Task Load Estimation and Mediation Using Psychophysiological Measures

Rahul Rajan
ECE, Carnegie Mellon University
Pittsburgh, PA, USA
rahulraj@cmu.edu

Ted Selker
CITRIS, University of California, Berkeley
Berkeley, CA, USA
ted.selker@gmail.com

Ian Lane
LTI, Carnegie Mellon University
Pittsburgh, PA, USA
lane@cmu.edu

ABSTRACT
Human performance falls off predictably with excessive task difficulty. This paper reports on a search for a task load estimation metric. Of the five physiological signals analyzed from a multitasking study, only pupil dilation measures correlated well with real-time task load. The paper introduces a novel task load estimation model based on pupil dilation measures. We demonstrate its effectiveness in a multitasking driving scenario. Autonomous mediation of notifications using this model significantly improved user task performance compared to no mediation. The model showed promise even when used outside in a car. Results were achieved using low-cost cameras and open-source measurement tools lending to its potential to be used broadly.

ACM Classification Keywords
H.5.m. Information Interfaces and Presentation (e.g. HCI): Miscellaneous

Author Keywords
Cognitive Load; Divided Attention; Psychophysiological Measures; Pupil Dilation; Considerate Systems

INTRODUCTION
People use ubiquitous computing platforms like the mobile phone to access information and interact with others, even when they are socially or physically engaged. People, however, have finite mental resources and can only process a limited amount of information without degradation of task performance. Despite this being the case, there is an increasing trend towards computers being proactive and providing information to the user without being prompted. For cognitively challenging activities like driving, divided attention can have dire consequences. Thus, there is a need for systems to gauge the load on this mental resource in order to predict or preempt degradation in task performance, while interacting with a user.

IUI 2016, March 7–10, 2016, Sonoma, CA, USA.

http://dx.doi.org/10.1145/2856767.2856769

While progress has been made towards gauging this load, we are still a long way off from being able to measure it at a real-time fine-grained level. In the future, such capabilities might avert human mistakes in situations of divided attention. For instance, voice interaction might become the most efficient way for a user to interact with a system, when their manual and visual resources are already occupied. By using a rapid and fine-grained cognitive load measure, dialog or proactive agents would be able to track the ebbs and flows of the load being experienced by the user in real-time. This would allow it to preempt disfluencies and other irregularities in speech, as well as to time its responses and other actions, so as to prevent overloading the user. In the driving scenario, it has been shown that passengers adapt their conversation to the driving situation, which leaves the driver with more resources to perform the driving task when it gets difficult [5, 9]. Interactive agents should aim to emulate such considerate behaviors.

Cognitive load can be gauged by directly modelling the driver via psychophysiological measures, or by modelling driving context and its effect on the driver, or by jointly modelling both [18]. Compared to modelling the external driving context, less progress has been made in modelling the driver's internal state in order to identify when to interrupt them. One advantage of the internal state approach is the potential for these models to generalize to other domains. Modelling external context requires specific sensors and techniques to be considered for each domain separately. Furthermore, a physiological-based approach can be tuned for each user individually, as different users might experience external contexts differently. Recent advances in wearable technologies suggest that monitoring at least a few physiological signals in everyday life might become a feasible option.

In this paper, we evaluate several signals that might be used as part of a psychophysiological approach to gauging cognitive load. We wanted to capture the temporal aspects of divided attention — the transitions in load when addressing and recovering from interruptions, for example. At the same time, we wanted to facilitate data collection that was repeatable, with real-time performance measures that were responsive to task load. To meet these goals, we designed a multitasking scenario with a driving-like primary task, and an intermittent secondary task of attending and responding to notifications. As part of our initial explorations, we present results from two separate user studies. In the first set of results, pupil dilation measures

were used to build classification models that can detect which tasks the user is engaged in. We show how the performance of the model varies with changes in the modality and timing of notifications. We do this for each user, as well as across all users. To evaluate the feasibility of using such a model built on pupil dilation measures we conducted a second study. Here the classification model was used to autonomously mediate notifications to users in real-time. We demonstrate its effectiveness by analyzing user task performance with and without mediation. In the following sections, we provide background and discuss related work, before describing the two studies and their results in detail.

RELATED WORK

In cognitive psychology, there is a general consensus that people have limited and measurable cognitive capacities for performing mental tasks [26]. Furthermore, engaging in one mental task interferes with the ability to engage in other tasks, and can result in reduced performance on some or all of the tasks as a consequence [17]. To characterize the demand on these limited resources, psychologists have employed notions like cognitive load and mental workload, which gains definition through the experimental methods that are used to measure it [19].

Measuring Cognitive Load

Cognitive load can be assessed using data gathered from three empirical methods: subjective data using rating scales, performance data using primary and secondary task techniques, and psychophysiological data from sensors [29]. Self-ratings, being post-hoc and subjective in nature, tend to be inaccurate and impractical to use when automated and immediate assessment is required. Secondary task techniques are based on the assumption that performance on a secondary measure reflects the level of cognitive load imposed by a primary task. A secondary task can be as simple as detecting a visual or auditory signal, and can be measured in terms of reaction time, accuracy, and error rate. However, in contexts where the secondary task interferes with the primary task, physiological proxies that can measure gross reaction to task load are needed to assess cognitive load.

Psychophysiological Measures

Psychophysiological techniques are based on the assumption that changes in cognitive functioning cause physiological changes. An increase in cortical activity causes a brief, small autonomic nervous response, which is reflected in signals such as heart rate (HR) and heart rate variability (HRV) [11, 27, 36], electroencephalogram (EEG) [30, 36], electrocardiogram (ECG) [30], electrodermal activity (EDA) [14, 31], respiration [27], and heat flux [12], eye movements and blink interval [3, 14, 15, 36] and pupillary dilations. Our dataset includes most of these signals as well as additional signals that have been shown to be sensitive to affect like pulse transit time (PTT), facial electromyography (EMG) and skin temperature [22, 28].

In particular, brain activity as measured through event-related potentials using EEG, or as inferred from pupillary responses have received more attention recently because of their high sensitivity and low latency [1, 19, 24]. There has been very little work that correlates these measures with the other physiological measures, or demonstrates how to effectively align them. Furthermore, to the best of our knowledge this is the only work that has focused on tracking cognitive load that is rapidly and randomly changing, since we are interested in teasing out the dynamic nature of instantaneous cognitive load. Lastly, prior work has typically focused on cognitive load arising in single-task scenarios like document editing [15], and traffic control management [31]. In contrast, we employ an increasingly common multitasking scenario, aspects of which we briefly review below.

Multitasking Scenarios

In multitasking scenarios, the distribution of cognitive resources when engaged in two or more tasks is not very well understood. This makes it difficult to assess and predict workload that will be experienced by the user. Theories have been proposed to model how multiple tasks might compete for the same information processing resources [2, 35]. One widely used approach that has been shown to fit data from multitask studies is Wickens' multiple resource theory. This attempts to characterize the potential interference between multiple tasks in terms of dimensions of stages (perceptual and cognitive vs. spatial), sensory modalities (visual vs. auditory), codes (visual vs. spatial), and visual channels (focal vs. ambient) [35]. Performance will deteriorate when demand for one or more tasks along a particular dimension exceeds capacity.

In the case of driving and notification comprehension, both tasks compete for resources along the stages dimension. We would expect performance to deteriorate when there is an increased demand for the shared perceptual resources, i.e. when driving is hard and/or when the notification is difficult to comprehend. If the notification is visual, both tasks might also compete along the modality and visual channel dimensions. We would expect performance deterioration to be greater for visual notifications.

Driving and Language

Listening and responding to another person while driving a car has been widely studied, and has been shown to effect driving performance, particularly with remote conversants [21]. Passengers sitting next to a driver are able to adapt their conversation to the traffic situation, allowing the driver to focus on driving when it becomes difficult [5, 9]. These findings have motivated research towards building dialog systems that are situation-aware and interrupt themselves when required [20]. As mentioned earlier, the focus in most of this work is on monitoring the driving environment, and less on determining the cognitive load of the driver. Recently, there has been an interest in studying the effect that complex linguistic processing can have on driving using physiological measures of pupil dilation and skin conductance [8].

Mediation

Successful dual-task scenarios depend on the availability and requirements of cognitive resources for the secondary task given resource consumption by the primary task [34]. This

presents opportunities to increase people's ability to successfully handle interruptions, and prevent expensive errors. McFarlane's seminal work proposed four methods for coordinating interruptions, including immediate, negotiated, mediated and scheduled [25]. Mediation has been widely studied in the desktop computing domain [13, 16], but has not been adequately explored in post-desktop, mobile situations. We believe that ours is also the first study to present results on autonomous mediation using a psychophysiological measure.

STUDY 1: DYNAMIC TASK LOAD ESTIMATION

Our goals in this study were to:

- Collect a dataset[1] consisting of psychophysiological signals from users experiencing fluctuating task loads in a multitasking scenario
- Experiment with building models based on psychophysiological signals that can rapidly track cognitive load in a multitasking scenario
- Study the impact that the timing and modality of an intermittent secondary notification task has on the load experienced by the user

The appeal of testing in real world scenarios is transfer of results. But such experiments suffer from lack of repeatability and reproducibility because of the large number of variables involved. Repeating a route introduces landmarks that become familiar and changes performance. Events of interest happen unpredictably and infrequently. Establishing ground truth is also non-trivial and sometimes requires manual effort like scoring or annotating video recordings, etc. Experiments in the lab can circumvent a lot of these shortcomings by abstracting out the problem. A common practice in psychophysiology, however, is to have a control condition, followed by a test condition, with a rest period in between. As a consequence, the temporal aspects of psychophysiological signals as they fluctuate are lost.

In this study, we wanted the advantage of the lab setting, while still grappling with some of the complexity of real world data. In particular, we wanted to capture the temporal aspects of fluctuating psychophysiological signals while a participant is intermittently multitasking. For the primary task, we chose to use the established ConTRe (Continuous Tracking and Reaction) task [23], which provides a highly controlled yet unpredictable task load for the participant. This allows for consistent and replicable analysis. For the secondary task, the user was intermittently presented with notification prompts that they had to attend and respond to. The timing (mediated vs. non-mediated) and modality (audio vs. visual) of the notifications were treated as independent variables to investigate their impact on the load experienced by the user.

The study was setup so that the driving task would randomly switch between low and high workloads. This was done to simulate a typical driving scenario where drivers episodically experience high workload when they are entering/exiting highways, changing lanes, following navigation instructions, etc.

[1] will be hosted at http://rahulrajan.com/physio_data

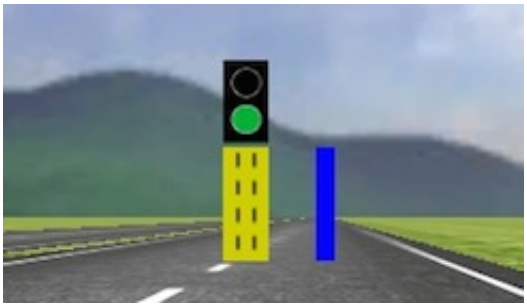

Figure 1. Screenshot of the ConTRe Task that displays the yellow reference cylinder with the traffic light on top, and the blue tracking cylinder.

In the mediated condition, the operator would send notifications only during the low driving workload condition, in contrast to their random delivery in the non-mediated condition. With regards to modality, audio notifications were delivered via speakers, and visual notifications through a heads-up display (HUD). The audio notifications were created using Apple's text-to-speech engine on OS X Yosemite (Speaking voice: Alex; Speaking rate: Normal). The HUD used was a Google Glass, which projects the screen at a working distance of 3.5 m, approximately 35° elevated from the primary position of the eye.

Design

The study was designed as a 2 (Audio/Visual *modes*) X 2 (Mediated/Non-mediated *conditions*) repeated-measures within-subjects study. To control for possible effects of order, the study was double counterbalanced for the *mode* and *condition* factors. Additionally, there was a baseline for both the low and high driving workload conditions.

Participants

20 people (10 male, 10 female) participated in our study, recruited through a call sent out to students selected randomly from a graduate school population. The mean age of the participants was 26.4 years, with a standard deviation of 2.7 years. Participants were rewarded with $40 gift cards for completing the study.

Tasks

We elaborate below on the design of the primary ConTRe task and the secondary notification task that make up the multitasking scenario.

Primary Task (T1): ConTRe

The ConTRe task comes as an add-on for OpenDS, an open-source driving simulator [23]. It is an abstracted and simplified task that is comprised of actions required for normal driving, i.e. operating the brake and acceleration pedals, as well as using the steering wheel. This focuses the user's task and simplifies the recording of tracking behavior. Fine-grained measures of performance on the primary task relative to the secondary task requests can be obtained, which is necessary for our investigation.

The vehicle in Figure 1 moves with a constant speed on a unidirectional straight road consisting of two lanes. The simulator shows two cylinders at a constant distance in front of

the vehicle: a yellow reference cylinder, and a blue tracking cylinder. The yellow reference cylinder moves autonomously and unpredictably. The lateral position of the blue tracking cylinder is controlled by the user through the use of the steering wheel. The cylinder moves left or right depending on the direction and angular velocity of the steering wheel, i.e the steering wheel controls the cylinder's lateral acceleration. The user's goal is to track the yellow reference cylinder, by overlapping it with the steering wheel-controlled blue cylinder, as closely as possible. Effectively, this corresponds to a task where the user has to follow a curvy road. For the low and high task load conditions, the lateral speed of the reference cylinder was set to values that were empirically determined to create low and high workloads for the user, respectively.

There is a traffic light with two colors placed on top of the yellow reference cylinder. The top light turns on red, whereas the bottom one turns on green. At any time, neither of the lights or only one is turned on. The red light requires that the user respond by depressing the brake pedal, while the green light corresponds to the accelerator pedal. This operates independently of the steering function. As soon as the user reacts to the light by depressing the correct pedal, the light turns off.

Secondary Task (T2): Notifications

The notification task was based on cognitive tests frequently used to measure working memory capacity [6]. Working memory has been purported to be involved in a wide range of complex cognitive behaviors, such as comprehension, reasoning, and problem solving as it is thought to reflect primarily domain-general, executive attention demands of the task [10]. In this work we do not aim to measure working memory per se, but instead want to measure the effect of engaging in a complex cognitive secondary task. Thus, we modify the cognitive tests for our purposes as described below.

In each condition, subjects were presented with a series of twenty items, which included ten equations and ten sentences taken from widely used complex span tasks [6] (see Table 1). The math equations and sentences are representative of the symbolic and verbal types of notifications, respectively, that users typically receive. Using standardized stimuli allows for consistency and replicability. Both types of notifications were randomly interspersed, so as to prevent the driver from getting into a rhythm of expecting either one. After the driver had read or listened to each item, they verbally indicated if the notification was *true* or *false*. Sentences are true when they are make sense, math equations are true when they are valid.

After each item, the subject was presented with an isolated letter, which represents something they had to remember from the notification. After two, three, or four items, the driving task was paused, and they were asked to recall the letters in sequence. This was done to mimic the behavior of drivers who usually attend to notifications while driving and respond to them, or perform other tasks that require more attention when stopped at a light, or while driving down a road with no noticeable gradient or curve at a constant speed [18].

Type	Notification
Math	$2/2+1=1$
Sentence	After yelling at the game, I knew I would have a tall voice

Table 1. Examples of the two types of notifications

Apparatus

The Robot Operating System (ROS Hydro) was used to synchronize signals from the different components of the experimental setup. This includes data from the simulator, physiological sensors, and the audio-visual feeds, all of which were being sampled at different frequencies, on separate machines. Each component publishes messages via ROS Nodes to the server, which synchronizes the data and writes it to disk. A Logitech camera, a mic, and audio mixer were used to capture audio-visual information. Participants controlled the simulator using a Logitech G27 Racing Wheel.

Physiological Sensors

Physiological signals were captured and recorded using the Biopac's BioNomadix monitoring devices for Electrocardiogram (ECG), Photoplethysmograph (PPG), Electromyogram (EMG), respiration, skin temperature, Electrodermal Activity (EDA), and Impedance Cardiography (ICG). Pupil dilation and eye gaze was captured using Pupil Pro hardware[2], which is a head-mounted mobile eye-tracking platform.

Since the hands of the participant were occupied for driving, we placed the PPG and EDA sensors on the participant's left toe & instep, respectively [32]. The facial EMG sensor was placed just above their left eyebrow to measure activation of the corrugator supercilii muscle, which is associated with frowning. Two skin temperature sensors were placed on the tip of the nose and on the left cheek. The ECG, impedance cardiography, and respiration sensors were placed in the default positions, i.e., on the chest and neck.

Methodology

Participants were guided through an informed consent process, followed by an overview of the study. They were aided through the process of having a number of sensors attached to their body for the purposes of recording their physiological responses. The participant was then seated in the simulator and given a demonstration of how visual notifications would be presented on the HUD, and audio notifications delivered over speakers.

The dataset collection was split in two parts: baseline and experimental. The baseline section records physiological measures for low and high driving workload separately. The experimental section records physiological measures for the multitasking scenario described in Section 3.3.

Baseline

Each participant was taken through a series of practice runs to get them comfortable with the primary driving task. When

[2] http://pupil-labs.com/pupil/

Physiological Measures		Performance Measures	
Raw	Derivative	ConTRe Task (T1)	Notification Task (T2)
Electrocardiogram (ECG)	Pulse Transit Time (PTT)	Steering Deviation	Sentence Response Time
Photoplethysmograph (PPG)	Inst. Heart Rate (IHR)	Acceleration Reaction Time	Sentence Accuracy
Impedance Cardiography(ICG)	SKT B − SKT A (SKT)	Acceleration Accuracy	Math Response Time
Respiration		Braking Reaction Time	Math Accuracy
Electrodermal Activity (EDA)		Braking Accuracy	Recall Accuracy
Skin Temp. Nose (SKT A)			
Skin Temp. Cheek (SKT B)			
Electromyography (EMG)			
Pupil Dilation			
Eye Gaze			

Table 2. Collection of measures available in the dataset.

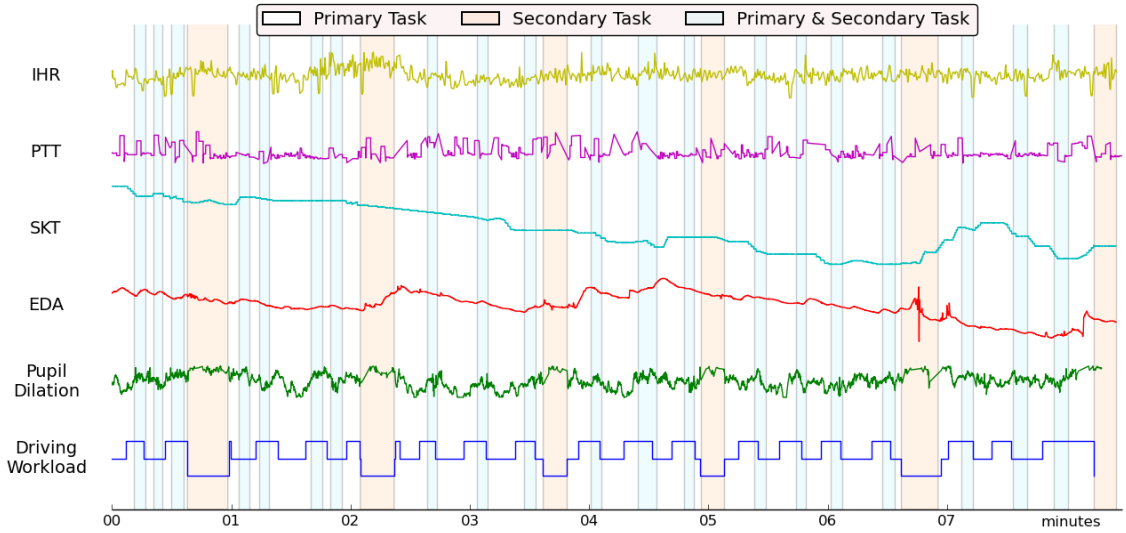

Figure 2. Example physiological measures collected during an audio non-mediated experimental condition. Driving workload is represented as a step function (1: High, 0: Low, -1: Pause). Colored regions delineate when the user was engaged in the primary driving task (T1; white regions), secondary notification task (T2; orange regions), or both (T1 & T2; blue regions).

done with the practice, the low benchmark was recorded using the low workload setting on the simulator. After a minute, they were asked to repeat a series of ten sentences that were read out to them, one-by-one, while they were still performing the primary ConTRe task. The same routine was performed to record the high benchmark using the high workload setting on the simulator.

Experimental

This was followed by another set of practice rounds that combined both the ConTRe task (with the randomly alternating workloads) and the notifications task. The notification task included a set of five items, three of which were equations, with the rest being sentences. This provided the participants with a sense of what to expect during the actual trials. The practice trials could be repeated if necessary. The participants then moved on to the experimental trials. Each participant had a total of four trials, one for each condition. The entire study lasted approximately 2 hours.

Data Processing

The dataset consists of a number of physiological and performance measures which are tabulated in Table 2. We recorded ten psychophysiological signals: EDA, EMG, skin temperatures (nose and cheek), four signals based on cardio-respiatory activity (ECG, PPG, ICG & respiration), and two based on eye activity (gaze and pupil dilation). Apart from the eye-based signals which were sampled at 30 Hz, the rest of the signals were sampled at 2000 Hz.

Three derivative signals were also calculated. Instantaneous heart rate (IHR) was obtained from the ECG signal using the BioSig library[3] which implements Berger's algorithm [4]. Pulse Transit Time (PTT) was obtained by calculating the difference in between the ECG R-wave peak time and the PPG peak time, which is the time it takes for the pulse pressure waveform to propagate through a length of the arterial tree. Difference in skin temperature (SKT) was also calculated by subtracting the temperature of the nose from that of the cheek.

[3] http://biosig.sourceforge.net/

The performance measures encompass both the primary driving task and the secondary notification task. Of interest are the reaction times and accuracies to the red and green light stimuli, and the steering deviation in tracking the reference cylinder. Also recorded are the performance measures for the secondary notification task as shown in Table 2.

Preprocessing & Labelling

In this exploration, 5 of the 13 psychophysiological signals collected were seen to be the easiest and most fruitful to analyze for dynamic task load modelling. They include IHR, PTT, SKT, EDA and pupil dilation (Figure 2). These signals were extracted from the collected data and down-sampled to 40 Hz (except for pupil dilation which remains at its original sampling rate of 30 Hz). Each signal was plotted, and thresholds were determined to filter out unlikely values (from movement artifacts, etc.). Data for each user was standardized (zero mean & unit variance), prior to which outliers that were more than three standard deviations from the average, were filtered out.

Two sets of labels are included in the dataset, a set each for the primary and secondary task. By syncing with the timestamps from both the task logs, we determined the precise primary and secondary task conditions that the participant was under for every physiological sample. The primary task labels denote if the participant is in the low, high, or paused driving workload condition (see Driving Workload in Figure 2). The logs from the secondary task allow us to determine the periods during which a participant was attending to a notification, i.e. blue regions in Figure 2. The orange regions signify the recall part of the secondary task, when the primary driving task was paused.

Feature Extraction

We derived a number of statistical features on the main signal ($x[n]$), the derivative signal ($x[n+1]-x[n]$), and the percentage change ($(x[n+1]-x[n])/x[n]*100$). These features include the mean, median, percentiles (10^{th}, 25^{th}, 75^{th}, 90^{th}), ranges (between min and max, 10^{th} and 90^{th} percentiles, and 25^{th} and 75^{th} percentiles), skewness, and standard deviation.

Features were extracted using a sliding window. To capture temporal properties, windows were overlapped, i.e. their step size was smaller than their length. Different window lengths and step sizes were considered. Specifically, the following pairs of window and step sizes (seconds) were analyzed: (7, 1), (5, 1), (3, 1) and (3, 0.25).

Modelling for Multitasking Scenario

Based on the insights from Wilkin's multiple resource theory described in the Related Work section, the loads that the primary and secondary tasks impose on the user are not mutually exclusive. Both tasks compete for resources along the stages dimension, and along the visual channel dimensions if the notifications are visual. Hence it would be more prudent for the classifier to make predictions on the load the user is under for both tasks separately and simultaneously, instead of attempting to make predictions on some notion of composite or absolute load. Thus, we can view this as a multi-label classification problem. In this formulation, each window is assigned two labels, where each label is drawn from the set of labels that corresponds to the primary and secondary tasks. Given the limited data to train the model on, we reduce our task to a multi-label binary classification problem, and ignore the specific states of the ConTRe (low/high workloads) and notification (attending/recall) tasks. Essentially, at this stage, we are simply trying to predict which tasks (T1 and/or T2) the user is engaged in by analyzing the psychophysiological data.

A sliding window is labelled as T1 if for the duration of a window the participant is only engaged in the primary ConTRe task. If the ConTRe task is paused, and the participant is engaged in recall, the window is labelled as T2. If the participant is attending to a notification while performing the ConTRe task, the window is labelled as both T1 and T2. For simplicity, the transitory unlabelled windows were discarded. The features of the remaining windows, and their corresponding multi-label assignments {T1,T2} were fed to a Random Forest classifier, which is an ensemble technique that learns a number of decision tree classifiers and aggregates their results. Models were built across all users, as well as for each user separately to account for individual differences in their psychophysiological response. To evaluate the classifier's performance, we used leave-one-user-out cross-validation for the population models, and 3-fold cross-validation for the individual user models.

The time it takes to comprehend a notification varies by participant. This creates variation in the number of driving and notification task labels generated per participant, which in turn results in a varying baseline accuracy for each user because of the class imbalance problem. Hence, instead of accuracy, we use the Area Under the Receiver Operating Characteristic Curve (ROC AUC) metric to evaluate the classifier, as it is insensitive to class imbalance. ROC curves show the trade-offs between higher sensitivity and higher specificity. Sensitivity refers to the correct detection of a condition or state when it is truly present. Specificity indicates the correct rejection of a state when it is truly not present. The area under the ROC curve is a measure of adequacy on both. Curves corresponding to random or chance classification of 50% would fall close to the diagonal, and result in an ROC AUC score of 0.5 regardless of class imbalance, while the most successful classifications would have an ROC AUC score close to 1.0.

Being a multi-label classification problem, the classifier outputs two probabilities simultaneously: one for the probability of the sample belonging to the primary task (T1), and another for the probability that the sample belongs to the secondary task (T2). We report the macro-averaged ROC AUC scores for the pair of labels, as a measure of how well the classifier is simultaneously able to predict both labels (T1$\vee$T2). We also report the ROC AUC score for each label, individually, to shed light on how accurately the classifier is able to identify each task.

Results

Of the five physiological signals analyzed, the pupil dilation measures were the only signal to yield results that were much better than random. For this reason, we only list and discuss results using the pupil dilation measures. For the four window and step size combinations considered, mean ROC AUC scores for the population and individual models are shown in Table 3.

Window, Step (s)	Population			Individual		
	T1∨T2	T1	T2	T1∨T2	T1	T2
7, 1	0.85	0.90	0.80	0.84	0.89	0.78
5, 1	0.84	0.89	0.78	0.83	0.88	0.78
3, 1	0.81	0.85	0.76	0.81	0.87	0.75
3, 0.25	0.80	0.86	0.75	0.80	0.86	0.74

Table 3. ROC AUC Scores for population and individual models using different window and step sizes

Condition	T1∨T2	T1	T2
Non-mediated			
Video	0.88	0.90	0.86
Audio	0.90	0.92	0.88
Overall	0.88	0.91	0.86
Mediated			
Video	0.82	0.89	0.76
Audio	0.81	0.88	0.74
Overall	0.81	0.89	0.74

Table 4. Population-based ROC AUC Scores under different timing and modality conditions.

A larger window size tends to provide better results, and this trend holds for both the individual and population models. The population scores are comparable to the average user scores, which tells us that the model based on pupil dilations is generalizable.

Table 3 also shows ROC AUC scores for predicting each label individually. The scores indicate that the models are better at identifying when the user is engaged in the primary driving task (T1) as compared to when the user is engaged in the secondary notification task (T2). This might be because of the differences in load induced by equation and sentence notifications, and also from the differences in the notifications being right or wrong. Our model doesn't account for these yet, but each can potentially be treated as a different class under a label in the multi-label framework.

We also compared how varying the independent variables of timing and modality impacted the ROC AUC scores. The results for these experiments are tabulated in Table 4. Only the analysis on the 7 second long windows are presented here as similar trends were observed for the other combinations. It is clear that mediating when notifications were sent had a larger effect than modality on the model's ability to identify the secondary task. To explain this, we must remember that in the mediated condition, the participant is sent notifications when they are in the low driving-workload state. The multiple resource theory predicts that the cognitive load on the user in this state (low driving-workload + notification) is similar to the cognitive load they experience when they are in the high driving-workload state. Thus, in the mediated condition windows where the user is driving with and without notification, i.e. windows labelled {T1} and {T1,T2}, look similar. In the non-mediated condition, this is not the case, as notifications are also delivered in the high driving-workload states. This allows the classifier to better identify the secondary task (T2) in the non-mediated condition.

STUDY 2: AUTONOMOUS MEDIATION

A classifier was built and tested for its ability to mediate notifications using real-time pupil dilatation measures. The classifier used was a modified version of the the model described above. It was trained on data from the non-mediated condition. During the pilot experiments, models using moving windows that were 5 seconds long with a step size of 1 second gave the most promising results.

The classifier gets standardized input from the pupil dilation data stream, and outputs a {T1,T2} classification every second. Since at this preliminary stage, we are only detecting what tasks the user is engaged in, we make the assumption that if the user is multitasking, i.e {T1,T2} = {1,1}, then the user is experiencing high load. If the user is only engaged in driving it outputs {1,0}.

The experimental setup is similar to the one used in the previous study with some modifications to the design and tasks. These changes are described in detail below.

Design

This study focused on audio notifications only. It was designed as a repeated measures within subject study with only one independent variable, i.e. non-mediated (control) vs. mediated (test) conditions. To control for possible effects of order the study was counterbalanced.

Participants

10 people (all male) participated in our study recruited through a call sent out to students selected randomly from a graduate school population.

Tasks

Since notifications are what the system needs to mediate, they could not be used as the secondary task (T2) that the classifier detects. We therefore increased task load in a different way by using a manual transmission-based gear changing task as the secondary task. The tasks were chosen so as to make it difficult for a user to perform perfectly on the primary and secondary tasks simultaneously.

Primary Task (T1): ConTRe

The primary task remains the same as in the first user study. The participant is engaged in an abstracted driving task, where they track a yellow cylinder with a steering wheel. The participant also has to simultaneously respond to red and green lights on the yellow cylinder by depressing the brake and accelerator pedals, respectively. The ConTRe task was set to alternate between periods of low and high workloads as described in the first study.

Secondary Task (T2): Gear Change

An LCD screen is placed in front of the simulator such that its contents are easily visible below the yellow and blue cylinders presented on the simulator screen. Numbers from 1–6 are

presented on the LCD screen, which correspond to the gears on the manual transmission gearbox which is included with the Logitech G27 Racing Wheel. The user was asked to shift to the right gear when the number changed on the screen. To create a high task load for the user, the gear number only changed when the ConTRe task was in its high load setting. The gear number was set to change every 1–3 seconds.

Mediated Task: Notifications

The notification task is a simplified version of the one used in the previous study. To create a continuous task scenario the pause and recall portion of the previous study was eliminated. In this study, notifications only consist of audio math and sentence prompts that the user responds to with a true or false.

Apparatus and Sensors

The apparatus used to conduct, synchronize and record the experiment was the same as before. Only audio notifications were presented to the user. As pupil dilation was the lone physiological measure of interest, the Pupil Pro headset was the only physiological sensor worn by the user.

Methodology

Participants were guided through an informed consent process, followed by an overview of the study. The participant was then seated in the simulator, and was asked to put on the Pupil Pro headset. They were instructed on how to perform the ConTRe task. Once comfortable with the task, the secondary gear changing task was introduced. After this the audio math and sentence notifications were demonstrated to the user. Once the user was familiar with all the tasks, a calibration step was performed to determine the parameters needed to standardize the data before classification. This step simply required the user to perform the ConTRe task in its low workload setting for 10 seconds. This was followed by two experimental trials. These included the test condition in which notifications were autonomously mediated based on task load, and the control condition in which notifications were randomly presented to the user regardless of task load.

Notifications were mediated by delaying them if they hadn't started playing. If they had started playing, and then the system detected that the task load on the user was high, the notification would cut off and repeat itself when the load on the user had reduced. A trigger-happy system that cuts off a notification every time a {1,1} is output by the classifier can be annoying to the user. For better user experience, notifications were mediated only when certain patterns of classification outputs were observed. Based on pilot studies, the protocol was set to delay or cut-off notifications anytime a pattern of either [{1,1}, {1,1}] or [{1,1}, {1,0}, {1,1}] classifications was output by the classifier. The system would then wait for a series of five {1,0} classifications before resuming delivery of notifications.

Measures

Quantitative performance data on primary, secondary, and mediated tasks were collected. From the primary ConTRe task, we collected the following: steering deviation, i.e. the difference in distance between the reference cylinder and the tracking cylinder (sampled at 570 Hz); reaction times to respond to the red and green lights, i.e. the amount of time from when the light went off to when the correct pedal was depressed; and the error rate of depressing the wrong pedal. These measures were automatically recorded by the simulator. An average of 23.8 and 13.7 acceleration stimulus points were presented to each user in the mediated and non-mediated conditions, respectively. Similarly, an average of 21.3 and 11.2 brake stimulus points were presented in the mediated and non-mediated conditions, respectively. Since notifications were being delayed in the mediated condition, these trials were longer that the non-mediated ones. For each user in each condition, the mean steering deviation, reaction times, and reaction errors were calculated.

Performance Measures	M	N	*p*
Primary Contre Task			
Steering Deviation (%)	22.0	23.1	.47
Accel Reaction Time (ms)	980	1014	.67
Brake Reaction Time (ms)	1117	1157	.47
Accel Response Error Rate	0.34	0.23	.07
Brake Response Error Rate	0.25	0.32	**.05**
Secondary Gear Task			
Attempts per stimulus	1.15	1.26	**.015**
Response Error Rate	0.22	0.31	**.05**
Mediated Notification Task			
Math Reaction Time (s)	2.02	2.30	.19
Sent. Reaction Time (s)	2.30	2.53	.32
Math Response Error Rate	0.08	0.08	.82
Sent. Response Error Rate	0.22	0.27	.33

Table 5. Mean performance measures of the primary, secondary and mediated tasks from both the mediated (M) and non-mediated (N) conditions, along with paired t-test two-tailed p-values.

From the secondary gear-changing task, the number of tries the user took to get to the right gear, and the number of times they didn't succeed in reaching the right gear were determined. The mean of these measures for each user in both conditions were then calculated. Per user, an average of 52.2 and 28.3 gear change requests were made in the mediated and non-mediated conditions, respectively.

For performance on the mediated notification task, the response times for the math and sentence prompts were computed. This is the time from when the notification was presented to the driver to when they respond to indicate true or false. The mean response times for notifications are then recorded for each user in every condition. The errors in the responses and the mean per user was also calculated for each condition. An average of 7.5 math and sentence prompts each, were presented to users in both conditions.

The outputs from the classifier, which occur every second, were also recorded for the mediated condition. These will be

System Stimulus	Sensitivity	Specificity	Accuracy
every *H* and *L*	40	72	61
H	90	19	56
HH or *HLH*	83	42	63
HHH	74	68	71
HHHH	58	82	70

Table 6. Evaluation of different types of stimulus to which a system could be designed to respond. The first two rows indicate the overly eager and cautious behaviors, respectively. The next three rows represent different patterns of classifier output.

analyzed to shed light on how the classifier's outputs could inform the system's mediation behavior.

Results

Below we report on results from the experiment. We look at mediation effects on each task by collectively analyzing their corresponding performance measures. Since this presents three sets of comparisons (one for each task), we use the Bonferroni adjusted alpha levels of .017 per test (.05/3) to control for Type I errors. To perform the analysis, we perform a multivariate ANOVA (MANOVA) on the performance measures from each task. As opposed to running multiple univariate F tests on each measure, MANOVA has the advantage of reducing the likelihood of a Type I error, and revealing differences not discovered by ANOVA tests [33]. We also analyze the classifier output with respect to task load, in order to shed light on how a system might mediate notifications more effectively.

Mediation Effects

The analysis of mediation effect on the primary ConTRe task using a repeated measures MANOVA showed no significant effect, $F(5,5)=1.44$, $p=.35$. The means for each of the five primary task measures in both conditions and the paired t-test two-tailed p-values are listed in Table 5.

For the secondary gear-changing task, a repeated measures MANOVA showed a significant effect, $F(2,8)=7.42$, $p=.015$. Further analysis of each of the dependent variables showed a significant difference in the mean number of tries the user took to get to the right gear between the mediated (M=1.15, SD=0.16) and non-mediated (M=1.26, SD=0.14) conditions, $t(9)$=-3.72 , p=.004. There was also a slightly significant difference in the failure rates between the mediated (M=0.22, SD=0.05) and non-mediated (M=0.31, SD=0.13) conditions, $t(9)$=-2.26 , p=.05. These are listed in Table 5.

A repeated measures MANOVA for the mediated notification task revealed no significant effect, $F(4,6)=0.98$, $p=.48$. The means for each of the four notification task measures in both conditions and the paired t-test two-tailed p-values are also listed in Table 5.

System Mediation Performance

Since the classifications are done on a sliding window, we can expect a lag from the onset of high task load to when the classifier output indicates so. Another reason for the delay in classifications might be that even though a high load is being imposed on the user, it might take a couple of seconds for them to experience it as such. To find the average delay, the cross-correlation between the alternating load conditions and time-shifted classification outputs was determined for multiple time shifts. Across users the average time-shift at which the cross-correlations were maximum was 4.9 s with a standard deviation of 1.44 s. For further analysis, this number was rounded up, and the classification outputs were time-shifted by 5 s for each user. For simplicity, we represent a {1,0} classifier output as *L* and a a {1,1} classifier output as *H*. The goal of this analysis is to get a sense of how well the classifier was detecting high load situations across users in the study, and how the system's mediation behavior could potentially be improved.

By being overly eager or overly cautious, a system can display two extremes in how it uses the classifier outputs to inform its mediation behavior. The eager system for example reacts immediately to every change in task load *L* and *H* being output by the classifier by playing or pausing a notification. We would expect the system to have high specificity, as it immediately changes its behavior based on classifier output. The cautious system also stops notifications immediately when high task load is sensed *H*, but continues to do so even if an *H* is followed by *L*s for a specified period of time. Thus it displays low specificity. Under the cautious behavior, a single *H* occurring during a high load section is considered as a true positive (correct classification). Conversely, a single *H* during a low load section is a false positive (incorrect classification). The system's sensitivity, specificity and accuracy are calculated by aggregating the true and false positives over the trials from all users. Results for the overly eager and cautious behaviors are shown in Table 6, along with a few intermediate behaviors which we describe next.

To trade-off between sensitivity and specificity, the system could be designed to mediate notifications only if it sees a particular pattern of classifier outputs. As described above, if a pattern occurs during a high load section it is marked as a true positive, and if it occurs during a low load section it is marked as a false positive. A few example patterns were evaluated, and their results are listed in Table 6. These include patterns such as [*H*,*H*] or [*H*,*L*,*H*] which reduces the sensitivity of the system to the classifier outputs, making it less cautious. This was also the pattern that was actually used by the system in the autonomous mediation study. We can reduce system sensitivity even further by having the system mediate notifications only when it sees [*H*,*H*,*H*] from the classifier. Table 6 also lists evaluation results when [*H*,*H*,*H*,*H*] is the pattern that the system responds to. In this way we get a sense of how the system's mediation behavior would have changed if the protocol was set to respond to different patterns of classifier outputs.

DISCUSSION

This paper presents a dataset of 13 psychophysiological signals to estimate cognitive load. These signals include ECG, PPG, ICG, Respiration, EDA, nose & cheek skin temperatures and the differences between them, EMG, pupil dilation, eye gaze, PTT and IHR (listed in Table 2). These were collected during

a dual-task user study that subjected a participant to a series of alternating low and high task loads. The study was designed in this way to mimic the fluctuating loads people experience while driving in the real world. The goal was to capture these fluctuations as reflected in the participants physiological responses.

The dual-task study consisted of a primary driving-like tracking and reaction task, and a secondary notification-based cognitive task. ConTRe was used as the primary task as it focuses on core driving skills and removes learnable contextual cues. This improves data and repeatability of experiment. Similarly, prompts frequently employed in complex span task experiments serve as the notifications presented to a participant. These represent the symbolic and verbal nature of notifications commonly received by people on their smartphones. In the study, the timing and modality of the notifications were treated as independent variables to understand their effects on cognitive load.

The dual-task scenario can be cast as a multi-label learning problem of the primary and secondary tasks. The approach succeeded at building classification models that distinguish whether the user is engaged in the primary task, the secondary task, or both. The model worked for each user and across all the participants. These models were built using statistical features derived from measurements of pupil dilation, which were fed to a random forest classifier. Our evaluations showed that the timing of the notifications has a larger effect on the load experienced by the user than the modality of notification delivery.

We evaluated the impact of this model in a separate real-time notification mediation study. The setup from the first study was altered to include a manual gear changing task instead of the notification itself. Pupil dilation data was streamed to the classifier which output a multi-label classification every second. In the test condition, the system would inhibit notifications if it believed that the user was simultaneously engaged in two tasks. In the control condition, notifications were delivered randomly. The effects of mediation were determined by analyzing the performance measures for each task. Mediation allowed users to reach the right gear (their secondary task) with less errors, and fewer number of attempts per gear change request. Notice that the gear-shifting task uses different perceptual and cognitive skills than the verbal notification task, which is what the model was trained on. Our model transferred and performed well on this mechanical performance stressing activity.

The system's mediation performance was evaluated using cross-correlation measures between the user task loads and time-shifted classifier outputs. We can interpret the results as there being an average lag of 4.9 s between the onset of high task load for the user, and when the system mediated notifications to them. System mediation behavior was also analyzed based on how it responds to different patterns of outputs from the classifier. The trade-off between the system's sensitivity and specificity was demonstrated for these different patterns.

Future work

There are a number of directions future work can take, and we briefly discuss a few here. First, our data analysis did not include moving windows over transitions from low to high workloads and back. We are optimistic that temporal models could be used to detect these transitions, reducing the lag in load detection. Second, improved measures of the load experienced by the users (ground truth) can be obtained by using a composite measure of the different task performance metrics. Reaction times can serve as more reliable proxies for cognitive load than externally imposed task load settings. Third, with more data we can make fine-grained estimations about user load within each task (for example, T1 = 0, 1, 2, 3, etc., based on difficulty of the primary driving task). Fourth, we should explore which physiological signals are more indicative of stress, and which are better suited for estimating cognitive load. Stress is likely to arise when failure at a task is coupled with feelings of lack of control, in situations where participants are evaluated by others [7]. We might hypothesize that stress is an affect. It ebbs and flows at a slower pace than cognitive load, which being reflective of the stages of mental processing, fluctuates more rapidly.

To show that pupil dilation measures can be robust we used an inexpensive off-the-shelf measuring technique. Prior work reports use of expensive eye-trackers with higher sampling rates for pupilometric measurements. Our study succeeded with a consumer webcam (Microsoft LifeCam HD-6000) that has a sampling rate of only 30 Hz. Even when tested outdoors in a car during the day, with no special attempt to control for ambient luminescence (apart from the initial calibration step), the system showed promising results in estimating task load through pupil dilation measures. More work could be done to refine the setup and understand the trade-offs between the fidelity of the equipment, environmental setup, and the robustness of results.

CONCLUSION

We show that technology can know when its appropriate to engage the user. This paper describes a system that can gauge the cognitive load using psychophysiological signals. We created a dataset of 13 physiological measures collected from participants in a multitasking study. They were asked to attend and respond to notifications while performing a primary driving task. We also collected performance measures on these tasks. Of the five most promising physiological measures analyzed, only pupil dilation reliably tracked task load in the near real-time five second range. We demonstrated the effectiveness of using pupil dilation measures for mediating task load in a second multitasking user study. Autonomous mediation of notifications significantly improved participant task performance. Cognitive load assessment is a rich area for exploration, and we hope to inspire other researchers to use our data set to further evaluate models of dynamic task load estimation. By enabling computers to interact appropriately and considerately, we can pave the way for future proactive computing scenarios.

REFERENCES

1. Pavlo Antonenko, Fred Paas, Roland Grabner, and Tamara van Gog. 2010. Using electroencephalography to

measure cognitive load. *Educational Psychology Review* 22, 4 (2010), 425–438.

2. Alan Baddeley. 2003. Working memory: looking back and looking forward. *Nature reviews neuroscience* 4, 10 (2003), 829–839.
3. Jackson Beatty and B Lucero-Wagoner. 2000. The pupillary system. *Handbook of psychophysiology* 2 (2000), 142–162. http://www.nrc-iol.org/cores/mialab/fijc/files/2003/090203_Pupillary_System_.pdf
4. Ronald D Berger, Solange Akselrod, David Gordon, and Richard J Cohen. 1986. An efficient algorithm for spectral analysis of heart rate variability. *Biomedical Engineering, IEEE Transactions on* 9 (1986), 900–904.
5. David Cohen, Akshay Chandrashekaran, Ian Lane, and Antoine Raux. 2014. The hri-cmu corpus of situated in-car interactions. *Proc. IWSDS* (2014), 201–212.
6. Andrew R a Conway, Michael J Kane, Michael F Bunting, D Zach Hambrick, Oliver Wilhelm, and Randall W Engle. 2005. Working memory span tasks: A methodological review and user's guide. *Psychonomic bulletin & review* 12, 5 (Oct. 2005), 769–786. DOI: http://dx.doi.org/10.3758/BF03196772
7. Dan Conway, Ian Dick, Zhidong Li, Yang Wang, and Fang Chen. 2013. The Effect of Stress on Cognitive Load Measurement. In *Human-Computer Interaction–INTERACT 2013*. Springer, 659–666.
8. V Demberg, E Kiagia, and A Sayeed. 2013. Language and cognitive load in a dual task environment. In *Proceedings of the 35th annual meeting of the cognitive science society (cogsci-13)*.
9. Frank A Drews, Monisha Pasupathi, and David L Strayer. 2008. Passenger and cell phone conversations in simulated driving. *Journal of Experimental Psychology: Applied* 14, 4 (2008), 392.
10. Randall W. Engle. 2002. Working memory capacity as executive attention. *Current Directions in Psychological Science* 11, 1 (Feb. 2002), 19–23. DOI: http://dx.doi.org/10.1111/1467-8721.00160
11. Tycho K. Fredericks, Sang D. Choi, Jason Hart, Steven E. Butt, and Anil Mital. 2005. An investigation of myocardial aerobic capacity as a measure of both physical and cognitive workloads. *International Journal of Industrial Ergonomics* 35, 12 (2005), 1097–1107. DOI: http://dx.doi.org/10.1016/j.ergon.2005.06.002
12. Eija Haapalainen, SeungJun Kim, Jodi F Forlizzi, and Anind K Dey. 2010. Psycho-physiological measures for assessing cognitive load. In *Proceedings of the 12th ACM international conference on Ubiquitous computing*. ACM, 301–310.
13. Eric Horvitz, Carl Kadie, Tim Paek, and David Hovel. 2003. Models of attention in computing and communication. *Commun. ACM* 46, 3 (March 2003), 52. DOI:http://dx.doi.org/10.1145/636772.636798
14. C.S. Ikehara and M.E. Crosby. 2005. Assessing Cognitive Load with Physiological Sensors. In *Proceedings of the 38th Annual Hawaii International Conference on System Sciences*. IEEE, 295a—-295a. DOI: http://dx.doi.org/10.1109/HICSS.2005.103
15. Shamsi T Iqbal, Piotr D Adamczyk, Xianjun Sam Zheng, and Brian P Bailey. 2005. Towards an index of opportunity: understanding changes in mental workload during task execution. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 311–320.
16. Shamsi T. Iqbal and Brian P. Bailey. 2010. Oasis. *ACM Transactions on Computer-Human Interaction* 17, 4 (Dec. 2010), 1–28. DOI: http://dx.doi.org/10.1145/1879831.1879833
17. Daniel Kahneman. 1973. *Attention and effort*. Citeseer.
18. SeungJun Kim, Jaemin Chun, and Anind K Dey. 2015. Sensors Know When to Interrupt You in the Car: Detecting Driver Interruptibility Through Monitoring of Peripheral Interactions. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 487–496.
19. Jeffrey Michael Klingner. 2010. *Measuring cognitive load during visual tasks by combining pupillometry and eye tracking*. Ph.D. Dissertation. Stanford University.
20. Spyros Kousidis, Casey Kennington, Timo Baumann, Hendrik Buschmeier, Stefan Kopp, and David Schlangen. 2014. A Multimodal In-Car Dialogue System That Tracks The Driver's Attention. In *Proceedings of the 16th International Conference on Multimodal Interaction*. ACM, 26–33.
21. Tate T Kubose, Kathryn Bock, Gary S Dell, Susan M Garnsey, Arthur F Kramer, and Jeff Mayhugh. 2006. The effects of speech production and speech comprehension on simulated driving performance. *Applied cognitive psychology* 20, 1 (2006), 43–63.
22. Changchun Liu, Karla Conn, Nilanjan Sarkar, and Wendy Stone. 2008. Online affect detection and robot behavior adaptation for intervention of children with autism. *Robotics, IEEE Transactions on* 24, 4 (2008), 883–896.
23. Angela Mahr, Michael Feld, Mohammad Mehdi Moniri, and Rafael Math. 2012. The ConTRe (Continuous Tracking and Reaction) task: A flexible approach for assessing driver cognitive workload with high sensitivity. *Adjunct Proceedings of the 4th AutomotiveUI.* (2012), 88–91.
24. Sandra P Marshall. 2007. Identifying cognitive state from eye metrics. *Aviation, space, and environmental medicine* 78, Supplement 1 (2007), B165—-B175.
25. Daniel McFarlane. 2002. Comparison of Four Primary Methods for Coordinating the Interruption of People in Human-Computer Interaction. *Human-Computer Interaction* 17, 1 (March 2002), 63–139. DOI: http://dx.doi.org/10.1207/S15327051HCI1701_2

26. George A Miller. 1956. The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological review* 63, 2 (1956), 81.

27. L J M Mulder. 1992. Measurement and Analysis-Methods of Heart-Rate and Respiration for Use in Applied Environments. *Biological Psychology* 34, 2 (1992), 205–236. <GotoISI>://A1992KA29300007

28. Calvin K L Or and Vincent G Duffy. 2007. Development of a facial skin temperature-based methodology for non-intrusive mental workload measurement. *Occupational Ergonomics* 7, 2 (2007), 83–94.

29. Fred Paas, Juhani E Tuovinen, Huib Tabbers, and Pascal W M Van Gerven. 2003. Cognitive load measurement as a means to advance cognitive load theory. *Educational psychologist* 38, 1 (2003), 63–71.

30. Kilseop Ryu and Rohae Myung. 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 (2005), 991–1009. DOI: http://dx.doi.org/10.1016/j.ergon.2005.04.005

31. Yu Shi, Technology Park, Natalie Ruiz, Ronnie Taib, Eric H C Choi, and Fang Chen. 2007. Galvanic Skin Response (GSR) as an Index of Cognitive Load. In *CHI EA '07 CHI '07 Extended Abstracts on Human Factors in Computing Systems*. ACM, 2651–2656. DOI: http://dx.doi.org/10.1145/1240866.1241057

32. Marieke van Dooren, J J G (Gert-Jan) de Vries, and Joris H Janssen. 2012. Emotional sweating across the body: Comparing 16 different skin conductance measurement locations. *Physiology & Behavior* 106, 2 (2012), 298–304. DOI: http://dx.doi.org/10.1016/j.physbeh.2012.01.020

33. Russell T Warne. 2014. A Primer on Multivariate Analysis of Variance (MANOVA) for Behavioral Scientists. *Practical Assessment, Research & Evaluation* 19, 17 (2014), 2.

34. Christopher D. Wickens. 2002. Multiple resources and performance prediction. *Theoretical Issues in Ergonomics Science* 3 (2002), 159–177. DOI: http://dx.doi.org/10.1080/14639220210123806

35. Christopher D Wickens. 2008. Multiple resources and mental workload. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 50, 3 (2008), 449–455.

36. Glenn F. Wilson. 2002. An Analysis of Mental Workload in Pilots During Flight Using Multiple Psychophysiological Measures. *The International Journal of Aviation Psychology* 12, 1 (2002), 3–18. DOI: http://dx.doi.org/10.1207/S15327108IJAP1201_2