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, IL 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 interactive, hierarchical tasks while mental workload was measured 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 boundaries higher in a task model and *less* at boundaries lower in the model, (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. From the results, we show how to map mental workload onto a computational Index of Opportunity that systems can use to better reason about human attention.

KEYWORDS
Attention, Interruption, Pupil size, Task models

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

INTRODUCTION
When interacting with applications, users often suffer *interruption overload*. E-mail notifications [18], instant messages [6], agent requests [24], and system alerts all contribute to this burgeoning epidemic of interruption that negatively affects almost every user. When a background application interrupts a user at an inopportune moment during task execution, the user performs tasks slower [2, 7, 26], commits more errors [23], makes worse decisions [32], and experiences more frustration, annoyance, and anxiety [1, 2, 36] 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 [12, 14, 15]. These systems compute the cost of interruption using external cues such as desktop activity, visual and acoustical analyses of the physical task environment, and scheduled activities of the user.

CHI 2005, April 2–7, 2005, Portland, Oregon, USA.

However, to compute a more accurate cost of interruption, systems need a direct measure of a user's mental workload [17]. Researchers have long argued that opportune moments for interruption occur during periods of low mental workload [2, 6, 9], and posit that these periods occur at subtask boundaries during task execution [27]. Interactive tasks, however, are composed of hierarchical patterns of goal formulation, execution, and evaluation, creating many levels of subtask boundaries in a task model [5]. Our work empirically shows how a user's mental workload changes during task execution, focusing on subtask boundaries, and shows how to map workload to an Index of Opportunity that systems can use to better reason about human attention.

We conducted a user study where 12 users performed two interactive, hierarchical tasks. While a user performed the tasks, 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, 10, 21]. To analyze response data, we developed and validated GOMS models for the tasks and precisely aligned pupillary response with the models.

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 model, (iv) mental 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 of workload be tightly coupled to a validated task model.

Our work contributes the first evidence showing how much mental workload changes at different levels of boundaries in a hierarchical task model. From our results, we develop an Index of Opportunity that maps mental workload - as measured by pupillary response – to a computational index that is sensitive to changes in mental workload at subtask boundaries. The index would be useful for systems that manage human attention – not only on the desktop but also in control rooms, aviation cockpits, and in-vehicle displays.

Also, there is rapidly growing interest in the use of mental workload to evaluate user interfaces [25]. By leveraging our

research method of aligning pupillary response data to validated task models, interface designers can better link periods of unacceptably high workload with specific tasks in an interface, targeting those tasks 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

An opportune moment for interruption is during a period of low mental workload, which has been supported by many empirical studies [1, 2, 6-8, 26, 32]. When a user is interrupted during a period of low mental workload, the interruption causes substantially less disruption than if it had occurred during a period of high workload [1, 2, 32]. The challenge is to understand when a user's mental workload changes during task execution.

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 [27]. Interactive tasks, however, are composed of hierarchical, recursive patterns of goal formulation, execution, and evaluation, creating many levels of subtask boundaries in a task model [5].

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

Reasoning About When to Interrupt Users

In [12, 14, 15], 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 task environment, and scheduled activities of the user to compute a 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 an accurate assessment of workload, systems can make poor decisions about when to interrupt a user. For example, in [12], researchers model periods of desktop inactivity as *better* for interruption than periods of activity. Miyata and Norman [27] argue, however, that inactivity is generally *worse* for interruption because those moments often represent periods of planning or evaluation, which can require more mental workload than subtask execution.

From our results, we show how to map mental workload onto a computational Index of Opportunity for interruption. By using the Index of Opportunity as part of a broader reasoning framework, a system can make a more accurate assessment of the cost of interrupting a user, 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 an effective and reliable measure of mental workload [10, 11, 20, 21, 28, 33], where increases in pupil size correlate with increases in mental workload. 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 [3]. UI researchers are already using pupil size to evaluate the mental workload imposed by user interface designs [25].

Iqbal et al. [17] showed that pupillary response correlates with the workload of *interactive* tasks and discovered that changes in workload seem to align well with the hierarchical model of the task being performed. Our current study seeks to better understand this relationship.

Physiological signals can provide systems with a continuous measure of the state of the user [29], e.g., to measure user stress [35]. To measure mental workload, eye movement, blink rate, and heart rate variance have been used [22, 30, 33]. However, pupil size offers an *immediate* measure of workload, which simplifies analysis of the response data.

Of course, we do not expect users to wear existing eye-tracking equipment while performing computing tasks. We believe that future technology will provide more cost-effective and less physically intrusive means to measure pupil size, e.g., eye trackers built into LCD monitors [34] or even eye glasses [31], thus justifying our use of pupil size.

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. While there is a need to cross-validate mental workload across multiple measures, 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 changes at subtask boundaries and how much that change differs at different levels of subtask boundaries in a hierarchical task model. Our study also investigates whether different types of subtasks induce different mental workload. Specifically, we designed our 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 a user's changing mental workload during task execution be effectively mapped to a computational index that systems can use to better reason about attention?

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 shown that gender does not influence a user's pupillary response to mental workload [3].

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

Tasks

For the study, we developed two tasks – a route planning task and a document editing task. We designed the tasks 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 interactive tasks, the sequence had to be controlled to align changes in mental workload to task execution. The lower-level subtasks within the route planning and editing tasks are representative of those within many interactive tasks, e.g., selection, memory store and recall, data entry, reasoning, comprehension and processing, and motor movements. Each task required about 5 minutes to perform, which is considerably longer than tasks used in many prior studies, e.g., [4, 16, 20, 21, 33]. This resulted in about 3,000 data points for each task.

In each task, we varied the mental workload among some subtasks that were repeated to later validate that changes in mental workload caused changes in pupillary response.

Route Planning Task

In the Route Planning task (Figure 1a), 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 balloon that appeared when the user moved the mouse over a route segment.

To perform the task, the user moved the mouse over a route segment in the map, committed the distance and fare information that appeared in the tooltip to memory, and entered the data into the corresponding row in the table. When the user moved the mouse away from the segment, the tooltip disappeared. 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 drop down lists, shown near the bottom of Figure 1a.

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

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

Document Editing Task

In the Document Editing task (Fig. 1b), a user was given a document annotated (highlighted) with three comments that varied in the complexity of the edit required. The document

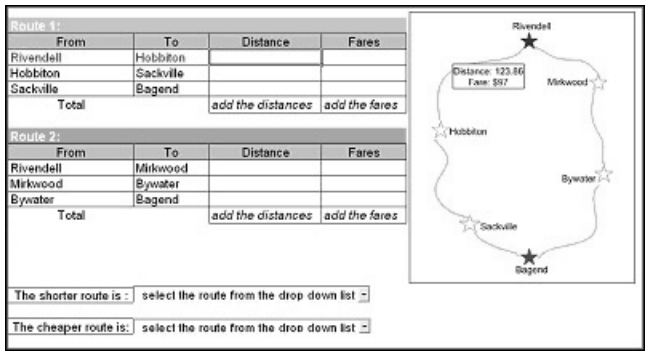

Figure 1a: The interactive route planning task. A user retrieves distance and fare information from the map, enters it into the tables on the left, adds the distances and fares, and selects the cheaper and the shorter of the two routes.

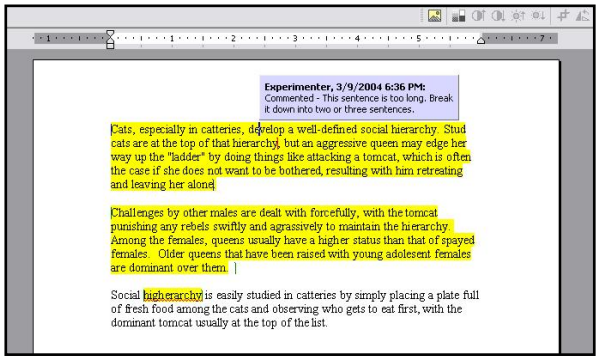

Figure 1b: The document editing task. A user edits the document based on the given comments. Once edited, a user saves the document to a specified location and file name.

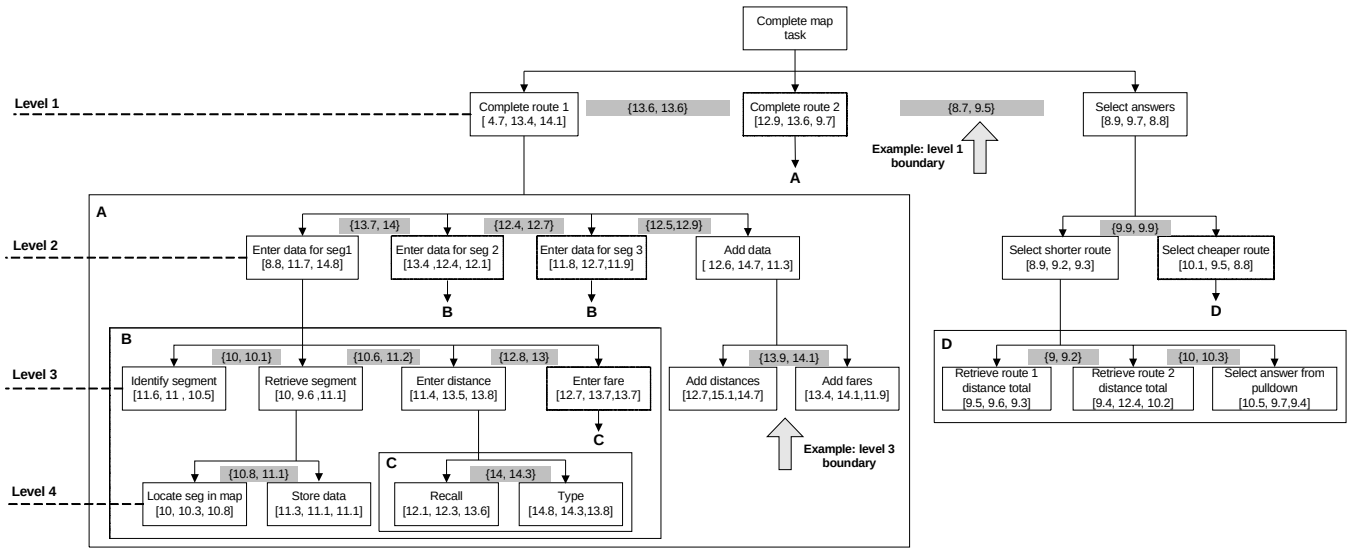

Figure 2: Validated GOMS model of the route planning task. The interior nodes represent goal nodes, the leaf nodes represent operators, and time moves from left to right. Regions A, B and C show parts of the task repeated elsewhere in the model. Within each subtask, we provide the [*beginning PCPS, average 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.

text was about the social hierarchy of a common household pet (cats). This topic was selected because we felt it would be interesting, familiar, and understandable to most users.

The user was instructed to edit the document according to each comment, which appeared as a tooltip when the user moved the mouse over the corresponding highlight. After reading a comment, the user located the text, made the appropriate edit, and repeated two more times. Once edited, the user saved the document to a specified directory with a specified file name, both were provided beforehand.

The main cognitive subtasks were understanding comments and the document text (Language Comprehension), making edits (Language Processing), and recalling the directory and file name (Recall). To vary mental workload, we varied the complexity of the required edits. The easy edit was to correct one misspelled word. The medium edit was to locate and correct two misspelled words. The difficult edit was to rephrase a sentence so that it was grammatically correct.

Procedure

Upon arrival at the lab, a user filled out a consent form, a questionnaire for background information, and received 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 relax and fixate on a blank screen for about 10 seconds. The user then performed the experimental task and was instructed to perform the task as quickly and as accurately as possible. The ordering of the tasks was counterbalanced. Pupil data was logged to time stamped files while a user's screen interaction was video recorded with eye gaze overlaid. Because the videos and pupil data received time stamps from the same clock, we could precisely align them. The user required about 5 minutes to perform each task and the entire experimental session lasted about 30 minutes.

Task Models and Validation

We performed a GOMS analysis to decompose each task into its component subtasks. In GOMS terminology [5], we started with the root goal, for example, 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 for the Route Planning task, reusing repetitive parts for brevity. 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, in Figure 2, consider the "Locate seg in map" and "Store data" subtasks at the left of level 4. 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 a subtask represents a store, recall, reasoning, language comprehension, language processing, or motor operator.

The GOMS models were developed in an iterative manner. For each task, we developed an initial GOMS model through our own analysis of its execution. Once defined, eight 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 from the user study. We compared the identified operator sequences to the leaves of our task model and refined it as necessary.

We validated the final task models by matching observable events (keyboard, mouse, and gaze) in the videos from the user study to the operators in the models. Gaze events were particularly helpful for matching cognitive operators. The number of error steps each user performed was counted. An error step was defined to be a 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. The average error rate for the Route Planning task was 2.81%, ranging from 0% to 5.66%, consistent with models validated in [5].

We repeated this procedure for the Document Editing task. The task model for this task had 5 levels and 41 nodes. The average error rate was 2.3%, ranging from 0% to 13.6%.

The GOMS models accurately reflected a user's execution of the tasks and enabled us to precisely align a user's pupillary response to the models. This was very challenging because each user performed the tasks at different speeds. Thus, we had to align the pupil data by meticulously synchronizing it to specific event points in each 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 subtask's time window. The time window varied according to the type and level of a subtask and ranged from about 24 msecs for the lower-level subtasks to about 63 seconds for the higher-level subtasks.

RESULTS

In this section, we first validate that changes in mental workload caused changes in pupillary response. Then, for both experimental tasks, 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.

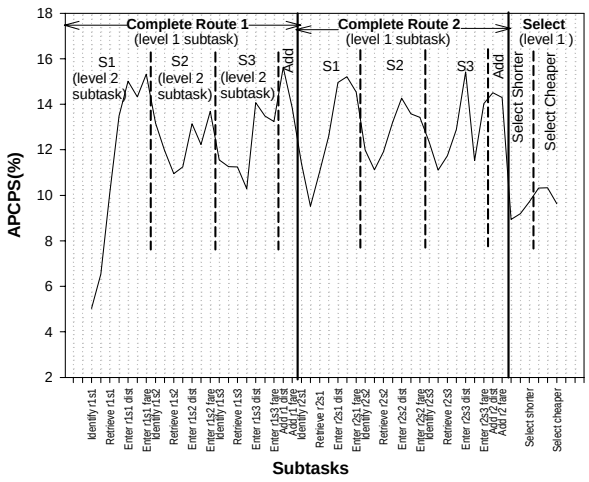

Figure 3: APCPS for the subtasks in the route planning task. Solid lines indicate level 1 boundaries and dashed lines indicate level 2 boundaries. The x-axis enumerates level 3 subtasks. Notice how the graph dips lower at level 1 boundaries than at level 2 boundaries – showing how mental workload decreases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

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.

Validation of Pupillary Response to Mental Workload

To validate that changes in mental workload caused changes in pupillary response, we compared pupillary response among the subtasks for which workload was manipulated. For route planning, we performed a one-way ANOVA with Load (fewest, middle, and most digits) as the factor on the PCPS of recall subtasks. Results showed that Load had a main effect on APCPS ($F(2,72)=17.363$, $p<0.028$), with higher load subtasks having a higher APCPS than lower load subtasks. For document editing, an ANOVA with Complexity (simple, medium, and difficult) as the factor on the edit subtasks showed that more difficult edits had higher APCPS than easier edits ($F(2,22)=3.404$, $p<0.05$). These results validate that changes in mental workload did cause changes in pupillary response.

Route Planning Task

Figure 3 shows the mean APCPS of the subtasks for the route planning task. Time moves from left to right and the vertical lines represent first and second level boundaries from the task model in Fig. 2. The rise and fall of the curve shows changing mental workload during task execution.

Mental workload during subtasks

To validate that cognitive subtasks induced increased mental workload on a user, we performed a one-sample t-test on the APCPS for the Store, Recall, and Reasoning subtasks. We found that the APCPS was greater than 0 across subtasks ($\underline{M}=12.7$, $\underline{SD}=7.3$, $\underline{t}(263)=28.25$, $\underline{p}<0.001$),

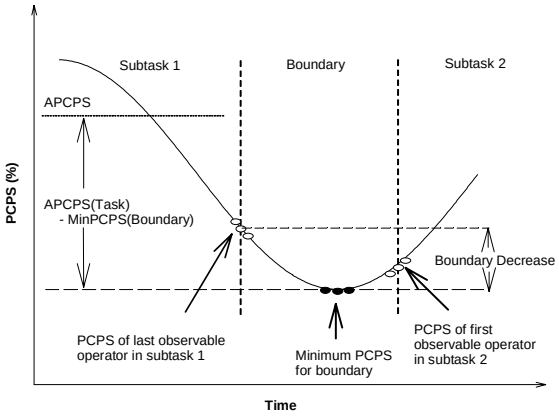

Figure 4: Illustration of metrics used in the analysis. The vertical dashed lines mark the last observable operator in subtask 1 and the first observable operator in subtask 2 (taken as the average of the three surrounding values) and define the boundary between the two subtasks. Our analysis compared the differences between the minimum PCPS at a boundary and both the PCPS of the last observable operator of its preceding subtask and the APCPS of its preceding subtask.

and the standardized effect size index d was 1.7, a high value. This represents a 12.7% increase over the baseline and shows that subtasks did impose increased workload on a user. We only used cognitive subtasks in our analysis since the relationship between cognitive effort and pupillary response is the one best established by prior work [3].

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

A one-way ANOVA with Level as the factor showed a main effect on APCPS ($F(1,262)=3.898$, $p<0.049$). Because the subtasks used in this comparison are the operators in the task model, each subtask existed at either level 3 or 4. Subtasks at level 3 had a higher APCPS than at level 4 (difference was 2.3 percentage points). This difference may be attributed to the cognitive demands of the subtasks rather than their level as level 3 contained all the reasoning (more cognitively demanding) subtasks. Our results show that subtasks induce increased mental workload and different types of subtasks induce different workload on a user.

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. For each boundary, we computed the *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 workload.

A one-sample t-test showed that Boundary Decrease was greater than 0 across all subtasks ($M=0.23$, $SD=2.17$, $t(611)=2.67$, $p<0.008$). The standardized effect size index d was 0.1. This shows that mental workload decreases at a subtask boundary, but the decrease is small on average.

One reason for the small effect size was that the lowest level boundaries showed little or no decrease in PCPS. PCPS likely did not decrease at these boundaries because the adjacent subtasks were short (about 200 msec), rapid, and closely related. We reran the one sample t-test for Boundary Decrease, but excluded level 4 samples. Results showed a stronger effect ($M=0.54$, $SD=2.15$, $t(395)=4.995$, $p<0.001$), with an improved d of 0.25. As level 3 and then level 2 samples are removed, results show increasingly stronger effects. This implies that changes in workload are meaningful down to the level of boundary just above the elementary operators in a 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 analysis showed that Boundary Decrease at level 2 was greater than at level 4 (difference was about one percentage point, $p<0.001$) and that Boundary Decrease at level 3 was greater than at level 4 (difference was about 0.8 percentage points, $p<0.001$). This along with the previous result shows that workload decreases *more* at higher level boundaries in a task model and *less* at lower level boundaries in the model (see Fig. 3). A plausible 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 [27].

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

Mental workload during 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 may execute in parallel [19], 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=0.28, SD=3.43, 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 lower level pairs not differing as much as higher level pairs. Excluding level 4 pairs, for example, the same test shows a larger difference among paired samples (M=0.97, SD=3.57, t(395)=5.4, p<0.001). The effect size d was 0.3, showing a marked improvement.

A one-way ANOVA showed that Level had a main effect on the paired samples (F(3,608)=19.677, p<0.001). Post hoc tests showed that differences at level 1 were marginally greater than level 2 (about 1.9 percentage points, p<0.056), were greater than at level 3 (about 2.13 percentage points, p<0.014) and were greater than at level 4 (about 3.9 percentage points, p<0.001). Differences at level 2 were greater than at level 4 (about 2 percentage points, p<0.001) and differences at level 3 were greater than at level 4 (about 2 percentage points, p<0.001). This and the previous result further shows that workload decreases more at boundaries higher in a model and less at boundaries lower in the model.

We not only found that workload changed between levels in the task model, but also that workload changed *within* the same level in the task model. For example, the APCPS over the level 1 boundaries differed (t(14)=4.23, p<.001) with a maximum difference of about 3 percentage points. 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.

Document Editing Task

Figure 5 shows the mean APCPS of the subtasks for the document editing task, analogous to Figure 3.

Mental workload during subtasks

We performed a one sample t-test for the APCPS induced by the Language Comprehension, Language Processing and Recall subtasks. These were the observable subtasks and existed only at Levels 2, 3 and 5. APCPS was greater than 0 across subtasks (M=6.72, SD=6.47, t(299)=17.98, p<0.001) with a standardized effect size d=1.0, a high value. This shows a 6.72% increase over the baseline level, meaning that subtasks did induce mental workload on a user, but not as much as in the route planning task.

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

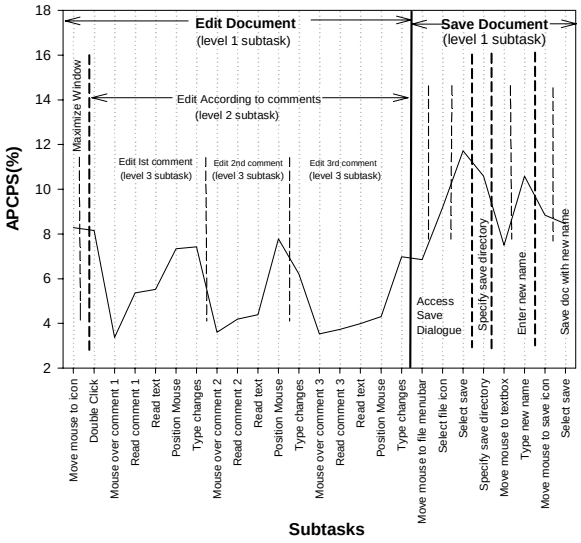

Figure 5: APCPS for subtasks in the document editing task. Solid lines indicate level 1, heavier dashed lines indicate level 2, and lighter dashed lines indicate level 3 boundaries. The x-axis enumerates the observable operators.

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

Mental workload at subtask boundaries

A one-sample t-test showed that Boundary Decrease was greater than 0 across all subtasks (M=0.81, SD=1.93, t(299)=7.271, p<0.001) with an effect size d=0.42. Level (1-5) had a main effect on Boundary Decrease (F(4, 295)= 8.043, p<0.001). Post hoc tests showed that Boundary Decrease at level 1 was greater than level 3 (p<0.013) and level 5 (p<0.004). Boundary Decrease at level 2 was greater than level 3 (p<0.001), level 4 (p<0.033) and level 5 (p<0.001). This shows that mental workload decreases more at boundaries higher in a model and less at boundaries lower in the model, consistent with the Route Planning task.

Mental workload during subtasks vs. subtask boundaries

A paired samples t-test for each pair of minimum PCPS within a subtask boundary and the APCPS of its preceding subtask execution showed no difference (M= -0.103%, SD=3.88%, t(299)= -0.461, p<0.645). As before, there were few and only very small decreases at the lowest level boundaries. Excluding levels 4 and 5 pairs, for example, analysis now showed a difference among the pairs (M=1.1, SD=4.1, t(155)=3.343, p<0.001) with an effect size d=0.27.

An ANOVA showed that Level had a main effect on the paired samples (F(2,153)=3.502, p<0.033). Differences at level 1 were greater than at level 3 (about 2.5 percentage points, p<0.027). Differences at level 1 tended to be greater

than level 2 (about 1.92 points) and differences at level 2 tended to be greater than at level 3 (about 0.54 points), although significance was not reached. These results further support that workload decreases *more* at boundaries higher in a model and *less* at boundaries lower in the model.

FINDINGS

From the user study, we found that:

- *Different types of subtasks impose different workload on a user*. Our results show that some types of subtasks induce more workload than others. For the route planning task, for example, the Reasoning subtasks induced more workload than the Store or Recall subtasks. For the document editing task, Language Processing subtasks induced more workload than Comprehension subtasks.
- *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 a subtask boundary.
- *Mental workload decreases more at boundaries higher in a task model and less at boundaries lower in the model*. We compared the minimum PCPS at a subtask boundary to both the last PCPS and the APCPS of the preceding subtask across levels of a task model. In both cases, the difference between pairs was larger at boundaries higher in a model and smaller at boundaries lower in the model. Our results provide the first evidence demonstrating this effect. Existing eye tracking systems can reliably measure changes at higher levels and many changes at lower levels of a task model. Advances in technology will continue to provide more precise measures of pupil size.
- *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 example, for levels 1 and 2 of the route planning task, we found that the change in workload differed within the level. Our 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. Our results empirically demonstrate, however, 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 more effective decisions about when to interrupt.
- *Effective understanding of why changes in mental workload occur requires that the measure be tightly coupled to a validated task model*. To make better sense of user's pupillary response, we validated task models and then overlaid the models onto the pupillary response curve, as exemplified in Figures 3 and 5. Our research method would be useful for interface designers seeking to use mental workload (pupil size) to evaluate alternative designs. By aligning pupillary response to validated task models, designers can better link periods of unacceptably high workload to specific tasks in an interface, and then target those tasks 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 decreases at subtask boundaries, an attention manager can and *should* perform fine-grained temporal reasoning about when to interrupt a user engaged in a task. Deferring the delivery of information or its attentional cue - 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 prior empirical results showing that fine-grained temporal manipulation of an interruption can cause dramatic differences in task performance, error rate, and reported levels of frustration, annoyance, and anxiety [1, 2].

In controlled settings, an attention manager could use pupillary response to learn how opportune different boundaries or other moments in a task are for interruption. When illumination cannot be controlled or the use of eye tracking systems is not possible or desirable, an attention manager could use *pre-defined* moments for interruption for common or frequent tasks. These moments would be identified from workload-aligned task models developed *a priori* in controlled settings and then used by an attention manager for the same tasks in uncontrolled settings. While this process may not allow for individual tailoring, it would eliminate the need for users to have eye tracking equipment.

TOWARDS AN INDEX OF OPPORTUNITY

Because interpretation of raw PCPS data is difficult, we show how to map PCPS data onto an easily interpreted, configurable, and computationally convenient scale, called an Index of Opportunity (IOP). The IOP is an index that maps pupillary response to a discrete, 20-point scale that indicates how opportune a particular moment is for an interruption. On the scale, '1' indicates the least opportune moment for an interruption while '20' indicates the most opportune moment. Each successive bin is assumed to represent a meaningful decrease in mental workload, meaning that an interruption would result in less disruptive impact. In determining the number of bins for the IOP, we wanted the scale to be sensitive to changes in mental workload at subtask boundaries, yet not be so fine-grained that it became an uninformative replacement for raw PCPS.

The IOP was developed using PCPS data from both of our experimental tasks to increase robustness. To compute the number of bins, we divided the span of the 95% range of values centered about the median (accounting for most of the PCPS data) by the lower end of the 95% CI for the average decrease in PCPS at a subtask boundary. This would make the index sensitive to most changes in PCPS at a subtask boundary. The span of the 95% range of values

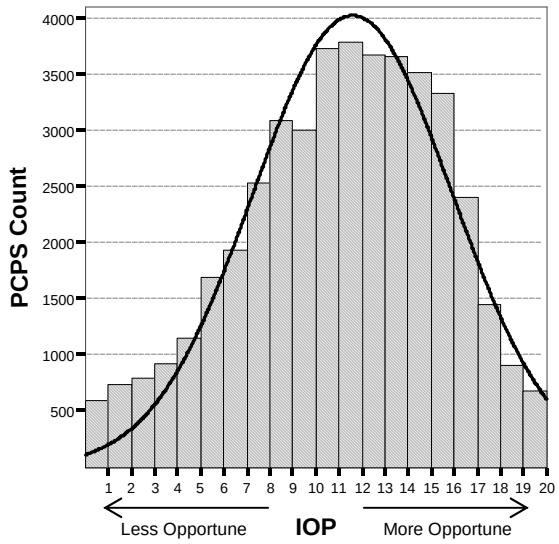

Figure 6: Histogram of IOP values mapped from Equation 1 using PCPS data from both tasks. The distribution is near normal and users were more often in the middle than at either end of the scale. Each PCPS count is a 100ms sample.

centered about the median was 30, ranging from [-4, 26]. The 95% CI for the average decrease in PCPS at a subtask boundary was 1.7±0.2. Using the lower end of the CI (~1.5), we divided the span, 30, by the average decrease, 1.5, to give 20 bins. This is the minimum number of bins such that the index is still sensitive to most changes in PCPS at subtask boundaries. A PCPS value can thus be mapped to the IOP [1, 20] using the linear function:

$$(1) \quad IOP = \left\lfloor 19 * \left[1 - \frac{PCPS_{current} - PCPS_{baseline}}{PCPS_{highest} - PCPS_{baseline}} \right] \right\rfloor + 1$$

Because results from our study showed that different users had different changes in PCPS, i.e., there was an effect of Subject, one can manipulate the high and baseline values in the mapping to configure it for specific users and tasks. This can be achieved by analyzing a specific user's PCPS for several tasks over a period of time. This would be most appropriate in a domain where a user performs critical tasks under different workload conditions, e.g., air traffic control.

We judged the quality of the mapping by its ability to produce a near-normal distribution of IOP values. Figure 6 shows a histogram of IOP values mapped from the pooled PCPS data. It shows that the mapping does indeed produce a near-normal distribution (skewness=-0.347, kurtosis=-0.448) and intuitively reflects that a user is more often in the middle of the IOP scale than at either end of it.

Although our mapping fits the data well, further research is needed to precisely determine the endpoints for each bin and validate that successive bins cause less disruption. Our mapping, however, provides an important first step toward an index that is sensitive to meaningful changes in workload during task execution, which a less sensitive scale, e.g., a scale of 'low', 'medium', and 'high', would otherwise miss.

In an attention manager, the IOP can serve either as a direct measure of the cost of interruption or as part of a broader reasoning framework that considers additional dimensions such as cognitive resource conflicts between the primary and interrupting task, or the presentation style, urgency, or relevancy of an interruption. The IOP, for example, could serve as an evidence variable in a Bayesian network [13]. A system could also model temporal patterns of IOP values to develop a more robust sense of availability that is less sensitive to transitory changes in IOP.

FUTURE WORK

For future work, we intend to:

- *Validate a mapping from PCPS to a scale of disruption*. Our IOP transforms ranges of PCPS into discrete bins, and assumes that successive bins have successively lower costs of interruption. To validate the mapping, we want to conduct user studies where we manipulate workload using different tasks, interrupt a user, and measure the disruptive impact. From the results, the IOP can be refined to better reflect a scale of disruption.
- *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 interruption*. The effects of interruption are typically measured using external measures such as task time, error rate, and subjective ratings. By using pupil size to measure changes in mental workload due to interruptions, we may further understand their disruptive effects.

CONCLUSION

To contribute to systems that reason about human attention, our work empirically demonstrates how a user's mental workload changes during task execution. Results show that mental workload decreases at subtask boundaries and decreases more at boundaries higher in a task model and less at boundaries lower in the model. This contributes further theoretical understanding of how workload changes during task execution, and helps systems identify more appropriate moments for interruption. Also, by leveraging our research method of aligning pupillary response to validated task models, designers can better link periods of unacceptably high workload to specific tasks in an interface and target them for re-design. We show how to map mental workload onto a computational index that is sensitive to changes in workload at subtask boundaries. By using the index in a broader reasoning framework, systems can make more effective decisions about when to interrupt users.

REFERENCES

1. Adamczyk, P.D. and B.P. Bailey. If Not Now When? The Effects of Interruptions at Various Moments within Task Execution. *CHI*, 2004, 271-278.
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. *INTERACT*, 2001, 593-601.
3. Beatty, J. Task-Evoked Pupillary Responses, Processing Load, and the Structure of Processing Resources. *Psychological Bulletin*, *91* (2), 276-292, 1982.
4. Bradshaw, J.L. Pupil Size as a Measure of Arousal During Information Processing. *Nature*, *216*, 515-516, 1967.
5. Card, S., T. Moran and A. Newell. *The Psychology of Human-Computer Interaction*. Lawrence Erlbaum Associates, 1983.
6. Cutrell, E., M. Czerwinski and E. Horvitz. Notification, Disruption and Memory: Effects of Messaging Interruptions on Memory and Performance. *INTERACT*, 2001, 263-269.
7. Czerwinski, M., E. Cutrell and E. Horvitz. Instant Messaging and Interruption: Influence of Task Type on Performance. *Proc. OZCHI*, 2000, 356-361.
8. Czerwinski, M., E. Cutrell and E. Horvitz. Instant Messaging: Effects of Relevance and Timing. In *People and Computers XIV: Proceedings of HCI*, 2000, 71-76.
9. Gillie, T. and D. Broadbent. What Makes Interruptions Disruptive? A Study of Length, Similarity, and Complexity. *Psychological Research*, *50*, 243-250, 1989.
10. Hess, E.H. and J.M. Polt. Pupil Size in Relation to Mental Activity During Simple Problem Solving. *Science*, *132*, 1190-1192, 1964.
11. 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.
12. Horvitz, E. and J. Apacible. Learning and Reasoning About Interruption. In *Proceedings of the Fifth ACM International Conference on Multimodal Interfaces*, 2003, 20-27.
13. 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. *Proc. Uncertainty in Artificial Intelligence*, 1998, 256-265.
14. Horvitz, E., A. Jacobs and D. Hovel. Attention-Sensitive Alerting. *Proc. Uncertainty in Artificial Intelligence*, 1999, 305-313.
15. 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. *CHI*, 2003, 257-264.
16. Hytintk, J., J. Tommola and A. Alaja. Pupil Dilation as a Measure of Processing Load in Simultaneous Interpretation and Other Language Tasks. *The Quarterly Journal of Experimental Psychology*, *48A* (3), 598-612, 1995.
17. Iqbal, S.T., X.S. Zheng and B.P. Bailey. Task-Evoked Pupillary Response to Mental Workload in Human-Computer Interaction. *CHI*, 2004, 1477-1480.
18. 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.
19. John, B., A. Vera, M. Matessa, M. Freed and R. Remington. Automating CPM-Goms. *CHI*, 2002, 147-154.
20. Juris, M. and M. Velden. The Pupillary Response to Mental Overload. *Physiological Psychology*, *5* (4), 421-424, 1977.
21. Kahneman, D. Pupillary Responses in a Pitch-Discrimination Task. *Perception & Psychophysics*, *2*, 101-105, 1967.
22. 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.
23. 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.
24. Maes, P. Agents That Reduce Work and Information Overload. *Communications of the ACM*, 37 (7), 30-40, 1994.
25. Marshall, S.P. New Techniques for Evaluating Innovative Interfaces with Eye Tracking. *UIST*, 2003, Keynote Talk.
26. McFarlane, D.C. Coordinating the Interruption of People in Human-Computer Interaction. *INTERACT*, 1999, 295-303.
27. 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.
28. 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.
29. Picard, R.W. *Affective Computing*. MIT Press, 1997.
30. Rowe, D.W., J. Sibert and D. Irwin. Heart Rate Variability: Indicator of User State as an Aid to Human-Computer Interaction. *CHI*, 1998, 480-487.
31. Shell, J.S., T. Selker and R. Vertegaal. Interacting with Groups of Computers. *CACM*, *46* (3), 40-46, 2003.
32. 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.
33. Takahashi, K., M. Nakayama and Y. Shimizu. The Response of Eye-Movement and Pupil Size to Audio Instruction While Viewing a Moving Target. *Proc. Eye Tracking Research & Applications*, 2000.
34. Tobii-Systems. *http://www.tobii.se/*,
35. Wilson, G.M. and M.A. Sasse. The Head or the Heart?: Measuring the Impact of Media Quality. *CHI*, 2000, 117-118.
36. Zijlstra, F.R.H., R.A. Roe, A.B. Leonora and I. Krediet. Temporal Factors in Mental Work: Effects of Interrupted Activities. *Journal of Occupational and Organizational Psychology*, *72*, 163-185, 1999.