# Task-Evoked Pupillary Response to Mental Workload in Human-Computer Interaction

**Shamsi T. Iqbal[♦], Xianjun Sam Zheng[†] and Brian P. Bailey[♦]**
Department of Computer Science[♦] and Beckman Institute[†]
University of Illinois
Urbana, IL 61801, USA
(siqbal, xzheng, bpbailey)@uiuc.edu

## Abstract

Accurate assessment of a user's mental workload will be critical for developing systems that manage user attention (interruptions) in the user interface. Empirical evidence suggests that an interruption is much less disruptive when it occurs during a period of lower mental workload. To provide a measure of mental workload for interactive tasks, we investigated the use of task-evoked pupillary response. Results show that a more difficult task demands longer processing time, induces higher subjective ratings of mental workload, and reliably evokes greater pupillary response at salient subtasks. We discuss the findings and their implications for the design of an attention manager.

**Categories & Subject Descriptors:** H.5.2 [**Information Interfaces and Presentation**]: User Interfaces – *Evaluation and Methodology*; H.1.2 [**Models and Principles**]: User/Machine Systems – *Human Information Processing.*

**General Terms:** Design; Experimentation; Human Factors.

**Keywords:** Attention; Interruption; Mental Workload; Pupil Size; Task Models; User Studies.

## INTRODUCTION

Productive interaction between humans and computers requires that a user must effectively manage her attention among the applications that are competing for it. A poorly timed notification (interruption) due to instant messages, incoming email, or system alert can disrupt a user's task performance [3, 6] and emotional state [1, 3, 13].

An attractive solution is to develop an attention manager that manages where in a user's task sequence an application can gain user attention. Empirical studies show that a less disruptive moment for an interruption is during a period of low mental workload in a user's task sequence [6, 7, 11]. Thus, a significant challenge in developing an attention manager is to develop a reliable measure of mental workload for a dynamic task environment such as the desktop interface. Although pupil size is known to correlate well with the mental workload for discrete, non-interactive tasks [10], we investigated how well pupil size correlates with the mental workload demanded by *interactive* tasks representative of daily computer-based tasks that users often perform.

In our study, a user performed an easier and more difficult task from several task categories and we measured pupil size using a head-mounted eye-tracker. We used task completion time and subjective ratings of difficulty to validate the mental workload imposed by the tasks. Our results show that a more difficult task demands longer processing time, induces higher subjective ratings of mental workload, and reliably evokes greater pupillary response at corresponding subtasks than a less difficult task. We discuss our empirical findings and their implications for the design of an attention manager.

## RELATED WORK

### Mental Workload Assessment Techniques

Mental workload can be assessed with a number of techniques, including task performance on primary and secondary tasks [11], subjective ratings, and physiological measures (pupil size, heart rate, EEG) [8]. We believe that pupil size is the most promising single measure of mental workload because it does not disrupt a user's ongoing activities, provides real-time information about the user's mental workload, and is less intrusive than other physiological measures such as heart rate or EEG.

### Task-Evoked Pupillary Response

The correlation of pupil size with mental workload has long been supported [4, 9, 10]. Research has shown that pupil dilations occur at short latencies following the onset of a task and subside quickly once the task is completed. More importantly, the magnitude of the pupillary dilation appears to be a function of processing load, or the mental effort required to perform the cognitive task. Note that other than the task factor, some environmental factors (e.g. ambient illumination and the near reflex), or emotional states may also induce pupillary response, producing changes of pupil size. Nonetheless, Beatty [4] has shown that task-evoked pupillary response uniquely reflects the momentary level of processing load and is not an artifact of non-cognitive confounding factors. In fact, the task-evoked pupillary response has been widely used as a tool to investigate various aspects of human information processing, such as perception, memory,

*CHI 2004*, April 24–29, 2004, Vienna, Austria.
ACM 1-58113-703-6/04/0004.

reasoning and reading [4]. However, whether pupil size can provide a real-time measure of mental workload for more natural interactive tasks in human-computer interaction requires further investigation.

## USER STUDY

We conducted a user study to answer two main questions:

- How well does pupil size correlate with the mental workload that tasks from the same task category impose on a user?
- Is this correlation pattern consistent across several categories of primary task?

### Experimental Design

The study was a 4 Task Category (Object manipulation, Reading comprehension, Mathematical reasoning and Searching) x 2 Difficulty (Easy and Difficult) repeated measures within-subjects design.

### Equipment

As a user performed tasks, we recorded his pupil data using a head-mounted SR Inc., Eyelink II eyetracker with a 250 HZ sample rate and 0.005 degree spatial resolution.

### User and Tasks

Twelve users (6 female) volunteered in the user study. The average age of the users was 23.7 years (SD = 3.23).

An important part of the study was to identity a meaningful set of task categories representative of daily work tasks. We determined task categories from a literature review, an informal questionnaire to several users, our own experience, and the consideration of user time for the study. Four task categories were developed, each with two difficulty levels (easy vs. difficult):

- *Reading Comprehension.* A user read a given text and answered questions. The easier task belonged to grade 9 level and the more difficult task belonged to grade 17.
- *Mathematical Reasoning*. A user performed math calculations. For the easier task, a user had to mentally add two four digit numbers and select the correct answer from a list of three options. For the more difficult task, a user had to mentally add 4 five-digit numbers, retain the result in memory, and decide whether the result exceeded a given number.
- *Searching*. A user searched for a product from a list of similar products according to specified constraints. For the easier task, a user had to find the product from a list of seven products according to one constraint, e.g., the cheapest camera. For the more difficult task, a user had to identify the product using three constraints, e.g., the cheapest 3MP camera with 3X digital zoom.
- *Object Manipulation.* A user had to drag and drop email messages into appropriate folders. The user was given a list of emails, four folders, and classification rules. For the easier task, the rule was simple and specific, such as using the size of the email (1K, 2K, or 3K) in the list. For the more difficult task, the rules were less specific, such as the use of topics (travel, course related, fun and humor, announcements). The user had to make a judgment using the information provided in the email.

Tasks are typically completed through patterns of goal formulation, execution and evaluation, where higher level goals are repeatedly decomposed into simpler, lower level goals [5]. Although each of our tasks had a single high level goal, we were unsure about the level down to which we could still detect changes in mental workload through the use of pupil size. Thus, we also wanted to explore how the goal formulation-execution-evaluation pattern relates to changes in mental workload.

### Procedure

Upon arrival at the lab, the user filled out a background questionnaire and received general instructions for the study. Then, the user was set up with the eye-tracker and went through a calibration process. The user had to perform 8 tasks – one easy and one difficult for each of the 4 categories. At the beginning of each task category, the user was presented with specific instructions to that category and a practice trial to become familiar with the task. The baseline pupil size was collected by having the users fixate on a blank screen for 10 seconds. Then the user was presented with the actual task. After completing each task category, the user was asked to rate difficulty on a 1-5 scale (5 = very difficult, and 1= very easy). The presentation order of task category and tasks within each category were randomized. The users were instructed to perform the tasks as quickly and as accurately as possible. The system logged task performance and we video recorded the screen interaction for later analysis.

### Measurements

A user's subjective rating and task completion time for each task were collected to validate the mental workload associated with each task. The user's pupil data (eye movement information) as well as the user's on-screen activities were recorded separately. These two data sets were synchronized based on correlating timestamps.

For each user, we computed the *percentage change in pupil size* (PCPS), which is the measured pupil size at each task instant minus the baseline size, divided by the baseline. The average PCPS from the beginning to end of each task was used as the task-evoked pupillary response.

## RESULTS

A 4 (Category) x 2 (Difficulty) repeated ANOVA was performed on the collected data.

### Validation of Mental Workload

We used task completion time (Figure 1) and users' subjective ratings (Figure 2) to validate the mental workload imposed by the tasks. An ANOVA on task completion time showed that Category (F(3,33)=30.067,

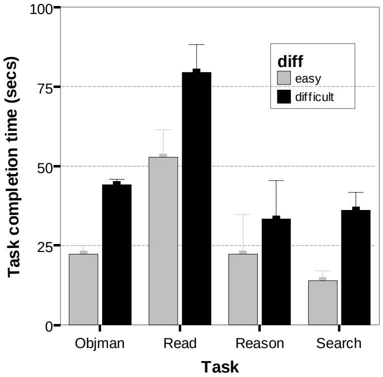

**Figure 1. Average task completion time for each easy and difficult task. Error bars show 95% CI of mean.**

p<0.0005) and Difficulty (F(1,11)=190.933, p<0.0005) had a significant main effect. Users spent more time on the difficult task than the easy task and post-hoc comparisons showed that this effect existed for all but the Reasoning tasks. The interaction between Category and Difficulty (F(3,33)=6.751, p<0.001) was significant, mainly due to the Reading category.

An ANOVA on user ratings showed a significant main effect of Difficulty (F(1,11)=34.912, p<0.0005), with higher ratings for the more difficult task in each category. An interaction between Category and Difficulty was detected (F(3,33)=4.121, p<0.014), mainly due to the easier task in the object manipulation category.

### Analysis 1: Effects of Mental Workload on PCPS

We performed an ANOVA on average PCPS from the beginning to the end of each task as the dependent variable. The results showed a main effect of Category (F(3,33)=4.743, p<0.007). Surprisingly, there was no significant effect of Difficulty (F(1,11)=3.12,p<0.105). See Figure 3. Even a planned t-test comparison between easy and difficulty level on different task categories only revealed a significant difference in the average PCPS between the easy and difficult search task (p<0.025). This is inconsistent with our expectations and with the task performance results and subjective ratings.

We postulated that except for the search task, none of the other tasks had a sustained mental effort throughout the task. These tasks were more hierarchical in that each high level goal could be decomposed into salient lower level goals. This suggests that for hierarchical tasks of short duration, there are periods of lower mental workload and periods of higher mental workload. Averaging PCPS over the entire task negates periods of higher mental workload and therefore may not show a significant increase.

However, if we can separate lower vs. higher periods of mental workload for tasks in the same category, e.g., motor movements such as drag and drop in the object manipulation task), then we can compare mental workload between similar subtasks and these finer comparisons may show differences in mental workload.

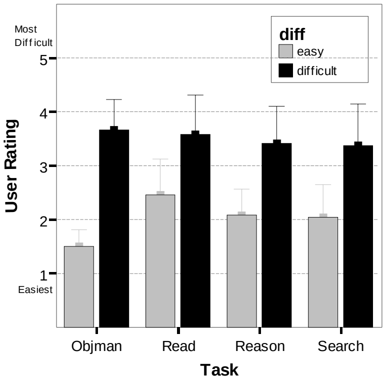

**Figure 2. Average user rating for each task. Error bars show 95% CI of mean. Bars show means.**

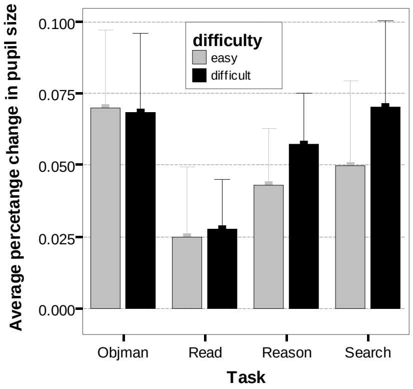

**Figure 3. Average PCPS for each task. Error bars show 95% CI of mean.**

To investigate this hypothesis, we performed a GOMS decomposition for the tasks with the most pronounced hierarchical structure – the object manipulation tasks.

### Analysis 2: Task Decomposition

We performed a GOMS analysis and decomposed the object manipulation task into ten lower-level subtasks that we refer to as L1 subtasks, see Figure 4. Except for the first two L1 subtasks, the remaining eight L1 subtasks were the same and formed the core of the task. We validated the GOMS model by comparing the interaction sequences in the videos with the GOMS model. There was about 95% conformance between them.

We further decomposed each L1 subtask into two more subtasks – a cognitive subtask and a motor subtask - that we refer to as L2 subtasks. In the cognitive subtask, a user was reasoning about the destination folder. In the motor subtask, the user dragged and dropped the object into the target folder. The workload to drag and drop an item should be similar across the easy and difficult tasks. The cognitive subtask, however, should differ in complexity across the easy and difficult task. Therefore we filtered out the motor subtasks and again compared the average PCPS between the easy and difficult tasks. Because of the filtering, only the cognitive subtasks were compared.

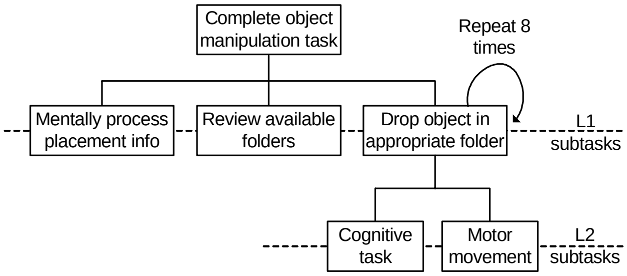
Complete object manipulation task
Repeat 8 times
Mentally process placement info
Review available folders
Drop object in appropriate folder
L1 subtasks
Cognitive task
Motor movement
L2 subtasks

**Figure 4. Partial GOMS analysis for object manipulation.**

This time, however, an ANOVA showed that Difficulty had a main effect on PCPS (F(1,11)=0.058, p<0.046). This suggests that mental workload varies across subtasks in a hierarchical task and that pupil size correlates well with those changes. This implies that a user may experience changes in cognitive load at task boundaries and that the amount of change may differ among those boundaries. Although this was claimed in [12], our analysis has provided supporting empirical evidence.

## DISCUSSION

From analysis of our empirical results, we learned that:

- *A hierarchical task imposes varying mental workloads*. Although measures such as task completion time and user ratings provide an overall measure of workload for a task, they do not reflect the changes in workload that a user experiences throughout the task, and these changes are meaningfully different in task execution.
- *Pupil size correlates well with cognitive load for interactive tasks*. As expected, a user's pupil size increased at the onset of a task and decreased back to baseline levels at the end of a task. For a sustained effort task such as the search task, pupil size correlated well with the difficulty of the overall task.
- *A hierarchical task requires varying mental workloads and pupil size correlates well with those changes*. Once we used just the cognitive subtasks to compare pupil size, we detected differences between the easier and more difficult tasks in each task category. Most striking, however, is that when users' changes in pupil size are overlaid on the task model, the rise and fall of the pupillary response graph matches very well with the onset and completion of the subtasks.

Our results suggest there may be meaningful periods of cognitive shift between subtasks where the user completes one subtask and begins activating goals for the next [2]. Mental workload is expected to be low at these transition periods and we plan to further investigate how significant these shifts are, how well pupil size can detect them, and how opportune they are for an interruption.

The findings from this study have implications for the design of an attention manager, which must balance a user's need for minimal disruption with an application's need to effectively deliver information. Past research has shown that periods of lower mental workload provide better moments for an interruption than periods of higher mental workload. Based on our results, pupil size can provide a reliable measure of mental workload for interactive tasks required by the attention manager.

Because there is strong evidence showing that a user's mental workload changes among subtasks in a hierarchical task structure, an attention manager can perform fine-grained temporal reasoning (at the subtask level) about when to interrupt a user engaged in the task. Furthermore, if the attention manager can record observations of mental workload in a user task model, it could *forecast* a user's mental workload, enabling the system to better reason about when to interrupt the user.

## REFERENCES

1. Adamzcyk, P.D. and B.P. Bailey. If Not Now When? The Effects of Interruptions at Various Moments within Task Execution. *CHI*, 2004, to appear.

2. Altmann, E.M. and J.G. Trafton. Memory for Goals: An Activation-Based Model. *Cognitive Science*, *26*, 39-83, 2002.

3. 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 Interact*, 2001, 593-601.

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

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. *Proceedings of Interact*, Tokyo, Japan, 2001, 263-269.

7. Gillie, T. and D. Broadbent. What Makes Interruptions Disruptive? A Study of Length, Similarity, and Complexity. *Psychological Research*, *50*, 243-250, 1989.

8. Hicks, T.G. and W.W. Wierwille. Comparison of Five Mental Workload Assessment Procedures in a Moving-Base Driving Simulator. *Human Factors*, *21* (2), 129-143, 1979.

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

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

11. McFarlane, D.C. Coordinating the Interruption of People in Human-Computer Interaction. *Proceedings of Interact*, 1999, 295-303.

12. 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, Hillsdale, 1986.

13. 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.