2022

Toward a Real-Time Index of Pupillary Activity as an Indicator of Cognitive Load

Gavindya Jayawardena
Old Dominion University

Yasith Jayawardana
Old Dominion University

Sampath Jayarathna
Old Dominion University

Jonas Högström

Thomas Papa

See next page for additional authors

Original Publication Citation

Jayawardena, G., Jayawardana, Y., Jayarathna, S., Högström, J., Papa, T., Akkil, D., Duchowski, A. T., Peysakhovich, V., Krejtz, I., Gehrer, N., & Krejtz, K. (2022). Toward a real-time index of pupillary activity as an indicator of cognitive load. *Procedia Computer Science*, *207*, 1331-1340. https://doi.org/10.1016/j.procs.2022.09.189

Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 207 (2022) 1331–1340

Procedia Computer Science

www.elsevier.com/locate/procedia

26th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES 2022)

Toward a Real-Time Index of Pupillary Activity as an Indicator of Cognitive Load

Gavindya Jayawardena[a], Yasith Jayawardana[a], Sampath Jayarathna[a], Jonas Högström[b], Thomas Papa[b], Deepak Akkil[b], Andrew T. Duchowski[c], Vsevolod Peysakhovich[d], Izabela Krejtz[e], Nina Gehrer[f], Krzysztof Krejtz[e]

[a]*Old Dominion University, Norfolk, VA, USA*
[b]*Tobii AB, Danderyd, Sweden*
[c]*Clemson University, Clemson, SC, USA*
[d]*ISAE-SUPAERO, Université de Toulouse, Toulouse, France*
[e]*SWPS University of Social Sciences & Humanities, Warsaw, Poland*
[f]*University of Tübingen, Tübingen, Germany*

Abstract

The Low/High Index of Pupillary Activity (LHIPA), an eye-tracked measure of pupil diameter oscillation, is redesigned and implemented to function in real-time. The novel Real-time IPA (RIPA) is shown to discriminate cognitive load in re-streamed data from earlier experiments. Rationale for the RIPA is tied to the functioning of the human autonomic nervous system yielding a hybrid measure based on the ratio of Low/High frequencies of pupil oscillation. The paper's contribution is drawn from provision of documentation of the calculation of the RIPA. As with the LHIPA, it is possible for researchers to apply this metric to their own experiments where a measure of cognitive load is of interest.

Keywords: pupillometry; eye tracking

* Gavindya Jayawardena
E-mail addresses: gavindya@cs.odu.edu (Gavindya Jayawardena)., yasith@cs.odu.edu (Yasith Jayawardana)., sampath@cs.odu.edu (Sampath Jayarathna)., Jonas.Hogstrom@tobii.com (Jonas Högström)., thomas.papa@tobii.com (Thomas Papa)., deepak.akkil@tobii.com (Deepak Akkil)., duchowski@clemson.edu (Andrew T. Duchowski)., vsevolod.peysakhovich@isae.fr (Vsevolod Peysakhovich)., ikrejtz@swps.edu.pl (Izabela Krejtz)., nina.gehrer@uni-tuebingen.de (Nina Gehrer)., kkrejtz@swps.edu.pl (Krzysztof Krejtz).

* Corresponding author. Tel.: +0-000-000-0000 ; fax: +0-000-000-0000.
E-mail address: author@institute.xxx

1877-0509
10.1016/j.procs.2022.09.189

1. Introduction & Background

Eye movement recording, including pupil diameter as a matter of course, has been extensively used in Human Computer Interaction (HCI) and offers the possibility of understanding how information is perceived and processed by users. Hardware developments have allowed gaze recording to enter common usage as a control modality (e.g., available as standard *Eye control* setting in Windows 10). However, the connection between real-time eye movement measures and cognitive state has been largely under-exploited in HCI [34].

In particular, cognitive load, as classically introduced by Sweller [31, 32] and its indication through gaze-based metrics, has generated considerable interest in a variety of human factors settings, including linguistics tasks, reading, visualization [33], among many others [25, 3, 15, 12, 26, 18, 20, 36]. As noted by Fridman et al. [10], the breadth and depth of the published work highlights the difficulty of identifying useful measures of cognitive load that do not interfere with or influence behavior. Apart from gaze-based metrics, various other measures include the Paas scale, the NASA Task Load Index (NASA-TLX), the Cognitive Load Component Survey, the SOS scale, and others [13], although some of these are largely subjective. Gaze-based metrics are attractive due to their objective quality, but often require fairly structured experimental design, not to mention use of an eye tracker which may include restricted procedures, e.g., reduced head movement, specialized stimulus, etc. Other less restrictive but still gaze-related metrics can include blink rate and blink entropy [5], metrics which can rely on less specialized hardware such as webcams.

A specific need identified by Cho [5] is *continuous*, or real-time, detection of cognitive load. According to Hutt et al. [16], while real-time gaze-based models of attention are increasingly prevalent, apart from blink rate, closed-loop attentionaly-aware technologies are still sparse. They showed a gaze-based means of detection of *mind wandering* using a commercially available off-the-shelf eye tracker. They used 57 global eye movement features computed from the time series of fixations and saccades to characterize general gaze patterns independent of the displayed content. Interestingly, this approach relies on a 18-30 second time window for mind-wandering detection. While this approach can be considered continuous, a 30 second history (e.g., buffer) introduces a fairly significant delay in estimation of real-time cognitive state. Real-time objective measures of attentional states are therefore essential. In the present paper, the goal is to use a one second buffer to estimate cognitive load through pupil diameter oscillation.

With the aim of moment-to-moment assessment of cognitive load, Duchowski et al. [8] introduced the Low/High Index of Pupillary Activity, or LHIPA, a wavelet-based algorithm inspired by Marshall's [2000, 2002] Index of Cognitive Activity (ICA), and Duchowski et al.'s [2018] implementation of their IPA. All three, ICA, IPA, and LHIPA, analyzed pupil diameter oscillation, reflecting moment-to-moment pupil diameter changes as captured in the frequency spectrum by the wavelet transform. IPA sensitivity was reported under restricted (fixed-gaze) conditions during performance of mental arithmetic, while the LHIPA was shown to distinguish cognitive load during execution of an n-back (1- and 2-back) task with gaze fixed at specific screen coordinates. The LHIPA also yielded significant results in a less restrictive eye-typing task where the eyes were free to move to effect key entry.

However, both the IPA and LHIPA were analyzed off-line, with the entirety of the pupil diameter signal available to the Discrete Wavelet Transform (DWT). This is important, as the longer the signal duration, the greater the number of frequency bands available for analysis (specifically $\log_2 n$ where n is the signal length). Moreover, even if the trial length was short (e.g., two seconds), because the analysis was carried out off-line, no indication was made regarding its computational latency. Because the DWT is a pyramidal transform, it relies on successive (and hence sequential) application of scaling ($\phi_{j,k}$) and wavelet ($\psi_{j,k}$) functions. Using filters such as *symlet-16* wavelets (of length 32), means that at least $2\times32\times.5\log_2(n)$ operations are required to compute the LHIPA. For a signal of $n = 1024$ (just over 2 seconds captured at 500 Hz) would require 320 operations which is likely to incur some delay. Therefore, the wavelet approach is considered inappropriate for real-time use. In this paper, we propose a mathematically similar operation based on cleverly tuned Savtizky-Golay smoothing and differentiating filters to effect a similar low- to high-frequency ratio computed by the LHIPA, but in a fraction of the computational demand.

2. Implementation of the Real-Time IPA

Implementation of the RIPA follows Duchowski et al.'s [2018, 2020] IPA and LHIPA, both inspired by the ICA [22, 4]. All three metrics were based on wavelet analysis.

Although all three of the ICA, IPA, LHIPA metrics showed moderately good results under varying conditions (e.g., fixed-gazed number counting, n-back, eye-typing), wavelet analysis is not well suited for real-time implementation. The problem with wavelets is that to analyze the input signal at low resolution for the low- to high-frequency (LH/HF) ratio of the LHIPA, successive applications of the scaling and wavelet functions are used *sequentially*, which is slow. For real-time use, the Savitzky-Golay filter can be used to obtain both low- and high-frequency analysis of the signal directly, and *in parallel*. To better explain the perhaps mathematically nuanced but critically important underlying processes, both types of analyses are given below for completeness.

2.1. Wavelet Analysis of the Pupil Diameter

Mathematically, the DWT shares its analytical characteristics with the Fourier Transform (FT), and offers spectral analysis of a given input signal $x(t)$. The FT requires the entire length of the signal in order to model the signal by sine and cosine basis functions at different wavelengths. The DWT does something similar, but enjoys spatio-temporal processing advantages due to its so-called *compact support* meaning that the wavelets themselves are Finite Impulse Response (FIR) filters of typically short length (as opposed to the infinite sines and cosines used by the FT). The Wavelet Browser[1] lists several popular wavelet "families", e.g., Haar, Daubechies, Symlets, and others, where each can be inspected and where filter coefficients can readily be obtained. For example, the Symlet family of filters contains filters of varying length, ranging from 2 upwards.

Short filter lengths are essential for real-time implementation. Because of their short filter length, it would appear that wavelets would be suitable for real-time use. However, in the case of LHIPA, where the ratio of the low- and high-frequency filter responses is needed, the low-frequency response from the DWT is only available at higher levels of its multiresolution pyramid (the DWT is known as a *pyramidal* transform). These lower levels of the transform are produced by the DWT's cascaded use of its scaling function, $\phi_{j,k}$ in combination with the wavelet function $\psi_{j,k}$. The DWT is therefore a sequential process: first, approximate the input signal $x(t)$ with the scaling function $\phi_{j,k}$: $x_\phi^{j-1}(t) = \sum_k \phi_k^j x_\phi^j(2t+k)$, then, analyze the signal with the wavelet $\psi_{j,k}$: $x_\psi^{j-1}(t) = \sum_k \psi_k^j x_\phi^j(2t+k)$, and repeat at successive resolution levels j, where integral powers of 2 are used in the decomposition of the wavelet. The importance of the use of integral powers of 2 are evident when indexing wavelet coefficients at different scales, i.e., when obtaining the pointwise low/high coefficient ratio.

For the high frequency component, Duchowski et al. [8] chose $j = 1$ and for the low, $j = \frac{1}{2}\log_2(n)$, the mid-level frequency octave with $\log_2(n)$ number of octaves. The LF/HF ratio is thus: $x_\psi^{1/2\log_2(n)}(t)/x_\psi^1(2^{1/2\log_2(n)}t)$ where $2^{1/2\log_2(n)}t$ in the denominator is the scale factor of the index into the high frequency signal when iterating over the (shorter) low frequency component. Remember that the length of the wavelet coefficient signal (array) at each lower resolution level is half that of the previous higher level of resolution. Hence, while iterating over the lower frequency signal, the index must be multiplied by raising 2 to the power of frequency octave level being processed.

Note that the LHIPA is a ratio of low to high frequency, with high frequency response expected to increase with cognitive load (task difficulty), thus LHIPA decreases with increased cognitive load, the reverse of the IPA response. Notice also that the filtering operation (i.e., convolution) by $\phi_{j,k}$ is needed at every step because analysis of the signal at increasing levels of resolution j is performed on subsequent approximations $x_\phi^j(t)$ to the original input $x^0(t)$.

Since the LHIPA was found to be a more robust indicator of cognitive load to the IPA, why is the DWT not appropriate for its real-time implementation? The reason is because the LHIPA used a mid-level resolution level (known as octave) for the low-frequency numerator. This means that to obtain this signal, earlier combinations of the scaling and wavelet functions must have been employed. Even with short wavelet filters, using them in successive sequential operations is analogous to extending the scaling function $\phi_{j,k}$ in length so that it acts as a longer, smoothing filter. And so the question then is whether a filter can be used to directly obtain a similar low-frequency response without having to execute the underlying sequential convolutions involved in the DWT.

The key observation required to bypass the DWT is that in effect all that the DWT is doing in computing the wavelet coefficients is merely differentiating the signal at successive levels of approximation (resolution). Hence, instead of the DWT, is there another filter that can be used to differentiate the input signal and at the same time be tuned to

[1] `http://wavelets.pybytes.com`

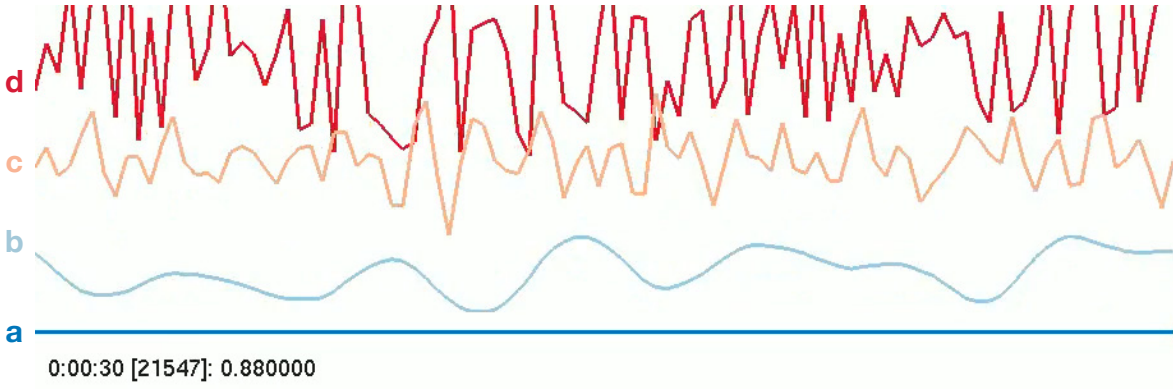


Fig. 1: Savitzky-Golay filter responses: (a) the raw pupil diameter signal (scaled to fit), (b) the low frequency response ($g^{\triangledown}$), (c) the high frequency response ($g^{\triangle}$), (d) the low- to high-frequency ratio $g^{\triangledown}/g^{\triangle}$. Text at bottom left indicates the real time, number of pupil diameter elements processed thus far, and the real-time (normalized) RIPA (see text, §2.4).

different frequencies? The Savitzky-Golay filter is such a candidate as it performs both smoothing and differentiating operations in one step, depending on which parameters one chooses for the filter coefficients.

For the last stage of IPA and LHIPA computations, sharp points of variation are found in either signal via modulus maxima detection, then universal thresholding is used to include (IPA) or exclude (LHIPA) outliers. Instead of 'hard' thresholding for IPA, 'less' thresholding is used for LHIPA wherein data is decimated if above the threshold while lesser values pass untouched. Unlike the IPA, smaller LHIPA response is expected for greater cognitive load. In both cases, the number of remaining coefficients is counted to produce either the IPA or LHIPA. A similar counting operation is performed following Savitzky-Golay filtering, as detailed below.

2.2. Savitzky-Golay Signal Differentiation

In part due to Andersson et al. [1] and Nyström and Holmqvist [24], the Savitzky-Golay filter has become very popular for velocity-based eye movement *event detection* [7], i.e., detection of saccades and consequently fixations in the eye movement signal. Why is this so? Early on, Duchowski et al. [6] showed that velocity-based eye movement analysis could easily be implemented with very short filters, of length 7, 5, and even 2, where filters of length 2 ("2-tap") were sufficient to differentiate the eye movement position signal to compute velocity. Indeed, the simplest 2-tap velocity filter with normalized coefficients $\{-1/\sqrt{2}, 1/\sqrt{2}\}$ is (coincidentally) the Haar wavelet. However, short filters are more susceptible to noise, hence some element of signal smoothing is often needed to ameliorate filter response, e.g., either smoothing the signal beforehand (what the wavelet scaling function effectively does), or cleverly incorporating the smoothing step within the differentiation (what the Savitzky-Golay filter does).

In their seminal paper, Savitzky and Golay [29] gave a method of data smoothing and differentiation by a simplified least-squares polynomial approximation. They showed that fitting a polynomial to the input and evaluating the result at a single point within the approximation interval is equivalent to convolution with a fixed impulse response. They also showed how this least-squares technique can be applied for determining the signal derivative.

Least-Squares Smoothing & Differentiation. A Savitzky–Golay filter determines the best mean-squares fit of a polynomial of degree n through $2m+1$ data samples ($n < 2m+1$). For a sequence of samples $x(t)$, and $2m+1$ filter coefficients a_k of a polynomial p, centered at $t=0$, are obtained by: $p(t) = \sum_{k=0}^{n} a_k t^k = a_0 + a_1 t^1 + \ldots + a_n t^n$ with first derivative: $dp(t)/dt = a_1 + 2a_2 t + 3a_3 t^2 + \ldots + n a_n t^{n-1} = \sum_{k=1}^{n} k a_k t^{k-1}$ where $t \in [-m, m]$ and m is referred to as the "half width" of the approximation interval.

The least-squares criterion necessitates that the sum of squares of the differences between the observed samples $x(t)$ and their approximations $p(t)$ be a minimum over the given interval: $\xi_n = \sum_{t=-m}^{m} (p(t) - x(t))^2 = \sum_{t=-m}^{m} \left(\sum_{k=0}^{n} a_k t^k - x(t)\right)^2$ Minimization of the mean-squared approximation error ξ_n, is done in the same manner for any other group of $2m+1$ input samples centered on $t=0$.

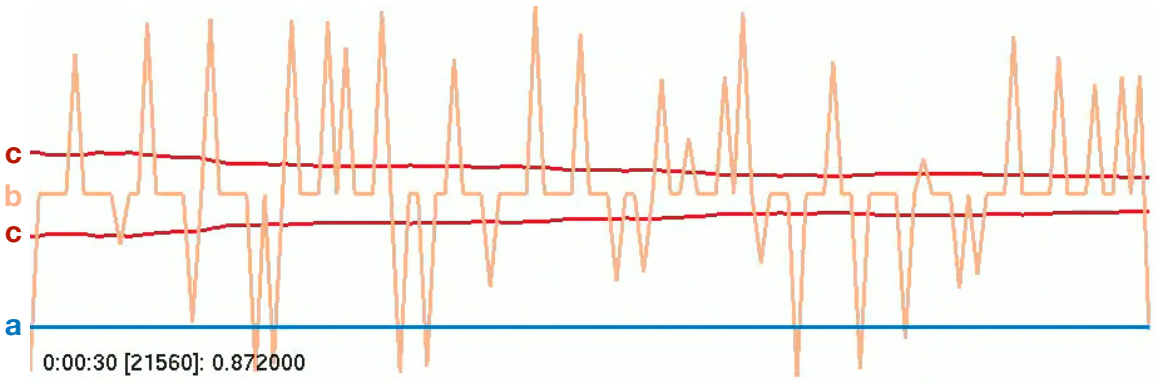


Fig. 2: Visualization of the real-time RIPA: (a) the raw pupil diameter signal (scaled to fit), (b) the modulus maxima of the low- to high-frequency ratio, (c) the adaptive threshold used to detect the RIPA, with bounds above and below the modulus maxima. Text at bottom left indicates the real time, number of pupil diameter elements processed thus far, and the real-time (normalized) RIPA (see text, §2.4).

Optimal coefficients of the polynomial are found by partially differentiating ξ_n with respect to each of the $n+1$ unknowns and setting the corresponding derivative to 0. Interchanging the order of the summations gives a set of $n+1$ equations in $n+1$ unknowns which are known as the normal equations for the least-squares approximation.

Coefficients are computed by expressing the normal equations in matrix from [30]. The $(2m+1)\times(n+1)$ matrix $\mathbf{A}=\{\alpha_{t,i}\}$ is called a design matrix for the approximation, with matrix elements $\alpha_{t,i}=t^i$, $-m\le t\le m$, and $i=0,1,\ldots n$. The transpose of $\mathbf{A}$ is $\mathbf{A}^T=\{\alpha_{i,t}\}$ and the product matrix $\mathbf{B}=\mathbf{A}^T\mathbf{A}$ is an $(n+1)\times(n+1)$ symmetric matrix with elements $\beta_{i,k}=\sum_{t=-m}^{m} t^{i+k}=\beta_{k,i}$ for $i=0,1,\ldots n$ and $k=0,1,\ldots n$, which are the coefficients for the normal equations.

Defining the vector of polynomial coefficients by $\mathbf{a}=[a_0,a_1,\ldots,a_n]^T$ and the vector of input samples by $\mathbf{x}=[x(-m),\ldots,x(-1),x(0),x(1),\ldots,x(m)]^T$, the normal equations can be rewritten in matrix form: $\mathbf{B}\,\mathbf{a}=\mathbf{A}^T\mathbf{A}\,\mathbf{a}=\mathbf{A}^T\mathbf{x}$ with solution: $\mathbf{a}=(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{x}=\mathbf{H}\,\mathbf{x}$ that gives the sought polynomial coefficients. Since the output for the input samples centered on $t=0$ is a_0, only this coefficient is needed. Coefficient a_0 is the 0^{th} row of the matrix $\mathbf{H}$. Matrix $\mathbf{H}$ is a $(n+1)\times(2m+1)$ matrix, and can be written as $\mathbf{H}=(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T$ which only depends on n and m and is independent of the input samples. Therefore, the same weighting coefficients are obtained for any $2m+1$ input samples, yielding least-squares minimization as a shift-invariant discrete convolution.

Savitzky-Golay Filter Properties. Of specific interest to using the Savitzky-Golay filter for the purpose of signal differentiation at different frequencies, to mimic the effect of wavelet analysis, is to examine the filter properties, e.g., noting the filter's impulse response. In particular, the nominal normalized cutoff (3 dB down) frequency $f_c=\omega_c/\pi$, depends on both the implicit polynomial order n and the length of the impulse response $(2m+1)$. To quantify the frequency-domain behavior of Savitzky-Golay filters, Schafer [30] computed impulse responses for various combinations of n and m for $0\le\omega\le\pi$. Short filters e.g., with $m<6$, are preferred for real-time usage, and cutoff frequencies for filters with $m=2,3,\ldots,6$, for $n=2,4,6,8,10$, see Schafer [30]. For longer filters, specifically for filters with $m\ge 25$ and $n<m$, Schafer provides a linear model of filter response, $f_c=n+1/3.2m-4.6$, $m\ge 25$, $n<m$, and points out that selection of a given cutoff frequency f_c through selection of different combinations of n and m is an often overlooked feature of the Savitzky-Golay filters.

Indeed, the filter cutoff frequency selection feature is the key to mimicking the low- to high-frequency ratio obtained by the wavelet analysis of LHIPA. To roughly approximate the ratio computed by the wavelet expression given above, the goal is to compute the ratio of responses from two Savitzky-Golay filters computed in parallel, $\dot{x}(t)=\sum_{k=-m}^{m} g_k^{\triangledown}(t-k)\bar{x}_k(t)/\sum_{k=-m}^{m} g_k^{\triangle}(t-k)x_k(t)$ where $x(t)$ is the raw data, $\bar{x}(t)$ is low-pass filtered data, processed with a Savitzky-Golay smoothing filter, $g(t)=\sum_{k=0}^{n} a_k t^k$, are the Savitzky-Golay differentiation filters with $-m\le n\le m$, and $g^{\triangledown}$ and $g^{\triangle}$ are each tuned to the low- and high-frequency cutoffs, respectively, by careful selection of polynomial order n and filter length m. For example, selecting $m=6$, $n=2$ yields cutoff frequency $f_c=0.165$ and thus a lowpass response for $g^{\triangledown}$, and selecting $m=6$, $n=10$ yields cutoff frequency $f_c=0.681$ and thus a highpass response for $g^{\triangle}$. The

output signal $\dot{x}$ can then be subject to similar modulus maxima detection as performed in computation of the LHIPA but at a reduced computational cost.

Visualization of the filter responses ($g^{\triangledown}$, $g^{\triangle}$), and the low- to high-frequency ratio ($g^{\triangledown}/g^{\triangle}$) is given in Fig. 1 captured during real-time playback of a previously recorded signal.

2.3. Usage of Ring Buffers to Effect Real-Time Filtering

The obvious advantage of offline analysis is the availability of data in its entirety. For example, a 2-second eye-tracking trial, sampled at 500 Hz yields 1,000 data points that can be analyzed over the whole course of time that the data represents. Real-time analysis, in contrast, relies on a temporary buffer that is necessarily of limited length, e.g., a double-ended queue (deque or ring buffer) serves the purpose.

The real-time nature of RIPA relies on several ring buffers, one each to hold: (1) raw data, $x(t)$, (2) low-pass filtered data, $\bar{x}(t)$, (3) low frequency response, $g^{\triangledown}(t)$, (4) high frequency response, $g^{\triangle}(t)$, (5) low- to high-frequency ratio, $\dot{x}(t) = g^{\triangledown}/g^{\triangle}(t)$, as above, (6) modulus maxima of the low- to high-frequency ratio, $m(t)$, (7) an approximation to the universal threshold, $\lambda(t)$. Buffers holding raw data, $x(t)$, low-pass filtered data, $\bar{x}(t)$, low frequency response, $g^{\triangledown}(t)$, high frequency response, $g^{\triangle}(t)$, and the low- to high-frequency ratio, $\dot{x}(t) = g^{\triangledown}/g^{\triangle}(t)$, are computed as above.

The modulus maxima of the low- to high-frequency (LF/HF) ratio, $m(t)$, is computed similarly as per Duchowski et al. [2018, 2020], with one exception. For each point in the LF/HF ring buffer, the modulus maximus is either $m(t) = \text{sgn}(\dot{x}(t))\|\dot{x}(t)\|$ if $|\dot{x}(t)|$ is larger than its two closest neighbors, and strictly larger than at least one of them [21], i.e., $|\dot{x}(t-1)| \leq |\dot{x}(t)| \geq |\dot{x}(t+1)|$ and $|\dot{x}(t-1)| < |\dot{x}(t)|$ or $|\dot{x}(t)| > |\dot{x}(t+1)|$ where $\text{sgn}(\dot{x}(t))$ is used to preserve the sign of the LF/HF ratio, or $m(t)$ is set to 0, see Fig. 2. The use of $\text{sgn}(\cdot)$ is the exception to what was done by Duchowski et al. [8] since it preserves the positive and negative direction of the LF/HF ratio. This is mainly used for visualization purposes, where two symmetrical thresholds are rendered as the lower and upper threshold bounds.

LHIPA relied on computation of the universal threshold to preserve wavelet coefficients [22], i.e., $\lambda = \hat{\sigma}\sqrt{2\log_2 n}$ where n is the signal length (wavelet array or in the present case the ring buffer), and $\hat{\sigma}$ is the standard deviation of the modulus maxima, $m(t)$. The universal threshold is often used in orthonormal regression [11], e.g., at multiple scales of the wavelet transform. In the present case since in effect there is only one level (octave) of the signal, a simpler threshold can be used, namely, $\lambda = \omega\hat{\sigma}$ where $\hat{\sigma}$ is the standard deviation of the modulus maxima $m(t)$ as before, and $\omega = (1/0.6745)^2$ is a scalar set to approximate the wavelet universal threshold as per He et al. [14]. Using $\langle m(t)\rangle$ to represent the median of the modulus maxima, RIPA is obtained as the number of occurrences of the low- to high-frequency ratio signal $\dot{x}$ outside the bounds defined by the median ± threshold λ, or more formally, with $|\cdot|$ indicating absolute value, $C(\dot{x}) = \sum_{t=1}^{n} \delta(|\dot{x}(t)|, |\dot{x}(t)| > \langle m(t)\rangle + \lambda)$ over the length of the buffer n, and $\delta(i, j)$ is the Kronecker delta defined to be 1 if $i = j$ and 0 otherwise. For visualization purposes, symmetrical thresholds are set with respect the running median of m, as shown in Fig. 2: $\lambda^{+} = \langle m\rangle + \lambda$, $\lambda^{-} = \langle m\rangle - \lambda$.

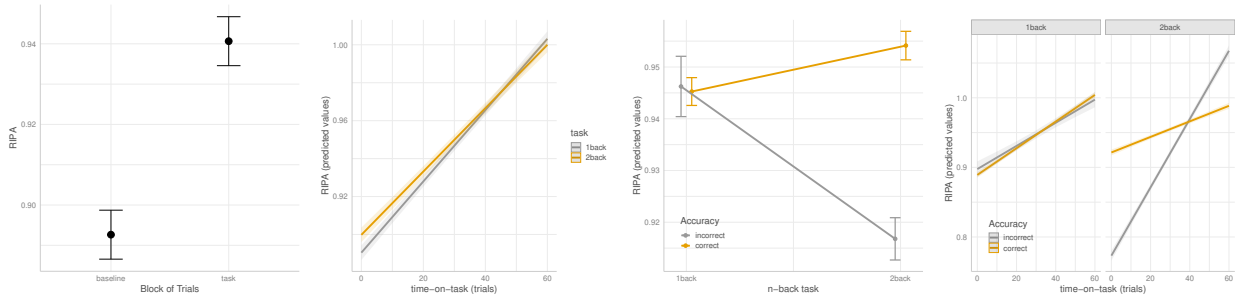
The length of all buffers is the same and is set arbitrarily, but can reasonably be set to 1 second worth of data, e.g., to a length of 250 for a signal sampled at 250 Hz. Incidentally, using buffers holding 1 second's worth of data will yield RIPA as the frequency of occurrence (of pupil oscillation) per second, similar to what iBoehm-Davis et al. [4] used in their computation of the Index of Cognitive Activity (ICA), suggesting that doing so provides a common basis for comparison between individuals, groups of individuals, single and multiple events.

2.4. RIPA Normalization

RIPA, $C(\dot{x})$, responds similarly to LHIPA, i.e., it decreases with increased pupillometric oscillation (presumed to indicate increased cognitive load). The inverse relation of RIPA to cognitive load is unwieldy. However, because RIPA is clearly bound by the length of the buffer, meaning that acting as a counter, it can range from either the minimum (0) to a bounded maximum ($|\dot{x}|$). Therefore, RIPA can be normalized and inverted as well. That is, $\hat{C}(\dot{x}) = 1 - C(\dot{x})/n$ where $\hat{C}$ is normalized, $\hat{C} \in [0, 1]$, and directly proportional to pupil diameter oscillation, and hence cognitive load.

3. Empirical Validation

To evaluate RIPA and compare it to the earlier LHIPA response, data from Duchowski et al. [8] was re-analyzed using RIPA. To simulate real-time processing, the data was re-streamed using StreamingHub as per Jayawardana et



(a) Trial blocks. (b) Time-on-task. (c) Accuracy. (d) Time-on-task × accuracy.

Fig. 3: Differences in mean RIPA response between baseline and tasks: (a) depending on trial block; (b) depending on time-on-task; (c) depending on accuracy of responses; and (d) depending on time-on-task moderated by accuracy.

al. [17]. StreamingHub reads data from the original file into memory (in this instance CSV data) and streams the data according to information contained in Findable, Accessible, Interoperable, and Reusable, or FAIR [35], metadata, which stipulates the streaming rate. When streaming, the RIPA simulation accepts a new pupil diameter sample at the specified streaming rate. This sample is appended to the ring buffer which is then convolved with the low- and high-frequency response filters $g^{\triangledown}$, $g^{\triangle}$), to compute the the low- to high-frequency ratio ($g^{\triangledown}/g^{\triangle}$) for the given timestamp.

Experimental Procedure. Introduced by Kirchner [19], the n-back task was used to establish a connection between task difficulty and active working memory [31]. Duchowski et al. [8] used the n-back task with a gaze fixation point shown at five (5) different locations on the screen to evaluate potential effects of the distortion of the apparent off-axis pupil. Their experiment followed the n-back experimental protocol described by Appel et al. [2], who used the ICA [22] as a feature of their cross-subject classification of cognitive load. The n-back task, from which data is re-analyzed here, consisted of a randomly generated sequence of letters from the set $L=\{C, F, H, S\}$ shown one after another, each for a duration of 0.5 seconds with an inter-stimulus duration of 1.5 seconds. Each letter and inter-stimulus period thus constituted a 2 second trial. Within each trial, participants were asked to state whether the currently shown letter was the same as the one n trials before ($n=\{1, 2\}$) by pressing one of two keyboard buttons.

Participants completed five blocks of each of the 1- and 2-back task. Each of the five blocks displayed the target stimulus (letter) at one of five positions on the screen: center, top-left, top-right, bottom-right, or bottom-left. Each of the 1- and 2-back task was preceded by a baseline task with a fixation cross positioned at the point where the stimulus letters would appear and lasted for 30 seconds.

With the exception of the center block, the remaining 4 blocks were at 10° visual angle from center along each of the x- and y-axes. The order of the 1- and 2-back tasks was counterbalanced. Within the tasks, the order of the 5 locations was randomized. Each block consisted of 60 trials, lasting 2 minutes.

Participants. The dataset consisted of $N=18$ participants (15 M, 3 F) with ages in range [21:29] years old ($M=22.77$, $SD=2.26$). All participants had normal or corrected to normal vision.

Apparatus. The experimental setup consisted of an SR Research Eye-Link 1000 eye tracker to record eye movements binocularly at a sampling rate of 1000 Hz. Each participant's head had been stabilized with a chin rest during the experimental procedure. Visual stimuli were displayed on a 24-inch wide computer screen with 1920×1200 resolution. The experimental laboratory had been devoid of windows limiting the amount of ambient light during the study.

Results. The analysis of RIPA has two interrelated aims. First, we replicate results concerning LHIPA from Duchowski et al. [8] showing that RIPA is sensitive to the difficulty of n-back tasks. The n-back tasks require more cognitive resources than the baseline task, reflected in larger RIPA response for n-back tasks than baseline.

Second, we extend the results to dynamical analysis of real-time changes in cognitive load, hypothesizing that (a) RIPA response increases with time-on-task reflecting an increase in cognitive load; (b) correct answers to the n-back

task are more cognitively demanding thus yielding greater RIPA response for correct rather than incorrect answers, especially in the more demanding 2-back task.

We first tested whether RIPA is sensitive to the n-back tasks in comparison to baseline with a paired sample t-test. The test showed that the difference between RIPA scores from baseline ($M = 0.89, SD = 0.01$) and n-back ($M = 0.94, SD = 0.01$) blocks of trials was statistically significant, $t(18) = 11.74$, $p < 0.001$, $d = 2.693$.

To test hypotheses (a) and (b) we analysed the sensitivity of RIPA to differentiate between n-back tasks (1-back vs. 2-back), cognitive exhaustion expected with time-on-task, and accuracy of answers in the n-back tasks. To meet these aims a Linear Mixed Model (LMM) was used with RIPA as the dependent variable. The model at level 1 included only random slopes for n-back tasks nested within subjects. The second level model included the same random effects structure and full factorial structure for fixed effects of time-on-task, task type, and accuracy of answers (coded as 0 and 1). For accuracy of answers, a wrong answer was treated as the base and for task type, the 1-back task was treated as the base. The models at level 1 and level 2 were fit only for task block of trials because during baseline participants were not providing any answers.

The difference between models at level 1 and 2 was statistically significant, $\chi^2(7) = 4148.20$, $p < 0.001$. The full model showed a good fit to the data *Pseudo* $R^2 = 0.17$ with $AIC = -47437.22$ and $BIC = -47349.32$, revealing a significant main effect of time-on-task and interaction effect of time with task type. As expected, the longer time-on-task the higher the RIPA response in both task types (1-back and 2-back). Analysis of the interaction term showed that the slope for the 2-back task was slightly but significantly steeper ($est. = 0.003$, $z.ratio = 55.98$, $p < 0.001$) than for the 1-back task ($est. = 0.018$, $z.ratio = 21.44$, $p < 0.001$), see Fig. 3(b). The model also showed a significant interaction of task and response accuracy. Pairwise comparisons of estimated means showed that the correct answers yielded a significantly higher RIPA response than incorrect answers but only in the 2-back task, see Fig. 3(c).

The model also revealed a three-way interaction of time-on-task, accuracy of responses, and task type. Decomposing this interaction effect, we compared pairwise simple slopes of RIPA vs. time for correct and incorrect answers separately for 1-back and 2-back tasks. Analysis showed that the slope for incorrect answers was significantly ($z = 32.07$, $p < 0.001$) steeper ($est. = 0.005$, $z.ratio = 46.60$, $p < 0.001$) than the slope for correct answers ($est. = 0.001$, $z.ratio = 20.80$, $p < 0.001$) but only for participants executing the 2-back task. For the 1-back task, both slopes for incorrect ($est. = 0.002$, $z.ratio = 10.44$, $p < 0.001$) and correct answers ($est. = 0.002$, $z.ratio = 38.13$, $p < 0.001$) were positive and did not differ significantly ($z = 1.51$, $p = 0.13$), see Fig. 3(d).

4. Discussion

In line with expectations, statistical findings replicated results given earlier [8]. We observed greater RIPA response during execution of n-back tasks than during baseline (when not performing a task requiring cognitive resources). Also in line with predictions on the dynamics of cognitive load, we observed an increased RIPA response to time-on-task in both 1-back and 2-back tasks. Moreover, in the 2-back task, RIPA response was greater than in the 1-back task from the very first trials. Results also indicated greater RIPA response from correct n-back trials rather than from incorrect trials. This difference was visible only during the 2-back task and not during the 1-back task. One plausible explanation for this moderation effect is that the 1-back tasks were too easy and not cognitively demanding. The other plausible explanation is that the 2-back tasks were cognitively demanding to the extent that for incorrect answers cognitive disengagement was observed rather than cognitive load.

Although we did not obtain straightforward results that RIPA is sensitive to n-back task difficulty, we did observe that the n-back task type moderated the interaction of response accuracy with time-on-task. In both n-back tasks RIPA response increased with time-on-task regardless of answer accuracy, but this increase was significantly steeper for incorrect answers during the 2-back (more difficult) tasks. RIPA response rose dramatically during the time-on-task. This may be due to the way RIPA loads the ring buffer during initial stages of real-time use (see below). Further studies are required to explore this functionality. The steep slope of RIPA response when answering incorrectly in the 2-back task may also be an effect of increasing frustration and stress. Current studies demonstrate that pupil diameter may be a reliable and sensitive psycho-physiological marker for increased stress [27, 28].

5. Limitations & Future Work

The present approach and its analysis are subject to two main limitations. First, the filtering approach is sensitive to specific tuning parameters. Second, analysis of the data is given in the aggregate, lacking real-time empirical evidence, although recorded data was analyzed in a streamed manner simulating real-time. Both limitations are detailed below.

The filtering approach relies on fine-tuning of several parameters, including queue (buffer) length, and filtering threshold. On the one hand, if the buffer length is too short, the distinction between task and baseline may be lost. On the other hand, if the buffer length is too long, processing time is increased. Setting buffer length to the sampling period appears to produce good results (for the given data set) and allows real-time processing but may be too slow for visualization. The choice of threshold appears to scale the RIPA response uniformly, e.g., a smaller threshold increases RIPA response. A good approach to fine-tuning of the parameters is to characterize the data *a priori*, as was done using via sinusoidal approximation, so that RIPA response can be gauged given *idealized* conditions first.

6. Conclusion

A new eye-tracked measure of pupil diameter oscillation is introduced, the Real-time Index of Pupillary Activity (RIPA), based on a ratio of several finite-impulse response filters, designed to function in real-time. RIPA is shown to discriminate between a low- and high-frequency sinusoidal ideal signal and between empirically collected baseline and cognitive load. As with earlier wavelet-based implementation of the IPA and LHIPA, due to its shorter filter lengths, RIPA is better suited for real-time usage whenever a pupillometric measure of cognitive load is needed.

References

[1] Andersson, R., Nyström, M., Holmqvist, K., 2010. Sampling frequency and eye-tracking measures: how speed affects durations, latencies, and more. Journal of Eye Movement Research 3, 1–12.

[2] Appel, T., Scharinger, C., Gerjets, P., Kasneci, E., 2018. Cross-subject Workload Classification Using Pupil-related Measures, in: Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications, ACM, New York, NY. pp. 4:1–4:8. URL: http://doi.acm.org/10.1145/3204493.3204531, doi:10.1145/3204493.3204531.

[3] Bailey, B.P., Iqbal, S.T., 2008. Understanding Changes in Mental Workload During Execution of Goal-directed Tasks and Its Application for Interruption Management. ACM Trans. Comput.-Hum. Interact. 14, 21:1–21:28. URL: http://doi.acm.org/10.1145/1314683.1314689, doi:10.1145/1314683.1314689.

[4] Boehm-Davis, D.A., Gray, W.D., Adelman, L., Marshall, S., Pozos, R., 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.

[5] Cho, Y., 2021. Rethinking Eye-Blink: Assessing Task Difficulty through Physiological Representation of Spontaneous Blinking, in: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, Association for Computing Machinery, New York, NY. URL: https://doi.org/10.1145/3411764.3445577, doi:10.1145/3411764.3445577.

[6] Duchowski, A., Medlin, E., Cournia, N., Gramopadhye, A., Melloy, B., Nair, S., 2002. 3D Eye Movement Analysis for VR Visual Inspection Training, in: ETRA '02: Proceedings of the 2004 symposium on Eye tracking research & applications, ACM, New Orleans, LA. pp. 103–110,155.

[7] Duchowski, A.T., 2017. Eye Tracking Methodology: Theory & Practice. 3rd ed., Springer-Verlag, Inc., London, UK.

[8] Duchowski, A.T., Krejtz, K., Gehrer, N.A., Bafna, T., Bækgaard, P., 2020. The Low/High Index of Pupillary Activity, in: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, Association for Computing Machinery, New York, NY. p. 1–12. URL: https://doi.org/10.1145/3313831.3376394, doi:10.1145/3313831.3376394.

[9] Duchowski, A.T., Krejtz, K., Krejtz, I., Biele, C., Niedzielska, A., Kiefer, P., Raubal, M., Giannopoulos, I., 2018. The index of pupillary activity: Measuring cognitive load vis-à-vis task difficulty with pupil oscillation, in: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, ACM, New York, NY. pp. 282:1–282:13. URL: http://doi.acm.org/10.1145/3173574.3173856, doi:10.1145/3173574.3173856.

[10] Fridman, L., Reimer, B., Mehler, B., Freeman, W.T., 2018. Cognitive Load Estimation in the Wild, in: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, ACM, New York, NY. pp. 652:1–652:9. URL: http://doi.acm.org/10.1145/3173574.3174226, doi:10.1145/3173574.3174226.

[11] Giacobino, C., Sardy, S., Diaz-Rodriguez, J., Hengartner, N., 2017. Quantile universal threshold. Electronic Journal of Statistics 11, 4701–4722. URL: https://doi.org/10.1214/17-EJS1366, doi:10.1214/17-EJS1366.

[12] Haapalainen, E., Kim, S., Forlizzi, J.F., Dey, A.K., 2010. Psycho-physiological measures for assessing cognitive load, in: Proceedings of the 12th ACM International Conference on Ubiquitous Computing, ACM, New York, NY. pp. 301–310. URL: http://doi.acm.org/10.1145/1864349.1864395, doi:10.1145/1864349.1864395.

[13] Haynes, C.C., Ericson, B.J., 2021. Problem-Solving Efficiency and Cognitive Load for Adaptive Parsons Problems vs. Writing the Equivalent Code, in: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, Association for Computing Machinery, New York, NY. URL: https://doi.org/10.1145/3411764.3445292, doi:10.1145/3411764.3445292.

[14] He, C., Xing, J., Li, J., Yang, Q., Wang, R., 2015. A New Wavelet Threshold Determination Method Considering Interscale Correlation in Signal Denoising. Mathematical Problems in Engineering 2015. URL: https://doi.org/10.1155/2015/280251, doi:10.1155/2015/280251.

[15] Hollender, N., Hofmann, C., Deneke, M., Schmitz, B., 2010. Integrating cognitive load theory and concepts of human-computer interaction. Computer in Human Behavior 26, 1278–1288.

[16] Hutt, S., Krasich, K., Brockmole, J.R., D'Mello, S.K., 2021. Breaking out of the Lab: Mitigating Mind Wandering with Gaze-Based Attention-Aware Technology in Classrooms, in: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, Association for Computing Machinery, New York, NY. URL: https://doi.org/10.1145/3411764.3445269, doi:10.1145/3411764.3445269.

[17] Jayawardana, Y., Jayawardena, G., Duchowski, A.T., Jayarathna, S., 2021. Metadata-Driven Eye Tracking for Real-Time Applications, in: Proceedings of the 21st ACM Symposium on Document Engineering.

[18] Jiang, X., Atkins, M.S., Tien, G., Bednarik, R., Zheng, B., 2014. Pupil responses during discrete goal-directed movements, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM, New York, NY. pp. 2075–2084. URL: http://doi.acm.org/10.1145/2556288.2557086, doi:10.1145/2556288.2557086.

[19] Kirchner, W.K., 1958. Age differences in short-term retention of rapidly changing information. Journal of Experimental Psychology 4, 352—358. doi:10.1037/h0043688.

[20] Lyu, Y., Luo, X., Zhou, J., Yu, C., Miao, C., Wang, T., Shi, Y., Kameyama, K.i., 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, ACM, New York, NY. pp. 857–866. URL: http://doi.acm.org/10.1145/2702123.2702399, doi:10.1145/2702123.2702399.

[21] Mallat, S., Hwang, W.L., 1992. Singularity Detection and Processing with Wavelets. IEEE Transactions on Information Theory 38, 617–643.

[22] Marshall, S.P., 2000. Method and Apparatus for Eye Tracking Monitoring Pupil Dilation to Evaluate Cognitive Activity. US Patent No. 6,090,051.

[23] Marshall, S.P., 2002. The Index of Cognitive Activity: Measuring Cognitive Workload, in: Proceedings of the 7^{th} Human Factors Meeting, IEEE.

[24] Nyström, M., Holmqvist, K., 2010. An adaptive algorithm for fixation, saccade, and glissade detection in eyetracking data. Behaviour Research Methods 42, 188–204.

[25] Oviatt, S., 2006. Human-centered Design Meets Cognitive Load Theory: Designing Interfaces That Help People Think, in: Proceedings of the 14th ACM International Conference on Multimedia, ACM, New York, NY. pp. 871–880. URL: http://doi.acm.org/10.1145/1180639.1180831, doi:10.1145/1180639.1180831.

[26] Palinko, O., Kun, A.L., 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, ACM, New York, NY. pp. 413–416. URL: http://doi.acm.org/10.1145/2168556.2168650, doi:10.1145/2168556.2168650.

[27] Pedrotti, M., Mirzaei, M., Tedesco, A., Chardonnet, J.R., Merienne, F., Benedetto, S., Baccino, T., 2014. Automatic stress classification with pupil diameter analysis. International Journal of Human-Computer Interaction 30, 1–17. doi:10.1080/10447318.2013.848320.

[28] Ren, P., Barreto, A., Huang, J., Gao, Y., Ortega, F.R., Adjouadi, M., 2014. Off-line and on-line stress detection through processing of the pupil diameter signal. Annals of biomedical engineering 42, 162–176.

[29] Savitzky, A., Golay, M.J.E., 1964. Smoothing and differentiation of data by simplified least squares procedures. Analytical Chemistry 36, 1627–1639. URL: http://pubs.acs.org/doi/abs/10.1021/ac60214a047.

[30] Schafer, R.W., 2011. What Is a Savitzky-Golay Filter? [Lecture Notes]. IEEE Signal Processing Magazing 28, 111–117. URL: http://dx.doi.org/10.1109/MSP.2011.941097, doi:10.1109/MSP.2011.941097.

[31] Sweller, J., 1988. Cognitive load during problem solving: Effects on learning. Cognitive science 12, 257–285.

[32] Sweller, J., 1994. Cognitive load theory, learning difficulty, and instructional design. Learning and Instruction 4, 295–312.

[33] Wang, X., Bylinskii, Z., Castelhano, M., Hillis, J., Duchowski, A., 2021. EMICS'21: Eye Movements as an Interface to Cognitive State. Association for Computing Machinery, New York, NY. CHI EA '21, pp. 1–6. URL: https://doi.org/10.1145/3411763.3441357, doi:10.1145/3411763.3441357.

[34] Wang, X., Bylinskii, Z., Castelhano, M., Hillis, J., Duchowski, A.T., 2020. EMICS'20: Eye Movements as an Interface to Cognitive State, in: Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems, Association for Computing Machinery, New York, NY. pp. 1—-4. URL: https://doi.org/10.1145/3334480.3381062, doi:10.1145/3334480.3381062.

[35] Wilkinson, M.D., Dumontier, M., Aalbersberg, I., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J., da Silva Santos, L.B., Bourne, P.E., Bouwman, J., Brookes, A.J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C.T., Finkers, R., Gonzalez-Beltran, A., Gray, A.J.G., Groth, P., Goble, C., Grethe, J.S., Heringa, J., 't Hoen, P.A.C., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S.J., Martone, M.E., Mons, A., Packer, A.L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S.A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M.A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J., Mons, B., 2016. The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 3. URL: https://doi.org/10.1038/sdata.2016.18, doi:10.1038/sdata.2016.18.

[36] Yuksel, B.F., Oleson, K.B., Harrison, L., Peck, E.M., Afergan, D., Chang, R., Jacob, R.J.K., 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.