Psycho-Physiological Measures for Assessing Cognitive Load

Eija Haapalainen[1,2], SeungJun Kim[2], Jodi F. Forlizzi[2], Anind K. Dey[2]

[1]Intelligent Systems Group
University of Oulu, Finland
eija.haapalainen@ee.oulu.fi

[2]Human-Computer Interaction Institute
Carnegie Mellon University, USA
{sjunikim, forlizzi, anind}@cs.cmu.edu

ABSTRACT
With a focus on presenting information at the right time, the ubicomp community can benefit greatly from learning the most salient human measures of cognitive load. Cognitive load can be used as a metric to determine when or whether to interrupt a user. In this paper, we collected data from multiple sensors and compared their ability to assess cognitive load. Our focus is on visual perception and cognitive speed-focused tasks that leverage cognitive abilities common in ubicomp applications. We found that across all participants, the electrocardiogram median absolute deviation and median heat flux measurements were the most accurate at distinguishing between low and high levels of cognitive load, providing a classification accuracy of over 80% when used together. Our contribution is a real-time, objective, and generalizable method for assessing cognitive load in cognitive tasks commonly found in ubicomp systems and situations of divided attention.

Author Keywords
Cognitive load, divided attention, interruption, psycho-physiological measurement, elementary cognitive task

ACM Classification Keywords
H5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

General Terms
Design, Human Factors

INTRODUCTION
Advances in computer technologies have improved people's multi-tasking performance. However, human attention is a finite resource [44] and the benefit of being able to process multiple streams of information comes with a cost. Cognitive demands and limitations will ebb and flow in situations of divided attention, due to an interruption of a primary task, or engaging in dual- (or multi)-tasking, making the prediction of when information can be attended to particularly hard. For example, in the context of an interruption, attention switches from one task to another, whether the interruption is relevant or a distraction. Consider, for example, a navigation display that is deemed useful, annoying, or even dangerous as it continually delivers information to the driver, or attending to an information stream on a mobile device while walking, driving, or listening to a lecture.

The ubicomp community can benefit greatly from learning the most salient human measures of cognitive load. Such an understanding can help designers and developers gauge when and how to best communicate information, particularly with the focus in ubicomp on proactively and seamlessly providing the right information at the right time. Presenting information at the wrong time can drastically increase one's cognitive demands, can have negative impacts on task performance and emotional state, and in extreme cases, even be life threatening [26, 45]. Additionally, how information is designed and presented can help or hinder our ability to resume a task that has been interrupted, and to provide more information about the context of interruption [14, 24, 35].

While research advances have been made in sensing context to determine when an individual is interruptible, or on monitoring the interaction between human and interrupter to understand the cost of interruption [19, 20, 26], much more needs to be understood about how cognitive load factors into contexts of multitasking and interruption. Determining both what the right information is and when the right moment to present it are still open problems in ubicomp research. The reason may be the lack of generalized methods for detecting a user's cognitive load fluctuation. However, it could also be the case that ubicomp solutions aim for the most minimal types of interventions, with the goal of interrupting users for shorter periods of time, resulting in a greater number of attention switches. Regardless, what is still needed is a mechanism that monitors a person's internal state as shaped by task or tasks, interruption, and aspects of context. However, conventional methods for assessing cognitive load have yet to attain this. Most current methods are either *post-hoc* subjective assessments of cognitive load, or are often not sensitive to changes in cognitive load.

UbiComp '10, Sep 26–Sep 29, 2010, Copenhagen, Denmark.

In this paper, we collect data from a range of psycho-physiological sensors to explore which ones are useful for assessing cognitive load. As an initial undertaking, we focus on obtaining a method for the objective and real-time assessment of cognitive load while performing tasks that focus on visual perception and cognitive speed. We chose this problem space because it is relevant to many ubicomp contexts of handling interruptions, multitasking and divided attention in the real world — for example, driving and reading a secondary display, attending a meeting receiving a text message, working and observing an ambient display or walking and accessing a mobile device, to name a few.

To derive a measure of cognitive load in this context, we present a variety of stimuli that occupy elementary cognitive processes, which help us to understand divided attention issues. We sense a number of psycho-physiological responses to those stimuli, and use them to determine which responses are most predictive of cognitive load. Our result is an initial model for detecting the degree of cognitive load a user is experiencing in tasks that require visual perception and cognitive speed.

In creating this model, we address the following questions:

- Can we acquire a real-time, objective measure of cognitive load by examining a wide range of psycho-physiological sensor streams for tasks that leverage visual perception and cognitive speed?
- How do the psycho-physiological signals that produce the best cognitive load classifiers differ across individuals? *i.e.*, can a single model or single set of signals be used across all individuals?

To address these questions, we explore the predictability of cognitive load based on psycho-physiological measurements. To do so, we design a user study in which we present six elementary cognitive tasks, each of which is manipulated to induce either high or low cognitive load. We use four sensor devices to measure psycho-physiological responses from twenty participants to these induced loads. We collected time-on-task performance data, and subjective rating of the difficulty of the task. Analyzing this data plus the physiological measures, we found that across all participants, the median heat flux and electrocardiogram median absolute deviation measurements were the most accurate at distinguishing between low and high levels of cognitive load, providing a classification accuracy of over 80% when used together. Our contribution is a real-time, objective, and generalizable method for accurately assessing cognitive load in cognitive tasks commonly found in ubicomp systems and situations of divided attention.

RELATED WORK

Our literature review spans the domain of cognitive capabilities and cognitive load, assessment methods and tools, and psycho-physiological responses. Our goal was to assess the literature to ascertain the best approach for determining a measure of cognitive load.

Cognitive load

Cognitive load is defined as a multidimensional construct representing the load that a particular task imposes on the performer [32, 33]. This also refers to the level of perceived effort for learning, thinking and reasoning as an indicator of pressure on working memory during task execution [53]. This measure of mental workload represents the interaction between task processing demands and human capabilities or resources [16, 49].

Subjective rating-based methods (self-reporting)

Both subjective and objective methods have been used to assess a user's cognitive load. We first discuss the subjective approaches. A number of studies have found that *post-hoc* self-reports of cognitive load are a relatively reliable method for assessing mental effort [34]. In fact, the most commonly used assessment for cognitive load is the subjective NASA task load index (TLX) tool [17]. Despite widespread use of the NASA TLX, other studies do not consider the self-reports to be reliable indicators of cognitive load [*e.g.*, 28]. The subjective, post-hoc nature of this assessment approach can make it difficult to apply in ubicomp systems where automated and immediate assessment is often crucial.

Task Performance-based methods

While less commonly used, a more objective assessment of cognitive load is to measure task performance. Primary task measurements use the performance on a primary, focal task, while secondary task measurements use the performance on a secondary task that is performed simultaneously with the primary task [34]. In this approach, the variation of reaction performance represents the variation in cognitive load. However, the user must be subjected to enough of the task for which the performance is being measured, in order to use this assessment technique. This may not always be viable in a ubicomp setting where users are switching frequently between primary tasks, and multitasking. As well, this approach is not always sensitive to small differences in cognitive load: *i.e.*, if a user performs two tasks equally well, the perceived cognitive load will be identical, although the actual load may not be.

Combinational methods

A few researchers have attempted to integrate behavioral models into a performance model. [15, 36, 37, 41, 42]. This integration can help predict the performance effect of, for example, different phone dialing interfaces, and driving steering tasks. While these approaches are very promising, they require the creation of a sophisticated task model using, for example, ACT-R or GOMS, that is specific to the task being studied. Instead, we are interested in a more generalized method for assessing cognitive load.

Physiological response-based methods

In this work, we apply a psycho-physiologically-based assessment approach, to address the issues with the previously discussed approaches: need in-the-moment, automatic assessment of cognitive load, and to assess load even when no change in task performance can be detected,

for a variety of tasks without significant customization. While typically not used outside a laboratory environment, cognitive load has also been assessed by measuring changes in psycho-physiological signals [34, 23]. This approach is based on evidence that varying task difficulty influences psycho-physiological signals such as pupillary responses, eye movements and blink interval [3, 21, 22, 52], heart rate (HR) and heart rate variability (HRV) [10, 29, 52], electroencephalogram (EEG or brainwave levels) [23, 52], electrocardiogram (ECG) [23], galvanic skin response (GSR) [21, 43], and respiration [29].

Our approach provides an opportunity to objectively detect small variations in cognitive load, in real-time, as desired by ubicomp systems. As we are interested in identifying a generalized mechanism for assessing cognitive load, we will stimulate that load using tasks that leverage basic cognitive processes related to visual perception and cognitive speed. While the earlier psycho-physiological studies provide a solid base to build from, none of them have focused on such basic processes. Instead, they have used applied tasks such as document editing [22], simulated public speaking [10], and traffic control management [43]). We are unable to use their results directly because they leverage combinations of cognitive processes, whereas we are interested in cognitive load responses to individual processes. Instead, for the ubicomp domain, we build significantly on this previous work in using psycho-physiological signals but by using cognitive tasks appropriate to ubicomp, that is, tasks that leverage visual perception and cognitive speed.

ELEMENTARY COGNITIVE TASKS

In measuring psycho-physiological signals to obtain a measure of cognitive load, our approach is to present a variety of stimuli, in the form of *elementary cognitive tasks*. An elementary cognitive task (ECT) refers to any of a range of basic tasks which require only a small number of mental processes and which easily specify correct outcomes [4]. Most ECTs designed in the field of psychology have been used to demonstrate individual differences (or personal characteristics) between more than two participant groups (*e.g.*, patients *vs*. health-controlled people, younger people *vs*. elder people) [1, 4, 38, 39, 40].

In this study, we focus on 'visual perception' and 'cognitive speed' among the human cognitive abilities addressed in [4, 27]. These abilities highly engage spatial orientation or spatial attention [11], which are highly leveraged in today's world of location-based services, situations of divided attention, and ubicomp applications where you may be attending to one activity (*e.g.*, crossing the street) and are either interrupted by incoming information (*e.g.*, text from a friend or ad from a nearby store) or seeking information (*e.g.*, search for information on a car that just drove past).

Based on a review of a number of cognitive factors to assess the elementary cognitive abilities, we identified the major discriminable first-order factors in the areas of visual perception or 'major spatial factors' [25] and cognitive speed. These factors are 'flexibility of closure' (CF), 'speed of closure' (CS) and 'perceptual speed' (PS). We note that the mental processes related to these three cognitive factors highly associate with handling interruptions, the execution of dual-task processing (*e.g.*, way finding requiring spatial attention switching or cognitive mapping) or integration tasks (*e.g.*, comparing the appearance of an ambient display with a mental legend that indicates its meaning) [13, 50, 51]. These activities are common in a wide variety of ubicomp applications and the cognitive factors generalize well to the ubicomp domain.

Flexibility of closure (CF) refers to the ability to keep one or more definite configurations in mind so as to make identifications in spite of perceptual distraction [12]; it also refers to the ability to hold a given visual percept or configuration in mind so as to disembed it from other well-defined perceptual material [6]. Speed of closure (CS) is the ability to combine disconnected, vague, visual stimuli into a meaningful whole [11]; it also refers to the ability to unify an apparently disparate perceptual field into a single percept [4]. As an example, these two cognitive processes operate when we identify an incomplete picture (CS) or detect a reference pattern (a figure, object, word, or sound) that is hidden in other distracting materials (CF). Perceptual speed (PS), also known as 'inspection time' [31, 35], is the cognitive ability to quickly and accurately find target information in literal, digital or figural forms, make comparisons, and carry out other very simple tasks involving perception [12]. Most of the ECTs for this factor arrange one or more visual stimuli, and ask a participant to compare a presented object with a remembered object [4].

In the design of our experiment, we leverage these three basic cognitive factors to assess how cognitive load changes as the task difficulty changes. In our experiment, we focus more heavily on aspects of perceptual speed. Issues of divided attention and interruptions decrease our perceptual capacity, and, in particular, most critical incidents in mobile contexts come from delayed inspection time (*i.e.*, slowed perceptual speed) because one's attention was not switched to the appropriate stimulus in time. Perception and reaction time to stimuli necessarily precedes how to interpret the visual organization of the stimuli. Therefore, in this study, we employ more ECTs for the 'perceptual speed' factor (4) than the other factors (1 each). We discuss our exact experimental design in the following section.

EXPERIMENTAL DESIGN

Participants and tasks

To minimize the confounds of age-related decline in cognitive abilities, we recruited twenty younger participants (younger than 35) with normal or corrected-to-normal vision and hearing (age range/mean/SD: 19-34/25.15/4.45, gender: female 25% and male 75%). They performed six elementary cognitive tasks (ECTs) for 14 minutes 49 seconds on average (net time-on-task). They were asked to

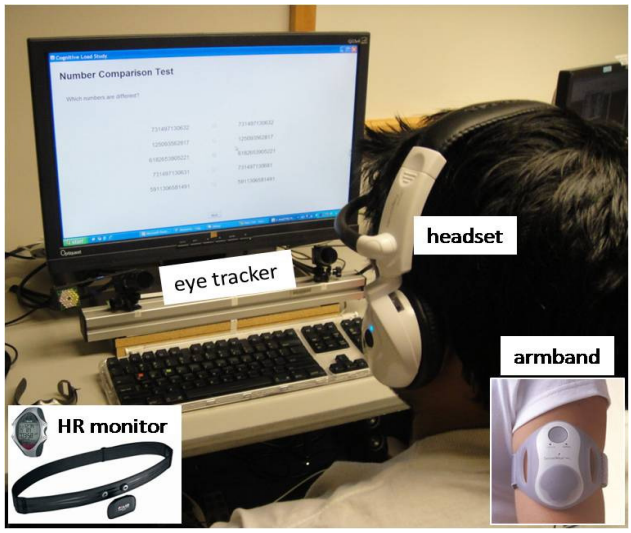

Figure 1. Experiment setup and sensor devices

wear three sensor devices and execute all the tasks in front of two cameras for gaze tracking installed at the bottom of the screen (screen size: 47.2cm×29.5cm, resolution: 1280×768) (see Figure 1). The participants used a mouse and a keyboard to answer the screen-based ECT questions.

For each ECT, two sets of questions were shown to the participant in a random order. One of the sets contained questions of lower difficulty level (inducing a lesser degree of cognitive load) while the other was comprised of more difficult questions (inducing a greater degree of cognitive load). We piloted the two sets for each ECT with 10 individuals not participating in the actual study, and collected subjective ratings of difficulty for each question from these participants, to validate that the two sets had distinguishable differences in difficulty.

After each ECT question set, the participant was asked to give his/her subjective rating of the difficulty of the task, for a total of 12 ratings (6 ECT types × 2 question sets with low/high difficulty levels). As a final step, a questionnaire was given to collect each participant's demographics, to verify that they did not partake in any activities prior to participating in the experiment that could impact their results (*e.g.*, smoking, drinking coffee or other caffeinated drinks, performing any strenuous exercise), and to get a self-report on any impairments and their mental and physical wellbeing. After the questionnaire, subjects were compensated $15 US for their time.

Testbed

A Java-based application was implemented for presenting the six ECTs to subjects. We counterbalanced both the order of the ECT question types and the difficulty of the question sets for each type using a Balanced Latin Square design [5]. For each question set, the subject was given 3 minutes to review the question slides and answer the questions. If this set time was exceeded, the subject was automatically directed to a task difficulty rating slide and the test continued with the next set of questions. Before

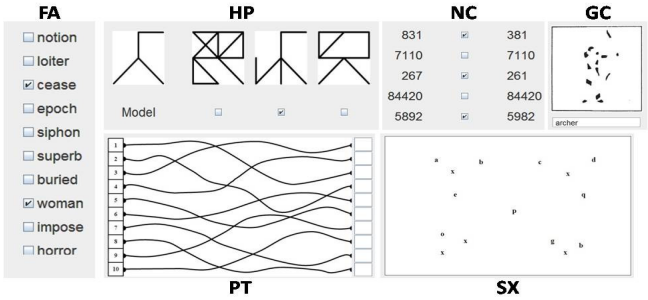

Figure 2. Six elementary cognitive tasks (ECTs)

each question set, the subject was asked to close his/her eyes for a brief period of mental relaxation. The application logs the subjects' answers and ratings along with a time stamp and current question set information (type and difficulty level), so that the performance (task completion time and number of correct answers) can be analyzed.

Six ECTs

As stated earlier, we selected six ECTs that mapped onto the 3 contextual factors (speed of closure, flexibility of closure and perceptual speed) identified earlier. The ECT contents and scoring methods used, originated from conventional ECTs based in psychology and cognitive science [6, 7, 8, 40] and were adapted to allow manipulation for task difficulty. We now describe the ECTs we presented to our subjects.

ECT1 - GC (Gestalt Completion) test

This test measures the 'speed of closure (CS)' factor [51]. The subject was asked to look at an incomplete line drawing and try to identify it [6, 7] (see Figure 2, GC). For each level of difficulty, 5 unique images were presented, with the complexity of the images higher in the high level of difficulty than in the low level.

ECT2 - HP (Hidden Pattern) test

This test measures the 'flexibility of closure (CF)' factor. Each subject was shown a model image, in the form of a line drawing, and a row of comparison images [6, 7] (see Figure 2, HP). The task was to see whether the model image was hidden in the composition of other comparison images. Task difficulty is increased by adding more distractive branches to the images (*i.e.*, making the images more complex). For each level of difficulty, five model images, each with eight comparison images, were presented.

ECT3 - FA (Finding A's) test

This test (along with the next three) measures the perceptual speed (PS) of a participant, in finding the letter 'a' in presented words [6, adapted from Thurstone's Letter A]. In this test, the participant was asked to find five words containing the letter 'a' on a page full of words (see Figure 2, FA). The length of the words was used as the criterion of difficulty where the two sets of questions contained words of length 3-5 and 7-9 letters, respectively. For each level of difficulty, two questions with 40 words in each to review were presented.

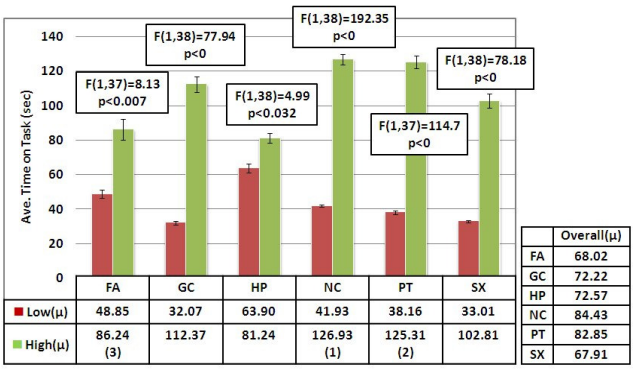

Figure 3. Average Time-on-Task (sec) *vs.* Task Difficulty (*Low/High*). The number of participants who did not finish a task within the time limit is in parentheses.

ECT4 - NC (Number Comparison) test

This is a test to find out how quickly the participant can compare two numbers and decide whether or not they are the same [6, 7] (see Figure 2, NC). Difficulty was manipulated by increasing the number of digits in each number and the number of digits that participants have to compare to identify the first differing digit (with an assumption that most people read numbers from left to right). For each level of difficulty, four questions with five pairs of numbers each were presented to the participant to review.

ECT5 - PT (Pursuit) test

This test measures how well participants can visually track irregularly curved overlapping lines from numbers on the left side of a rectangle to letters on the opposite side [40]. The participant is asked to trace each line from its beginning to its end with only his/her eyes (see Figure 2, PT). Task difficulty was controlled by manipulating the number of times the lines crossed each other, the length of the lines and whether backward tracking was required (*i.e.*, necessary for participants to gaze backward in the direction toward the starting point). For each level of difficulty, one question with ten curves was presented.

ECT6 - SX (Scattered X's) test

The goal of this test is to find the letter 'x' on screens containing random letters [40] (see Figure 2, SX). In this test, the overall number of letters, the proximity between the letters (how crowded) and the existence of similarly-shaped or rotated letters on the page are all used as criteria of difficulty. For each level of difficulty, participants were given 4 screens of letters to review.

Validation of task difficulty

To ensure that our manipulation of high and low cognitive load worked, and that our study design was valid, we performed a series of validation steps. For each of our six ECTs, we wanted to verify that the two sets of questions presented actually had two distinguishable levels of difficulty, high and low. In addition to our pilot test mentioned earlier, we also validated the task difficulty using participants' performance on the tasks and their subjective ratings of task workload.

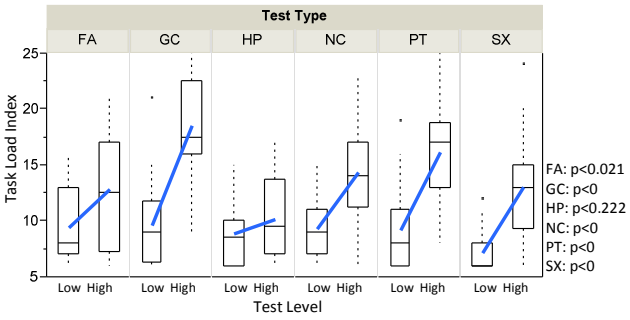

Figure 4. Average Task Load Index *vs.* Test Level (*Low/High*)

Task Performance (Time-on-Task)

We use participant performance to determine the difference in how much participants struggled to solve each set of questions. Most of the question sets were designed to keep participants engaged in solving the questions, rather than immediately giving up. As there was greater variation in 'time-on-task', the time a participant needed to complete a set of questions, than in accuracy, we focus our validation on 'time-on-task'. (Note, however, that the low variability in accuracy means that a typical task-performance accuracy assessment would not have been useful for assessing cognitive load in our experimental setting.)

Rating (Task Load Index)

After each set of questions, participants completed subjective workload assessments. For this, we used the NASA-TLX (Task Load Index) [17], in which the participant provided subjective ratings along six subscales: Mental Demands, Physical Demands, Temporal Demands, Own Performance, Effort Level and Frustration Level. Each subscale was a five-point Likert scale, ranging from Very Low (1) to Very High (5). We added the subjective ratings of the six factors to create an integrated load index.

Validation Results

We tested both time-on-task and participants' ratings of the question types to validate our manipulation of task difficulty in each ECT. Our results demonstrate that our manipulation is valid.

- Task performance

Figure 3 compares the average time-on-task for both levels of difficulty for each of the ECTs. It illustrates that there was a significant difference in the time-on-task between the low and high difficulty question sets for each ECT ($p<0.05$). On average, the participants spent 74.7 seconds answering each set of the questions; taking an average of 43.0 seconds for the easier questions, and an average of 105.8 seconds for the harder questions.

- Subjective rating

Figure 4 compares the average subjective rating for both levels of difficulty for each of the ECTs. There was a significant difference in the participant rating of the low and high difficulty question sets ($p<0.05$), for each of the ECTs except for the HP test. Although participants took

significantly different amounts of time to complete the low and high difficulty question sets for HP ($p<0.05$), their subjective ratings were indistinguishable.

Based on the combination of the pilot test results, the performance analysis and the subjective analysis, we conclude that our manipulation of low and high difficulty question sets was successful. Next, we describe the sensing devices we used to measure psycho-physiological signals while the various ECT stimuli were presented.

Psycho-Physiological sensors

In this study, we used four sensor devices – a contactless eye tracker, BodyMedia armband, wireless EEG headset, and a wireless heart rate monitor – to measure the psycho-physiological signals from our participants during task execution. Three computers (main tester, eye tracking system, headset reader) were used to collect data and had their clocks synchronized to allow for data integration.

Contactless eye tracker

Earlier work has shown the value of tracking eye movements and changes in pupil size as measures of cognitive load [3, 21, 22, 52]. We used a SmartEye 5.5.2 eye tracking system (http://www.smarteye.se) to detect and record the pupillometry (change in pupil size) of participants. The system is comprised of two cameras (Sony XC-HR50 with 12 mm lenses) and two Infrared (IR) flashes. The eye tracking system was calibrated for each participant, through a standard eye profiling task.

ECG-enabled armband

The BodyMedia SenseWear Pro3 armband (http://www.bodymedia.com) was used to collect a number of psycho-physiological responses that previous work showed to be valuable for measuring cognitive load, including electrocardiograms (ECG), and galvanic skin response (GSR) [21, 23, 43]. The armband was placed on the participants' left arm and two cables were plugged into the two conductive electrodes for measuring ECG, which were adhered above the clavicle and to the top-center of the back of the left arm (triceps). The device was used to collect galvanic skin response (GSR), heat flux (rate of heat transfer) and median absolute deviation (MAD – measure of variability) of the ECG.

Wireless EEG headset

As earlier work showed a correlation between electroencephalogram (EEG) or brainwave signals and cognitive load [23, 52], we used a NeuroSky mindset kit (http://www.neurosky.com) to extract, filter and amplify EEG signals and convert this information into two mental state outputs (attention and meditation). The brainwave signals provided by the headset are the raw EEG signal and the band powers: delta (1-3 Hz), theta (4-7 Hz), low alpha (8-9 Hz), high alpha (10-12 Hz), low beta (13-17 Hz), high beta (18-30 Hz), low gamma (31-40 Hz) and high gamma (41-50 Hz). The participants were asked to adjust ear cup sensors (ground and reference points) and a forehead-sensor arm (the primary electrode) to make skin contact with their left ear and forehead, respectively. A Bluetooth-based data logger was used to collect the signals and to verify signal strength and connectivity.

Wireless HR monitor

Finally, HR and HRV were shown to have value in assessing cognitive load [10, 29, 52], so we used a Polar RS800CX HR monitor (http://www.polar.fi/en) to collect interbeat interval (IBI) information with an accuracy of 1 ms. The device is comprised of a wireless transmitter attached to an elastic strap worn around the chest of the participant and a wrist worn training computer that stores the collected data.

Data analysis

We now discuss our approach for analyzing the psycho-physiological data for creating models of cognitive load.

Data

We recorded six psycho-physiological signals with the four devices. These were the interbeat interval signal measured with the HR monitor, galvanic skin response mean (32 Hz), heat flux mean (1 Hz) and ECG MAD information (32 Hz) measured with the armband, pupil diameter (60 Hz) measured by the eye tracker and EEG (128 Hz) measured with the headset. In addition, the headset gave eight power values (1 Hz) and two mental state outputs (1 Hz) derived from the raw signal. Examples of each of the signals are shown in Figure 5.

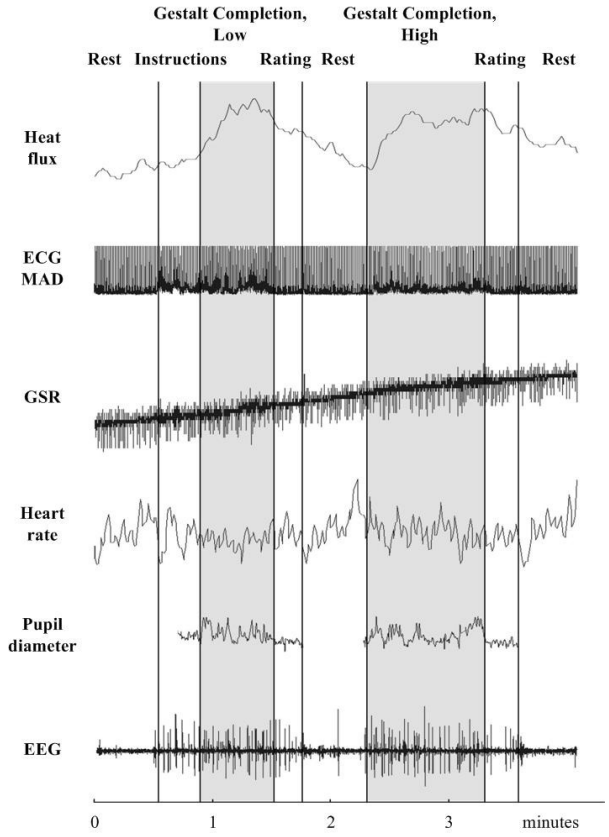

Figure 5. Example psycho-physiological signals collected during the Gestalt Completion test (low and high difficulty).

Because of technical difficulties with the measurement devices and anomalies in the HR data from participants 1 and 9, the EEG measurements from participant 8, and the GSR, heat flux and ECG MAD measurements from participant 20, these measurements were not used in the analysis. Additionally because of an error in the implementation of the Java-based test-bed, the question sets of lower difficulty for two of the subjects (11 and 12) in the PT and FA tests, respectively, were not presented to the subject. These question types were excluded from the analysis of the data from those participants.

Preprocessing

Before analysis, the heart rate IBI data was preprocessed by removing outliers falling outside the range of 35-155 bpm (387-1714 ms). The GSR values were observed to have an increasing trend at the beginning of each measurement caused by the properties of the measurement device. This trend was removed and the lowest and highest 0.1 percent of values from each participant were excluded as outliers.

Features

The level of cognitive load (low *vs.* high) was modeled using features derived from non-overlapping segments of psycho-physiological sensor data corresponding to the different questions in the ECT tests. Because the question sets in the Pursuit test were comprised of only one question each, the data corresponding to these segments was divided into two parts to increase the number of samples available for the modeling. Altogether *51* statistical features were calculated from the psycho-physiological signals measured with the four devices.

The mean, variance and median of pupil diameter, GSR, heat flux, ECG MAD, 8 EEG power values and two mental state outputs (attention and meditation) were calculated. Spectral power was also calculated from the raw EEG signal on five bands (delta 0-4 Hz, theta 4-7 Hz, alpha 8-12 Hz beta 12-30 Hz and gamma over 13 Hz) to compare to the values calculated by the EEG headset. Average powers for each of these were used as features. Two HRV features (standard deviation of IBIs (SDNN) and the root mean square of the difference of successive IBIs (RMSS)) and the mean and variance of HR were derived from the HR data.

Modeling

We then evaluated the performance of each of the features in assessing cognitive load. Because of individual differences in the levels of psycho-physiological responses to cognitive load, each participant was modeled individually. For each question type, the data from the separate questions were classified into one of two classes representing the two difficulty levels. Classification was performed based on one feature alone, using a Naïve Bayes classifier. We used a leave-one-out validation approach between the questions in each question type to calculate the average classification accuracy for the question type. Data from all but one of the questions was used to train the classifier and the data from the remaining question was used to evaluate the classification accuracy. This was repeated for all the questions in turn and the accuracy for the question type was defined as the average of these accuracies (*i.e.*, if a question type presented 5 questions to the user, we averaged over the 5 leave-one-out results). Because the difficulty levels in two different question types were unlikely to correspond to each other (*e.g.*, high difficulty questions for the Finding A's and the Pursuit tasks would not necessarily induce the same amount of cognitive load in a participant), only data from the same question type was used in the classification. The overall classification accuracy of the feature was then calculated as the average accuracy over the 6 question types. This was performed for each participant and each feature in turn.

RESULTS

The best feature and the corresponding classification accuracy for each participant are shown in Table 1. The results show that for each participant, a feature that discriminates the two classes with a high accuracy was found. Most of the best features were calculated from either the heat flux measurement or the ECG MAD signal.

Table 1 also presents the classification results for the three groups of tests targeting the 'speed of closure' (SC), 'flexibility of closure' (FC) and 'perceptual speed' (PS) factors contributing to cognitive load. On average, we

Sb#	Avg%	CS%	CF%	PS%	Best feature
1	82.5	100.0	70.0	81.3	median: heat flux
2	86.7	90.0	80.0	87.5	median: ECG MAD
3	86.7	100.0	70.0	87.5	mean: heat flux
4	74.0	70.0	50.0	83.3	median: ECG MAD
5	81.7	80.0	60.0	87.5	median: EEG power low beta
6	76.3	60.0	60.0	84.4	mean: EEG attention
7	83.3	100.0	100.0	75.0	median: heat flux
8	80.4	70.0	100.0	78.1	median: heat flux
9	86.3	90.0	90.0	84.4	median: EEG power high beta
10	87.0	60.0	100.0	91.7	median: heat flux
11	92.5	100.0	100.0	87.5	median: ECG MAD
12	75.5	90.0	50.0	79.2	variance: GSR
13	78.3	90.0	30.0	87.5	mean: ECG MAD
14	80.8	60.0	100.0	81.3	median: pupil diameter
15	82.5	100.0	70.0	81.3	median: ECG MAD
16	81.3	50.0	100.0	84.4	median: ECG MAD
17	88.3	80.0	100.0	87.5	variance: ECG MAD
18	89.2	80.0	80.0	93.8	mean: heat flux
19	94.0	100.0	70.0	100.0	mean: heat flux
20	76.3	50.0	70.0	84.4	variance: EEG power theta

Table 1. The best feature for each participant and the corresponding classification accuracy: overall, speed of closure (CS), flexibility of closure (CF), perceptual speed (PS).

Sensor	Feature	Avg%	CS%	CF%	PS%
Heat flux	median	**76.1**	73.7	67.4	76.6
ECG	median: MAD	**71.4**	80.5	74.7	68.3
EEG	median: attention	60.2	67.4	61.6	54.7
HR	mean	58.7	67.8	57.2	57.0
Pupillometry	median	57.4	69.6	57.0	53.8
GSR	variance	53.7	58.4	50.0	50.9

Table 2. Average classification results of the best features from each sensor stream over all participants: overall, speed of closure (CS), flexibility of closure (CF), perceptual speed (PS).

Features	Avg%	CS%	CF%	PS%
median: heat flux + median: ECG MAD	81.1	83.7	83.1	81.0

Table 3. Classification accuracy of the two best features from the BodyMedia SenseWear Pro3 armband combined. The result is calculated as an average over all the participants.

succeeded in modeling cognitive load related to each of the factors. The accuracies for the CF factor are somewhat inferior to the results for the other two factors and also the variation in these accuracies is larger. However, this follows logically from the fact that the two levels of difficulty in the Hidden Pattern test targeting this factor did not differ significantly in the analysis of the subjective ratings.

In Table 2 the average classification results over all 20 participants are presented, using models created with the best feature from each sensor stream. Here again, the features that perform the best are based on either the heat flux measurement or the ECG MAD signal. The classification performance of all the other features is clearly inferior.

The two best features (median of heat flux and median of ECG MAD) were then used together to classify the levels of cognitive load. The average classification accuracy across participants is shown in Table 3. The result (81.1%) is higher than the accuracy of using any single feature alone. This combination of features performed equally well in each of the test categories targeting different factors of cognitive load.

The above accuracy results come from individual models created for each participant. We also attempted to find a single model that would have been able to accurately discriminate different levels of cognitive load across participants. However, due to individual differences between participants, we have not yet been able to do so.

DISCUSSION

In this study, our goal was to establish a method for evaluating the cognitive load of a participant during the execution of elementary cognitive tasks. In particular, we examined tasks that focus on visual perception and cognitive speed since these factors are relevant to many ubicomp contexts. We targeted the three major discriminable first-order factors in these areas: 'speed of closure' (CS), 'flexibility of closure' (CF) and 'perceptual speed' (PS). These factors were chosen because the cognitive processes they are associated with also appear in situations of divided attention.

We evaluated the usefulness of a wide range of psycho-physiological signals in assessing cognitive load in six different elementary cognitive tasks. Four of the tests were chosen to address the PS factor while each of the other two tests targeted one of the other factors, SC and FC. Our results demonstrate that, for each participant, a psycho-physiological signal was found that can be used to accurately discriminate (74%) tasks of low and high level of difficulty, and following that, levels of low and high cognitive load in participants. Across all the participants, the heat flux and ECG MAD measurements were shown to be the best indicators of differences in cognitive load. When combined, a classification accuracy of 81.1% was achieved.

In addition to the overall accuracy in evaluating cognitive load in the six tests, we also examined the validity of our results with respect to each of the three factors (SC, FC, and PS) contributing to cognitive load. We demonstrated our ability to model each of these factors with equally high accuracy ($\geq$ 81.0%). Therefore, our results are potentially very generalizable to different tasks inducing cognitive load. As well, the average length of a data segment that we used for classification was only 23.7 seconds, which means that a real world implementation would react in close to real-time, to changes in a user's cognitive load.

Our results show great potential in being able to obtain a real-time, objective and generalizable measure of cognitive load. The two psycho-physiological signals identified as most valuable in assessing cognitive load, heat flux and ECG MAD, can both be measured with an armband that is very easy to use and wear in everyday settings, can be hidden under clothes, and is non-invasive. Therefore, integration of information about the user's cognitive load within other ubicomp applications and systems is certainly feasible.

While the same psycho-physiological signals produced the best results, for the most part, across all the participants, we were not able to build a single model based on these signals that had high classification accuracy. Instead, individual models were created for each user, based on the same features. At least for now, this means that when our cognitive load assessment is integrated into ubicomp systems, there must be a short training period to create a personalized cognitive load classifier for each user.

Our findings differ notably from the previous studies in using physiological signals in modeling cognitive load discussed in the related work section. In the previous studies, all the measurement signals we used had been found to contain information relevant to assessing cognitive load. In our study, however, only the heat flux and ECG

measurements produced accurate results. One reason might lie in the differences in the types of tasks the participants were asked to perform while the signals were measured. Different or more difficult tasks that call on different cognitive capabilities might induce cognitive load that manifests itself in different ways. It is also possible that the GSR sensor, located on the participant's arm, was not sensitive enough. In other studies, more accurate finger sensors have been used. The poor performance of the pupil size measurement might be caused by the eye tracker occasionally not being able to track a participant's gaze, which caused some data loss.

Our tasks, which focused on different levels of difficulty in elementary cognitive tasks did not result in useful changes in the majority of the signal streams used. However, our physiological measurements confirmed that ECTs can be used in interruption (or divided attention or dual-task) studies as reliable stimuli that induce different amounts of cognitive load.

CONCLUSION

Cognitive demands and limitations ebb and flow in situations of divided attention; much more needs to be understood about the limits of human attention and the best ways to provide information to support it. As a first step to remedy this situation, we collected data from multiple sensors and compared their ability to assess cognitive load. We focused on visual perception and cognitive speed-focused tasks that leverage cognitive abilities common in ubicomp applications. We targeted three major factors in these areas: speed of closure, flexibility of closure and perceptual speed. Data collected from multiple sensors showed that across all the participants, the median heat flux and ECG MAD measurements were the most accurate at distinguishing between low and high levels of cognitive load, providing a classification accuracy of over 80% when used together. In achieving this, we provide a real-time, objective, and generalizable method for assessing cognitive load in cognitive tasks commonly found in ubicomp systems and situations of divided attention. These results can therefore be applied to both the development and evaluation of ubicomp systems.

In continuing this work, we have a number of goals. First, we will collect more data, which will allow us to evaluate models with more features, and combine all features to increase classification accuracy. Second, we would like to identify a way to normalize the individual differences between participants, perhaps through the use of a baseline task, which would allow us to create a single model of cognitive load for all participants, making our contribution even more generalizable. Third, we will integrate our results into a real ubicomp system, such as a mobile location-based service or an ambient display, and evaluate the ability of our approach to characterize cognitive load in real world settings.

ACKNOWLEDGEMENTS

This work was generously supported by General Motors, the NSF and Quality of Life Engineering Research Center (EEEC-540865), the Fulbright Foreign Student Program, the Graduate School in Electronics, Telecommunications and Automation (GETA), the Emil Aaltonen Foundation and the Seppo Säynäjäkangas Foundation. We also thank Kimberley Nederlof who helped in collecting the data and all of our study participants.

REFERENCES

1. Ackerman, P.L. and Cianciolo, A.T. (2000) Cognitive, perceptual-speed, and psychomotor determinants of individual differences during skill acquisition. *Journal of Experimental Psychology: Applied*, 6(4):259-290.
2. Bailey, B.P., Konstan, J.A. and Carlis, J.V. (2001) The Effects of Interruptions on Task Performance, Annoyance, and Anxiety in the User Interface. In *Proc. Interact 2001*, 593-601.
3. Beatty, J. and Lucero-Wagoner, B. (2000) The pupillary system. In Cacioppo, J.T., Tassinary, L.G. and. Berntson, G.G. (Eds.), *Handbook of psychophysiology*, Cambridge University Press, Hillsdale, NJ, USA, 142-162.
4. Caroll, J.B. (1993) *Human Cognitive Abilities: A Survey of Factor-Analytic Studies*. Cambridge University Press, Cambridge, UK.
5. Edwards, A.L. (1962) *Experimental Design in Psychological Research*. Holt, Rinehart and Winston, New York, NY, USA.
6. Ekstrom, R.B., French, J.W., Harman, H.H. and Dermen, D. (1976) *Manual for KIT of factor-referenced cognitive tests*. Educational Testing Service, Princeton, NJ, USA.
7. Ekstrom, R.B., French, J.W. and Harman, H.H. (1979) Cognitive factors: Their identification and replication. *Multivariate Behavioral Research Monographs*, 79(2): 3-84.
8. Eliot, J. and Smith, I.M. (Eds). (1983) *An International directory of spatial tests.* NFER Nelson, Windsor, England.
9. Field, G. (1987) Experimentus Interruptus. *ACM SIGCHI Bulletin* 19(2): 42-46.
10. Fredericks, T.K., Choi, S.D., Hart J., Butt, S.E. and Mital, A. (2005) An investigation of myocardial aerobic capacity as a measure of both physical and cognitive workloads. *International. Journal of Industrial Ergonomics,* 35(12): 1097-1107.
11. French, J.W. (1951) The description of aptitude and achievement tests in terms of rotated factors. *Psychometric Monograph*, 5, University of Chicago Press, Chicago, IL, USA.
12. French, J.W., Ekstrom, R.B. and Price, L.A. (1963) *Manual and kit of reference tests for cognitive factors.* Educational Testing Service, Princeton, NJ, USA.
13. Golledge, R.G. (Ed.). (1999) *Wayfinding behavior: Cognitive mapping and other spatial processes*. Johns Hopkins University Press, Baltimore, MD, USA.
14. Grandhi, S. and Jones, Q. (2010) Technology-mediated interruption management. *International Journal of Human-Computer Studies*, 68(5): 288-306.
15. Gray, W.D., John, B.E. and Atwood, M.E. (1992) The Precis of Project Ernestine or an overview of a validation of GOMS. In *Proc. CHI 1992*, ACM Press, 307-312.
16. Hancock, P.A. and Chignell, M.H. (1986) Toward a Theory of Mental Work Load: Stress and Adaptability in Human-Machine Systems. In *Proc. IEEE SMC 1986*, 378-383.

17.Hart, S.G. and Staveland, L.E. (1988) Development of NASA-TLX (Task Load Index): results of empirical and theoretical research. In Hancock, P.A. and Meshkati, N. (Eds.), *Human Mental Workload*. Amsterdam, North-Holland, 139-183.

18.Healey, J.A. and Picard, R.W. (2005) Detecting stress during real-world driving tasks using physiological sensors. *IEEE Transactions of Intelligent Transportation Systems*, 6(2): 156-166.

19.Horvitz, E. and Apacible, J. (2003) Learning and Reasoning About Interruption. *Proc. ICMI 2003*, 20-27.

20.Hudson, S.E., Fogarty, J., Atkeson, C.G., Avrahami, D., Forlizzi, J., Kiesler, S., Lee, J.C. and Yang, J. (2003) Predicting Human Interruptibility with Sensors: A Wizard of Oz Feasibility Study. In *Proc. CHI 2003*, ACM Press, 257-264.

21.Ikehara, C.S. and Crosby, M.E. (2005) Assessing Cognitive Load with Physiological Sensors. *Proc. HICSS 2005*, 295-303.

22.Iqbal, S.T., Adamczyk, P.D., Zheng, X.S., and Bailey, B.P. (2005) Towards an index of opportunity: understanding changes in mental workload during task execution. In *Proc. CHI 2005*, ACM Press, 311-320.

23.Kilseop, R. and Rohae, M. (2005) Evaluation of mental workload with a combined measure based on physiological indices during a dual task of tracking and mental arithmetic. *International Journal of Industrial Ergonomics* 35(11): 991-1009.

24.Kreifeldt, J.G. and McCarthey, M.E. (1981) Interruption as a Test of the User-Computer Interface. In *Proc. Conference on Manual Control*, California Institute of Technology, JPL Publication, 81-95, 655-667.

25.Lohman, D.F. (1988) Spatial abilities as traits, processes, and knowledge. In Sternberg, R.J. (Ed.), *Advances in the psychology of human intelligence*, 40. Erlbaum, Hillsdale, NJ, USA, 181-248.

26.McFarlane, D.C. (1999) Coordinating the Interruption of People in Human-Computer Interaction. In *Proc. Interact 1999*, 295-303.

27.McGrew, K.S. (2009) CHC theory and the human cognitive abilities project: Standing on the shoulders of the giants of psychometric intelligence research. *Intelligence*, 37, 1-10.

28.Mital, A. and Govindaraju, M. (1999) Is it possible to have a single measure for all work?. *International Journal of Industrial Engineering Theory*, 6, 190-195.

29.Mulder, L.J.M. (1992) Measurement and analysis methods of heart rate and respiration for use in applied environments. *Biological Psychology* 34, 205-236.

30.Nasoz, F., Alvarez, K., Lisetti, L. and Finkelstein, N. (2004) Emotion recognition from physiological signals using wireless sensors for presence technologies. *Cognition, Technology and Work*, 6(1): 4-14.

31.Nettelbeck, T. and Lally, M. (1976) Inspection Time and Measured Intelligence. *British Journal of Psychology*, 67, 17-22.

32.Paas, F. and Merrienboer, J.V. (1993) The efficiency of instructional conditions: an approach to combine mental effort and performance measures. *Human Factors*, 35, 737-743.

33.Paas, F. and van Merriënboer, J.J.G. (1994) Instructional control of cognitive load in the training of complex cognitive tasks. *Educational Psychology Review*, 6, 51-71.

34.Paas, F., Tuovinen, J.E., Tabbers, H.K. and Van Gerven, P.W.M. (2003) Cognitive load measurement as a means to advance cognitive load theory. *Educational Psychologist,* 38(1): 63-71.

35.Petrill, S.A. and Deary, I.J. (2001) Inspection time and intelligence: Celebrating 25 years of research. *Intelligence*, 29(6): 441-442.

36.Ritter, F.E. and Avraamides, M.N. (2000) Steps towards including behavioural moderators in human performance models in synthetic environments. *Tech Report No. 2000-1*. Applied Cognitive Science Lab, Penn State University, PA, USA.

37.Ritter, F.E. and Young, R.M. (2001) Embodied models as simulated users: introduction to this special issue on using cognitive models to improve interface design. *International Journal of Human-Computer Studies*, 55(1): 1-14.

38.Salthouse, T.A. (1992) What do adult age differences in the Digit Symbol Substitution Test Reflect?. *Journal of Gerontology*, 47(3): 121-128.

39.Salthouse, T.A. (1996) The processing-speed theory of adult age differences in cognition. *Psychological Review*, 103(3): 403-428.

40.Salthouse, T.A. (2000) Aging and measures of processing speed. *Biological Psychology*, 54, 35-54.

41.Salvucci, D.D. (2001) Predicting the effects of in-car interface use on driver performance: An integrated model approach. *International Journal of Human-Computer Studies*, 55(1): 85-107.

42.Salvucci, D.D. (2006) Modeling Driver Behavior in a Cognitive Architecture. *Human Factors*, 48(2): 362-380.

43.Shi, Y., Ruiz, N., Taib, R., Choi, E. and Chen, F. (2007) Galvanic skin response (GSR) as an index of cognitive load. *Ext. Abstracts CHI 2007*, ACM Press, 2651-2656.

44.Simon, H.A. (1971) Designing organizations for an information rich world. In Greenberger, M. (Ed.), *Computers, communications, and the public interest*. Johns Hopkins University Press, Baltimore, MD, USA, 37-72.

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

46.Stramler, J.H. (1993) *The Dictionary for Human Factors / Ergonomics.* CRC Press, Inc., Boca Raton, FL, USA.

47.Street, R.F. (1931) A gestalt completion test: a study of a cross section of intellect. *Contributions to Education*, 481, Teacher's College, Columbia University, New York, NY, USA.

48.Treisman, A. and Gelade, G. (1980) A feature integration theory of attention. *Cognitive Psychology*, 12, 97-136.

49.Welford, A.T. (1986) Mental workload as a function of demand, capacity, strategy and skill. *Ergonomics*, 21, 151-176.

50.Wickens, C.D. (1991) Processing resources and attention. In *Multiple Task Performance*. D.L. Damos (Ed.), Taylor and Francis, Ltd., Bristol, UK, 3-34.

51.Wickens, C.D. and McCarley, J. (2008) *Applied attention theory*. Taylor and Francis, Boca Raton, FL, USA.

52.Wilson, G.F. (2002) An analysis of mental workload in pilots during flight using multiple psychophysiological measures. *International Journal of Aviation Psychology*, 12(1): 3-18.

53.Yin, B., Chen, F., Ruiz, N. and Ambikairajah, E. (2008) Speech-based cognitive load monitoring system. In *Proc. IEEE ICASSP 2008*, 2041-2044.