# Re-streamed Validation of Real-Time Cognitive Load Measurement

Andrew T. Duchowski[1][0000−0003−1681−7878], Samaneh Zamanifard[1][0000−0001−8736−1690], Sonja Cecchetti[2][0009−0008−2439−0361], Marco Cavallo[2][0000−0002−4784−0803], and Andy Edmonds[3][0000−0002−1447−1861]

[1] Clemson University, Clemson SC, USA
[duchowski, szamani]@clemson.edu
https://clemson.edu

[2] eCampus University, Novedrate, Italy
cecchettisonja@gmail.com, marco.cavallo@uniecampus.it
https://uniecampus.it

[3] Quora, CA, USA
andyed@gmail.com
https://quora.com

**Abstract.** Real-time measurement of cognitive load is important for interactive, adaptive systems where task difficulty could be adjusted to match the user's level of engagement. Although some attempts have been made towards windowed versions of earlier offline measures, to date no real-time measurement of cognitive load is as yet available. Here, three different interactive scenarios are used to evaluate performance of Duchowski's [14] recently introduced real-time cognitive load metric based on pupil oscillation in response to task difficulty. Although analysis re-streams offline data, results suggest that the approach is effective in distinguishing between high and low levels of cognitive load.

**Keywords:** Cognitive load · Real-time filtering · Human-computer interaction.

## 1 Introduction

Traditional approaches to cognitive load assessment rely on subjective instruments such as NASA-TLX [22], secondary-task paradigms [5], or batch analysis of physiological recordings [17]. While these methods have produced robust findings in controlled laboratory settings, they impose unacceptable latency for closed-loop adaptive systems: a measure computed in minutes or hours after an interaction cannot guide real-time interface adaptation. This limitation motivates the need for real-time or near-real-time cognitive-load measurement.

Cognitive load is not static across an activity; it can increase during complex subtasks, decrease after feedback, or spike when users encounter uncertainty, interruptions, or unfamiliar interface elements [3,34,44]. A single post-task score can indicate that a task was demanding, but it cannot identify when the demand occurred or which task event caused it. Real-time measurement is therefore important for adaptive systems, intelligent tutoring systems, human-computer interaction, safety-critical work, and immersive learning environments, where systems may need to adjust task difficulty, pacing, feedback, or interface complexity while the user is still engaged in the task. The development of a validated, low-latency physiological proxy for cognitive load, therefore, represents a key open problem in human-centered computing [12].

## 2 Background & Related Work

Metrics related to the measurement of cognitive load have traditionally been derived from pupillometry, where the Task-Evoked Pupillary Response (TEPR) was used to show pupil dilation in response to cognitively demanding tasks relative to baseline [1]. Typical tasks involved arithmetic [23]. Although TEPR-type metrics have endured, more recent approaches to cognitive load measurement have considered evaluation of microsaccade magnitude and amplitude [13,29] and pupil diameter oscillation [12,16]. However, although Duchowski [14] provided real-time implementation of their earlier Index of Pupillary Activity (IPA) And Low/High Index of Pupillary Activity (LHIPA), validation of its efficacy has mainly been restricted to simulation and analysis of data from one study involving the n-back and Color Visual Short-Term Memory (CVSTM) tasks.

### 2.1 Cognitive Load and the Need for Real-Time Measurement

Cognitive load refers to the mental effort required to process task-relevant information, maintain task goals, and execute cognitive operations under limited working-memory capacity [26,41]. In human-computer interaction, cognitive workload is an important construct for evaluating interactive systems because modern interfaces often require users to divide attention across multiple information sources, decisions, and interaction channels [28].

Prior work has explored several real-time indicators of cognitive load. Eye tracking and pupillometry are commonly used because pupil dilation, fixation patterns, blink behavior, and gaze dynamics can reflect changes in processing demand [2,10,20,31]. Eye tracking is a promising approach for real-time cognitive-load estimation because gaze and pupil data can be collected continuously during interaction. Pupil diameter has long been associated with mental effort. A foundational study showed that pupil diameter changes with memory load during short-term memory processing [27]. Subsequent eye-tracking research has used pupil diameter, fixation behavior, saccades, blink patterns, and related measures to infer attention, task difficulty, and cognitive processing [24,31]. Nevertheless, raw pupil diameter is difficult to interpret directly because it is affected by luminance, fatigue, arousal, individual differences, and measurement artifacts. These limitations motivate derived pupillometric measures that attempt to isolate cognitively meaningful components of pupil dynamics.

### 2.2 Pupillary Oscillation Metrics: IPA, LHIPA, RIPA, and LF/HF

The measurement of cognitive load through pupillary oscillation rests on a well-established physiological foundation. Pupil diameter is continuously modulated by the autonomic nervous system, producing spontaneous oscillations commonly described as pupillary hippus or pupillary unrest [4,40]. Turnbull et al. further examined the autonomic origin of pupillary hippus and found evidence that hippus is primarily linked to parasympathetic rather than sympathetic nervous system activity [43]. These spontaneous fluctuations make the pupil not only a light-regulating aperture, but also a dynamic physiological signal that can reflect arousal, autonomic regulation, and cognitive effort.

Frequency-domain analysis provides a way to decompose this signal into components that may respond differently to luminance and cognitive demand. Peysakhovich et al. analyzed task-evoked pupil responses using power spectral density (PSD) and defined a low-frequency (LF) band from 0 to 1.6 Hz and a high-frequency (HF) band from 1.6 to 4 Hz [38]. Their results showed that the LF/HF ratio of pupillary PSD was sensitive to cognitive load while being less affected by luminance than raw pupil diameter, making it promising for ecological workload measurement. This result is important because luminance sensitivity is one of the main obstacles to using pupil diameter directly as a cognitive-load measure outside tightly controlled laboratory conditions.

Duchowski's recent LF/HF formulation extends this frequency-domain approach by providing a first-principles treatment of pupillary oscillation as a real-time cognitive-load signal [14]. The method derives three computational approaches for estimating the LF/HF ratio: the Fast Fourier Transform (FFT), the Discrete Wavelet Transform (DWT), and a real-time Butterworth infinite impulse response (IIR) filter. The IIR implementation is particularly important for adaptive systems because it can operate on short streaming windows with approximately one second of latency, making it suitable for moment-to-moment interaction analysis than earlier metrics requiring longer temporal windows.

A key implication of the LF/HF approach is that the direction of change is task-dependent. Duchowski distinguishes between at least two task archetypes [14]. In computation-dominated tasks, such as mental arithmetic, both LF and HF amplitudes may increase, but LF increases disproportionately, producing an increase in the LF/HF ratio. In information-processing tasks, such as memory recall, HF activity may increase while LF remains comparatively stable, producing a decrease in the LF/HF ratio. Therefore, LF/HF should not be interpreted as a universal monotonic scale in which higher values always mean higher load. Instead, LF/HF changes must be interpreted in relation to task demands and, where possible, supported by convergent behavioral or physiological evidence.

## 3 Validation Across Diverse Contexts

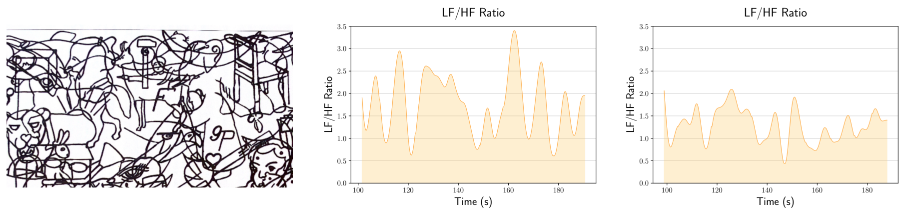
Three disparately varied scenarios are used to evaluate performance of the LF/HF real-time cognitive load filter: visual search of a jumbled image by two participant groups; those with (mild) Cognitive Impairment and Healthy Controls; interaction with an Artificial Intelligence agent, where the agent was set up to produce LLM-generated responses to either align or not align with students' learning styles; and perusal of web Search Engine Response Pages (SERPs). Analysis was performed offline but in a simulated real-time manner, i.e., re-streaming (or re-playing) of the data in a manner similar to Jayawardena et al. [25].

### 3.1 Data from Detection of Mild Cognitive Impairment

The first study is part of the CogET project and aims to evaluate whether eye movement parameters could serve as digital biomarkers for the early detection of cognitive decline [7,8].

**Experimental Design and Conditions.** The study was conducted in accordance with the Declaration of Helsinki, received approval from the Ethics Commission of eCampus University (Protocol No. 07/2024, 5 December 2024), and was pre-registered on the Open Science Framework (OSF). All participants provided written informed consent prior to participation. This cross-sectional study included two groups: healthy controls (HC) and individuals with cognitive impairment (CI). Participants were pre-screened and assigned to the HC or CI group based on a comprehensive battery of neuropsychological assessments. Eye movements were recorded using a Gazepoint GP3 remote eye tracker operating at a sampling rate of 60 Hz, with a manufacturer-reported spatial accuracy of $0.5°–1.0°$ of visual angle. The experiment was run on a Dell Inspiron 15 laptop equipped with an Intel Core i5-1235U processor, 16 GB of RAM, and Windows 11. The eye tracker was controlled via a custom PsychoPy script (Release 2023.1.3) [37]. Stimuli were shown on a 24-inch Dell HD monitor with a resolution of $1920\times1080$ pixels, positioned approximately 60–65 cm from the participants' eyes (mean distance 62.45 cm) and at a 45° angle, in accordance with the Gazepoint manual instructions. Participants were seated comfortably in front of the screen in a quiet, darkened room. Although no chin rest was used, they were instructed to minimize head movements throughout the recording session. Before data acquisition, a standard five-point calibration procedure was performed.

**Stimuli.** Participants were presented with a single visual stimulus for 90 s and instructed to perform a visual search task to find, recognize, and name the figures displayed on the screen [15]. The stimulus consisted of a black-and-white line drawing containing 50 overlapping figures, including animals, objects, letters, and numbers. The superimposition of the figures makes individual items difficult to identify, requiring participants to actively scan the entire visual field and segment target elements from the background. The image was adapted from the original stimulus of Rey [39], which included 72 figures; in the present version, 22 figures were removed, resulting in a final set of 50 figures. This type of stimulus is commonly used to assess figure-ground segmentation abilities [33] and has previously been employed in memory assessment protocols [6,36], see Fig. 1(a).




(a) jumbled image (b) LF/HF ratio (HC) (c) LF/HF ratio (CI)

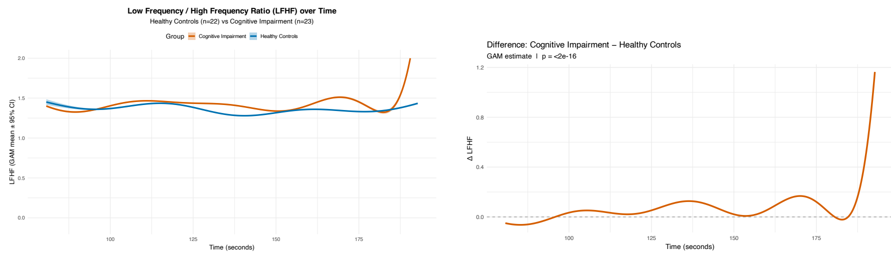
**Fig. 1.** The task featured (a) a jumbled image *Test Delle Figure Aggrovigliate* (translated to "Tangled Figure"), from Mondini and Mapelli © 2022 [35], used with permission. Example smoothed and normalized LF/HF ratio captured during each of the tasks performed by a (b) representative Healthy Control (HC), and (c) patient with (Mild) Cognitive Impairment (CI). In this instance, the HC appears to be more engaged over the jumbled image then the CI individual.

**Participants.** Participants were recruited from a nursing home in Cavallermaggiore, Cuneo, Italy. The study included 60 older adults, comprising 30 healthy controls (HC) and 30 individuals with cognitive impairment (CI). All participants underwent a neuropsychological assessment, including the Mini-Mental State Examination (MMSE) [19], a brief screening measure of global cognitive status with scores ranging from 0 to 30, and the Repeatable Battery for the Assessment of Neuropsychological Status (RBANS), administered to obtain a broader profile of cognitive functioning. Frontal-executive abilities were further assessed using the Frontal Assessment Battery (FAB) [11], which is sensitive to impairments in cognitive flexibility, inhibition, self-monitoring, and planning. Affective status was gauged using the Geriatric Depression Scale (GDS) [45], administered as an oral yes/no interview. Functional independence was assessed using the Barthel Index [32], which provides an index of autonomy and need for assistance in daily activities. When applicable, raw scores were corrected using the respective standardized and validated neuropsychological norms. Neuropsychological assessment was followed by a 30-minute eye-tracking protocol consisting of four tasks previously used in related studies [9].

**Results.** We used Generalized Additive Models (GAMs) to compare the temporal trajectory of the low-frequency to high-frequency (LF/HF) ratio between Healthy Controls ($n = 22$) and participants with Cognitive Impairment ($n = 23$) in a between-subjects design. Fifteen participants were excluded a priori because they failed pre-registered validation criteria (eye tracker calibration).

An analysis-of-deviance comparison of nested GAMs showed that allowing the smooth time course of LF/HF to differ between groups significantly improved model fit ($F(10.04, 206, 669) = 184.43, p < 0.01$).

Visualization of the estimated difference curve (see Fig. 2(b)) indicates that Healthy Controls exhibited a slightly lower LF/HF ratio (hence increased cog-




(a) LF/HF ratio. (b) LF/HF ratio difference.

**Fig. 2.** Butterworth filter response when inspecting the scrambled image. The LF/HF ratio decreases for (a) Healthy Controls suggesting greater cognitive load. The difference in the Generalized Additive Model (GAM) estimate is (b) significant. Increased cognitive load could suggest greater cognitive engagement.

nitive load) (see Fig. 2(a), peak difference $\approx 1.163$), although the absolute magnitude of the effect was small (mean $|\Delta| \approx 0.087$).

### 3.2 Data from Interaction with an Artificial Intelligence Agent

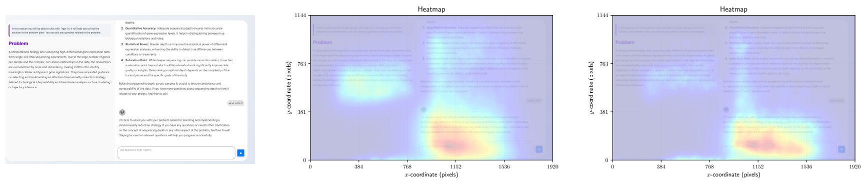
This study examined whether aligning responses generated by a Large Language Model (LLM) with students' learning styles influenced cognitive load during problem-solving interaction.

**Experimental Design and Conditions.** This study used a within-subjects experimental design. The independent variable was the response condition, with two levels: aligned and non-aligned. The primary dependent variable for the cognitive-load analysis was the LF/HF ratio, computed from the pupil-diameter signal captured during interaction. Each participant completed two tasks within the same self-selected topic: one in the aligned condition and one in the non-aligned condition. Participants' learning styles were assessed using the Index of Learning Styles (ILS) questionnaire [18]. The resulting ILS profile was used to prompt the LLM to generate responses aligned with each participant's identified learning style in the aligned condition.

In the aligned condition, the LLM generated responses tailored to the participant's ILS profile. Alignment was operationalized by adapting response aspects including representational format, level of abstraction, interaction style, and organizational structure, according to the participants' ILS profile.

In the non-aligned condition, the LLM generated responses not adapted to the participant's learning style and followed a generic explanatory format.

The order of the two conditions was counterbalanced across participants to reduce order and learning effects. Different but comparable problems were used across the two conditions to avoid repetition effects while maintaining similar task difficulty and structure. Across both conditions, the interface layout, topic domain, interaction modality, and task format were held constant.




(a) LLM stimulus. (b) aligned response. (c) not aligned response.

**Fig. 3.** Heatmaps showing gaze in the aggregate from a representative individual interacting with (a) the AI agent, (b) where responses were aligned or (c) not aligned. Gaze data, such as depicted by these heatmaps do not readily divulge any particular differences in visual attention between the two conditions.

Operationally, after completing the ILS questionnaire, participants completed an eye tracker calibration procedure and selected a topic of interest. They then completed two problem-solving tasks within that topic: one in the aligned condition and one in the non-aligned condition. During both tasks, participants interacted with the LLM while gaze and pupil data were recorded continuously.

**Stimuli.** The task stimuli consisted of problem-solving prompts presented to the LLM. During interaction, the LLM generated instructional responses intended to support students during problem-solving. Problem pairs were designed to require similar forms of reasoning, including interpreting the problem statement, consulting the LLM-generated explanation, and formulating follow-up questions or solution steps.

Participants completed two different problems within the same self-selected topic. Available topics included psychology, data science, and computer networks. The paired problems differed in surface content to avoid repetition effects, but were matched within topic for general structure, expected difficulty, and interaction demands.

All tasks were presented within the LLM interface, see Fig. 3(a), which was organized into three predefined functional regions corresponding to the AOIs used in the eye-tracking analysis. The operational definitions of these AOIs are provided in Table 1.

**Table 1.** Areas of Interest (AOIs) Used in the Eye-Tracking Analysis

| AOI | Operational definition |
|---|---|
| Problem box | Region containing the problem statement or task prompt. |
| Solution box | Region containing the LLM-generated explanation or response. |
| Text box | Region in which participants typed or reviewed their own prompts, questions, or solution steps. |

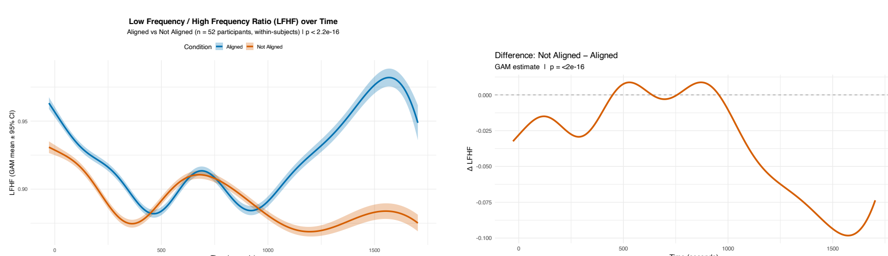
During the problem-solving tasks, eye movements and pupil diameter were recorded using a Gazepoint GP3 eye tracker operating at 60 Hz. Participants were seated approximately 60 cm from a $1920 \times 1080$ monitor. Before beginning the task, each participant completed a standard 9-point calibration procedure. Calibration quality was checked before data collection, and recalibration was performed when tracking quality was insufficient.

**Participants.** The study recruited 52 undergraduate students from a public university in the southeastern United States. Participants were reached through in-class announcements facilitated by the instructor. To ensure a qualified sample, participants had to be at least 18 years old, currently enrolled as undergraduate students, and had prior experience using LLMs in an educational context.

Participation was entirely voluntary, and students received extra credit as compensation. To maintain ethical standards and ensure there was no coercion, an alternative assignment of equivalent effort was provided to those who opted out or discontinued the study. This research was reviewed and approved by the university's Institutional Review Board (IRB).

**Results.** We examined the time course of the Low-Frequency/High-Frequency (LF/HF) ratio using Generalized Additive Mixed Models (GAMMs) in a within-subjects design. The model included fixed smooth functions of time that were allowed to differ by condition (aligned vs. not aligned) plus a random intercept for participant to account for repeated measures.

Model comparison via analysis of deviance (a direct GAMM analogue of ANOVA for model comparison) showed that the condition-specific curves significantly improved fit over a common curve ($F(10.17, 3, 106, 405) = 261.36, p < 0.01$). The estimated difference curve (see Fig. 4(b)) indicates that the non-aligned condition produced a modestly lower LF/HF ratio (hence increased cog-




(a) LF/HF ratio. (b) LF/HF ratio difference.

**Fig. 4.** Butterworth filter response during interaction with the LLM agent when responses were aligned or not aligned. The LF/HF ratio decreases when responses are not aligned (a) except for the central section during the trial. The difference in the Generalized Additive Mixed Model (GAMM) estimate is (b) significant. A decrease in the LF/HF ratio suggests increased cognitive load.

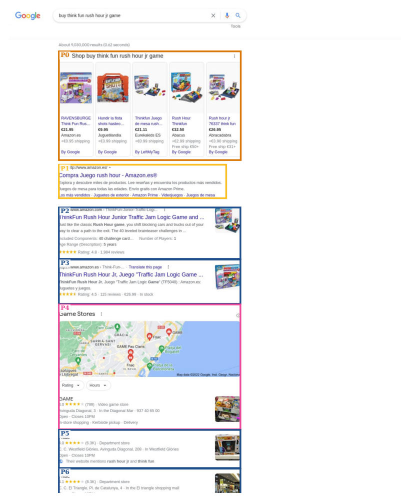



**Fig. 5.** Example AdSERP trial with typed-AOI bounding boxes overlaid: organic results, top-of-page advertisements, native advertisements, and content widgets along the main column. Right-rail content is excluded from the rank-position analysis.

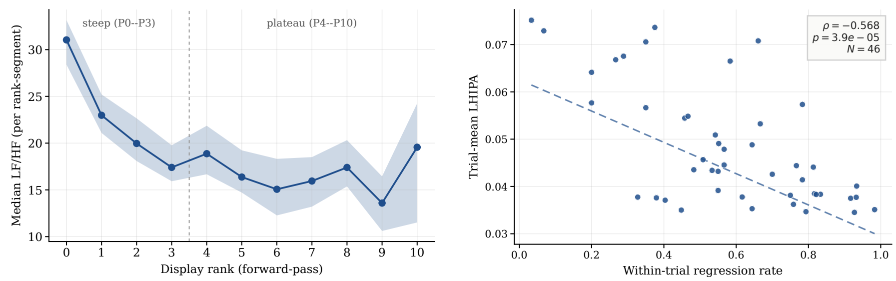
nitive load) for much of the task (see Fig. 4(a) peak difference $\approx -0.098$), although the absolute magnitude of the effect was small (mean $|\Delta| \approx 0.035$).

### 3.3 Data from Advertising Search Engine Results (Web) Pages

We reanalyzed a public dataset of mouse and eye movements on Web Search Engine Results Pages (SERPs) [30], see Fig. 5, applying the Butterworth IIR variant of the LF/HF ratio [14] to every visit to each *rank position* on each SERP, whether a first-pass evaluation or a later return.

**Experimental Design and Conditions.** The independent variable is the SERP rank that a participant chose to evaluate; rank position is an ordinal, participant-driven covariate rather than an experimenter-imposed condition. We computed one LF/HF value per rank-segment that contained at least one second of valid pupil samples (150 samples at the dataset's 150 Hz sampling rate) [14]. A high-water-mark classifier then *labeled* each segment as a first-pass visit or a later return, by replaying the fixation sequence and marking the first time each rank position was reached; subsequent analyses (the forward-pass rank gradient in §Results, and the first-vs-return contrast in Re-evaluation on return) split on this label rather than recomputing LF/HF.

**Stimuli.** Stimuli were 2,776 transactional product-purchase queries of the form "buy *brand product*," spanning 1,320 distinct brands. Each query was executed against a live Google SERP, and the resulting page was captured as a full-page




(a) LF/HF rank gradient.

(b) LHIPA $\times$ regression rate.

**Fig. 6.** Pupillary cognitive load at two granularities from the AdSERP corpus: (a) median Butterworth LF/HF per rank position during forward-pass scanning, with 95% bootstrap confidence intervals (2,000 iterations) shaded around the medians; the dashed line marks the steep (P0–P3) / plateau (P4–P10) boundary, with sample sizes from $N$ = 1,488 rank-segments at P0 down to $N$ = 109 at P10; (b) trial-mean LHIPA decreases with within-trial regression rate across participants ($\rho = -0.568$, $p < 0.01$, $N = 46$): high-regression participants carry higher cognitive load.

screenshot together with the SERP HTML [30]. Participants were instructed to click the result they would typically choose for product purchase within a one-minute window, extended to two minutes after a confirmation prompt; every trial ended with a click. Data were collected before Google's AI Overviews rollout.

The released dataset ships bounding boxes for advertisements only. To support per-position LF/HF analysis we extracted bounding boxes for organic results from the captured screenshots and joined them with semantic types parsed from the captured HTML, then assigned each card a display-order *rank position* along the main column. The rank scale covers organic results and main-column advertisements (top-of-page ads and native ads); content widgets (knowledge panels, image packs, "People also ask," top-places, related searches) and right-rail material are off-axis and excluded from the rank-position analysis.

**Participants.** The dataset contained 2,776 trials from 47 adult participants recorded on a Gazepoint GP3 HD desktop eye tracker at 150 Hz [24,30]. Of these, 2,719 trials yielded at least one rank-segment that satisfied the one-second LF/HF window requirement and were retained for analysis.

**Results — forward-pass rank gradient.** Per-position median LF/HF declined with rank position during forward-pass scanning of the SERP. A Spearman rank correlation across the nineteen position medians in the corpus confirmed a strong monotonic decline ($\rho = -0.749$, $p < 0.01$; Fig. 6(a)).

The decline split cleanly into a steep early phase (P0–P3) and a plateau (P4–P10). The steep phase was perfectly monotone ($\rho = -1.000$ across position medians, the mathematical limit). Pooled across raw rank-segments, a Mann-

Whitney comparison confirmed a sharp discontinuity between the two phases ($U$ = 4,728,133, $p < 0.01$; steep median 22.15, $N$ = 4,056 vs. plateau median 15.41, $N$ = 2,005). The plateau itself showed a directional but non-significant slope ($\rho = -0.393$, $p = 0.38$).

The same direction held at the single-trial level. Among the 1,074 trials with at least three valid rank-segments, the median per-trial slope was $\rho = -0.400$, with 62.0% of slopes negative. Restricting to trials with at least five or seven valid rank-segments yielded median slopes of $-0.200$ and $-0.161$, with 62.3% and 65.9% negative respectively.

We tested against three alternative hypotheses for the gradient. First, that it was an artefact of scroll-related segment selection: restricting to forward-pass segments visited before the participant's first significant downward scroll tightened the cross-position Spearman to $\rho = -0.857$ ($p < 0.05$, $N$ = 7 medians). Second, that the gradient reflected a small subset of participants who scrolled deeper than the rest: with each participant's contribution capped at ten segments per rank position, the full-range Spearman held at $\rho = -0.689$ ($p < 0.01$), a population-wide effect consistent with schema acquisition [42,44]. Third, that it indexed individual differences in evaluation strategy rather than the within-trial cognitive operation: with the 46 participants split by within-trial regression rate (median split), per-participant within-trial slopes were uncorrelated with regression rate ($\rho = -0.020$, $p = 0.90$). The load trajectory is therefore orthogonal to evaluation strategy.

The pupillary signal carries cognitive load at two granularities, and the right tool depends on the question. Per-segment LF/HF resolves the cost of evaluating each result candidate within a trial. The established trial-level LHIPA metric [12] resolves how much load a participant carries across the trial as a whole. Trial-mean LF/HF and LHIPA were anticorrelated as the polarity rule predicts ($\rho = -0.125$, $p < 0.01$, $N$ = 2,416). Across the 46 participants, trial-mean LHIPA was lower — i.e., load was higher — on participants with higher within-trial regression rates ($\rho = -0.568$, $p < 0.01$; Fig. 6(b)): the participants who deliberate more thoroughly across the result set also carry more load.

**Re-evaluation on return.** LF/HF rises on revisits, and it rises more on candidates the participant will reject than on the one they will choose. Across 46 participants and 2,646 paired (trial, position) records, return-visit LF/HF was higher than first-visit LF/HF on the same item, with median paired $\Delta = +6.31$, mean +12.55 (Wilcoxon $p < 0.01$; 59.8% of pairs positive). At the participant level the mean-of-means $\Delta$ was +10.73, with 80% of participants positive ($p < 0.01$).

Splitting paired records by whether the return is to the position eventually clicked sharpens the result. Returns to a candidate the participant rejects carried median paired $\Delta = +7.05$ ($n$ = 1,697, $p < 0.01$); returns to the candidate the participant will choose carried median paired $\Delta = +4.03$ ($n$ = 742, $p < 0.01$). Both rise above the first-visit baseline; the rise is larger on the rejected items. The cost is paid once per return, not amplified across successive returns to the same position: paired LF/HF on the second return did not differ from the first (median $\Delta = +0.03$, $n = 1{,}486$, Wilcoxon $p = 0.09$).

Most return visits began well before the click. The median visit began 10.6 s before the click, and only 5% began within 2 s of it. Revisit fixations are shorter than first-pass fixations, and revisit episodes contain fewer fixations than first-pass episodes. The window is brief, and the pupillary signal reads elevated load per second of dwell.

Baseline drift across trial time runs in the opposite direction. Across forward-pass-only segments, LF/HF at the latest forward-visited position was *lower* than at the earliest (median $\Delta = -1.97$, $p < 0.01$, $N = 1{,}241$). The elevation is also not a trial-difficulty artefact: 89% of trials (2,391 / 2,676) eventually regressed, and first-visit LF/HF on those trials did not separate from first-visit LF/HF on the 285 trials that never regressed (Mann-Whitney $p = 0.14$).

The elevation extends earlier in the trial. We classified every (trial, position) segment by replaying the fixation sequence: first visits before any regression in the trial (*first-pre*), first visits to a previously-unseen position after the trial had already regressed (*first-post*), and return visits (*return*). On a position the participant had not yet seen, LF/HF was higher when the trial had already begun regressing than when it had not (median across matched ranks P1–P4, $\Delta = +5.22$; pooled Mann-Whitney $p < 0.01$, $N_{\text{pre}} = 1{,}982$, $N_{\text{post}} = 3{,}904$).

Under the computation-dominated polarity rule, the return-visit elevation reads as *more* effort on the return than on the original encounter, consistent with re-application of an evaluation criterion that has sharpened across the trial [21]. By the time the user returns to a previously-seen result they have evaluated additional candidates and are comparing against a tighter standard. The asymmetry between rejected and chosen candidates fits this reading: rejecting a candidate against the tighter standard is the comparative work, while the candidate that survives that work is the one that will be clicked.

## 4 Conclusion

Three different interactive scenarios were used to evaluate performance of Duchowski's [14] recently introduced real-time cognitive load metric based on pupil oscillation in response to task difficulty. Results suggest that that the approach is effective in distinguishing between high and low levels of cognitive load. These findings indicate the potentially robust response of the IIR filter for cognitive load detection, carrying implications for interactive systems where task difficulty could be adjusted to match the user's level of engagement in real time. However, real-time evaluation of this approach is still needed and relegated to future work.

## References

1. Ahern, S., Beatty, J.: Pupillary Responses During Information Processing Vary with Scholastic Aptitude Test Scores. Science **205**(4412), 1289–1292 (1979) 2

2. Beatty, J.: Task-evoked pupillary responses, processing load, and the structure of processing resources. Psychological Bulletin **91**(2), 276–292 (1982). https://doi.org/10.1037/0033-2909.91.2.276 2
3. Boehm, U., Matzke, D., Gretton, M., Castro, S., Cooper, J., Skinner, M., Strayer, D., Heathcote, A.: Real-Time Prediction of Short-Timescale Fluctuations in Cognitive Workload. Cognitive Research: Principles and Implications **6**(1), 30 (2021). https://doi.org/10.1186/s41235-021-00289-y 1
4. Bouma, H., Baghus, L.C.J.: Hippus of the Pupil: Periods of Slow Oscillations of Unknown Origin. Vision Research **11**(11), 1345–1351 (1971). https://doi.org/10.1016/0042-6989(71)90016-2, https://pubmed.ncbi.nlm.nih.gov/5148578/ 3
5. Brunken, R., Plass, J.L., Leutner, D.: Direct measurement of cognitive load in multimedia learning. Educational Psychologist **38**(1), 53–61 (2003) 1
6. Carlesimo, Giovanni Augustoand Buccione, I., Fadda, L., Graceffa, A., Mauri, M., Lorusso, S., Bevilacqua, G., Caltagirone, C.: Normative data of two memory tasks: Short-Story recall and Rey's Figure. Nuova Rivista di Neurologia **12**, 1–13 (January 2002) 4
7. Cecchetti, S., Duchowski, A.T., Cavallo, M.: The CogET protocol: a pilot study of neuropsychological assessment with eye tracking for the early detection of cognitive impairment [OSF preregistration of study protocol] (October 2025). https://doi.org/https://doi.org/10.17605/OSF.IO/VG3TD, https://osf.io/vg3td/ 4
8. Cecchetti, S., Duchowski, A.T., Cavallo, M.: Eye-tracking metrics as a digital biomarker for neurocognitive disorders in multiple sclerosis: A scoping review. Brain Sci **15**(2) (2025). https://doi.org/10.3390/brainsci15020149 4
9. Cecchetti, S., Duchowski, A.T., Cavallo, M.: "See What I Mean?" Linguistic Digital Markers of Cognitive Decline in Nursing Home Residents. Journal of Integrative Neuroscience (2026), (In press.) 5
10. Chen, S., Epps, J.: Automatic classification of eye activity for cognitive load measurement with emotion interference. Computer Methods and Programs in Biomedicine **110**(2), 111–124 (2013). https://doi.org/10.1016/j.cmpb.2012.10.021 2
11. Dubois, B., Slachevsky, A., Litvan, I., Pillon, B.: The FAB: A Frontal Assessment Battery at bedside. Neurology **55**(11), 1621–1626 (2000). https://doi.org/10.1212/WNL.55.11.1621 5
12. Duchowski, A.T., Krejtz, K., Gehrer, N.A., Bafna, T., Bækgaard, P.: The Low/High Index of Pupillary Activity. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. p. 1–12. CHI '20, Association for Computing Machinery, New York, NY (2020). https://doi.org/10.1145/3313831.3376394 2, 11
13. Duchowski, A., Krejtz, K., Żurawska, J., House, D.: Using Microsaccades to Estimate Task Difficulty During Visual Search of Layered Surfaces. IEEE Transactions on Visualization and Computer Graphics **213** (2019). https://doi.org/10.1109/TVCG.2019.2901881 2
14. Duchowski, A.T.: Real-Time Cognitive Load Measurement of the Pupillary Oscillation. Proceedings of the ACM Computer Graphics and Interactive Techniques (PACMCGIT) **9**(2) (2026). https://doi.org/10.1145/3803537 1, 2, 3, 9, 12
15. Duchowski, A.T., Cavallo, M., Cecchetti, S.: Dynamical Time Course Analysis of Real-Time Gaze Metrics. In: Proceedings of the 2026 Eyes4Access Workshop: New Frontiers of Eye Tracking in Accessible Interaction and Communication. Eyes4Access @ ETRA '26, Association for Computing Machinery, New York, NY (2026) 4

16. Duchowski, A.T., Krejtz, K., Krejtz, I., Biele, C., Niedzielska, A., Kiefer, P., Raubal, M., Giannopoulos, I.: 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. pp. 282:1–282:13. CHI '18, ACM, New York, NY (2018). https://doi.org/10.1145/3173574.3173856 2
17. Fairclough, S.H.: Fundamentals of physiological computing. Interacting with computers **21**(1-2), 133–145 (2009) 1
18. Felder, R.M., Silverman, L.K.: Learning and teaching styles in engineering education. Engineering education **78**(7), 674–681 (1988) 6
19. Folstein, M.F., Folstein, S.E., McHugh, P.R.: "Mini-mental state": A practical method for grading the cognitive state of patients for the clinician. Journal of Psychiatric Research **12**(3), 189–198 (1975). https://doi.org/10.1016/0022-3956(75)90026-6 5
20. Gorin, H., Patel, J., Qiu, Q., Merians, A., Adamovich, S., Fluet, G.: A review of the use of gaze and pupil metrics to assess mental workload in gamified and simulated sensorimotor tasks. Sensors **24**(6), 1759 (2024). https://doi.org/10.3390/s24061759 2
21. Haider, H., Frensch, P.A.: Information reduction during skill acquisition: The influence of task instruction. Journal of Experimental Psychology: Applied **5**(2), 129–151 (1999). https://doi.org/10.1037/1076-898X.5.2.129 12
22. Hart, S.G., Staveland, L.E.: Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In: Advances in psychology, vol. 52, pp. 139–183. Elsevier (1988) 1
23. Hess, E.H., Polt, J.M.: Pupil Size in Relation to Mental Activity during Simple Problem-Solving. Science **143**(3611), 1190–1192 (March 1964) 2
24. Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H., van de Weijer, J.: Eye Tracking: A Comprehensive Guide to Methods and Measures. Oxford University Press, Oxford (2011) 2, 10
25. Jayawardena, G., Jayawardana, Y., Jayarathna, S., Högström, J., Papa, T., Akkil, D., Duchowski, A.T., Peysakhovich, V., Krejtz, I., Gehrer, N., Krejtz, K.: Toward a real-time index of pupillary activity as an indicator of cognitive load. Procedia Computer Science **207**, 1331–1340 (2022). https://doi.org/10.1016/j.procs.2022.09.189 4
26. Kahneman, D., Beatty, J.: Pupil diameter and load on memory. Science **154**(3756), 1583–1585 (1966) 2
27. Kahneman, D., Beatty, J.: Pupil diameter and load on memory. Science **154**(3756), 1583–1585 (1966). https://doi.org/10.1126/science.154.3756.1583 2
28. Kosch, T., Karolus, J., Zagermann, J., Reiterer, H., Schmidt, A., Woźniak, P.W.: A Survey on Measuring Cognitive Workload in Human-Computer Interaction. ACM Computing Surveys **55**(13s), 1–39 (2023). https://doi.org/10.1145/3582272 2
29. Krejtz, K., Duchowski, A.T., Niedzielska, A., Biele, C., Krejtz, I.: Eye tracking cognitive load using pupil diameter and microsaccades with fixed gaze. PLOS ONE **13**(9), 1–23 (September 2018). https://doi.org/10.1371/journal.pone.0203629 2
30. Latifzadeh, K., Gwizdka, J., Leiva, L.A.: A versatile dataset of mouse and eye movements on search engine results pages. In: Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '25, Association for Computing Machinery, New York, NY, USA (2025). https://doi.org/10.1145/3726302.3730325 9, 10

31. Mahanama, B.M.B., Jayawardana, Y.J.Y., Rengarajan, S.R.S., Jayawardena, G.J.G., Chukoskie, L.C.L., Snider, J.S.J., Jayarathna, S.J.S.: Eye movement and pupil measures: A review. Frontiers in Computer Science **3**, 733531 (2022). https://doi.org/10.3389/fcomp.2021.733531 2
32. Mahoney, F.I., W., B.D.: Functional evaluation: the Barthel Index. Maryland State Medical Journal **14**, 56–61 (1965) 5
33. Marchetti, C.: Abilità Visuo-Construtive e Memoria Spaziale Nel Disturbo Dello Spettro Dell'Autismo Senza Disabittà Intellettiva: La Figura Complessa Di Rey (2022), bachelor's Thesis, Università Degli Studi Di Padova 4
34. McLaughlin, A.C., Rogers, W.A., Fisk, A.D.: Feedback support for training: Accounting for learner and task. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting. vol. 52, pp. 2057–2061 (2008). https://doi.org/10.1177/154193120805202605 1
35. Mondini, S., Mapelli, D. (eds.): Esame Neuropsicologico Breve 3 (ENB-3). Raffaello Cortina Editore, Milano, Italy (2022), https://www.raffaellocortina.it/scheda-libro/autori-vari/esame-neuropsicologico-breve-3-9788832854831-3806.html 5
36. Mondini, S., Montemurro, S., Pucci, V., Ravelli, A., Signorini, M., Arcara, G.: Global Examination of Mental State: An open tool for the brief evaluation of cognition. Brain and Behavior **12**(8), e2710 (2022). https://doi.org/10.1002/brb3.2710 4
37. Peirce, Jonathan Gray, J.R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E., Lindeløv, J.K.: PsychoPy2: Experiments in behavior made easy. Behavior Research Methods **51**(1), 195–203 (2019). https://doi.org/10.3758/s13428-018-01193-y 4
38. Peysakhovich, V., Causse, M., Scannella, S., Dehais, F.: Frequency Analysis of a Task-Evoked Pupillary Response: Luminance-Independent Measure of Mental Effort. International Journal of Psychophysiology **97**(1), 30–37 (2015). https://doi.org/10.1016/j.ijpsycho.2015.04.019 3
39. Rey, A.: I disturbi della memoria ed il loro esame psicometrico. Edizioni OS (1968) 4
40. Stark, L., Campbell, F.W., Atwood, J.: Pupil Unrest: An Example of Noise in a Biological Servomechanism. Nature **182**(4639), 857–858 (1958) 3
41. Sweller, J.: Cognitive load theory, learning difficulty, and instructional design. Learning and Instruction **4**(4), 295–312 (1994). https://doi.org/10.1016/0959-4752(94)90003-5 2
42. Sweller, J.: Cognitive load theory, learning difficulty, and instructional design. Learning and Instruction **4**(4), 295–312 (1994) 11
43. Turnbull, P.R.K., Irani, N., Lim, N., Phillips, J.R.: Origins of Pupillary Hippus in the Autonomic Nervous System. Investigative Ophthalmology & Visual Science **58**(1), 197–203 (2017). https://doi.org/10.1167/iovs.16-20785 3
44. Wirzberger, M., Esmaeili Bijarsari, S., Rey, G.D.: Embedded interruptions and task complexity influence schema-related cognitive load progression in an abstract learning task. Acta Psychologica **179**, 30–41 (2017). https://doi.org/10.1016/j.actpsy.2017.07.001 1, 11
45. Yesavage, J.A., Brink, T.L., Rose, T.L., Lum, O., Huang, V., Adey, M., Leirer, V.O.: Development and validation of a geriatric depression screening scale: a preliminary report. Journal of Psychiatric Research **17**(1), 37–49 (1982). https://doi.org/10.1016/0022-3956(82)90033-4 5