The Index of Pupillary Activity

Measuring Cognitive Load *vis-à-vis* Task Difficulty with Pupil Oscillation

Andrew T. Duchowski
School of Computing
Clemson University
duchowski@clemson.edu

Krzysztof Krejtz[1,2] **& Izabela Krejtz**[1]
[1]SWPS University of Social Sciences & Humanities
[2]Ulm University
{kkrejtz|ikrejtz}@swps.edu.pl

Cezary Biele & Anna Niedzielska
Interactive Technologies Laboratory
OPI-PIB
{cbiele|aniedzielska}@opi.org.pl

Peter Kiefer[3]**, Martin Raubal**[3] **& Ioannis Giannopoulos**[3,4]
[3]Institute of Cartography and Geoinformation, ETH Zürich
[4]Vienna University of Technology
{pekiefer|mraubal}@ethz.ch, igiannopoulos@geo.tuwien.ac.at

ABSTRACT
A novel eye-tracked measure of the frequency of pupil diameter oscillation is proposed for capturing what is thought to be an indicator of cognitive load. The proposed metric, termed the Index of Pupillary Activity, is shown to discriminate task difficulty *vis-à-vis* cognitive load (if the implied causality can be assumed) in an experiment where participants performed easy and difficult mental arithmetic tasks while fixating a central target (a requirement for replication of prior work). The paper's contribution is twofold: full documentation is provided for the calculation of the proposed measurement which can be considered as an alternative to the existing proprietary Index of Cognitive Activity (ICA). Thus, it is possible for researchers to replicate the experiment and build their own software which implements this measurement. Second, several aspects of the ICA are approached in a more data-sensitive way with the goal of improving the measurement's performance.

Author Keywords
pupillometry; eye tracking; task difficulty

ACM Classification Keywords
H.1 Models and Principles: User/Machine Systems; J.4 Computer Applications: Social and Behavioral Sciences

INTRODUCTION
Systems that can detect and respond to their users' cognitive load have the potential to improve both users' experiences and outcomes in many domains: students and teachers, drivers, pilots, and surgeons may all benefit from systems that can detect when their jobs are too hard or easy and dynamically adapt the difficulty [3, 20, 41, 71, 11]. Key to this functionality is the ability to accurately estimate a person's cognitive load without distracting them from their tasks.

CHI 2018, April 21–26, 2018, Montréal, Québec, Canada.
Copyright © 2018 ACM ISBN 978-1-4503-5620-6/18/04...$15.00.
https://doi.org/10.1145/3173574.3173856

Estimation of human workload is couched in Cognitive Load Theory (CLT) [65]. Because CLT aims to model cognitive aspects of human behavior, it is relevant to several Human-Computer Interaction (HCI) research areas, including human-centered design, human cognition modeling, usability, and learning systems (e.g., e-learning) [48, 24]. Estimating the user's workload is helpful for many situations where people interact with computing devices or machines [20]. Minimizing cognitive load is suggested as an integral part of human-centered design [10]. Pfleging et al. [53] and Palinko and Kun [50] provide notable examples related to HCI, including automotive and online learning domains. Bailey and Iqbal [3] show how moment-to-moment detection of mental workload can help reduce the interruption cost of notifications when performing interactive tasks such as driving. Other important applications include surgery [28, 29] and flight safety [52].

Cognitive Load Theory can play an important role in the design of interactive systems as it can guide designers of such systems to avoid overloading users. For example, Yuksel et al. [71] devised an interactive music learning interface that adapts to the user's level of cognitive load as measured by functional near-infrared spectroscopy (fNIRS). They note, however, that reliable measurement of cognitive load is the weak link between CLT and HCI. Other physiological measures include heart rate variability (HRV), electrodermal activity (EDA, previously galvanic skin response (GSR)), photoplethysmogram-based stress induced vascular index (sVRI), and blink rate [9]. With the exception of blink rate, all of these methods are invasive, relying on physical contact with the user. A non-invasive, reliable measure of cognitive load is thus highly desirable.

Of the three predominant cognitive load measurement methods in CLT studies, namely self-reporting, the dual-task paradigm, and physiological measures [71], eye tracking, of the latter type, offers the greatest potential for delivering a non-invasive estimate of cognitive load (for an excellent recent review of psychophysiological measures with a focus on HCI, see Cowley et al. [11]). Measurement of gaze for estimating cognitive

load holds promise, as eye movements are both known to correlate with cognitive activity and likely to become more widely available to computer systems as gaze tracking tools are developed for both commercial-grade web cameras and for devices that support augmented and virtual reality interfaces. Indeed, because of the long-standing association of pupil dilation with cognitive load [1], eye trackers have received a good deal of attention recently as they record pupil diameter as a matter of course. However, a commonly-proposed metric for estimating cognitive load, pupil diameter, suffers from severe practical limitations related to factors unrelated to cognitive load, notably ambient light [6] and camera angle, manifested by off-axis distortion of the imaged pupil as captured by the eye tracker [46]. Whether an eye tracker is appropriate for measuring cognitive load is thus worthy of investigation. Using a high-speed eye tracker, we develop a mathematical estimate of cognitive load based on pupil oscillation, then study the feasibility and accuracy of its use.

Paper Overview and Contributions

We briefly review Cognitive Load Theory and focus on its historical connection to pupil diameter prior to the use of an eye tracker. We conclude the review of CLT by summarizing the relationship between eye-tracked fixations and cognitive activity. In the section on related work, we then summarize the primary method of cognitive load estimation by an eye tracker: computation of the averaged difference in pupil diameter with respect to an (averaged) baseline measurement. We then highlight the chief technical limitations of such baseline-related pupillometric measures owing to ambient illumination and off-axis distortion. We list several compensatory approaches that readers may not be aware of. We then focus on a wavelet-based estimate of the frequency of pupil diameter oscillation (known as *hippus* or *pupil unrest*), popularized by Marshall [43] as the Index of Cognitive Activity (ICA). This moment-to moment pupil diameter measurement is an alternative to pupillometric baseline-related difference measures. We review the ICA and then provide full documentation of our version's implementation, termed the Index of Pupillary Activity (IPA). We then present an experiment where we show how the IPA discriminates between task difficulty (*vis-à-vis* cognitive load).

Because the ICA procedure is not fully documented, no independent verification of the method appears to exist (even though it is implemented for many eye trackers). The contribution of the current paper is twofold: First, in contrast to the proprietary ICA, we provide full documentation of how to calculate our proposed measurement, which can be considered as an alternative to the existing Index of Cognitive Activity. Thus, it is possible for researchers to replicate our experiment and build their own software which implements this measurement tool. Second, we approach several aspects of the ICA procedure in a more data-sensitive way with the goal of improving the sensitivity of the measurement.

BACKGROUND: COGNITIVE LOAD MEASURES

We start with a brief, chronological summary of the origins of pupillometric measures related to elicited mental activity. Pupillometric measures of cognitive load are couched in Cognitive Load Theory (CLT) [64]. Sweller [65] introduced a computational model of cognitive load based on a production system and advocated a dual-task paradigm as a means of its performance-based (but indirect) measurement. Sweller's goal was to explain how individuals acquire and store information, thus linking CLT to the use of short and long term memory. It is understood that CLT recognizes the concept of cognitive load as a crucial factor in the learning of complex cognitive tasks—see Paas et al. [49] for a review of CLT and its components. Of particular relevance to the present work is estimation of cognitive load through measurement of pupil diameter.

Cognitive Load and Pupil Diameter

One of the most popular measures to assume indication of cognitive load is pupil diameter. This assumption can be traced back to Hess and Polt [23], who demonstrated correlation between pupil dilation and problem difficulty, i.e., showing that pupil size increases with problem difficulty. In a follow-on study to Hess and Polt's, Kahneman and Beatty [33] suggested that pupil diameter provided a "very effective index of the momentary load on a subject as they perform a mental task." Ahern and Beatty [1] referred to the metric as Task-Evoked Pupillary Response (TEPR). In their review of TEPR, Beatty and Lucero-Wagoner [6] noted that "it has long been recognized that a relationship exists between cognitive load and pupil diameter". Generally, more difficult problems evoke larger pupillary dilations, suggesting a relationship between problem difficulty and task-evoked activation. Differences in TEPR are thought to reflect differences in central, rather than peripheral, brain processes. TEPR is due to the psychosensory stimulus itself, producing the observed pupillary dilation also known as the *pupillary* (or *psychosensory* or *dilation*) *reflex*, with no differences observed in the pupils' *light reflex* [6]. An extensive review of TEPR is given by Beatty [5].

Cognitive Load and Eye-tracked Measures

Eye trackers were not initially used for pupillometric analysis. Early studies of task-evoked pupillary response relied on the use of specialized pupillometers to measure pupil diameter. Traditional eye tracking metrics related to cognitive load have implicated fixations, e.g., their number and duration. Fixation duration as an indicator of task difficulty was recognized early on by Fitts et al. [17], who noted fixation duration as an indicator of difficulty of information extraction and interpretation (task difficulty in essence). Jacob and Karn [27] observed that prior to the 1970s, psychologists who studied eye movements generally avoided cognitive factors (e.g., learning, memory, workload, etc.). Work on the relationship between fixations and cognitive activity began with improved eye tracking technology, resulting in rudimentary models based on fixations. Just and Carpenter [30] suggested that during cognitive tasks such as mental rotation, sentence verification, and quantitative comparison, fixation duration is proportional to the duration of underlying cognitive operations. They later posited their *eye-mind assumption*, which states that the eye remains fixated on the stimulus so long as it is being processed [31]. Other fixation-related measures e.g., in the context of reading, include total fixation time and number of regressions [2].

Debue and van de Leemput [12] suggest that eye-related measures have become one of the most cost-effective of physio-

logical methods for monitoring user attention, processing demands, and mental workload. They do not, however, advocate eye-related measures exclusively, noting the importance of subjective ratings (e.g., the NASA Task Load Index, or NASA-TLX, an assessment of perceived workload), performance-based measures, and physiological measures. Eye movement metrics include fixation durations and saccade length, e.g., longer fixations and shorter saccades may suggest increased cognitive load (e.g., focal attention [67, 37]). Positional eye movements during fixation known as microsaccades have also been suggested as potential indicators of task difficulty (increased task difficulty is reflected by reduced microsaccade rate and increased microsaccade magnitude) [61]. Blink rate and pupillary response are also implicated [10].

RELATED WORK: EYE TRACKING THE PUPIL

Because of the recent proliferation of eye trackers, due to their improvement in accuracy and reduction in cost, interest has turned to these devices for estimation of cognitive load, or at least task difficulty, via measurement of pupil diameter, which eye trackers report as a matter of course [54, 9, 10, 53].

The general approach to cognitive load estimation with eye-tracked pupil diameter data relies on measurement relative to a baseline. Measurement of the change in pupil diameter in relation to its baseline is performed due to the assumed correspondence between its *tonic* and *phasic* components. TEPR is assumed to correspond to the pupil's phasic response, while the baseline measurement is assumed to correspond to the pupil's tonic response, its sustained component of pupillary response [52]. The pupil's phasic response refers to a transient component, expressed as dilation relative to the baseline. Numerous examples of eye-tracked baseline-related pupil diameter measurements exist, focusing either on inter- [25, 35, 38, 34], or intra-trial baseline differences [5, 54, 36, 10, 29].

Baseline-Related Pupillometric Measures: Problems

One problem with eye-tracked baseline-related measures is the pupil's sensitivity to illumination levels found in the given visual stimulus. Some studies fail to report illumination measurements although luxmeters are not particularly expensive. Often it is simply assumed that pupil diameter (or more correctly relative pupil diameter) is representative of cognitive load regardless of the nature of the stimulus.

An additional problem is that pupil diameter, as measured by an eye tracker, undergoes significant variation upon movement of the eye. This is because from the eye tracker camera's (usually fixed) perspective, the pupil appears as an ellipse when the eye is rotated away from the camera's visual axis. The distortion has been modeled empirically by Mathur et al. [46] as a function of the cosine of the viewing angle θ (in degrees), i.e., $y(\theta) = R^2 \cos([\theta + 5.3]/1.121)$, where $R^2 = 0.99$, and y is the viewing-angle-dependent ratio of the ellipse major and minor axes. When off-axis, the apparent dimension of the pupil can be diminished by as much as 12% potentially impacting pupil diameter measurement and interpretation. Baseline-related difference measurements should therefore calibrate pupil diameter when looking at the screen center, with pupil diameter adjusted by a factor proportional to the angle (θ) that the eye is rotated away from center. The idea is that off-axis pupil diameter measurements should be compared with those of the pupil at center, to compensate for the off-axis distortion.

Reports of eye-tracked pupil diameter, e.g., by researchers (e.g., Chen and Epps [9]), or by eye tracking manufacturers, often do not consider off-axis compensation. One eye tracking manufacturer does report the problem in their manual, warning that pupil size may be affected by up to 10% by pupil position due to optical distortion of the cornea of the eye and camera-related factors [62]. In fact, it is suggested in this manual that if research using pupil size is to be performed, the subject should not move their eyes during trials. This clearly poses a problem not only in interpretation of changing pupil diameter, especially when one considers that reported changes are often quite small (e.g., tenths of millimeters), but also for experimental design where restricting gaze position to a central target borders on being highly impractical (not to mention extremely limited in terms of ecological validity).

Eye-tracked measures of the pupil hold promise, so long as effects of illumination and off-axis image distortion are taken into account, i.e., when the eye is free to move [46]. An example of compensating for off-axis position of the eye is given by Hayes and Petrov [22], who incorporate Mathur et al.'s [46] empirically derived foreshortening and suggest using $\omega_0 = \omega / \sqrt{0.992 \cos([\theta + 5.3]/1.121)}$ where ω_0 denotes the angle subtended by the pupil diameter in the baseline configuration with ω the pupil's apparent angle.

Susceptibility to luminance could be handled by modeling the brightness induced pupil diameter change as a function of the intensity in the foveal neighborhood around gaze position, as shown by Raiturkar et al. [56]. Similarly, Palinko and Kun [50] show that it is possible to separate the effects of illumination and visual cognitive load on pupil diameter by cleverly subtracting the averaged pupil diameter difference from baseline trials where illumination is purposefully varied. Alternatively, the pupil diameter signal can be transformed to the frequency domain, e.g., either via the Low Frequency/High Frequency (LF/HF) ratio [52], or our IPA. One advantage of the IPA is that in its reliance on the Discrete Wavelet Transform, it offers analysis at multiple frequency scales whereas Fourier transform-based techniques such as the LF/HF do not.

Our tenet is that eye-tracked baseline-related pupil measures are problematic due to luminance and camera angle. We provide an alternative metric based on *relative moment-to-moment* pupil size, inspired by Marshall's [44] proprietary (closed-source) Index of Cognitive Activity. We hypothesize that our Index of Pupillary Activity is sensitive and directly proportional to task difficulty.

The Index of Cognitive Activity (ICA)

As an alternative to pupillometric baseline-related measures, the Index of Cognitive Activity (ICA) is an instantaneous measure of pupil diameter, i.e., a measure of the fluctuation of the diameter, not of the difference relative to a baseline. Said another way, the ICA is a measure of the rate of change of pupil diameter, and not a difference between averages (e.g., as detailed by Chen and Epps [10], among others). What is

important is the moment-to-moment change in pupil diameter, regardless of gaze position.

The pupil of the human eye continuously undergoes small fluctuations in area, even in steady illumination—this is known as *pupillary hippus* or *pupil unrest* [63]. Marshall [44] notes that in the presence of effortful cognitive processing, the pupil responds rapidly with a reflex reaction (the *psychosensory* or *dilation reflex*). At the same time, the pupil responds with a reflex reaction to light changes (i.e., the *light reflex*). Marshall [44] developed the ICA based on the assumption that an increase in the appearance of abrupt discontinuities in the signal created from continuous recording of the pupil diameter are representative of increased cognitive load.

The ICA is claimed to successfully separate the light reflex from the dilation reflex, hence how it is computed, and its purported advantages over baseline-related pupil diameter measures, are worth reviewing. Details about the ICA's computation unfortunately refer to an unpublished manuscript. Still, some details can be found in a patent [43] and a report to the U.S. Air Force Office of Sponsored Research (AFOSR) [7].

According to Boehm-Davis et al. [7], because the ICA always reflects the same ratio—the frequency of occurrence per second—it provides a common basis for comparing individuals, groups of individuals, single events, and multiple events. They note that it is useful to examine the average ICA across the entire time period. Typical index values range from 0-20 Hz, with low values reflecting little cognitive effort and high values indicating strong cognitive effort [45]. Bartels and Marshall [4] claim the ICA is reliable across hardware platforms and sampling rates. An example of ICA usage can be found in its analysis in a driving simulator study [59].

Apart from our implementation of the IPA, we are not aware of any other attempts at replication of the ICA. In effect, the ICA has apparently gained adoption as an indicator of cognitive load without independent verification. Inclusion of the ICA module by eye tracking vendors in their own software is evidence of this adoption. Its commercial offering has led to user adoption, as indicated by an online search for publications utilizing the method, but its implementation details are still hidden due to its proprietary nature. The IPA is thus our offering of a similar but different and fully detailed alternative.

IMPLEMENTATION OF THE IPA

Prior to implementation of our Index of Pupillary Activity (IPA), eye movement data is first extracted in a pre-processing step to remove data 200 ms before the start of, and 200 ms following the end of a blink, as identified by the eye tracker, following Engbert and Kliegl [14]. After this pre-processing step, we then compute the IPA.

The IPA is a wavelet-based algorithm inspired by Marshall's [43] Index of Cognitive Activity. Our approach differs in certain key aspects from Marshall's patent, however, namely in choice of wavelet, use of the modulus maxima, and a different thresholding approach.

Computation of both the ICA and IPA relies on wavelet decomposition of the pupil diameter signal, $x(t)$, and its wavelet analysis [43, 7]. Marshall suggests that one needs to locate peaks in the wavelet detail (coefficient) signal to localize significant changes; she suggests doing so following "de-noising" of the coefficient signal via minimax thresholding (using hard thresholding, as found in Matlab code given in the patent). The choice of wavelet is important, e.g., for a 60 Hz signal, the Daubechies-4 wavelet is recommended, since it is of length 8 utilizing a $8\times16=134$ ms sampling window, whereas the Daubechies-16 wavelet (with 32 coefficients) is recommended for data sampled at 250 Hz (utilizing a comparable $32\times4=128$ ms sampling window). Each of the ICA and IPA is then computed as the frequency (per second) of abrupt discontinuities detected in the signal [44].

We now provide details of the computation of the IPA through wavelet analysis, which relies on selection of a mother *wavelet function* $\psi_{j,k}(t)$ expressed by

$$\psi_{j,k}(t) \;=\; 2^{j/2}\psi(2^j t-k), \quad j,k\in\mathbf{Z}, \tag{1}$$

with wavelet dilation and translation parameters j,k, respectively. The process of wavelet analysis of the signal $x(t)$ then proceeds via the wavelet transform, where the dyadic wavelet $\psi_{j,k}(t)$ generates a dyadic series representation of $x\in L^2(\mathbf{R})$: $x(t)=\sum_{j,k=-\infty}^{\infty} c_{j,k}\psi_{j,k}(t),\ j,k\in\mathbf{Z}$, with wavelet coefficients $\{c_{j,k}\}$ given by the integral transform (see Graps [19] or Duchowski [13] for a review):

$$\begin{aligned} c_{j,k} &= 2^{j/2}\int_{-\infty}^{\infty} x(t)\overline{\psi(2^j t-k)}dt,\ x\in L^2(\mathbf{R}),\ j,k\in\mathbf{Z}, \\ &= \{W_\psi x(t)\}(j,k) \;=\; \langle x(t),\psi_{j,k}(t)\rangle. \end{aligned} \tag{2}$$

At the heart of this process is *multiresolution* signal analysis, which necessarily involves the use of a *scaling function*, denoted by $\phi_{j,k}(t)$. In fact, multiresolution analysis starts with a clever choice of $\phi_{j,k}(t)$. The scaling function is very similar in nature to the wavelet in that it also spans the same subspace as $\psi_{j,k}(t)$. It is chosen to satisfy continuity, smoothness, and tail requirements, and, most importantly, the family $\phi_{j,k}(t-k), k\in\mathbf{Z}$ forms an orthonormal basis for the multiresolution *reference* space (see Vidaković and Müller [68] for an introduction to multiresolution analysis). The scaling function is also a compactly supported function, defined as

$$\phi_{a,b}(t)=\frac{1}{\sqrt{a}}\phi(\frac{t-b}{a}), \quad a>0, b\in\mathbf{R},$$

where again a,b are the dilation and translation parameters. As for the wavelet function, integral powers of 2 are used where the scaling function is obtained by a binary dilation (2^j), and a dyadic translation ($k/2^j$) of a single function ϕ. That is, a,b are chosen as for the wavelet function, and the scaling function becomes (c.f. (1))

$$\phi_{j,k}(t)=2^{j/2}\phi(2^j t-k), \quad j,k\in\mathbf{Z}.$$

In practice, the Discrete Wavelet Transform (DWT) is used to analyze the signal at multiple levels of resolution. Given an n-length discrete function at the j^{th} level of resolution, $x^j(t)=x^j_\phi(1),x^j_\phi(2),\ldots,x^j_\phi(n)$, the decomposition relations

of the function are:

$$x_{\phi}^{j-1}(t) = \sum_k h_k x_{\phi}^{j}(2t+k), \;\; x_{\psi}^{j-1}(t) = \sum_k g_k x_{\phi}^{j}(2t+k),$$

where $\{h_k\}, \{g_k\}$ are one-dimensional low- and high-pass wavelet filters. This gives the discrete wavelet transform:

$$\{Wx(t)\}(j-1) = x_{\phi}^{j-1}(1), x_{\psi}^{j-1}(2), \ldots, x_{\psi}^{j-1}(n). \qquad (3)$$

With this multiresolution decomposition of the original function $x(t)$, level j (or octave) can be chosen arbitrarily, giving a progressively smoother approximation of $x(t)$ along with corresponding wavelet coefficients (akin to residuals) at level j, expressed respectively by $x_{\phi}^{j-1}(t)$ and $x_{\psi}^{j-1}(t)$.

We use wavelet analysis of the pupil diameter signal at the second level of resolution, as high-frequency oscillations are likely to reflect the high frequency changes associated with pupillary hippus [6].

The last stage of IPA computation is performed by hard thresholding (decimation of) wavelet coefficients and counting up those that remain. Decimating wavelet coefficients below threshold leaves single peaks in the resulting signal. Hard thresholding, as suggested by Marshall [43], however, while reducing the number of events in the signal, does so in a rather uninformed way, without considering the information contained within the signal. A more meaningful approach is to first seek sharp points of variation in the signal, e.g., edges, since these events indicate where abrupt changes in pupil diameter occur. For the IPA, these events are found by detecting the local maxima of the wavelet modulus. Choosing the resolution at which these modulus maxima are identified will select the rate of pupil diameter oscillation considered most interesting (we choose the second resolution level, as noted above).

Sharp variation points are detected by finding the local maxima of the modulus $|\langle x(t), \psi_{j,k}\rangle|$ (i.e., the modulus of (2)). At each scale j, local modulus maxima are found where $|\langle x(t), \psi_{j,k}\rangle|$ is larger than its two closest neighbors, and strictly larger than at least one of them [42]. That is, modulus maxima are located at scale j and location (t_0) if:

$$|\langle x(t_0-1), \psi_{j,k}\rangle| \;\leq\; |\langle x(t_0), \psi_{j,k}\rangle| \;\geq\; |\langle x(t_0+1), \psi_{j,k}\rangle|$$

and

$$\begin{cases} |\langle x(t_0), \psi_{j,k}\rangle| > |\langle x(t_0-1), \psi_{j,k}\rangle|, & \text{or} \\ \\ |\langle x(t_0), \psi_{j,k}\rangle| > |\langle x(t_0+1), \psi_{j,k}\rangle|. \end{cases}$$

The modulus maxima of the wavelet transform at scale j and location (t_0) are strict local maxima of the modulus on the right or the left of location t_0. Following modulus maxima detection, instead of minimax thresholding, as suggested by Marshall [43], we threshold the wavelet modulus maxima coefficients via "universal thresholding", defined as $\lambda_{univ} = \hat{\sigma}\sqrt{2\log n}$ with $\hat{\sigma}$ the standard deviation of the noise [26].

Unlike Marshall, we use *symlet*-16 wavelets instead of Daubechies wavelets, and use the periodic DWT implemented in Python's `pywt` module.[1] As Marshall suggests, we then count the number of remaining coefficients following modulus maxima detection and universal thresholding to produce the IPA as a frequency count of coefficients per second. Marshall suggests low counts reflect little cognitive effort while high counts indicate strong cognitive effort.

[1] `http://pywavelets.readthedocs.io`

```
import math, pywt, numpy as np

def ipa(d):
  # obtain 2-level DWT of pupil diameter signal d
  try:
    (cA2,cD2,cD1) = pywt.wavedec(d,'sym16','per',level=2)
  except ValueError:
    return

  # get signal duration (in seconds)
  tt = d[-1].timestamp() - d[0].timestamp()

  # normalize by 1/2^j, j=2 for 2-level DWT
  cA2[:] = [x / math.sqrt(4.0) for x in cA2]
  cD1[:] = [x / math.sqrt(2.0) for x in cD1]
  cD2[:] = [x / math.sqrt(4.0) for x in cD2]

  # detect modulus maxima, see Listing 2
  cD2m = modmax(cD2)

  # threshold using universal threshold λ_univ = σ̂ sqrt(2 log n)
  # where σ̂ is the standard deviation of the noise
  λ_univ = np.std(cD2m) * math.sqrt(2.0*np.log2(len(cD2m)))
  cD2t = pywt.threshold(cD2m,λ_univ,mode="hard")

  # compute IPA
  ctr = 0
  for i in xrange(len(cD2t)):
    if math.fabs(cD2t[i]) > 0: ctr += 1
  IPA = float(ctr)/tt

  return IPA
```

Listing. 1. IPA implementation.

Python implementations of the IPA and of the modulus maxima detection are given in Listings 1 and 2, respectively.

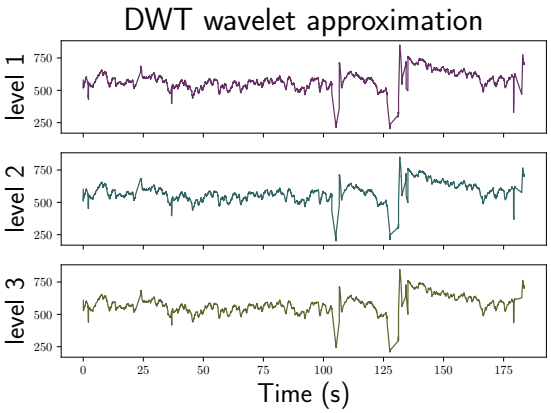
Example output is given in Figures 1(e) and 1(f), showing thresholded modulus maxima (at level 2), whose frequency per second is compared (in the aggregate) under different task difficulties. Thresholded modulus maxima are obtained from the wavelet coefficients in Figures 1(c) and 1(d), which can be thought of as residuals of the wavelet approximation to pupil diameter at three levels of resolution, in Figures 1(a) and 1(b).

EXPERIMENT

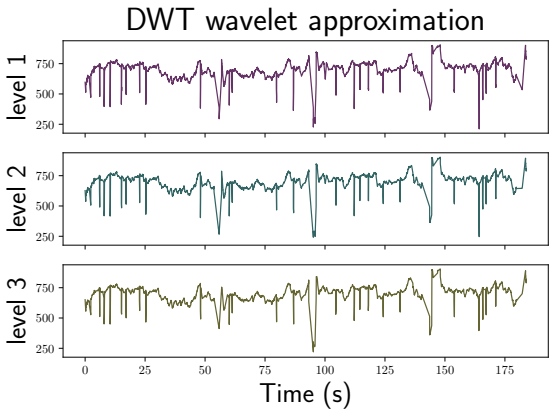
To evaluate the IPA, we conducted an eye tracking study, replicating Siegenthaler et al.'s [61] experimental design. Details of our study methodology are given below, including experimental design with independent and dependent measures, procedure, participants, equipment, and analyses. Our study hypothesis was that the IPA would be sensitive and directly proportional to task difficulty *vis-à-vis* cognitive load.

Experimental Design and Factors

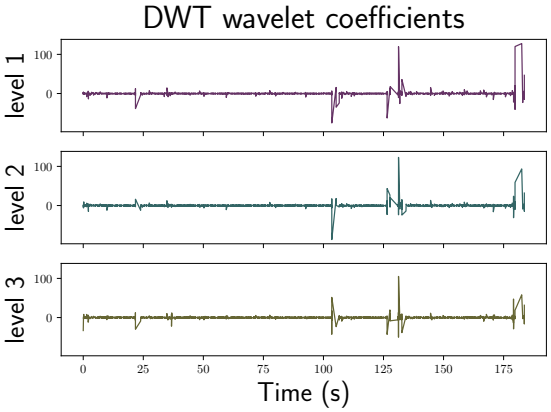
Following Siegenthaler et al. [61], the present study was a 3×6 within-subjects eye tracking experiment. The first fixed factor was task type (Difficult vs. Easy vs. Control). In the Difficult and Easy tasks, participants were asked to perform difficult and easy mental calculations, while in the Control task, they were not asked to perform any mental calculations



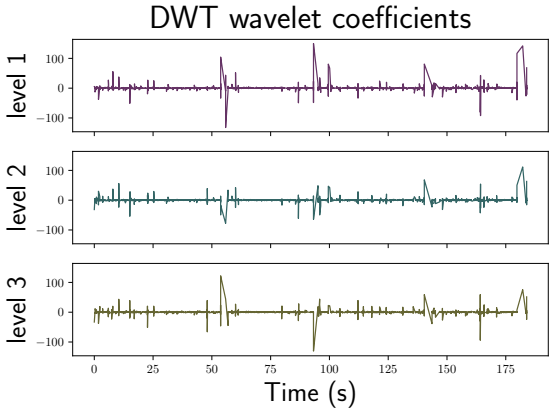
(a) Easy trial wavelet approximation $x_{\phi}^{j-1}(t)$.



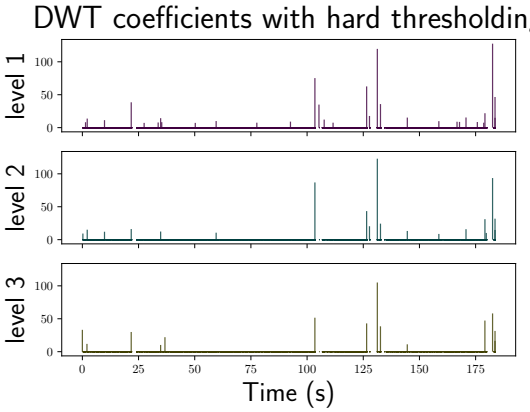
(b) Difficult trial wavelet approximation $x_{\phi}^{j-1}(t)$.



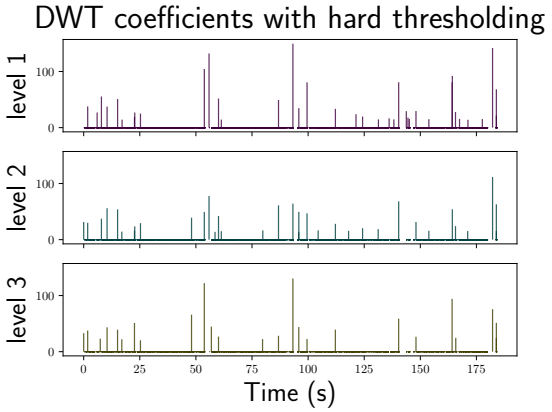
(c) Easy trial wavelet coefficients $x_{\psi}^{j-1}(t)$.



(d) Difficult trial wavelet coefficients $x_{\psi}^{j-1}(t)$.



(e) Easy trial wavelet thresholding $|\langle x(t), \psi_{j,k} \rangle|$.



(f) Difficult trial wavelet thresholding $|\langle x(t), \psi_{j,k} \rangle|$.

Figure 1. Representative 3-level wavelet decomposition of a single participant's pupil diameter when conducting Easy and Difficult trials. Compare in particular (e) with (f): a larger number of ticks per second is thought to indicate increased workload.

```
import math, pywt, numpy as np

def modmax(d):
  # compute signal modulus
  m = [0.0]*len(d)
  for i in xrange(len(d)):
    m[i] = math.fabs(d[i])

  # if value is larger than both neighbours, and strictly
  # larger than either, then it is a local maximum
  t = [0.0]*len(d)
  for i in xrange(len(d)):
    ll = m[i-1] if i >= 1 else m[i]
    oo = m[i]
    rr = m[i+1] if i < len(d)-2 else m[i]
    if (ll <= oo and oo >= rr) and (ll < oo or oo > rr):
      # compute magnitude
      t[i] = math.sqrt(d[i]**2)
    else:
      t[i] = 0.0

  return t
```

Listing. 2. Modulus maxima detection.

at all (see Experimental Procedure below). Six blocks of trials within the experimental procedure constituted six levels of the second fixed factor, termed Time-On-Task.

Working Memory Capacity (WMC). Each participant's WMC was treated as a controlled independent variable, measured with the Digit SPAN task (DSPAN) using both *Forward* and *Backward* assessment versions adopted from Woods et al. [70]. The last length of a correctly recalled numerical sequence (before making two consecutive errors) is used as an indicator of a participant's WMC. We used the mean value of the two-error maximum length DSPAN from *Backward* and *Forward* instantiations as a covariate in the statistical analysis of eye movement measures (see below for details).

Self-assessed cognitive load. We used the Raw NASA Task Load Index [21] (NASA-TLX) as a dependent measure of self-reported cognitive load. We used the following NASA-TLX items: *mental demand*, *physical demand*, *temporal demand*, *performance*, and *effort*. We dropped the *frustration* item as we deemed it irrelevant to the task. The TLX questionnaire was scaled from 1 ("Very Low") to 21 ("Very High").

Self-assessed emotional valence. The Self-Assessment Manikin (SAM) [8] was used to evaluate participants' arousal and emotional valence after each task. Participants responded to two questions regarding arousal and emotional valence assessing them on visual scales (ranging from 1 to 9) by moving a visual slider with the computer mouse. We skipped the dominance assessment deeming it irrelevant to the study.

Experimental Procedure

After signing a consent form, participants completed an online demographic questionnaire using *LimeSurvey* [39]. Next, each participant completed the DSPAN assessment, consisting of 14 trials. In each trial a participant saw a sequence of digits (starting with 3 digits), each presented for 1 second. After seeing the sequence, the participant was asked to recall the digit sequence (in the same order in the *Forward* assessment and in the reverse order in *Backward* assessment) by typing the sequence into a text box presented on the screen. Given a correct response, the digit sequence was extended by 1 digit in the next trial. Given an incorrect response, the length of the next sequence was kept the same.

After finishing with the DSPAN, participants sat at the eye tracker (an SR Research EyeLink 1000) with their head stabilized by a chin rest. After making sure participants felt comfortable with their body and head position, a 5-point eye tracker calibration was performed. Experimental tasks started when the average calibration error was lower than 0.5° visual angle (as measured by SR Research software).

The experimental procedure followed that of Siegenthaler et al. [61], described here for completeness. Three types of number counting trials, Difficult, Easy, and Control, were grouped into 6 blocks, giving 18 trials total. Each block started with the Control trial, followed by the Easy and Difficult trials in counterbalanced order, see Table 1. Between each block, participants were asked to take a short break lasting 2–5 minutes; they were not allowed to start the next block until at least 2 minutes had elapsed.

Each trial started with an instruction screen and included a break at the end of each of the six blocks (see Table 1). In the Difficult trials, participants were asked to mentally count backwards, as fast and accurately as possible, in steps of 17 starting at one of the following 4-digit numbers drawn randomly from this set: {1375, 8489, 5901, 5321, 4819, 1817}.

The Easy and Control trials were constructed similarly to Difficult trials, but differed in task performance and initial instructions. In the Easy tasks, participants were instructed to mentally count forward, as fast and accurately as possible, in steps of 2 starting at one of the following 3-digit numbers drawn randomly from this set: {363, 385, 143, 657, 935, 141}. In the Control trials, participants were asked just to gaze at the fixation point with no mental task assigned.

During each trial, participants were prompted four times to enter their current number in a text box on the screen. A limit of 9 seconds was given for providing the entry. Three prompts appeared at random times during each trial, and the fourth at the very end of the trial. The gap between prompts was a minimum of 15 seconds and a maximum of 80 seconds.

After each trial, the NASA-TLX and SAM evaluations were conducted (18 evaluations in all). Each trial lasted 3 minutes.

When performing the mental calculations, participants were asked to gaze at the fixation point appearing at screen center. Whenever their gaze shifted 3° visual angle away from the fixation point a warning beep sounded.

Table 1. Schematic representation of the experiment. following Siegenthaler et al. [61]. Block order was randomized for each participant.

Block	Trials (Tasks)			
1	Control	Easy	Difficult	Break
2	Control	Difficult	Easy	Break
3	Control	Difficult	Easy	Break
4	Control	Easy	Difficult	Break
5	Control	Easy	Difficult	Break
6	Control	Difficult	Easy	Break

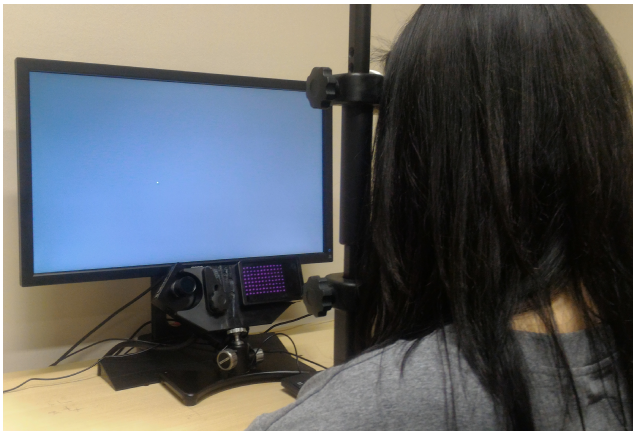


Figure 2. Eye-tracking apparatus with chin rest.

Response accuracy. We adopted the analytical procedure used by Siegenthaler et al. [61] to check response accuracy. First, the difference between the starting number or previously entered number and the present response was calculated. Correct responses in the Easy tasks were defined as any positive even difference. Correct responses in the Difficult tasks were defined as any negative difference divisible by 17. Correct responses in the Control tasks were defined as any three-digit numbers entered. We imposed a minimum performance criterion, requiring a minimum of 4 out of 24 correct answers in the Difficult tasks. Based on this criterion, the data of one participant, who only scored 3 correct answers in all of the Difficult tasks, were removed from further analyses. Additionally, if the number of correct responses in all of the Difficult tasks and the majority of responses in the Easy tasks was exactly 1, we treated such cases as a misunderstanding of the task.

Participants

Volunteers ($N = 17$) for the study were recruited verbally and by social media. Due to problems with eye tracker calibration or misunderstanding of the task, data from 4 were discarded giving a final sample of $N = 13$ (7 M, 6 F with ages in range [20:40] years old, $M = 29.77$, $SD = 7.15$). All were right-handed with normal, uncorrected vision.

Experimental Setting and Apparatus

An SR Research EyeLink 1000 eye tracker was used to record eye movements binocularly at a sampling rate of 500 Hz. Each participant's head was stabilized with a chin rest during the entire experimental procedure, see Figure 2. Eye tracker accuracy is reported by the manufacturer as 0.25–0.5° visual angle on average. Wang et al. [69] corroborate this accuracy measurement via root-mean-squared analysis. van der Geest and Frens [66] found the EyeLink's horizontal × vertical precision to be $0.98° \times 1.05°$ visual angle.

The experimental procedure was controlled by a personal computer connected to the eye-tracking computer. Visual stimuli were displayed on a computer screen with 1920×1080 resolution. The procedure was written in Python with the use of the PsychoPy package [51]. Responses made by participants were performed on a standard numerical keyboard connected to the stimuli presentation computer and placed at the side of the participant's dominant hand. A laptop was used for conducting the DSPAN assessment, written and conducted with Millisecond Inc.'s Inquisit 4 Lab software.

The experimental laboratory was devoid of windows limiting the amount of ambient light during the study. Ambient luminance in the laboratory was 520 lux. Luminance was measured at 120-130 lux at the computer screen with the fixation point target present at screen center during the main part of the procedure. Screen luminance was slightly higher when showing instructions, the NASA-TLX, and SAM measures (150 lux).

RESULTS

There are two main criteria of any measure's validity: internal and external validity, also referred to as reliability and sensitivity, respectively. A useful measure of cognitive load should be sensitive to both between-task and within-task variability as well as between-subjects differences [32]. First, we report the results of internal validity (reliability). Second, we present results focusing on external validity (sensitivity) reflected in the ability to distinguish between task difficulty within sequential blocks of trials representing Time-On-Task.

Statistical Analyses

Internal validity was assessed with Cronbach's α. To evaluate the influence of task difficulty and Time-On-Task on dependent variables, two-way (3×6) within-subjects Analyses of Covariance (ANCOVAs) were used, where task difficulty (Control vs. Easy vs. Difficult) and Time-On-Task (block of trials from 1 to 6) served as independent factors. The analyses of covariance were followed by pairwise comparisons with HSD Tukey correction when needed.

The literature points to the moderating role of working memory capacity in the relation between cognitive load and eye-related measures. For example, Granholm et al. [18] showed that pupillary response increases with increasing task demand until cognitive resources are exceeded, at which point pupillary response then begins to decline. Thus, we used working memory capacity as a covariant variable in our statistical analyses.

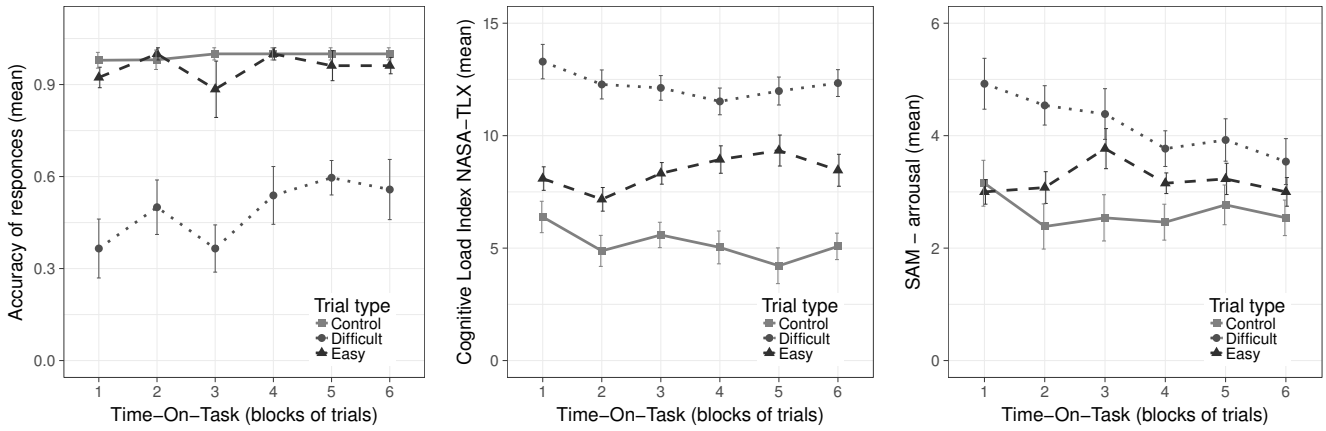
We used parametric ANCOVA despite the the IPA showing skewed distributions deviating from normality. However, ANCOVA (and ANOVA) is relatively robust to violation of the normality assumption [40, 58]. Also, ANCOVA allows for full design analyses which is the most appropriate option for hypothesis testing. All ANCOVA results are reported with main effect size (η^2). All analyses were conducted in R [55].

Reliability of Measures

Comparison of Cronbach's α (see Table 2) shows good reliability of the IPA and NASA-TLX self-reported measures. An $\alpha \geq 0.80$ is assumed to be acceptable [47]. The IPA showed excellent reliability ($\alpha > 0.90$) in the Difficult task.

Table 2. Internal consistency per trial given by Cronbach's α.

	Task		
Variable	**Control**	**Easy**	**Difficult**
NASA-TLX	0.83	0.87	0.77
IPA	0.89	0.82	0.91



(a) Proportion of correct responses. (b) NASA-TLX. (c) SAM arousal.

Figure 3. Manipulation check: response accuracy and self-assessed evaluation of effort. Means are plotted per trial over Time-On-Task, with error bars representing $\pm$ 1 SE for the means. With respect to the Difficult task, significantly lower mean response accuracy in (a) shows its relative difficulty; significantly higher mean NASA-TLX scores in (b) show its perceived relative difficulty; and significantly higher mean SAM arousal scores in (c) show its perceived relative greater emotional response.

Table 3. Descriptive statistics per trial: means and SE (in parentheses).

	Task		
Variable	**Control**	**Easy**	**Difficult**
Correct responses (proportion)	0.98 (0.02)	0.97 (0.02)	0.49 (0.04)
SAM valence	5.73 (0.11)	5.67 (0.12)	5.08 (0.13)
SAM arousal	2.64 (0.18)	3.21 (0.13)	4.18 (0.20)
NASA-TLX	5.19 (0.33)	8.39 (0.30)	12.26 (0.31)
IPA	0.19 (0.008)	0.19 (0.008)	0.22 (0.07)

Experimental Manipulation Check

Effectiveness of the experimental manipulation was examined via response accuracy (proportion of correct responses), Raw TLX, and both SAM scales, see Table 3. As expected, ANCOVA of response accuracy revealed a main effect of task, $F(2,22) = 64.65, p < 0.001, \eta^2 = 0.58$. Difficult tasks yielded significantly fewer correct responses than the Easy and Control tasks ($p < 0.001$), see Figure 3. The difference between the Control and Easy tasks was not significant ($p > 0.1$).

Analysis also revealed a significant main effect of Time-On-Task on accuracy, $F(5,55) = 2.42, p < 0.05, \eta^2 = 0.05$, see Figure 3(a) and compare mean accuracy between the first (M = 0.74, SE = 0.06) and last (M = 0.84, SE = 0.05) trial blocks. The interaction effect between WMC and Time-On-Task was significant, but weak, $F(2,2) = 4.76, p < 0.05, \eta^2 = 0.02$.

As expected, a two-way ANCOVA of the Raw NASA TLX scale revealed a significant main effect of task difficulty, $F(1.31, 14.46) = 46.09, p < 0.001, \eta^2 = 0.38$. Participants reported significantly higher cognitive load during the Difficult tasks than during the Easy and Control tasks ($p < 0.001$). Pairwise comparisons showed a statistically significant ($p < 0.01$) difference between the Easy and Control tasks, see Table 3.

ANCOVA of the Raw NASA TLX score also revealed a statistically significant interaction effect of task difficulty and Time-On-Task $F(4.64, 51.06) = 2.91, p < 0.05, \eta^2 = 0.02$, see Figure 3(b). Pairwise comparisons of means showed that in all blocks of trials the difference between Easy and Control tasks in the TLX score was significant ($p < 0.001$) with perceived higher workload for the Easy tasks in all but the first block of trials where the difference was not significant ($p > 0.1$). In all blocks of trials the differences between Difficult vs. Control and Difficult vs. Easy tasks were significant ($p < 0.001$) with greater cognitive load self-reported following Difficult tasks.

We expected participants would report higher arousal and lower emotional valence in the Difficult tasks. Both predictions were supported by two separate ANCOVAs of the SAM Valence and Arousal scales as dependent variables. The main effect of task difficulty on the SAM arousal scale, $F(1.46, 16.03) = 10.43, p < 0.01, \eta^2 = 0.14$, and subsequent post-hoc tests showed that, after the Difficult tasks, participants reported significantly higher emotional arousal than after the Control tasks ($p < 0.01$), see Figure 3(c). Similar but not statistically significant differences ($p = 0.052$) were recorded between the Difficult and Easy tasks. The difference between the Easy and Control tasks was not significant. ANCOVA also revealed a significant effect of WMC, $F(1,11) = 5.85, p < 0.05, \eta^2 = 0.21$. Participants with higher working memory capacity reported lower arousal than those with lower working memory capacity.

ANCOVA of SAM emotional valence also showed a main effect of task difficulty, $F(1.15, 12.67) = 8.86, p < 0.01, \eta^2 = 0.03$. Post-hoc analyses showed that when performing the Difficult tasks, participants evaluated their emotions significantly more negatively compared to the Easy and Control tasks ($p < 0.02$, $p < 0.01$, respectively), see Table 3.

Sensitivity of Pupillary Response to Task Difficulty

We hypothesized that the IPA should indicate differences in cognitive load, distinguishing between Difficult, Easy, and Control tasks. We expected the IPA to be significantly greater for the Difficult tasks. In order to test this hypothesis the IPA served as a dependent variable in a series of two-way within-

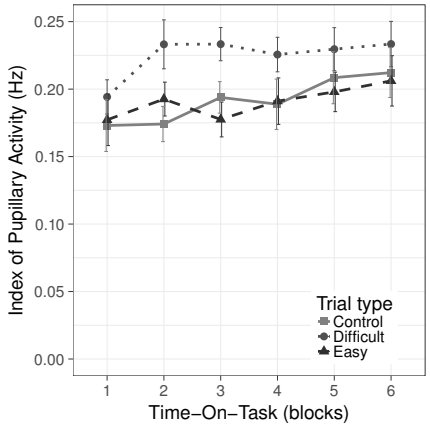


Figure 4. Pupil response to task difficulty, with mean IPA versus Time-On-Task; error bars represent ± 1 SE. Significantly greater IPA, especially in the 2nd and 3rd trials, shows relatively greater pupil oscillation.

subjects ANCOVAs with task difficulty and Time-On-Task as fixed factors. Working memory capacity served as a covariate.

Although the IPA failed the test for distribution normality, according to a one-sample Kolmogorov-Smirnov test, $D = 0.53$, $p < 0.001$, the IPA appears to increase significantly with task difficulty. ANCOVA of the IPA revealed a statistically significant main effect of task difficulty, $F(1.70, 18.73) = 3.73, p < 0.05, \eta^2 = 0.05$, see Figure 4. Pairwise comparisons, without correction, showed that the IPA differed significantly ($p < 0.05$) between the Difficult ($M = 0.22, SE = 0.07$) and each of the Easy ($M = 0.19, SE = 0.008$) and Control tasks ($M = 0.19, SE = 0.008$). The difference between the Easy and Control tasks was not significant ($p = 0.93$), see Table 3.

DISCUSSION

Results support our hypothesis of the IPA's sensitivity to task difficulty. Task difficulty itself was corroborated by measurement of accuracy (proportion of correct responses) and self-assessed evaluation of effort (NASA-TLX scores). Interestingly, analysis of the IPA's response to task difficulty did not reveal a significant effect of working memory capacity (WMC). According to the literature, WMC is related to performance of a variety of higher-order cognitive tasks e.g., reading comprehension, complex learning, and reasoning [15, 16]. Participants with high WMC are thus expected to experience lower cognitive load on difficult tasks than participants with low WMC. This may explain our observed relation between WMC and participants' response accuracy during performance of the mental arithmetic tasks. Participants with high WMC may have sufficient resources to handle cognitive task demands resulting in accurate performance. Meanwhile, cognitive task demands may exceed the resources of participants with low WMC causing degradation in performance. This would agree with Sweller's [65] original production-system model of cognitive load. Because WMC did not correlate significantly with the IPA, this may suggest that the IPA is sensitive to task difficulty independent of working memory capacity.

The IPA shows promise as an indicator of cognitive load, but requires further exploration. We used the second frequency octave in our wavelet analysis; different frequency resolutions remain to be examined. Future experiments are also needed to investigate the response of the IPA to eye movements, light conditions, as well as sampling rates.

IMPLICATIONS FOR INTERACTION DESIGN

There is a pressing need for a non-invasive measure of cognitive load, as it can guide designers of interactive systems to avoid overloading users. Examples of its use include a wide range of applications, including surgery [28], flight safety [52], human-centered design, human cognition modeling, usability, and learning systems (e.g., e-learning) [48, 24]. A reliable (real-time) measurement of cognitive load is sought.

Implications for interaction design are such that eye-tracked cognitive load measures should be used with care, especially those involving baseline differences of pupil diameter. Our study shows that pupillometric oscillation is indeed effective at distinguishing task difficulty, but our results are from an experiment where gaze was held fixed at screen center. Further testing of cognitive load measures is needed. Specifically, experiments must be carried out where the eye moves to controlled locations away from screen center.

CONCLUSION

Being able to distinguish a user's level of cognitive load has significant implications for design and/or evaluation of interactive systems. Measurement of cognitive load could allow a system to respond appropriately, modulating the level of task difficulty (e.g., as in e-learning systems [57]), or by adapting mission-critical systems to the user's cognitive state [60].

We reviewed Cognitive Load Theory and its connection to task-evoked eye movement measures, namely pupillary responses. We gave a novel method of estimating frequency of pupil oscillation termed the Index of Pupillary Activity. We have discussed the limitations of pupillometric measures and suggested measurement of pupil oscillation as an alternative for estimating task difficulty *vis-à-vis* cognitive load.

Acknowledgments

This work is supported in part by the U.S. National Science Foundation (grant IIS-1748380), and by the Swiss National Science Foundation (grant 200021_162886). We thank reviewers for their suggestions for improvement.

REFERENCES

1. Sylvia Ahern and Jackson Beatty. 1979. Pupillary Responses During Information Processing Vary with Scholastic Aptitude Test Scores. *Science* 205, 4412 (1979), 1289–1292.
2. Miyuki Azuma, Takehiro Minamoto, Ken Yaoi, Mariko Osaka, and Naoyuki Osaka. 2014. Effect of memory load in eye movement control: A study using the reading span test. *Journal of Eye Movement Research* 7, 5 (2014), 1–9.
3. Brian P. Bailey and Shamsi T. Iqbal. 2008. Understanding Changes in Mental Workload During Execution of Goal-directed Tasks and Its Application for Interruption Management. *ACM Trans. Comput.-Hum. Interact.* 14, 4,

Article 21 (Jan. 2008), 28 pages. DOI: http://dx.doi.org/10.1145/1314683.1314689

4. Mike Bartels and Sandra P. Marshall. 2012. Measuring Cognitive Workload Across Different Eye Tracking Hardware Platforms. In *ETRA '12: Proceedings of the 2008 Symposium on Eye Tracking Research & Applications*. ACM, Santa Barbara, CA.

5. Jackson Beatty. 1982. Task-Evoked Pupillary Responses, Processing Load, and the Structure of Processing Resources. *Psychological Bulletin* 91, 2 (1982), 276–292.

6. Jackson Beatty and Brennis Lucero-Wagoner. 2000. The Pupillary System. In *Handbook of Psychophysiology* (2nd ed.), John T. Cacioppo, Louis G. Tassinary, and Gary G. Bernston (Eds.). Cambridge University Press, 142–162.

7. Deborah A. Boehm-Davis, Wayne D. Gray, Leonard Adelman, Sandra Marshall, and Robert Pozos. 2003. *Understanding and Measuring Cognitive Workload: A Coordinated Multidisciplinary Approach*. Technical Report AFRL-SR-AR-TR-03-0407 (ADA417743); Grant #49620-97-1-0353. AFOSR, Arlington, VA.

8. Margaret M. Bradley and Peter J. Lang. 1994. Measuring emotion: The self-assessment manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry* 25, 1 (March 1994), 49–59. DOI:http://dx.doi.org/10.1016/0005-7916(94)90063-9

9. Siyuan Chen and Julien Epps. 2014a. Efficient and Robust Pupil Size and Blink Estimation from Near-Field Video Sequences for Human-Machine Interaction. *IEEE Transactions on Cybernetics* 44, 12 (2014), 2356–2367. DOI:http://dx.doi.org/10.1109/TCYB.2014.2306916

10. Siyuan Chen and Julien Epps. 2014b. Using Task-Induced Pupil Diameter and Blink Rate to Infer Cognitive Load. *Human-Computer Interaction* 29, 4 (2014), 390–413. DOI:http://dx.doi.org/10.1080/07370024.2014.892428

11. Benjamin Cowley, Marco Filetti, Kristian Lukander, Jari Torniainen, Andreas Henelius, Lauri Ahonen, Oswald Barral, Ilkka Kosunen, Teppo Valtonen, Minna Huotilainen, Niklas Ravaja, and Giulio Jacucci. 2016. The Psychophysiology Primer: A Guide to Methods and a Broad Review with a Focus on Human-Computer Interaction. *Foundations and Trends in Human-Computer Interaction* 9, 3-4 (2016), 151–308. DOI: http://dx.doi.org/10.1561/1100000065

12. Nicolas Debue and Cécile van de Leemput. 2014. What does germane load mean? An empirical contribution to the cognitive load theory. *Frontiers in Psychology* 5 (2014), 1099–1–1099–12.

13. Andrew T. Duchowski. 1997. *Gaze-Contingent Visual Communication*. Ph.D. Dissertation. Texas A&M University, College Station, TX.

14. Ralf Engbert and Reinhold Kliegl. 2003. Microsaccades uncover the orientation of covert attention. *Vision Research* 43 (2003), 1035–1045.

15. Randall W. Engle. 2002. Working Memory Capacity as Executive Attention. *Current Directions in Psychological Science* 11, 1 (2002), 19–23. DOI: http://dx.doi.org/10.1111/1467-8721.00160

16. Randall W. Engle and Michael J. Kane. 2003. Executive Attention, Working Memory Capacity, and a Two-Factor Theory of Cognitive Control. *Psychology of Learning and Motivation* 44 (2003), 145–199. DOI: http://dx.doi.org/10.1016/S0079-7421(03)44005-X

17. Paul M. Fitts, Richard E. Jones, and John L. Milton. 1950. Eye Movements of Aircraft Pilots During Instrument-Landing Approaches. *Aeronautical Engineering Review* 9, 2 (1950), 24–29.

18. Eric Granholm, Robert F. Asarnow, Andrew J. Sarkin, and Karen L. Dykes. 1996. Pupillary responses index cognitive resource limitations. *Psychophysiology* 33, 4 (1996), 457–461. DOI: http://dx.doi.org/10.1111/j.1469-8986.1996.tb01071.x

19. Amara Graps. 1995. An Introduction to Wavelets. *IEEE Computational Science & Engineering* (1995), 50–61.

20. 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 (UbiComp '10)*. ACM, New York, NY, 301–310. DOI: http://dx.doi.org/10.1145/1864349.1864395

21. Sandra G. Hart. 2006. Nasa-Task Load Index (NASA-TLX); 20 Years Later. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 50. 904–908.

22. Taylor R. Hayes and Alexander A. Petrov. 2016. Mapping and correcting the influence of gaze position on pupil size measurements. *Behavior Reearch* 48 (2016), 510–527. DOI:http://dx.doi.org/10.3758/s13428-015-0588-x

23. Eckhard H. Hess and James M. Polt. 1964. Pupil Size in Relation to Mental Activity during Simple Problem-Solving. *Science* 143, 3611 (March 1964), 1190–1192.

24. Nina Hollender, Cristian Hofmann, Michael Deneke, and Bernhard Schmitz. 2010. Integrating cognitive load theory and concepts of human-computer interaction. *Computer in Human Behavior* 26, 6 (2010), 1278–1288.

25. Jukka Hyönä, Jorma Tommola, and Anna-Mari Alaja. 1995. Pupil Dilation as a Measure of Processing Load in Simultaneous Interpretation and Other Language Tasks. *The Quarterly Journal of Experimental Psychology* 48, 3 (1995), 598–612.

26. B. Ismail and Anjum Khan. 2012. Image De-noising with a New Threshold Value Using Wavelets. *Journal of Data Science* 10 (2012), 259–270.

27. Robert J. K. Jacob and Keith S. Karn. 2003. Eye Tracking in Human-Computer Interaction and Usability Research: Ready to Deliver the Promises. In *The Mind's Eye: Cognitive and Applied Aspects of Eye Movement Research*, Jukka Hyönä, Ralph Radach, and Heiner Deubel (Eds.). Elsevier Science, Amsterdam, The Netherlands, 573–605.

28. Xianta Jiang, M. Stella Atkins, Geoffrey Tien, Roman Bednarik, and Bin Zheng. 2014. Pupil Responses During Discrete Goal-directed Movements. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '14)*. ACM, New York, NY, 2075–2084. DOI:http://dx.doi.org/10.1145/2556288.2557086

29. Xianta Jiang, Bin Zheng, Roman Bednarik, and M. Stella Atkins. 2015. Pupil responses to continuous aiming movements. *International Journal of Human-Computer Studies* 83 (2015), 1–11.

30. Marcel Adam Just and Patricia A. Carpenter. 1976. Eye Fixations and Cognitive Processes. *Cognitive Psychology* 8, 4 (October 1976), 441–480.

31. Marcel A. Just and Patricia A. Carpenter. 1980. A theory of reading: From eye fixations to comprehension. *Psychological Review* 87, 4 (July 1980), 329–354.

32. Daniel Kahneman. 1973. *Attention and effort*. Prentice-Hall Englewood Cliffs, NJ.

33. Daniel Kahneman and Jackson Beatty. 1966. Pupillary Diameter and Load on Memory. *Science* 154 (1966), 1583–1585.

34. Peter Kiefer, Ioannis Giannopoulos, Andrew Duchowski, and Martin Raubal. 2016. Measuring Cognitive Load for Map Tasks through Pupil Diameter. In *Proceedings of the Night International Conference on Geographic Information Science (GIScience 2016)*. Springer International Publishing.

35. Jeff Klingner, Rakshit Kumar, and Pat Hanrahan. 2008. Measuring the Task-Evoked Pupillary Response with a Remote Eye Tracker. In *ETRA '08: Proceedings of the 2008 Symposium on Eye Tracking Research & Applications*. ACM, New York, NY, 69–72. DOI: http://dx.doi.org/10.1145/1344471.1344489

36. Jeff Klingner, Barbara Tversky, and Pat Hanrahan. 2011. Effects of visual and verbal presentation on cognitive load in vigilance, memory, and arithmetic tasks. *Psychophysiology* 48, 3 (2011), 323–332. DOI:http://dx.doi.org/doi:10.1111/j.1469-8986.2010.01069.x

37. Krzysztof Krejtz, Andrew Duchowski, Izabela Krejtz, Agnieszka Szarkowska, and Agata Kopacz. 2016. Discerning Ambient/Focal Attention with Coefficient $\mathscr{K}$. *Transactions on Applied Perception* 13, 3 (2016).

38. Jan-Louis Kruger, Esté Hefer, and Gordon Matthew. 2013. Measuring the Impact of Subtitles on Cognitive Load: Eye Tracking and Dynamic Audiovisual Texts. In *Proceedings of the 2013 Conference on Eye Tracking South Africa (ETSA '13)*. ACM, New York, NY, 75–78. DOI:http://dx.doi.org/10.1145/2509315.2509333

39. LimeSurvey Project Team / Carsten Schmitz. 2015. *LimeSurvey: An Open Source survey tool*. LimeSurvey Project, Hamburg, Germany. http://www.limesurvey.org

40. Lisa M. Lix, Joanne C. Keselman, and H. J. Keselman. 1996. Consequences of Assumption Violations Revisited: A Quantitative Review of Alternatives to the One-Way Analysis of Variance F Test. *Review of Educational Research* 66, 4 (1996), 579–619. DOI: http://dx.doi.org/10.3102/00346543066004579

41. Yongqiang Lyu, Xiaomin Luo, Jun Zhou, Chun Yu, Congcong Miao, Tong Wang, Yuanchun Shi, and Ken-ichi Kameyama. 2015. Measuring Photoplethysmogram-Based Stress-Induced Vascular Response Index to Assess Cognitive Load and Stress. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. ACM, New York, NY, 857–866. DOI: http://dx.doi.org/10.1145/2702123.2702399

42. Stephane Mallat and Wen Liang Hwang. 1992. Singularity Detection and Processing with Wavelets. *IEEE Transactions on Information Theory* 38, 2 (March 1992), 617–643.

43. Sandra P. Marshall. 2000. Method and Apparatus for Eye Tracking Monitoring Pupil Dilation to Evaluate Cognitive Activity. US Patent No. 6,090,051. (18 July 2000).

44. Sandra P. Marshall. 2002. The Index of Cognitive Activity: Measuring Cognitive Workload. In *Proceedings of the 7th Human Factors Meeting*. IEEE.

45. Sandra P. Marshall. 2007. Identifying Cognitive State from Eye Metrics. *Aviation, Space, and Environmental Medicine* 78, 5, Seciton II (May 2007), B165–B175(11). Supplement 1.

46. Ankit Mathur, Julia Gehrmann, and David A. Atchison. 2013. Pupil shape as viewed long the horizontal visual field. *Journal of Vision* 13, 6 (2013), 1–8.

47. Jum C. Nunnally and Ira H. Bernstein. 1994. *Psychometric Theory* (3rd ed.). McGraw-Hill Inc., New York, NY.

48. Sharon Oviatt. 2006. Human-centered Design Meets Cognitive Load Theory: Designing Interfaces That Help People Think. In *Proceedings of the 14th ACM International Conference on Multimedia (MM '06)*. ACM, New York, NY, 871–880. DOI: http://dx.doi.org/10.1145/1180639.1180831

49. 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.

50. Oskar Palinko and Andrew L. Kun. 2012. Exploring the Effects of Visual Cognitive Load and Illumination on Pupil Diameter in Driving Simulators. In *Proceedings of the Symposium on Eye Tracking Research and Applications (ETRA '12)*. ACM, New York, NY, 413–416. DOI:http://dx.doi.org/10.1145/2168556.2168650

51. Jonathan W Peirce. 2007. PsychoPy–Psychophysics Software in Python. *Journal of neuroscience methods* 162, 1 (2007), 8–13.

52. Vsevolod Peysakhovich. 2016. *Study of pupil diameter and eye movements to enhance flight safety*. Ph.D. Dissertation. Université de Toulouse, Toulouse, France.

53. Bastian Pfleging, Drea K. Fekety, Albrecht Schmidt, and Andrew L. Kun. 2016. A Model Relating Pupil Diameter to Mental Workload and Lighting Conditions. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM, New York, NY, 5776–5788. DOI: http://dx.doi.org/10.1145/2858036.2858117

54. Tepring Piquado, Derek Isaacowitz, and Arthur Wingfield. 2010. Pupillometry as a Measure of Cognitive Effort in Younger and Older Adults. *Psychophysiology* 47, 3 (2010), 560–569. DOI:http://dx.doi.org/doi:10.1111/j.1469-8986.2009.00947.x

55. R Development Core Team. 2011. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/ ISBN 3-900051-07-0.

56. Pallavi Raiturkar, Andrea Kleinsmith, Andreas Keil, Arunava Banerjee, and Eakta Jain. 2016. Decoupling Light Reflex from Pupillary Dilation to Measure Emotional Arousal in Videos. In *Proceedings of the ACM Symposium on Applied Perception (SAP '16)*. ACM, New York, NY, 89–96. DOI: http://dx.doi.org/10.1145/2931002.2931009

57. Gergely Rakoczi, Andrew Duchowski, and Margit Pohl. 2014. Designing Online Tests for a Virtual Learning Environment: Evaluation of Visual Behaviour at Different Task Types. In *Proceedings of the International Conference on Human Behavior in Design (HBiD)*. The Design Society, Ascona, Switzerland.

58. Emanuel Schmider, Matthias Ziegler, Erik Danay, Luzi Beyer, and Markus Bühner. 2010. Is It Really Robust? Reinvestigating the Robustness of ANOVA Against Violations of the Normal Distribution Assumption. *Methodology* 6, 4 (2010), 147–151. DOI: http://dx.doi.org/10.1027/1614-2241/a000016

59. Maximilian Schwalm. 2009. *Pupillometrie als Methode zur Erfassung mentaler Beanspruchungen in automotiven Kontext*. Ph.D. Dissertation. Universität des Saarlandes, Saarbrücken, Germany.

60. Yu Shi, Eric Choi, Ronnie Taib, and Fang Chen. 2009. Designing Cognition-Adaptive Human Computer Interface for Mission-Critical Systems. In *Information Systems Development: Towards a Service Provision Society*, George Angelos Papadopoulos, Wita Wojtkowski, Gregory Wojtkowski, Stanslaw Wrycza, and Joze Zupancic (Eds.). Springer Science+Business Media, LLC, New York, NY, 111–119. DOI: http://dx.doi.org/10.1007/b137171

61. Eva Siegenthaler, Francisco M. Costela, Michael B. McCamy, Leandro L. Di Stasi, Jorge Otero-Millan, Andreas Sonderegger, Rudolf Groner, Stephen Macknik, and Susana Martinez-Conde. 2014. Task difficulty in mental arithmetic affects microsaccadic rates and magnitudes. *European Journal of Neuroscience* 39, 1 (2014), 1–8.

62. SR Research Ltd. 2008. *EyeLink User Manual*. SR Research Ltd., 5516 Main St., Osgoode, ON, Canada K0A 2W0. Version 1.4.0.

63. Lawrence Stark, Fergus W. Campbell, and John Atwood. 1958. Pupil Unrest: An Example of Noise in a Biological Servomechanism. *Nature* 182, 4639 (1958), 857–858.

64. John Sweller. 1976. The Effect of Task Complexity and Sequence on Rule Learning and Problem Solving. *British Journal of Psychology* 67, 4 (1976), 553–558. DOI: http://dx.doi.org/10.1111/j.2044-8295.1976.tb01546.x

65. John Sweller. 1988. Cognitive load during problem solving: Effects on learning. *Cognitive science* 12, 2 (1988), 257–285.

66. Jos N. van der Geest and Maarten A. Frens. 2002. Recording eye movements with video-oculography and scleral search coils: a direct comparison of two methods. 114, 2 (2002), 185–195. DOI: http://dx.doi.org/10.1016/S0165-0270(01)00527-1

67. Boris M. Velichkovsky, Markus Joos, Jens R. Helmert, and Sebastian Pannasch. 2005. Two Visual Systems and their Eye Movements: Evidence from Static and Dynamic Scene Perception. In *CogSci 2005: Proceedings of the XXVII Conference of the Cognitive Science Society*. 2283–2288.

68. Brani Vidaković and Peter Müller. 1991. *Wavelets for kids: A tutorial introduction*. Technical Report. Duke University.

69. Dong Wang, Fiona B. Mulvey, Jeff B. Pelz, and Kenneth Holmqvist. 2017. A study of artificial eyes for the measurement of precision in eye-trackers. *Behavior Research Methods* 49, 3 (2017), 947–959. DOI: http://dx.doi.org/10.3758/s13428-016-0755-8

70. David L. Woods, Mark M. Kishiyama, E. William Yund, Timothy J. Herron, Ben Edwards, Oren Poliva, Robert F. Hink, and Bruce Reed. 2011. Improving digit span assessment of short-term verbal memory. *Journal of Clinical and Experimental Neuropsychology* 33, 1 (2011), 101–111. DOI: http://dx.doi.org/10.1080/13803395.2010.493149

71. Beste F. Yuksel, Kurt B. Oleson, Lane Harrison, Evan M. Peck, Daniel Afergan, Remco Chang, and Robert J. K. Jacob. 2016. Learn Piano with BACh: An Adaptive Learning Interface that Adjusts Task Difficulgty based on Brain State. In *Human Factors in Computing Systems: CHI '16 Conference Proceedings*. ACM, San Jose, CA.