Assessing real-time cognitive load based on psycho-physiological measures for younger and older adults

Eija Ferreira[1], Denzil Ferreira[1], SeungJun Kim[2], Pekka Siirtola[1], Juha Röning[1], Jodi F. Forlizzi[2], Anind K. Dey[2]

Department of Computer Science and Engineering
University of Oulu, Finland
{eija, dteixeir, pesiirto, jjr}@ee.oulu.fi

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

***Abstract*—We are increasingly in situations of divided attention, subject to interruptions, and having to deal with an abundance of information. Our cognitive load changes in these situations of divided attention, task interruption or multitasking; this is particularly true for older adults. To help mediate our finite attention resources in performing cognitive tasks, we have to be able to measure the real-time changes in the cognitive load of individuals. This paper investigates how to assess real-time cognitive load based on psycho-physiological measurements. We use two different cognitive tasks that test perceptual speed and visio-spatial cognitive processing capabilities, and build accurate models that differentiate an individual's cognitive load (low and high) for both young and older adults. Our models perform well in assessing load every second with two different time windows: 10 seconds and 60 seconds, although less accurately for older participants. Our results show that it is possible to build a real-time assessment method for cognitive load. Based on these results, we discuss how to integrate such models into deployable systems that mediate attention effectively.**

Keywords—cognitive load; psycho-physiological measures; elementary cognitive tests; attention; interruption; interfaces

I. INTRODUCTION

Cognitive load is a complex concept that is often not well defined [1]. In the field of human-computer interaction (HCI), Oviatt [2] has defined it as the mental resources a person has available for solving problems or completing tasks at a given time; it depends on the amount of information or number of elements that need to be processed simultaneously. The actual amount of cognitive load experienced by a person is influenced by the tasks, individual differences, and social and environmental factors [2]. We are increasingly exposed to a deluge of information every day. Given that human attention is a finite resource, the balance of our cognitive load or attention demands can easily fluctuate.

While driving, for example, we can experience attention interference due to an expected or unexpected interruption (*e.g.*, the need to change lanes or a pedestrian crossing in front of a driver, respectively). There is also increased cognitive load when switching our attention between virtual/information spaces and physical spaces (*e.g.*, using a navigation display while driving) or between two distinct user interfaces (*e.g.*, using a smart phone and a laptop together) [3]. While these moments are often short-lived, they can have significant impact on a person's ability to attend to the task at hand.

In order to help users effectively manage their attention and cognitive effort, systems can provide appropriate support for these moments (*e.g.*, providing a driver with information about available parking spaces in a congested urban area) and can avoid providing inappropriate distractions at these moments (*e.g.*, presenting the driver with an advertisement for cheap gas). To remedy this situation, future systems must be able to provide the right cognitive aid at the most appropriate time.

One's cognitive ability changes with age [4]. As we age, deficits in cognitive abilities increase, and more support is needed [5]. Older adults have difficulty focusing on relevant information, and are prone to distractions. Also mental processing and reaction time become slower with age [6]. However, the magnitude and speed of these changes also vary widely from person to person [4]. In order to design usable and accessible systems for older adults, we need to first understand if cognitive load assessment tools that are used for younger adults will work for older adults. This is particularly challenging, given changes in psycho-physiological abilities that occur as a part of aging [7].

Especially in situations where cognitive load is high (*e.g.,* interruption, divided attention, multitasking), perceptual speed and visio-spatial cognitive processing capabilities [8] are very important. *Perceptual speed*, also known as 'inspection time', 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 [9]. *Visio-spatial cognitive processing capabilities* refer to our ability to perceive visual stimuli and both spatially and cognitively integrate them with what we have seen previously. Understanding these capabilities is important for determining when and how to present information to individuals. Doing so at the wrong time can dramatically increase one's cognitive demands, can have negative impacts on task performance and emotional state, and, in extreme cases, even be life-threatening [10]. Given the transient nature of these situations, it is clear that a real-time method of measuring cognitive load is required.

There are reliable methods for assessing cognitive load, based on task performance and subjective ratings [11,12]. However, due to their *post-hoc* (measured after a completed experience) and *static* (measured at a single point in time) nature, these methods are inappropriate for measuring variations in cognitive load over a continuous time frame.

In addition, one's performance may not always change, even though their cognitive load varies considerably.

Our previous research has shown that a psycho-physiological response-based assessment method can sensitively detect time-based variations in cognitive load, which could be effective over a number of cognitive processes, including memory, perception, and spatial processing [13]. Such a system could respond with an appropriate signal to help mediate a shortfall in a particular cognitive process. However, our previous work only supported coarse-grained assessment (up to minutes) instead of real-time, and has focused solely on assessment for young adults.

In order to determine how to respond to the temporal and subtle changes of cognitive load, it is necessary to measure the cognitive load of individuals *in real-time* and *in-situ*. With a real-time, objective measure, we can develop novel interfaces that react to users' cognitive load. In this paper we will particularly focus on a method of measurement that is noninvasive, inexpensive, easy-to-use and accessible to as wide a range of users as possible by using low-cost, off-the-shelf psycho-physiological wearable sensors.

Accordingly, this paper explores the following questions:

- Can we acquire a real-time measure of cognitive load for both younger and older adults by examining psycho-physiological sensor streams?
- Can we use the same sensors to measure cognitive load in older adults as for young adults?

To address these questions, we derived a real-time measure of cognitive load using two elementary tasks that measure perceptual speed and visio-spatial cognitive processing, good indicators of attention demand. We manipulated the levels of difficulty in each elementary task, and presented these tasks to 13 younger participants and 17 older participants over the age of 65. We used four sensor devices to measure their psycho-physiological responses. We validated the induction of different levels of cognitive load by using assessment methods based on time-on-task and subjective ratings. Through this effort, we developed individual models for measuring cognitive load in real-time (every second) for young and older adults. We present these models along with a discussion of how to integrate our models into a deployable attention management system.

II. Related work

A. Cognitive load in HCI and UI design

In Human-Centered Design, Cognitive Load Theory (CLT) [14], one of the most important theories in educational psychology, has been applied to the designing of interfaces that effectively minimize cognitive load [2]. According to CLT, decreasing cognitive load associated with an interface frees people's available intellectual resources for their main task [2]. CLT has been used, for example, to design educational interfaces that effectively minimize students' load and to optimize their learning efficiency by matching the learning task with their mental capabilities [15]. It has also been applied in traffic control [16] and in safety critical applications [17].

In HCI, a real-time technique for assessing cognitive load is needed to minimize users' cognitive load to navigate an interface, and to dynamically adapt to users' current cognitive state, resources and context. For example, in computer-based healthcare environments, UIs must be designed with consideration of requirements, cognitive capabilities and limitations of the end users, and adapt to their information needs [18]. However, the older the users, the slower they are to react to or realize that an update occurred in the user interface [19].

Users' context is important, especially in multimodal interaction. From [20] guidelines, we should maximize human cognitive and physical abilities, by supporting dynamic adaptivity. In cars, CoDrive [21] successfully adapted the navigation system by replacing the map from the display with arrows when drivers' workload was higher.

B. Cognitive load assessment methods

Performance-based methods for assessing cognitive load are frequently employed in dual-task settings. These methods examine how a participant's responses deteriorate (*e.g.*, lag in reaction time or increase in errors) when using finite cognitive resources to perform two or more tasks. This objective approach has been demonstrated to have a strong link with cognitive load [12]. However, it is less sensitive to subtle differences in cognitive load and can only be measured after a task is complete and not during the task itself when the result may be the most useful.

Subjective rating-based methods such as the NASA TLX [11] use participants' own judgment of their task execution efforts. This approach is applied *post-hoc* (after task completion), is reliable and is non-intrusive to the task performed [12]. However, it is inappropriate in assessing dynamic changes in cognitive load and less promising for automated or immediate assessment. Also, even when users struggle to complete a task in a timely fashion, they may self-report the task as low workload, if they believe they did not make any errors [22].

Psycho-physiological response-based methods can also be used to assess cognitive load. Some believe that this approach can more sensitively assess load over a continuous time frame [12], allowing for the detection of changes in cognitive load even when no deterioration in task performance is demonstrated. The approach supports assessment for tasks that leverage major cognitive processes such as perception, memory or reasoning [23].

Accordingly, we choose psycho-physiological measurements as our assessment method. However, given its sensitive nature, we cannot rely on a small number of physiological measures that have been demonstrated in the literature as being useful for assessing cognitive load (*e.g.*, gaze information [24,25], heart rate [25,26], electroencephalography - the electrical activity of the brain (EEG) [27,28], electrocardiography - the electrical activity of the heart (ECG) [26], galvanic skin response (GSR) [25,26], breathing rate (BR) [26], and skin temperature [26]. In our previous work [13] we compared features calculated from different psycho-physiological signals for assessing cognitive

load. However, those results were limited to models of young adults based on one single feature and supported assessment of cognitive load on a granularity that ranged from seconds to minutes.

The information used for cognitive load assessment impacts the temporal granularity achieved. Most commonly, coarse granularities, such as 2 minutes [25] or 5 minutes [29], have been used in studies considering psycho-physiological data only. Others have combined psycho-physiological signals with performance data [26,30] to achieve assessment granularities of 10-30 seconds. Studies considering multi-channel EEG measurements have experimented with shorter window sizes, such as between 2-120 seconds with a 32-channel EEG cap for working memory load classification [28], and 5 or 10 seconds with a 6 lead EEG for mental workload assessment [31]. However, these experiments used wet electrodes requiring proper skin preparation and conductive gel application, and hence are impractical for naturalistic use. Knoll *et al*. [32] used a 14-channel dry electrode EEG headset for cognitive load assessment at a granularity of 1.5 seconds. All of these studies suggest that relatively high classification accuracies can be attained even with shorter windows, but Grimes *et al.* [28] demonstrate that there exists a tradeoff between window size and classification accuracy. In this study, we took the approach of using an easy-to-use, low-cost and off-the-shelf one-channel EEG device together with other minimally obtrusive wearable sensor devices.

Both individual and cross-user models for psycho-physiological data have been previously implemented. Solovey *et al*.'s recent results [30] demonstrate potential of building models based on heart rate and skin conductance measurements that work across individuals to identify elevated cognitive workload levels while driving. However, Grimes *et al.* [28] concluded that, because of individual differences in EEG characteristics, cross-user models only appear to be reliable when data is averaged over long time periods, whereas individual models are needed to assess user state in real-time. Their results emphasize the importance of individual feature selection, as this accounts for most of the individual differences, similar to Wilson *et al.* [31]. Moreover, Mehler *et al.* [25] emphasize that individuals differ in the extent to which they show reactivity across different physiological measures.

C. Cognitive aging and psycho-physiological changes

There are only a few studies where the use of a psycho-physiological assessment method has been explored for older adults' cognitive load [25,30,33] and, to the best of our knowledge, no studies where implementation aspects of real-time models have been compared between age groups. However, numerous efforts have been devoted to addressing cognitive decline of older adults [4,34], investigating psycho-physiological changes [7,34] and cognitively assisting their everyday tasks [35].

It is widely agreed that different cognitive variables have different patterns of relations with age [4]. Perceptual speed and episodic memory are known to decline with increased age [4,34]. There are also age-related differences in tasks involving working memory, attention, task switching and interruption, and older adults have a generally slower processing speed [34].

III. Experimental Methods

A. Participants

We recruited 30 participants split across two age groups: 13 younger participants with their ages ranging from 18-30 (M=22.9; SD=3.9), including 6 males and 7 females, and 17 older participants with ages ranging from 65 to 88 (M=74.3; SD=5.7), including 6 males and 11 females, with normal or corrected-to-normal vision. One of the older participants (#3) suffered from a neurological disorder (hand tremor). Participants were recruited through study fliers, advertising at a lifelong learning institute at two local universities and a local center for behavioral decision research. They were compensated $20 US for their time.

B. Elementary Cognitive Tests (ECTs)

Many studies in the area of cognitive aging have made use of elementary cognitive tasks (ECTs). An ECT refers to any of a range of basic tasks that require only a small number of mental processes and that easily specify correct outcomes [8]. ECTs have been widely and reliably employed to compare individual differences between two population groups of interest (*e.g.*, young vs. older adults, patients vs. health-controlled people) [8,36], and to compare cognitive factors associated with age-related trends [37]. Different levels of task complexity have also been used in some cognitive aging studies. For example, Salthouse *et al*. [6] showed that higher task complexity has a greater impact on older adults' task performance impairment than younger adults'. In our study, we also make use of ECTs that incorporate differing task complexity.

We presented two different types of ECTs to the participants: the Pursuit Test (PT), in which subjects determine where a line begins and ends, and the Scattered X's test (SX), in which subjects locate all the X's in a display (Fig. 1). These tests measure perceptual speed and visio-spatial cognitive processing capabilities that were identified from the fields of psychology and cognitive science [38]. The tests were displayed on a computer screen (Fig. 2) and participants answered them using a mouse and a keyboard. After finishing each question, they clicked on a 'next' button positioned at the bottom of the screen to be directed to the next question.

For each of the two tests, we prepared two sets of questions that were more and less difficult, that we expected would induce different levels of cognitive load on the participants, for details see [13]. To confirm this, we piloted the question sets before the actual study with 10 individuals not participating in the study.

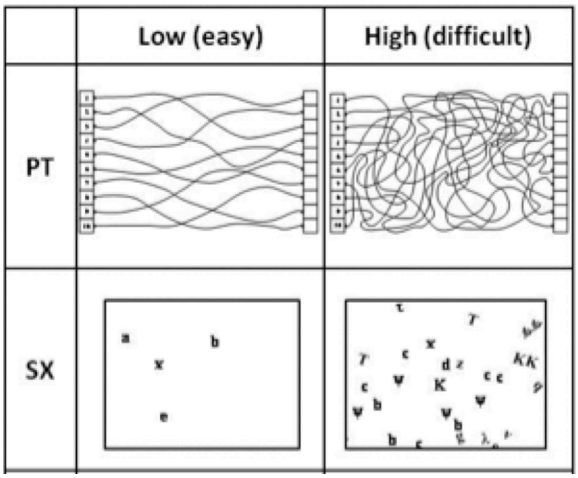

Fig. 1. Two elementary cognitive tasks (ECTs) with two task difficulty levels (low and high).

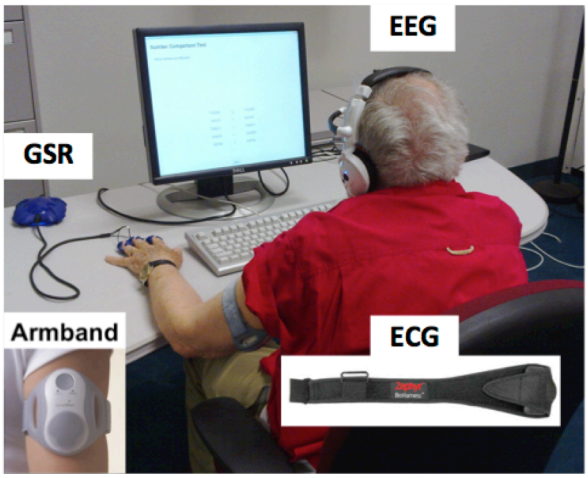

Fig. 2. Experimental setup with four sensor devices.

C. *Psycho-physiological sensors*

We measured participants' psycho-physiological responses with four sensor devices (Fig. 2): a wireless ECG monitor (Bioharness BT) that also records heart rate and breathing rate, an armband that measures heat flux (rate of heat transfer on the skin) (SenseWear Pro3), a wireless EEG headset (NeuroSky Mindset) with one dry electrode located on the participant's forehead (position Fp1, as defined by the 10-20 system [39]), and a GSR finger sensor (LightStone). As mentioned before, these signals have been shown to have value in assessing cognitive load in previous studies.

We selected these sensors because of their ease of use and relative non-invasiveness as well as their low cost and off-the-self availability. While most of the previous studies have used high-end EEG caps with a high number of channels (*e.g.*, 32 channels by Grimes *et al.* [28]), low-cost EEG has previously been successfully used for task classification (a 2-channel EEG system) by [40] and cognitive load assessment (a 14-channel headset) by [32].

Right before the onset of the data collection, the ECG monitor and the armband internal clocks were synchronized with the computer running the study software and logging the headset and GSR data. The data transmission times were assumed to be so short that they would not affect our analysis at a 10-second granularity.

D. *Protocols and Procedures*

Participants filled out a participation consent form after a brief introduction to the study. We helped them put on each of the sensor devices and then provided them with descriptions of each of the ECTs and how to answer the questions using the mouse and keyboard. During this time, the sensors took baseline readings (allowing for initiation time and time for participants to feel at ease). To minimize distractions and noise in the sensor readings, we ensured that there was no ambient noise and asked participants to keep their non-dominant hand on the desk, maintaining a stable pose. These tasks took approximately 10 minutes.

The ECTs began after a 90-second relaxation period, during which participants were asked to close their eyes for mental relaxation. The study consisted of three blocks of questions, each block containing one question set for each of the difficulty levels for each of the two ECTs (3 blocks × 2 test types × 2 difficulty levels). The duration of the question sets in the first block was 4 minutes, 3 minutes in the second, and 2 minutes in the last one. The participants were asked to answer as many questions as they could during each question set at a good pace but without rushing. They were not informed about how much time they had left. The durations of the question sets were kept short to maintain the participants' interest in solving the tasks throughout the set, while, at the same time, maximizing the amount of data we would have to build our models. After each question set in the first block, the participant was automatically directed to a task difficulty rating slide, where we used the NASA TLX (Task Load index) method [11]. A 30-second relaxation time was given between question sets. The order of the ECT question types and the two difficulty question sets were counterbalanced.

E. *Validation of different cognitive load induction*

We validated that the two sets of questions we designed actually induced distinguishable levels (*i.e.*, low and high) of cognitive load. For this, we examined two measures: 1) Time on task and 2) Taskload index (NASA TLX).

1) *Time on Task*

For both young and older participants the differences in time on task between the two task difficulty levels were significant (One Sample t-test, Young: $t(12) = 8.36$, $d = 2.32$ (PT), $t(12) = 10.28$, $d = 2.85$ (SX), $p < .0001$; Older: $t(16) = 10.35$, $d = 2.51$ (PT), $t(16) = 18.07$, $d = 3.15$ (SX), $p < .0001$). Older participants consistently took longer to complete tasks. On average, the young participants took 33.1 seconds (low) and 94.8 seconds (high) to complete a PT task. Older participants took longer, 59.8 seconds (low) and 160.7 seconds (high). On the SX task, young participants took 7.7 seconds (low) and 22.1 seconds (high), while older participants took 10.0 seconds (low) and 35.1 seconds (high).

We then inspected how both age groups experienced the difference between the two difficulty levels in each task. There was no significant difference for the PT task, where the young participants took on average twice as long to solve the difficult questions (high) than the easy questions (low), and the older participants took 1.7 times longer (Welch Two Sample t-test, $t(26.50) = 1.18$, $p = .25$, $d = .43$). On the other hand, for the SX task, the corresponding numbers are 2.0 and 2.7, a significant difference ($t(27.51) = -2.24$, $p = .03$, $d = .78$).

2) Subjective Rating (NASA TLX)

The difference between easy and difficult level TLXs significantly differed for both young and older participants (Young: $t(12) = 7.09$, $d = 1.97$ (PT), $t(12) = 6.41$, $d = 1.78$ (SX), $p < .0001$; Older: $t(16) = 7.46$, $d = 1.81$ (PT), $t(16) = 5.85$, $d = 1.42$ (SX), $p < .0001$). On average, the younger participants rated the task load of PT tasks to be 2.5 (low) and 3.3 (high), whereas the older participants rated 2.6 (low) and 3.9 (high) on a scale of 1 to 5. In the SX task the young participants' ratings were 2.0 (low) and 3.0 (high) and the older participants' ratings 2.0 (low) and 2.9 (high).

Across the age groups, the difference in how the participants experienced the difference between the two difficulty levels was significant for the PT task ($t(26.25) = -2.35$, $p = .03$, $d = .81$), where the younger participants' difference in ratings was, on average, 0.78 and the older participants' difference was 1.26. The corresponding differences for the SX task, 1.03 and .91, were not significant ($t(27.14) = .54$, $p = .60$, $d = .20$).

These two measures validated that our participants experienced a significant difference in difficulty between the two difficulty levels. Regardless of participants' experience with computers, both age groups experienced the difference in difficulty between the two question sets similarly. Therefore, we confirm that our ECTs with different difficulty levels can reliably induce distinguishable levels of cognitive load from both younger and older participants, and that we can reliably use them in our psycho-physiological assessment.

IV. Data analysis

A. Data

Seven psycho-physiological signals were measured with the four sensors: average heat flux recorded by the armband (at a sampling rate of 32 Hz); raw EEG signal measured by the headset (128 Hz); raw ECG signal (250 Hz), breathing rate (1 Hz), breathing wave amplitude (1 Hz), and heart rate measured by the ECG monitor; and GSR measured by the finger sensor (30 Hz). The headset also gave 8 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, as well as two mental state outputs: attention and meditation, calculated at 1 Hz. We extracted R-R intervals from the raw ECG signal using a peak detection algorithm [41]. As a pre-processing step, the raw GSR and heat flux signals were convoluted with a Bartlett window to smoothen and differentiated to remove trends, such as an increasing level of GSR during the course of the experiment.

We encountered some challenges with noisy or missing sensor data, especially with the EEG and ECG sensors. The poor quality of the EEG signal may be caused by poor contact of the sensor/ground/reference electrodes to a participant's skin, motion of the participant, environmental electrostatic noise, or non-EEG biometric noise (*i.e.*, EMG, ECG, EOG, and others) [42]. Other researchers have reported on technical problems with a similar headset [43]. Sources of recording noise in the ECG can include artifacts caused by movement of the electrode away from the contact area on the skin or EMG noise due to muscle contractions under the sensor surface [44]. Particularly, ECG measurements from the older participants had irregularities that may have influenced the extraction of R-R intervals. For each participant, missing or noisy measurements (EEG signals flagged with the headset's poor signal quality metric or ECG data with notable recording noise) are presented in Table I (missing or '○', respectively). Previous work found EEG [27,28] and ECG [13,26] to be good indicators for cognitive load and therefore we expect them to perform well on real-time cognitive load assessment. Participants who had missing or noisy ECG and EEG data were excluded from the analysis (2 young subjects and 5 older subjects, highlighted in grey in Table I). In addition, the heat flux signal was missing from one participant due to a device malfunction and the GSR measurement had to be dropped from 5 participants because of incomplete readings. This left us with 11 young and 12 older participants with whom we continued our analysis.

B. Feature extraction

Cognitive load models were built at two different granularities based on features derived from 10-second and 60-second segments of raw measurement data. For both feature sets, statistical features were calculated on sliding windows with a step of one second. Five seconds of data from the beginning of each question set shown to the participant was excluded from the analysis to allow for a changing cognitive load level. These granularities were selected on one hand to minimize the time lag when performing the assessment in real-time and on the other hand to compare how a longer window size would affect the detection accuracy.

In total, 128 features were extracted from the signals. The mean, median, variance, standard deviation, 10^{th}, 25^{th}, 75^{th} and 90^{th} percentile, interquartile range (IQR), root mean square of successive differences (RMSSD), mean of the absolute values of the first (MAFD) and second (MASD) differences [45], mean crossing rate and the difference of the last second mean and the first second mean of the window (end-start difference) were calculated from the heat flux, GSR and the R-R interval signal. Further, correlation and standard deviation of successive differences (SDNN), relative occurrence of successive differences exceeding 20ms (pNN20) and 50ms (pNN50) [46], and mean peak amplitude were calculated from the R-R data. In addition to these time domain HRV features, no frequency domain HRV features were extracted from the R-R intervals because of the short duration of the data segments used. According to [47], at least 2 minute but preferably 5 minute segments of R-R data are needed to calculate the spectral components. The count, maximum amplitude, mean amplitude, mean duration and area under skin

conductance response (SCR) occurrences were extracted from the GSR measurement. The SCR occurrences were detected using an algorithm that locates zero-crossings in the differentiated GSR signal (adapted from [29]). The heart rate, breathing rate and breathing wave amplitude were described by seven features: minimum, maximum, mean, median, variance, standard deviation and end-start difference. The mean, median, variance, standard deviation and end-start difference were used to summarize the 8 EEG power values and 2 mental state outputs. For comparison, the power was also calculated from the raw EEG data on five commonly used bands: delta (1-4 Hz), theta (4-7 Hz), alpha (8-12 Hz), beta (12-30 Hz) and gamma (30-50 Hz).

TABLE I. SENSOR SIGNAL QUALITY PER PARTICIPANT.

#	Heat flux	GSR	EEG	ECG, BR, HR
Younger				
1	●	●	●	○
2			●	●
3	●		○	
4	●	●		●
5	●	●	●	●
6	●	●		●
7	●	●	●	○
8	●	●	●	○
9	●	●	●	○
10	●	●	●	○
11	●	●		●
12	●	●	●	○
13	●	●		○
Older				
1	●	●		○
2	●	●	●	●
3	●	●		●
4	●	●		●
5	●	●	●	●
6	●	●	●	●
7	●	●		○
8	●	●	●	●
9	●	●	●	●
10	●	●	●	○
11	●	●		●
12	●		●	○
13	●	●	●	●
14	●	●		○
15	●			○
16	●		●	●
17	●	●		○

● – Good ○ – Poor – Missing

All the features were normalized to equalize their importance. The feature data from each participant in each question type (three sets with low difficulty level and three sets with high difficulty level) was scaled linearly so that the 5th and 95th percentiles of each of the features met the range [0,1].

C. Modeling

Quadratic discriminant analysis (QDA) was used to classify the feature values calculated on the sliding windows. A block cross-validation scheme similar to the one recommended by Grimes *et al.* [28] was adopted to avoid temporal dependence of the data segments and distortion of the results. The three data sets corresponding to the three blocks of the study design were used to train the models, select subsets of the original set of 128 features that would give the highest accuracy in measuring cognitive load, and to simulate a real-time system to provide an estimate of how the model would perform on a previously unseen set of data, respectively. Hence, mutually

TABLE II. COGNITIVE LOAD ASSESSMENT ACCURACIES FOR YOUNGER AND OLDER PARTICIPANTS IN THE TWO ECT'S. ASSESSMENT WAS PERFORMED AT A ONE-SECOND FREQUENCY.

	PT		SX	
#	*10s*	*60s*	*10s*	*60s*
Younger				
1	30%	53%	44%	66%
2	71%	**100%**	**88%**	39%
4	**77%**	**100%**	70%	50%
5	**94%**	**86%**	**100%**	**100%**
6	50%	50%	62%	**89%**
7	**76%**	**100%**	**76%**	69%
8	49%	54%	63%	31%
9	**76%**	53%	63%	**85%**
10	**79%**	74%	45%	**86%**
11	74%	**95%**	**98%**	**100%**
12	52%	**100%**	**92%**	**100%**
Avg	66%	**79%**	73%	74%
Older				
2	57%	**91%**	63%	66%
3	62%	**79%**	63%	**79%**
4	53%	**81%**	**77%**	**77%**
5	62%	**100%**	62%	52%
6	**88%**	**98%**	55%	50%
8	49%	**85%**	**77%**	**100%**
9	69%	**95%**	**90%**	58%
10	54%	50%	50%	24%
11	72%	**99%**	61%	**99%**
12	51%	53%	**80%**	**100%**
13	72%	**100%**	50%	50%
16	**80%**	**96%**	50%	28%
Avg	64%	**86%**	65%	65%

exclusive data sets were always used for training, model selection, and testing of our models. The simple algorithm of selecting the three best individual features was used for the feature subset selection. More sophisticated methods were also tested, but because of overfitting, this simple method proved the most efficient for the task.

A model with individual feature selection was trained for each participant's data for each ECT task to distinguish the two difficulty levels. We implemented the models both on the feature set with the 10-second time window and the set with the 60-second time window.

V. Results

Cognitive load assessment accuracies for the young and older participants for the two ECT's are presented in Table II. Classification accuracy is computed as the percentage of 10- or 60-second windows classified correctly. Very high accuracies are achieved for many of the young participants but with high variation. The rationale for the models not working for all participants might include noise in the measurement data or individual differences in how the changes in cognitive load manifest themselves in psycho-physiological signals. Participants might also have experienced the two levels of task difficulty differently, even though we did not find clear evidence of this in the subjective ratings and the task performance. Also our decision to fix the number of features at three might explain the inferior accuracies for some of the participants. This parameter value resulted in the best results for most of the participants but some of them would have benefitted from a higher number of features in the model.

The accuracies for the longer windows are generally better than for the shorter windows in the PT task, but in the SX task the assessment performance is equal at both granularities. It is worth noticing here that even the 10-second assessment is able to differentiate the two levels of cognitive load at a very high accuracy for some of the participants. The results are particularly good for the young participants 5, 11 and 12. On the other hand, our models did not work for young participants 1, 6, and 8. However, participant 6 was missing EEG data, and participants 1 and 8 had poor ECG, HR and BR data (all measured with the same sensor). The results for the older participants are very similar and the 60-second granularity results are comparable to those of the young adults. However, for the older adults, the difference in accuracy between the two granularity levels is greater than for the younger participants. This might be a consequence of the psycho-physiological changes related to aging [7] resulting in a slower reaction time to changes in cognitive load. Our models did not work well for older participants 2, 5, 10 and 13. We suspect that the reason for this is that some older adults do not express cognitive function with a high enough signal through psycho-physiological responses.

The real-time functioning of our model is demonstrated in Fig. 3 where estimations of a participant's cognitive load levels are provided once a second on the 10-second granularity level during execution of the two difficulty levels of the PT task. We can see the fluctuations of the predicted cognitive load especially during the low cognitive load task variant. When designing the experiment, we had to assume that our task difficulty levels would uniformly fix the participant's cognitive load for the entire duration of the question set. Although unrealistic in practice, this assumption provided ground truth for our models. Our validation of different cognitive load induction shows this was true on average. However, it may not be true for all of the 10-second data segments for which our model gave an estimate of cognitive load. In some cases, we may have actually classified the load correctly even though it may not match the label we have assigned based on average task difficulty.

We then analyzed the features selected for each of the models at the 10-second granularity. The relative count of times a feature was selected into the set of three best features normalized by the number of participants in the age group, from whom that sensor stream was available, is shown in Fig. 4. For this we only considered the participants and tasks that

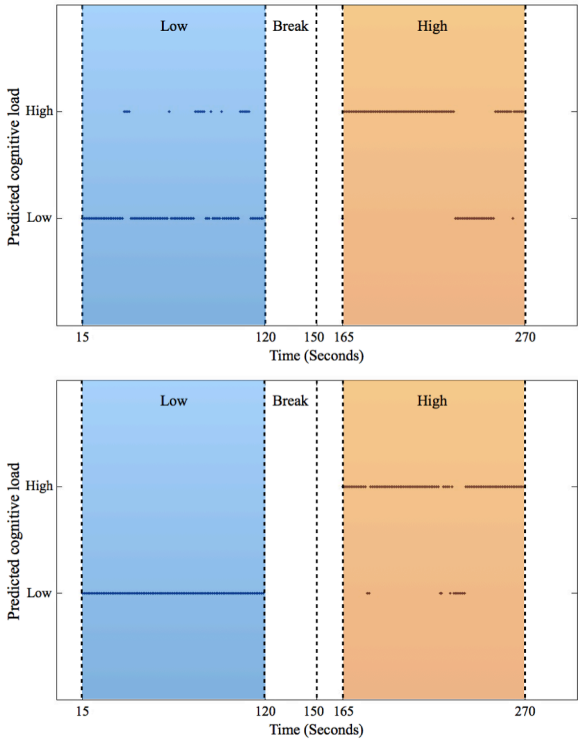

Fig. 3. Real-time prediction of cognitive load level for PT task, 10-second granularity. Top: young participant 10 (79% accuracy). Bottom: young participant 5 (94% accuracy).

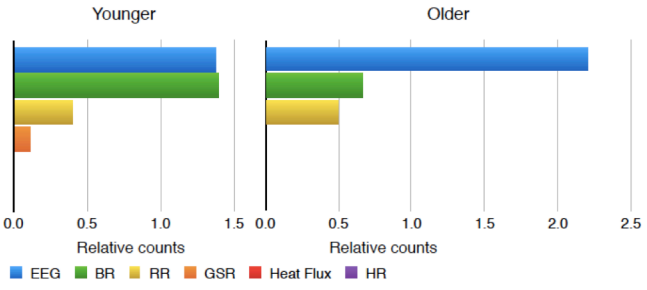

Fig. 4. Relative count of the times a feature from each sensor stream was selected into the models (both tasks, 10-s).

had an accuracy over 75%. The most often selected features for both the young and the older participants originate from the EEG and BR signals. The EEG signal was more important and the BR measurement less important for the older than for the younger participants. R-R signals are also fairly well represented among the most common features for both age groups, whereas GSR features were seldom used for the younger participants and never for the older participants.

When looking at the features selected for the three above mentioned young participants 5, 11 and 12, we notice that for participant 5, five of the six selected features for the two task types were calculated from the EEG signal and one from the RR intervals; for participant 11, five of the features originated from the BR signal and one from the RR signal, and for participant 12 the feature sets selected consisted of four BR features and two EEG features.

VI. Discussion

In this study, we investigated how to build a real-time system based on psycho-physiological measurements to assess the cognitive load of a person while he/she is executing tasks of different difficulty. We especially examined the applicability of our approach for two different age groups, young and older (over the age of 65) adults. We used three separate data sets to train the models, to select the features to be used and to simulate the real-time functioning of our models on previously unseen data.

A. Model performance

We built models for each ECT task, where we were able to accurately (on average 64-73%) and quickly (time-scale of 10 seconds) discriminate the two levels of task difficulty based on psycho-physiological measurements. Models for both age groups performed roughly the same (Table II). As expected, our results for the longer time-scale of 60 seconds were better (65-86%). With these results, we satisfy our first question: we are able to create an accurate real-time assessment tool for measuring cognitive load for both young and older adults.

B. Differences between the young and the older participants

Our second question of interest was whether we could use the same set of sensors to measure cognitive load for young and older adults. We found that the sensors were the same: for both age groups, EEG and BR proved most informative of cognitive load. Overall, the consistency across sensors means that it will be easier to deploy a system for both young and older adults because the hardware aspects of the system would be the same. The EEG measurement appeared to be even more informative of cognitive load for older participants while the BR signal was more valuable for the younger participants. The fact that several of our older participants had missing EEG data might explain some of the less accurate results for older participants. The importance of the EEG recording for the older adults is an interesting finding considering that many of the cognitive deficits of normal aging involve dysfunction in the prefrontal cortex [7], where the electrode of our EEG headset is connected.

C. Comparison with previous work

In addition to considering two different age groups, a topic that has not yet been widely researched, the main contribution of this work is our real-time models of cognitive load. The time granularity of 10 seconds, shown to result in good classification accuracies for most of our participants, has previously been achieved by Solovay *et al.* [30] in a study where driver workload was classified at a 10 to 30-second granularity based on HR and skin conductance measurements. In their study, however, the HR data was manually reviewed and edited for artifacts and anomalies whereas we used our data without manual correction. Our modeling accuracies are comparable with their individual models with varying time granularity where accuracies of 75%, on average, were achieved.

Other studies, where the assessment of cognitive load has been based on multi-channel EEG measurements, have reached higher accuracies and finer time granularities (*e.g.*, Grimes *et al*. [28] with a 32-lead EEG gap, and Knoll *et al.* [32] with a 14-channel EEG headset). In this study, however, we wanted to test if acceptable results could be obtained with easy-to-use, low-cost and off-the-shelf sensors. Despite the challenges we had with the quality of some of the recorded data, the EEG and BR measurements turned out to be very valuable for assessing cognitive load. These results are different from our previous study [13] where ECG and heat flux were found most indicative of cognitive load. However, in the previous study the assessment was performed over a longer time period (up to minutes).

D. Cognitive Interfaces

Young participants complete the PT and SX tasks faster than older participants, by almost twofold (see *Measure 1 – Time on task*). Both tasks measure perceptual speed and visio-spatial cognitive processing capabilities [38]. Our findings support Trewin *et al.*'s [19] results that increased age results in increased UI reaction time. Furthermore, it is clear that unpredictable autonomous interface adaptations do in fact reduce a system's usability and learnability [48]. With our models, however, an interface could be designed to gracefully degrade the interface update rate, and use prominent UI highlights to compensate for users' higher cognitive load, in real-time, therefore minimizing older adults' cognitive load when they need to find or use information.

E. How to build a system to assess cognitive load?

We implemented our models at two different time-scales, 10 seconds and 60 seconds, that both resulted in high accuracies for the majority of our participants. In practice, our subject-specific models require a short period of training before the model can be put to use. This training period, consisting of data collection and labeling, feature extraction and model specification, can be close to automated and accomplished in a matter of minutes as only a few minutes of training data is needed for our models. In this study, we conducted all the above-mentioned steps in a single session. However, to optimize the performance of the models at a later occasion without remodeling or calibration, the sensors could be disconnected and reconnected between the steps.

After the training period, to use the model to obtain real-time estimations of cognitive load, the small set of selected feature values can be calculated continuously once per second based on the previous 10 or 60 seconds of measurement data, depending on the granularity chosen, and the model can be applied causing no noticeable delay to the detection system. The computational requirements of our system are no greater than those of an average processor. After the initialization phase, the model could even be implemented on an embedded system.

Our method works for both for young and older adults. However, quality of data collection is crucial for both age groups, and makes taking the measurements out of the laboratory more challenging. Particularly for older adults, changes in psycho-physiological signals, for example lower heart rate variability, as well as other medical conditions, can affect the quality of data.

To build and integrate these models into deployable systems for managing attention, let us assume that a field-based task comprises of a number of sub-tasks, which roughly match the elementary cognitive tasks presented earlier. Performance data such as time on task or subjective ratings of difficulty could be collected for each user for the ground truth of cognitive load. We could then collect sensor data by simply using the sensors that provided the most information gain about cognitive load in this study: the EEG collected by a 1-channel EEG headset and the BR recorded by an easy-to-wear chest band sensor. Alternatively, we could also perform the same lab study comparing a large number of sensors described in this paper, and use the sensors and features that provide the best model for each user. New models for each field-based task can then be generated. Cognitive load can be determined at a fine-grained temporal level and a system designer can then determine how to respond to the changes in cognitive load. For example, when cognitive load is high, foveal visual signals could be minimized; and peripheral auditory signals could communicate the onset of new information.

F. Limitations

Despite our best efforts, we must acknowledge the limitations of studying cognitive load with a sample of 30 participants and that our findings might not be indicative of a larger sample. The assessment accuracies of our models had a large variability from person to person. In addition to possible problems with the quality of the data we collected, it might also be the case that cognitive load does not manifest itself in the same way in the psycho-physiological signals for all individuals.

VII. CONCLUSION & FUTURE WORK

In this paper, we have focused on developing a psycho-physiological response-based assessment for measuring cognitive load more sensitively. We presented cognitive tasks that test perceptual speed and visio-spatial cognitive processing capabilities and collected corresponding psycho-physiological data to assess cognitive load for both younger and older adults. We successfully built accurate models that can assess whether an individual is experiencing high or low levels of cognitive load in real-time, for both young and older adults. Our results showed that models for both age groups included the same sensors and performed roughly the same, although the more fine-grained assessment worked for fewer older participants.

In our future work, we will further study the performance of our models in situations where the user's cognitive load fluctuates more quickly and will tune the assessment to work at an even more fine-grained temporal level by using the latest off-the-shelf sensors. We expect that these sensors will improve the data quality for real-time cognitive load modeling hence allowing us to extend this approach to models that distinguish more than two levels of difficulty. We will explore the possibility of building general population models across all individuals within an age group and for all age groups (*e.g.*, children, young and older adults), and deploy these models in interfaces that will help manage user attention according to users' cognitive load.

REFERENCES

[1] T. Jong, "Cognitive load theory, educational research, and instructional design: some food for thought," *Instr. Sci.*, vol. 38, no. 2, pp. 105–134, Aug. 2009.

[2] S. Oviatt, "Human-Centered Design Meets Cognitive Load Theory : Designing Interfaces that Help People Think," pp. 871–880, 2006.

[3] C. Ho and C. Spence, "Assessing the effectiveness of various auditory cues in capturing a driver's visual attention.," *J. Exp. Psychol. Appl.*, vol. 11, no. 3, pp. 157–74, Sep. 2005.

[4] T. A. Salthouse, *Major issues in cognitive aging*. Oxford; New York: Oxford University Press, 2010.

[5] F. I. M. Craik, "Memory changes in normal aging," *Curr. Dir. Psychol. Sci.*, vol. 3, no. 5, pp. 155–158, 1994.

[6] T. A. Salthouse, D. R. Mitchell, E. Skovronek, and R. L. Babcock, "Effects of adult age and working memory on reasoning and spatial abilities.," *J. Exp. Psychol.-Learn. Mem. Cogn.*, vol. 15, no. 3, pp. 507–516, May 1989.

[7] M. Wang, N. J. Gamo, Y. Yang, L. E. Jin, X.-J. Wang, M. Laubach, J. a Mazer, D. Lee, and A. F. T. Arnsten, "Neuronal basis of age-related working memory decline," *Nature*, vol. 476, no. 7359, pp. 210–213, Aug. 2011.

[8] J. B. Carroll, *Human Cognitive Abilities: A Survey of Factor-Analytic Studies*. Cambridge, MA, USA: Cambridge University Press, 1993.

[9] J. W. French, *The Description of Aptitude and Achievement Tests in Terms of Rotated Factors*. Chicago, IL, US: University of Chicago Press, 1951, p. 278.

[10] D. C. McFarlane, "Coordinating the Interruption of People in Human-Computer Interaction," in *The 13th International Conference on Human-Computer Interaction*, 1999, vol. 1, no. Ntsb 1988, pp. 295–303.

[11] S. Hart and L. Staveland, "Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research," in *Human mental workload*, P. A. Meshkati and N. Hancock, Eds. Amsterdam: North Holland Press, 1988, pp. 239–250.

[12] J. Cegarra and A. Chevalier, "The use of Tholos software for combining measures of mental workload: toward theoretical and methodological improvements," *Behav. Res. Methods*, vol. 40, no. 4, pp. 988–1000, Nov. 2008.

[13] E. Haapalainen, S. Kim, J. F. Forlizzi, and A. K. Dey, "Psycho-physiological measures for assessing cognitive load," in *Proceedings of the 12th ACM International Conference on Ubiquitous Computing*, 2010, pp. 301–310.

[14] J. Sweller, "Cognitive Load During Problem Solving: Effects on Learning," *Cogn. Sci.*, vol. 12, no. 2, pp. 257–285, Apr. 1988.

[15] F. Paas, J. E. Tuovinen, H. Tabbers, and P. W. M. Van Gerven, "Cognitive Load Measurement as a Means to Advance Cognitive Load Theory," *Educ. Psychol.*, vol. 38, no. 1, pp. 63–71, Mar. 2003.

[16] G. Camp, F. Paas, R. Rikers, and J. van Merrienboer, "Dynamic problem selection in air traffic control training: a comparison between

performance, mental effort and mental efficiency," *Comput. Human Behav.*, vol. 17, no. 5–6, pp. 575–595, Sep. 2001.

[17] P. R. Cohen and D. R. McGee, "Tangible Multimodal Interfaces for Safety-critical Applications," *Commun. ACM*, vol. 47, no. 1, pp. 41–46, 2004.

[18] V. L. Patel and a W. Kushniruk, "Interface design for health care environments: the role of cognitive science.," *Proc. AMIA Symp.*, pp. 29–37, Jan. 1998.

[19] S. Trewin, B. John, J. Richards, D. Sloan, V. Hanson, R. Bellamy, J. Thomas, and C. Swart, "Age-specific predictive models of human performance," in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst, Extended Abstracts*, p. 2267, 2012.

[20] L. M. Reeves, J.-C. Martin, M. McTear, T. Raman, K. M. Stanney, H. Su, Q. Y. Wang, J. Lai, J. A. Larson, S. Oviatt, T. S. Balaji, S. Buisine, P. Collings, P. Cohen, and B. Kraal, "Guidelines for multimodal user interface design," *Commun. ACM*, vol. 47, no. 1, p. 57, Jan. 2004.

[21] M. Hoedemaeker and M. Neerincx, "Attuning In-Car User Interfaces to the Momentary Cognitive Load," in *Foundations of Augmented Cognition*, Springer Berlin Heidelberg, 2007, pp. 286–293.

[22] A. Mital and M. Govindaraju, "Is it possible to have a single measure for all work?," *Int. J. Ind. Eng.-Theory*, vol. 6, no. 3, pp. 190–195, 1999.

[23] J. Beatty, "Task-evoked pupillary responses, processing load, and the structure of processing resources," *Psychol. Bull.*, vol. 91, no. 2, pp. 276–292, Mar. 1982.

[24] J. Klingner, R. Kumar, and P. Hanrahan, "Measuring the task-evoked pupillary response with a remote eye tracker," in *Proceedings of the 2008 symposium on Eye tracking research & applications*, 2008, pp. 69–72.

[25] B. Mehler, B. Reimer, and J. F. Coughlin, "Sensitivity of Physiological Measures for Detecting Systematic Variations in Cognitive Demand From a Working Memory Task: An On-Road Study Across Three Age Groups," *Hum. Factors*, vol. 54, no. 3, pp. 396–412, Apr. 2012.

[26] A. Koenig, D. Novak, X. Omlin, M. Pulfer, E. Perreault, L. Zimmerli, M. Mihelj, and R. Riener, "Real-time closed-loop control of cognitive load in neurological patients during robot-assisted gait training," *IEEE Trans. Neural. Syst. Rehabil. Eng.*, vol. 19, no. 4, pp. 453–64, Aug. 2011.

[27] P. Antonenko, F. Paas, R. Grabner, and T. Gog, "Using Electroencephalography to Measure Cognitive Load," *Educ. Psychol. Rev.*, vol. 22, no. 4, pp. 425–438, Apr. 2010.

[28] D. Grimes, D. S. Tan, S. E. Hudson, P. Shenoy, and R. P. N. Rao, "Feasibility and pragmatics of classifying working memory load with an electroencephalograph," *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, pp. 835–844, 2008.

[29] J. a. Healey and R. W. Picard, "Detecting Stress During Real-World Driving Tasks Using Physiological Sensors," *IEEE Trans. Intell. Transp. Syst.*, vol. 6, no. 2, pp. 156–166, Jun. 2005.

[30] E. T. Solovey, M. Zec, E. A. Garcia Perez, B. Reimer, and B. Mehler, "Classifying driver workload using physiological and driving performance data," *Proc. SIGCHI Conf. Hum. Factors Comput. Syst*, pp. 4057–4066, 2014.

[31] G. Wilson and C. Russell, "Real-time assessment of mental workload using psychophysiological measures and artificial neural networks," *Hum. Factors*, vol. 298, no. 074, 2003.

[32] A. Knoll, Y. Wang, F. Chen, J. Xu, and N. Ruiz, "Measuring cognitive workload with low-cost electroencephalograph," in *Proc. IFIP TC 13 Int. Conf. Human-Computer Interaction*, pp. 568–571, 2011.

[33] P. W. M. Van Gerven, F. Paas, J. J. G. Van Merrienboer, H. G. H. G. Schmidt, and J. J. J. G. G. Van Merriënboer, "Memory load and the cognitive pupillary response in aging," *Psychophysiology*, vol. 41, no. 2, pp. 167–174, Mar. 2004.

[34] C. Grady, "The cognitive neuroscience of ageing.," *Nat. Rev. Neurosci.*, vol. 13, no. 7, pp. 491–505, Jul. 2012.

[35] M. Angevaren, G. Aufdemkampe, H. J. J. Verhaar, A. Aleman, and L. Vanhees, "Physical activity and enhanced fitness to improve cognitive function in older people without known cognitive impairment," *Cochrane Database Syst. Rev.*, no. 2, p. Art. No.: CD005381, Jan. 2008.

[36] P. L. Ackerman and A. T. Cianciolo, "Cognitive, perceptual-speed, and psychomotor determinants of individual differences during skill acquisition," *J. Exp. Psychol.-Appl.*, vol. 6, no. 4, pp. 259–290, 2000.

[37] E. M. Tucker-Drob and T. A. Salthouse, "Adult age trends in the relations among cognitive abilities," *Psychol. Aging*, vol. 23, no. 2, pp. 453–460, Jun. 2008.

[38] J. Eliot and I. Smith, *An international directory of spatial tests*. Windsor. Berks: The NFER-NELSON Publishing Company Ltd., 1983.

[39] H. H. Jasper, "The 10-20 electrode system of the international federation," *Electroencephalogr. Clin. Neurophysiol.*, vol. 10, pp. 371–375, 1958.

[40] J. C. Lee and D. S. Tan, "Using a low-cost electroencephalograph for task classification in HCI research," *Proc. 19th Annu. ACM Symp. User interface Softw. Technol.*, p. 81, 2006.

[41] S. Chernenko, "ECG processing - R-peaks detection," *Librow TM*, 2012. [Online]. Available: www.librow.com. [Accessed: 14-Feb-2013].

[42] NeuroSky, "ThinkGear API and Reference Manual." NeuroSky, 2009.

[43] G. Rebolledo-mendez, I. Dunwell, E. A. Martínez-mirón, and F. Liarokapis, "Assessing NeuroSky's Usability to Detect Attention Levels in an Assessment Exercise," pp. 1–10.

[44] G. M. Friesen, T. C. Jannett, M. a Jadallah, S. L. Yates, S. R. Quint, and H. T. Nagle, "A comparison of the noise sensitivity of nine QRS detection algorithms," *IEEE Trans. Biomed. Eng.*, vol. 37, no. 1, pp. 85–98, Jan. 1990.

[45] R. W. Picard, E. Vyzas, and J. Healey, "Toward machine emotional intelligence: analysis of affective physiological state," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 10, pp. 1175–1191, 2001.

[46] J. Mietus, C. Peng, and I. Henry, "The pNNx files: re-examining a widely used heart rate variability measure," *Heart*, vol. 88, pp. 378–380, 2002.

[47] M. Malik, J. T. Bigger, A. J. Camm, R. E. Kleiger, A. Malliani, A. J. Moss, and P. J. Schwartz, "Heart rate variability: Standards of measurement, physiological interpretation, and clinical use," *Eur. Heart J.*, vol. 17, no. 3, pp. 354–381, 1996.

[48] T. F. Paymans, J. Lindenberg, and M. Neerincx, "Usability trade-offs for adaptive user interfaces," in *Proc. Int. Conf. Intelligent user interfaces*, p. 301, 2004.