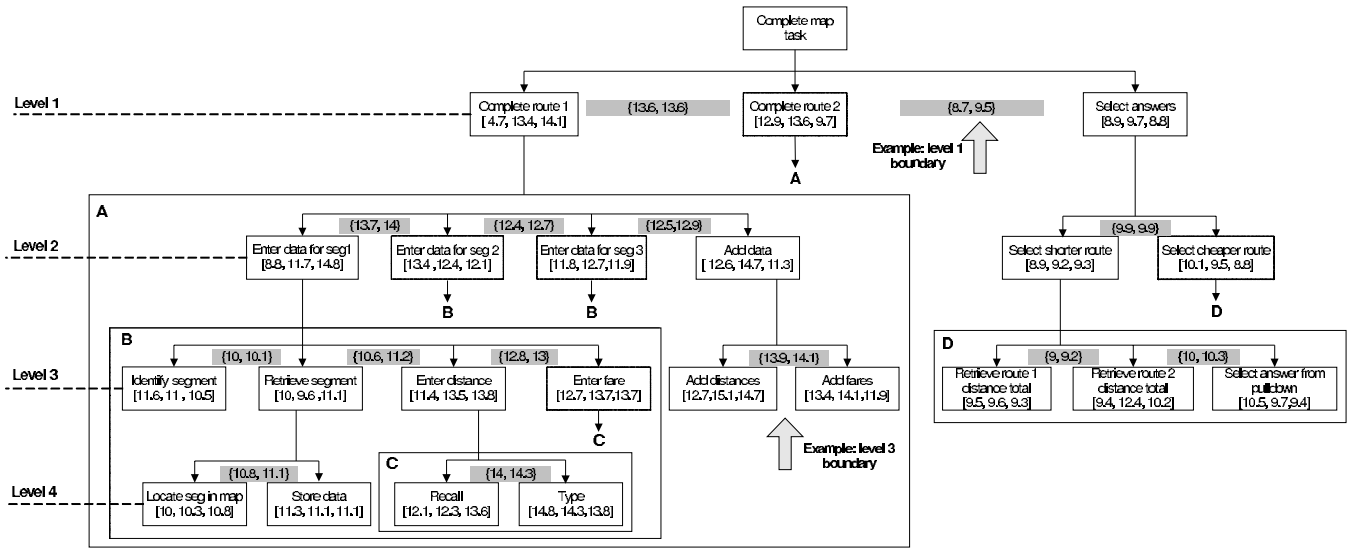

Figure 2: Validated GOMS model of the route planning task. The interior nodes represent goal nodes, the leaf nodes represent operators, and time moves from left to right. Regions A, B and C show parts of the task repeated elsewhere in the task. Within each subtask, we provide the [*beginning PCPS, average PCPS, last PCPS*] for that subtask. Each shaded area between subtasks indicates a boundary and contains the [*minimum PCPS, average PCPS*] across the boundary. The example level 3 boundary shows that the APCPS drops from 15.1 during the preceding subtask to a minimum of 13.9 within the boundary. The example level 1 boundary shows that the APCPS drops from 13.6 during the preceding subtask to a minimum of 8.7 within the boundary. All numbers shown are percents.

Procedure

Upon arrival at the lab, a user filled out a consent form, a questionnaire for background information, and received instructions for the task. 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 a practice task. Just before the user performed the experimental task, we collected baseline pupil size by having the user relax and fixate on a blank (white) screen for about 10 seconds. The user then performed the experimental task. The user was instructed to perform the task as quickly and as accurately as possible. Pupil and eye movement data were logged to time stamped files. We video-taped a user's screen interaction for later analysis. Because the videos and pupil data received time stamps from the same clock, we could precisely align them. The user required about 5 minutes to perform the task and the entire experimental session lasted about 20 minutes.

Task Model and Validation

We performed a GOMS analysis to decompose the task into its component subtasks. In GOMS terminology [21], we started with the root goal - to perform the route planning task - and then recursively decomposed the root goal into its component subgoals and operators. The decomposition continued until there was no observable or meaningful separation between operators.

Figure 2 shows the task model, reusing repetitive parts for clarity. The full task model has 4 levels and 81 nodes. The leaves of the model represent operators, the interior nodes represent subgoals, and the root represents the main task goal. The term *subtask* refers to any node in the task model. The term *subtask boundary* refers to the period between execution of consecutive subtasks. We define *level of boundary* between two consecutive subtasks to be 1 + the depth of their shared ancestor subtask in the model. For example, consider the "Locate seg in map" subtask and the "Store data" subtask shown at the left of level 4 in Figure 2. When a user completes the "Locate seg in map" subtask and moves to the "Store data" subtask, 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 the "Recall" subtask, this defines a level 3 boundary, since the depth of their shared ancestor subtask ("Enter data for seg 1") is (1 +) 2. Finally, *subtask type* refers to whether an operator (subtask) represents a store, recall or computation operator.

The GOMS model was developed in an iterative manner. We developed an initial GOMS model through our own analysis of the task's execution. Once defined, three people who did not participate in the user study were asked to view a video of the expected task execution and identify an observable sequence of operators. This video was recorded prior to and independent of the user study. We compared the identified operator sequences to the leaves of our task model and refined it as necessary.

We validated our final model by comparing the interaction videos from the user study to the model. The number of error steps each user performed was counted. An error step was defined to be any deviation from the prescribed operator sequence. If the user committed an error, each

action after that step would also count as an error until the user again performed a step in the prescribed sequence, from which point the analysis continued.

Across users, the average error rate was 2.81%, ranging from 0% to 5.66%, which is consistent with models validated in [6]. Thus, the GOMS model accurately reflects a user's execution of the task and enabled us to precisely align each user's pupillary response to the task model. This was very challenging because each user performed the task at a different speed. Thus, we had to align the response data by meticulously synchronizing the pupil data to specific event points in the task model.

Measurements

To measure changes in mental workload, we calculated the percentage change in pupil size, referred to as *PCPS*. This value was calculated by subtracting the baseline pupil size from each measured pupil size and then dividing the result by the baseline. We use PCPS to minimize the pupillary response differences among individual users, which is consistent with prior work [3]. The term *APCPS* refers to the *average* PCPS over a time window.

RESULTS

In this section, we discuss how much mental workload different types of subtasks induce on a user, how much a user's mental workload changes at subtask boundaries, and how much a user's mental workload differs between subtask execution and subtask boundaries. Figure 3 shows the pupillary response curve for a user in the study. The rise and fall of the curve shows changes of a user's mental workload during execution of the task.

As we discuss results, the reader should keep in mind that small changes in pupillary response represent meaningful changes in a user's mental workload and that there is an upper bound on how much a user's pupil will increase due solely to the effects of increased mental workload.

Mental workload at subtasks

To validate that each cognitive subtask - on average - did indeed induce increased mental workload on a user, we performed a one-sample t-test on the APCPS for the Store, Recall and Computation subtasks. We found that the APCPS was greater than 0 across subtasks (M=0.127, SD=0.073, $t(263)$=28.25, $p<0.001$). This represents a 12.7% increase over the baseline level. The standardized effect size index d was 1.7, a high value. This shows that the subtasks did impose a measurable increase in mental workload on a user. We did not include other types of subtasks, such as motor or perceptual subtasks because the perceptual and motor systems typically interact with but are distinct from the cognitive system [6].

We performed a one-way ANOVA with Type as the factor. Type had a main effect on APCPS ($F(2,261)$=3.247, $p<0.04$). Post hoc Bonferroni tests showed that computation subtasks induced more mental workload than store subtasks (difference was 1.7 percentage points, with $p<0.037$). This shows that certain types of subtasks (Computation) induce more mental workload than others (Store) while other subtask types induce similar workload (Store and Recall).

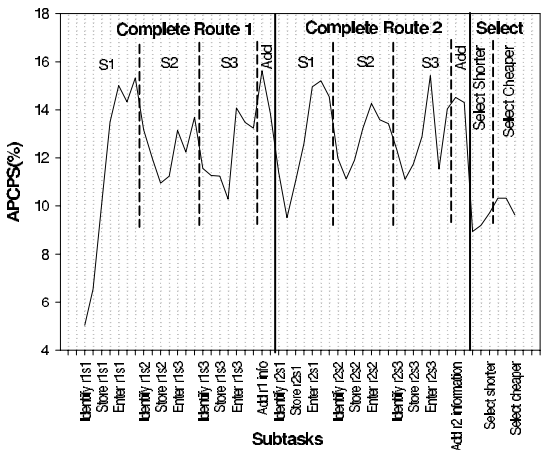

Figure 3: The graph shows the APCPS for each subtask in the model. Solid lines indicate level 1 boundaries and dashed lines indicate level 2 boundaries. The x-axis enumerates the level 3 subtasks. Notice how the graph dips lower at level 1 boundaries than at level 2 boundaries – showing how mental workload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

We performed a one-way ANOVA with Level as the factor. Because the subtasks used in this comparison represent the operators in the task model, each subtask existed at either level 3 or 4. Level had a main effect on APCPS ($F(1,262)$=3.898, $p<0.049$). Level 3 subtasks had a higher APCPS than level 4 subtasks (difference was about 2.3 percentage points). This difference in APCPS, however, may be attributed to the cognitive demands of the subtasks rather than their level as level 3 contained all the computation (more cognitively demanding) subtasks.

Our results show that the subtasks induced increased mental workload on a user and that different types of subtasks induced different mental workloads. This implies that a system can mitigate the disruptive effects of an interruption by deferring the interruption until a user is executing a lower workload subtask. This does not, however, require the system to defer the interruption for an extended period of time, waiting just a few seconds may result in considerable mitigation of disruption. This is consistent with our previous results on interruption [1, 2].

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 4. There was a clear boundary between consecutive subtasks at each level in the task model. For each boundary, we computed the

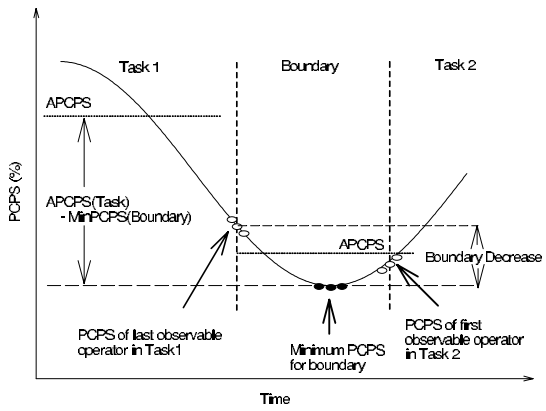

Figure 4: Illustration of metrics used in the analysis. Vertical lines drawn through the last observable operator in Task 1 and the first observable operator in Task 2 (taken as the average of the three surrounding points) define the boundary between Task 1 and Task 2. The horizontal lines show the APCPS for Task 1 and the boundary.

Boundary Decrease by subtracting the minimum PCPS (taken as the average of 3 values around the minimum to ensure support) within the boundary from the PCPS at the last observable operator in the preceding subtask (again, taken as the average of the 3 surrounding values) - just before the boundary occurred. Thus, a *positive* change in Boundary Decrease reflects a *decrease* in mental workload.

A one-sample t-test showed that Boundary Decrease was greater than 0 across subtasks (M=0.0023, SD=0.02165, $t(611)$=2.67, p<0.008). The standardized effect size index d was 0.1. This result shows that a user's mental workload does decrease at a subtask boundary, but the decrease is small on average.

One reason for the small effect size was that level 4 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 > 0, excluding the level 4 samples. Our results showed a stronger effect (M=0.005388, SD=0.0215, $t(395)$=4.995, p<0.001), with an improved d of 0.25. This implies that changes in mental workload are meaningful down to the level of boundary that exists just above the elementary operators in the task model.

A one-way ANOVA showed that Level had a main effect on Boundary Decrease ($F(3,608)$=8.037, p<0.001). Post hoc Bonferroni tests showed that Boundary Decrease at level 2 was greater than at level 4 (difference was about one percentage point, with p<0.001) and that Boundary Decrease at level 3 was also greater than at level 4 (difference was about 0.8 percentage points, with p<0.001). This result shows that a user's mental workload at a subtask boundary tends to decrease *more* at higher levels in the task model and *less* at lower levels in the model (Figure 3). A reasonable interpretation is that a user releases more cognitive resources when completing the final subtask of a larger goal chunk (higher in the model) than when completing the final subtask of a smaller goal chunk [32].

Boundary Decrease at level 1, however, did not differ from any other levels, although the trends in the means were in the expected direction. This may be due to the fact that level 1 boundaries had much fewer sample points than the other levels - the task model is wider at the lower levels than at the higher levels - resulting in a larger variance and limiting the power of the statistical test.

Mental workload at subtasks vs. subtask boundaries

In the prior analysis, we computed Boundary Decrease by subtracting the minimum PCPS during a subtask boundary from the last PCPS in the preceding subtask. From the pupillary response curve, we observed that the decrease in mental workload at a subtask boundary actually started *just before* the last measure in the preceding subtask. This is likely because the cognitive and motor systems execute mostly in parallel, but with cognitive function preceding motor function. To further investigate, we tested how the minimum PCPS at a boundary compared to the APCPS *over the execution* of the preceding subtask.

A paired samples t-test for each pair of minimum PCPS within a subtask boundary and the APCPS of its preceding subtask execution showed that the pairs differed (M=0.0028, SD=0.0343, $t(611)$=1.995, p<0.047), with PCPS at the boundary being less than the APCPS during subtask execution. The standardized effect size d was 0.08. Similar to the previous section, we found that the small effect size was partly due to the level 4 pairs not differing. Excluding level 4 pairs, we performed the same test and found a larger difference among the paired samples in levels 1-3 (M=0.0097, SD=0.0357, $t(395)$=5.94, p<0.001). The standardized effect size d was 0.3, which showed marked improvement. This again implies that changes in mental workload are meaningful down to the level of boundary that exists just above the elementary operators.

A one-way ANOVA showed that Level had a main effect on the pairs of the boundary minimum and the APCPS of the preceding subtask ($F(3,608)$=19.677, p<0.001). Post hoc Bonferroni tests showed that the difference at level 1 was marginally greater than at level 2 (about 1.9 percentage points, p<0.056), significantly greater than at level 3 (about 2.13 percentage points, p<0.014) and significantly greater than at level 4 (about 3.9 percentage points, p<0.001). The difference at level 2 was greater than at level 4 (about 2 percentage points, p<0.001) and the difference at level 3 was also greater than at level 4 (about 2 percentage points, p<0.001). This result further supports our previous finding that a user's mental workload at a subtask boundary tends to decrease *more* at higher levels in the task model and decrease *less* at lower levels in the task model.

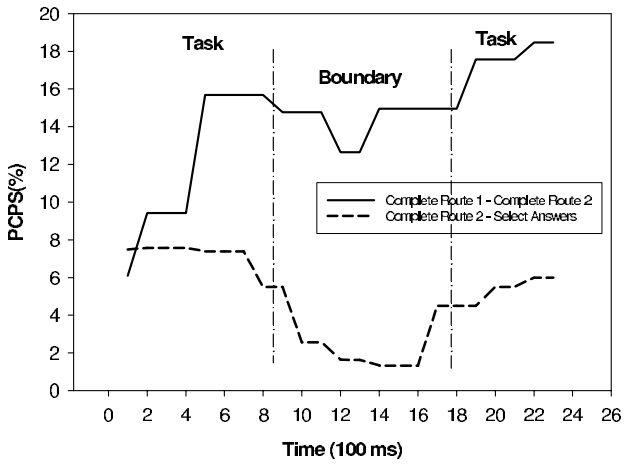

Figure 5: PCPS around two subtask boundaries. The solid curve shows the level 1 boundary between 'Complete route 1' and 'Complete route 2.' The dashed curve shows the level 1 boundary between 'Complete route 2' and 'Select answers.' The latter boundary shows a lower and more sustained drop than the first, indicating how changes in workload can differ at different boundaries. even within the same level.

We not only found that workload changed between levels in the task model, but we also found that workload changed *within* the same level in the task model. For example, we found that the APCPS over the two level 1 boundaries differed ($t(14)=4.23$, $p<.001$) with a maximum difference of about 3 percentage points. In Figure 5, we show the PCPS for the two level 1 boundaries for one user, which is representative of the summary values for these boundaries in Figure 2. We also found that the APCPS among level 2 boundaries differed ($F(3,33)=3.582$, $p<0.024$) with a maximum difference of about five percentage points.

Overall, results show that a system could use a task model alone to roughly infer *where* a user's mental workload may change during task execution. However, our results empirically demonstrate that a system requires a measure of mental workload to understand *how much* a user's mental workload changes at those points. Knowing how much a user's mental workload will change should enable a system to make better decisions about when to interrupt the user.

FINDINGS

From the user study, we found that:

- *Different types of subtasks within a task model impose different mental workloads on a user.* Our results show that Computation subtasks induce more mental workload on a user than Store or Recall subtasks. Thus, it is important to monitor a user's mental workload during subtask execution, as interrupting during different subtasks would likely cause different levels of disruption.
- *Mental workload decreases at subtask boundaries.* We compared the minimum PCPS at a subtask boundary to both the last PCPS measure in the preceding subtask as well as to the APCPS over the execution of the preceding subtask. From both perspectives, we found that a user's mental workload decreased at the subtask boundary.
- *Mental workload decreases more at boundaries higher in the task model and less at boundaries lower in the task model.* We compared paired differences between the minimum PCPS at a subtask boundary and the last PCPS measured in the preceding subtask across different levels in the task model. The difference between the pairs was greater at higher levels in the task model and smaller at lower levels in the task model. Although some researchers have roughly predicted this effect, we have provided the first empirical data supporting this effect and have used our quantitative results to develop a computational Index of Opportunity.
- *Mental workload changes among subtask boundaries within the same level of a task model.* We compared APCPS among subtask boundaries within the same levels of the task model. For levels 1 and 2, we found that the change in workload differed within the level. Although a system could use a task model alone to roughly infer *when* a user's mental workload may change during task execution, the system requires a tightly coupled measure of mental workload to understand *how much* the user's mental workload changes at those points.
- *Effective understanding of why changes in mental workload occur requires that the measure be tightly coupled to a validated task model.* While inspecting the pupillary response graph, much time was spent investigating distinctive patterns in the graph, we would often juxtapose the task model to make sense out of the patterns. We believe that this would also be useful when evaluating interfaces based on workload demand. For example, when measuring the workload induced by alternative interface designs, the measure must be tightly coupled to a task model to help understand why certain patterns of workload are occurring, and to understand which tasks and subtasks are inducing unnecessarily high load, and thus should be targeted for re-design.

Our findings have important implications for the design of a computational system - an *attention manager* - that reasons about when to interrupt a user. Because our results show that a user's mental workload changes among subtasks and generally decreases at subtask boundaries, an attention manager can and *should* perform fine-grained temporal reasoning (at the subtask level) about when to interrupt a user engaged in task execution. Deferring the delivery of information - even for a few seconds - until a user shows a lower mental workload can help mitigate the disruptive effects of interruption. This is consistent with our prior empirical results showing that fine-grained temporal manipulation of interruptions can cause dramatic differences in a user's task performance and levels of frustration, annoyance, and anxiety [1, 2].

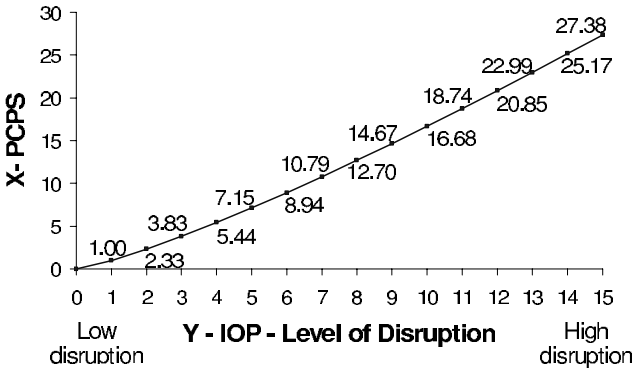

Figure 6: Graph of the mapping from PCPS to IOP. We invert the axes to better show boundaries of the IOP categories. A lower IOP indicates a less disruptive impact of an interruption.

Towards a Computational Index of Opportunity

Informed by our findings, we define an Index of Opportunity (IOP) that maps pupillary response (PCPS) to a 15-point scale of subjective disruption. The IOP provides an attention manager with a sense of how disruptive an interruption may be. On the scale, a '1' represents the lowest workload and indicates that an interruption would be least disruptive, or the *most* opportune moment for interruption. A '15' represents the highest workload and indicates that an interruption would be most disruptive, or the *least* opportune moment for interruption.

The IOP reflects the magnitude intensity impact on a user's mental workload in response to a stimulus, in this case an interruption. These kinds of stimulus/impact models are common in psychological studies of human perception [4], where the relationship between stimulus and impact is governed by a power-law equation known as Stevens' Law:

$$\Psi = c\Phi^r$$

Ψ is the subjective (perceived) magnitude of a stimulus, and Φ is the physical stimulus magnitude. c is a scaling constant and r is the power exponent. We equate Ψ to IOP, as it reflects the subjective disruption of an interruption, and we equate Φ to PCPS as it reflects the level at which the stimulus would occur. Ideally, Φ would relate to the objective magnitude of an interruption, but as no such metric is available, we suggest that an effective mapping would be to the current level of mental workload. From our results, we argue that in deciding the subjective impact of an interruption, it matters not only when in task execution the interruption occurs, but matters more at what particular level of workload the interruption occurs.

As mental workload dominates disruptive impact [4], our objective measure only considers the mental workload of the user, and not interruption presentation styles, potential cognitive resource conflicts of the primary and interrupting tasks, urgency, or relevancy. The IOP could be integrated with these and other factors in a broader reasoning system.

For our purposes the constant c is 1. We map IOP across 15 categories, anchoring Category 8 to the APCPS across users, which was 12.7%. We anchor Category 1 to 1% PCPS, reflecting minimal load as compared to the baseline. IOP Category 15 includes all PCPS values greater than the upper bound for Category 14 (>25.17%). We selected 15 categories because the resulting widths of the IOP categories allow for discrimination of subtask boundaries during task execution, i.e., drops at subtask boundaries are enough to influence the IOP measure. Also, the choice of 15 categories is enough to denote noticeable differences without being so fine grained as to simply be an uninformative replacement for a raw measure of PCPS. The upper bounds for the PCPS at high IOP values match well with the max PCPS values recorded in the study.

Although we use only the results from our user study to define the mapping, it can easily be configured for different groups of users and/or tasks. A designer may preset the APCPS for a task or the APCPS value could be computed over a time window and tuned by machine learning methods. However APCPS is determined, the r exponent is calculated by:

$$r = \frac{Log(CenterLevel)}{Log(APCPS)} = \frac{Log(8)}{Log(12.7)} = 0.818$$

The IOP at a particular moment is calculated by:

$$IOP = \lceil PCPS^r \rceil = \lceil PCPS^{0.818} \rceil$$

We judged the fitness of the model by its ability to produce a near-normal distribution of IOP values from users' PCPS values (shown in figure 6). The near-normal shape of the distribution is similar across users. The graph provides an intuitive sense of a user's interruptibility during task execution – with IOP falling more often in the middle than at either end of the interruptibility spectrum. Additionally, low IOP moments (IOP = 2, 3, 4, 5) are well distributed throughout task execution (27, 14, 53, and 29 are the counts by quartile, for figure 6).

After considering a number of possible models, we choose Stevens' Law as a starting point for several reasons. This particular model gives us the desired behavior of a small change at low values of PCPS leading to a comparatively large increase in IOP. At larger values of PCPS, progressively larger changes in PCPS are required to change IOP. This is consistent with human sensory mappings for sound, light, and touch, among others [4]. Importantly, it is grounded in an empirical model of the human sensory system and could be readily implemented in an attention manager.

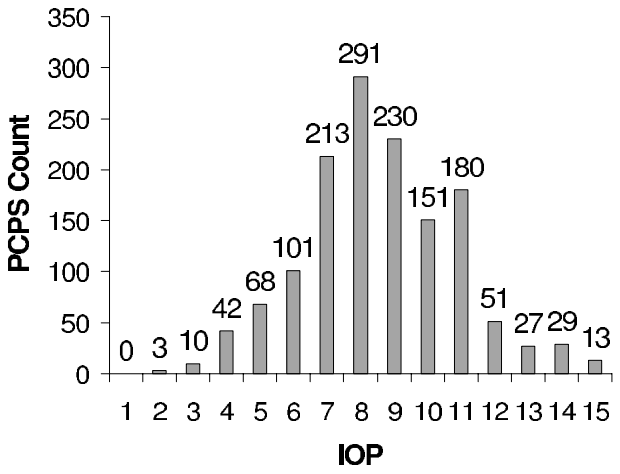

Figure 7: Distribution of IOP values for a user. Each PCPS value represents 100 msec of execution time. Graphs for other users produce a similar distribution.

Within an attention manager, the Index of Opportunity can serve either as a direct measure of the cost of interruption or as part of a broader reasoning model. For example, the index of opportunity could become an evidence variable in a Bayesian decision network [16]. A system could also observe temporal patterns of IOP values to develop a more robust sense of availability that is less sensitive to transitory changes in PCPS. By enabling an attention manager to include a direct measure of workload, the use of IOP may improve the decision quality of the system.

FUTURE WORK

For future work, we intend to:

- *Further validate our mapping from PCPS to IOP through empirical studies.* Although the presented mapping from pupillary response to IOP was informed by our findings, the mapping must be further validated through empirical studies. The studies would manipulate mental workload and use physiological responses such as startle-response [13] to measure disruption caused by an interruption. Because the IOP mapping is consistent with other stimulus/response mappings in human perception, however, we believe that the presented IOP mapping represents a reasonable first approximation.
- *Develop a tool that better supports analysis of pupillary response data for interactive tasks.* Software packages that ship with commercial eye trackers fall far short of what researchers need to analyze pupillary response data for interactive tasks. The software does not provide a tightly synchronized view among the user task model, video of onscreen interaction, and pupillary response. As a result, analysis of the data required tedious labor and complex macro writing. Tools that better support the analysis process could save researchers enormous effort.
- *Use mental workload to further measure the effects of interruptions.* The effects of interruptions are typically measured using logged task completion time and error data and self-report questionnaires. By using pupil size to measure changes in mental workload due to interruption, we may further understand their disruptive effects.

ACKNOWLEDGMENTS

This section has been cleared for the submission process.

REFERENCES

1. Adamczyk, P.D. and B.P. Bailey. If Not Now When? The Effects of Interruptions at Various Moments within Task Execution. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, 2004.
2. Bailey, B.P., J.A. Konstan and J.V. Carlis. The Effects of Interruptions on Task Performance, Annoyance, and Anxiety in the User Interface. *Proceedings of Interact*, Tokyo, Japan, 2001, 593-601.
3. Beatty, J. Task-Evoked Pupillary Responses, Processing Load, and the Structure of Processing Resources. *Psychological Bulletin*, *91* (2), 276-292,
4. Bernstein, D.A., A. Clarke-Stewart, E.J. Roy, T.K. Srull and C.D. Wickens. *Psychology*. Houghton Mifflin Company, Boston Toronto, 1994.
5. Boucsein, W. Psychophysiological Investigation of Stress Induced by Temporal Factors in Human-Computer Interaction. In Frese, M., et al. (eds.) *Psychological Issues of Human-Computer Interaction in the Work Place*, Elsevier, Amsterdam, 1987.
6. Card, S., T. Moran and A. Newell. *The Psychology of Human-Computer Interaction*. Lawrence Erlbaum Associates, 1983.
7. Cutrell, E., M. Czerwinski and E. Horvitz. Effects of Instant Messaging Interruptions on Computing Tasks. In *Extended Abstracts of Human Factors in Computing Systems*, 2000, 99-100.
8. Cutrell, E., M. Czerwinski and E. Horvitz. Notification, Disruption and Memory: Effects of Messaging Interruptions on Memory and Performance. *Proceedings of Interact*, Tokyo, Japan, 2001, 263-269.
9. Czerwinski, M., E. Cutrell and E. Horvitz. Instant Messaging: Effects of Relevance and Timing. In *People and Computers XIV: Proceedings of HCI*, 2000, British Computer Society, 71-76.
10. Dismukes, K., G. Young and R. Sumwalt. Cockpit Interruptions and Distractions. *ASRS Directline*, *10*, 1998.
11. Gillie, T. and D. Broadbent. What Makes Interruptions Disruptive? A Study of Length, Similarity, and Complexity. *Psychological Research*, *50*, 243-250, 1989.

12. Hess, E.H. and J.M. Polt. Pupil Size in Relation to Mental Activity During Simple Problem Solving. *Science*, *132*, 1190-1192, 1964.

13. Hillman, C.H., J.C. Cuthbert, J. Cauraugh, H.T. Schupp, M.M. Bradley and P.J. Lang. Psychophysiological Responses of Sport Fans. *Motivation and Emotion* (24), 13-28, 2000.

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

15. Horvitz, E. and J. Apacible. Learning and Reasoning About Interruption. In *Proceedings of the Fifth ACM International Conference on Multimodal Interfaces*, 2003.

16. Horvitz, E., J. Breese, D. Heckerman, D. Hovel and K. Rommelse. The Lumiere Project: Bayesian User Modeling for Inferring the Goals and Needs of Software Users. In *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence*, 1998, 256-265.

17. Horvitz, E., A. Jacobs and D. Hovel. Attention-Sensitive Alerting. In *Conference Proceedings on Uncertainty and Artificial Intelligence*, 1999, 305-313.

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

19. Iqbal, S.T., X.S. Zheng and B.P. Bailey. Task Evoked Pupillary Response to Mental Workload in Human-Computer Interaction. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, 2004, to appear.

20. Jackson, T.W., R.J. Dawson and D. Wilson. The Cost of Email Interruption. *Journal of Systems and Information Technology*, *5* (1), 81-92, 2001.

21. John, B.E. and D.E. Kieras. The Goms Family of User Interface Analysis Techniques: Comparison and Contrast. *ACM Transactions on Computer-Human Interaction*, *3*, 320-351, 1996.

22. Kahneman, D. Pupillary Responses in a Pitch-Discrimination Task. *Perception & Psychophysics*, *2*, 101-105, 1967.

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

24. Kreifeldt, J.G. and M.E. McCarthy. 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, 1981, 655-667.

25. Lieberman, H. Letizia: An Agent That Assists Web Browsing. In *Fourteenth International Joint Conference on Artificial Intelligence*, 1995, 924-929.

26. Maglio, P. and C.S. Campbell. Tradeoffs in Displaying Peripheral Information. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, 2000, 241-248.

27. Marshall, S.P. New Techniques for Evaluating Innovative Interfaces with Eye Tracking. In *Proceedings of the ACM Conference on User Interface Software and Technology*, 2003, Keynote Talk.

28. McCrickard, D.S., R. Catrambone, C.M. Chewar and J.T. Stasko. Establishing Tradeoffs That Leverage Attention for Utility: Empirically Evaluating Information Display in Notification Systems. *International Journal of Human-Computer Studies*, *58* (5), 547-582, 2003.

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

30. Miyata, Y. and D.A. Norman. The Control of Multiple Activities. In Norman, D.A. and Draper, S.W. (eds.) *User Centered System Design: New Perspectives on Human-Computer Interaction*, Lawrence Erlbaum Associates, Hillsdale, NJ, 1986.

31. Nakayama, M. and K. Takahashi. The Act of Task Difficulty and Eye-Movement Frequency for the Ocul-Motor Indices. In *Proceedings of Eye Tracking Research and Application*, 2002, 37-42.

32. Newell, A. and H.A. Simon. *Human Problem Solving*. Englewood Cliffs, NJ: Prentice-Hall, Inc., 1972.

33. Rowe, D.W., J. Sibert and D. Irwin. 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*, 1998, 480-487.

34. Shell, J.S., T. Selker and R. Vertegaal. Interacting with Groups of Computers. *CACM*, *46* (3), 40-46, 2003.

35. Speier, C., J.S. Valacich and I. Vessey. The Influence of Task Interruption on Individual Decision Making: An Information Overload Perspective. *Decision Sciences*, *30* (2), 337-360, 1999.

36. Takahashi, K., M. Nakayama and Y. Shimizu. 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 & Applications*, 2000.

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