A Comprehensive Analysis of the Influence of Cognitive Load on Physiological Signals in Virtual Reality

Jorge Pina[1]* Edurne Bernal-Berdun[1] Daniel Martin[1] Sandra Malpica[1] Carmen Real[1]
Alberto Barquero[1] Pablo Armañac-Julián[1 2] Jesus Lazaro[1 2] Alba Martín-Yebra[1 2]
Belen Masia[1] Ana Serrano[1]

[1]Universidad de Zaragoza - I3A [2]CIBER-BBN

Figure 1: We investigate the relationship between cognitive load and a number of objective and subjective measures while conducting visual search tasks in immersive environments. **Left**: We have devised an experiment in which participants are prompted with a visual search task and, in some cases, with an additional auditory task, which requires them to divert their cognitive resources. **Center**: We have measured several physiological signals during the task, such as electrocardiogram (ECG) or electrodermal activity (EDA), as well as collected task performance measures and subjective responses on workload. **Right:** We have exhaustively analyzed how markers derived from those signals, as well as their evolution over time, relate to increases in cognitive demand.

Abstract

The study of cognitive load (CL) has been an active field of research across disciplines such as psychology, education, and computer graphics and visualization for decades. In the context of Virtual Reality (VR), understanding mental demand becomes particularly relevant, as immersive experiences increasingly integrate multisensory stimuli that require users to distribute their limited cognitive resources. In this work, we investigate the effects of cognitive load during a search task in VR, combining objective and subjective measurements, including physiological signals and validated questionnaires. We designed an experiment in which participants performed a visual search task under two cognitive load conditions (either alone or while responding to a concurrent auditory task) and across two visual search areas (90° and 360°). We collected a rich dataset comprising task performance, eye tracking, electrocardiogram (ECG), electrodermal activity (EDA), photoplethysmography (PPG), and inertial measurements, along with subjective assessments (NASA-TLX questionnaires). Our analysis shows that increased cognitive load hinders visual search performance and affects multiple physiological markers, offering a solid foundation for future research on cognitive load in multisensory virtual environments.

Index Terms: Cognitive load, physiological signals, perception, visual search

*e-mail: jpinac@unizar.es

1 Introduction

Virtual Reality (VR) is becoming increasingly widespread and popular, driven by advances in hardware and software that have enabled its integration into a wide range of domains, including education, healthcare, entertainment, and professional training, among many others [2, 31]. As VR technologies continue to mature, there is a growing interest in developing experiences that leverage multiple sensory modalities (primarily visual, auditory, and haptic) to create richer and more compelling content [37, 40]. These multisensory inputs not only increase realism and immersion, but can also be used, for example, to improve spatial perception [13, 30] or guide attention [12, 16].

However, although the benefits of audiovisual integration are gathering attention, much remains untapped about the challenges and implications arising from combining audiovisual stimuli, especially in terms of mental demand and cognitive load. Cognitive load (CL) refers to the amount of mental resources that humans allocate to working memory at any given time [44, 45]. Sweller's theory [50] posits that mental resources are limited and that the cognitive load derived from the allocation of those resources can be distinguished in three categories: *intrinsic*, *germane* and *extraneous*. While the first two are considered inherent to the task or the user, respectively, *extraneous* cognitive load is considered external to those and is affected by the way the experience is conveyed. Some aspects of an experience can contribute to an increase in the mental resources required, making the task substantially more difficult. In fact, in many virtual experiences, users are exposed to multiple sources of stimuli that do not necessarily contribute to their main goal and can lead to an increase in extraneous cognitive load, which ultimately hinders user performance [37].

Measuring and quantifying cognitive load levels is an ongoing problem encompassing disciplines such as psychology [45, 51], education [9], and even computer graphics and visualization [3], for which no definite solution has yet been found. However, there seems to be a consensus that some changes in physiological signals could be highly correlated with cognitive load and its fluctuation, such as in eye tracking [33], electrocardiogram (ECG) [1], electrodermal activity (EDA) [19], or respiration [25], together with subjective measurements, such as the well-known NASA-TLX questionnaire [28]. However, while several recent works have explored the prediction [46] or classification [14] of cognitive load in immersive environments using selected physiological markers, comprehensive analyses that examine a broad set of signals jointly and account for their temporal evolution remain scarce. In this work, we present and analyze a novel dataset that, to the best of our knowledge, offers the most extensive combination to date of physiological signals for investigating cognitive load in a systematically designed VR scenario (see Fig. 1).

In particular, we focus on visual search, one of the most common tasks in VR [36], and analyze how task performance and physiological signals vary under different levels of cognitive load and across different search areas. We designed an experiment in which 36 participants performed a visual search task under two conditions: either alone (low cognitive load, LCL), or while concurrently performing an auditory task that required detecting and responding to specific stimuli (high cognitive load, HCL). The task was performed in two visual search area configurations: a narrow 90° search area and a wider 360° one. Throughout the experiment, we recorded a broad set of physiological signals, including eye tracking, ECG, EDA, and photoplethysmography (PPG), along with inertial measurements of head and body rotation, subjective assessments (NASA-TLX questionnaires), and task performance metrics.

From these physiological signals, we derived a set of physiological markers commonly used in the study of cognitive load, allowing us to study their behavior under varying conditions. We focused on the mean heart rate (HR), the mean respiration frequency (RF), the median of the EDA tonic component (EDA_{tonic}), and the median of the EDA phasic component (EDA_{phasic}). Our analysis revealed several key insights. First, we observed that high levels of cognitive load induced by audiovisual multitasking significantly impaired visual search performance. Second, physiological markers such as EDA_{tonic}, RF, and HR were found to correlate with cognitive load levels, with EDA_{tonic} exhibiting consistent behavior, while RF and HR appeared more susceptible to confounding factors such as movement during the search. Finally, we observed that HR, RF, and EDA_{phasic} varied noticeably over time, suggesting their sensitivity to temporal dynamics and experience duration. In contrast, EDA_{tonic} remained temporally stable, reinforcing its potential as a robust marker for sustained cognitive load.

In summary, our contributions include:

- A rich and diverse dataset collected during a visual search task in VR under both low and high cognitive load conditions (with and without a concurrent auditory task), encompassing physiological signals (including eye tracking, electrocardiogram (ECG), electrodermal activity (EDA), photoplethysmography (PPG), and respiration (RESP)), inertial measurements of body and head movement, subjective measurements of cognitive load (NASA-TLX questionnaire), and task performance.
- A detailed analysis of how cognitive load affects performance in visual search tasks.
- A comprehensive study of the relationship between cognitive load and multiple physiological markers, analyzed across different visual search areas.
- An exploration of the temporal evolution of these markers for each experiment condition and their dynamic correlation to cognitive load over time.

We believe this work provides a timely foundation for future studies investigating the role of physiological markers in assessing cognitive load within multisensory virtual environments.

2 Related Work

2.1 Cognitive Load

The study of cognitive load (CL) has been an active research area [18, 52, 53] since the publication of the seminal work from Sweller [50], which established the basis of the cognitive load theory. It defined CL as the amount of cognitive resources allocated to completing a specific task. Building upon it, different works have studied the influence of CL in task-oriented scenarios, across domains such as learning, memory, or navigation, among others, showing that high levels of cognitive load generally hinder task performance [4, 10, 47]. However, as posited by the Yerkes-Dodson law [58], and later revised in different scenarios [20, 22, 34, 48], performance follows an inverted U-shape, and some moderate workload can increase performance up to a peak. Numerous studies have emphasized the need to design interfaces and content to maintain CL at an "optimal" level to improve task performance and experience [29, 43]. Achieving this goal involves measuring and characterizing CL, which has no definite solution yet, and has been an active field of research for years. This becomes significantly more challenging in multisensory environments such as virtual reality (VR), where multiple sensory inputs simultaneously influence cognitive demands [17, 37, 54], complicating the estimation of the amount of cognitive resources occupied by each task at any given time.

2.2 Physiological Markers

Several works have studied the correlation between cognitive load (CL) and the fluctuation of different physiological markers, such as eye-tracking data [33] (including pupillometry [41], blinks [42], or fixations [21]), electrocardiogram (ECG) [1], electrodermal activity (EDA) [19], or respiration (RESP) [25], among many others. The autonomic nervous system (ANS) is a fundamental part of the nervous system and regulates most of the aforementioned involuntary physiological functions. It comprises two main branches: the sympathetic nervous system (SNS), which helps the body enter an alert state ("fight or flight" response), and the parasympathetic nervous system (PNS), which helps the body enter a calm state ("rest and digest"). These two branches typically operate in a dynamic balance, allowing the body to adapt to the environment.

The most used, non-invasive way to assess ANS function is through the analysis of heart rate (HR) and its variability (HRV). These measures can be derived from the ECG signal, which represents the electrical activity of the heart measured on the surface of the skin. Photoplethysmography (PPG), which captures variations in peripheral blood volume, can also enable access to the so-called pulse rate variability (PRV), which highly correlates with HRV [24], and other SNA markers [35]. Respiration is also driven by the ANS, which affects both breathing patterns and frequency. ECG-derived respiration (EDR) methods remain an unobtrusive way of estimating respiratory parameters, proving effective even in ambulatory conditions, avoiding the need for additional sensors [55]. EDA reflects changes in sweat gland activity modulated by SNS. Previous studies have identified EDA as a robust indicator of acute stress and cognitive load, particularly through decomposing EDA signals into phasic and tonic components [27]. Notably, the phasic component, indicative of discrete, sudden sympathetic activations, provides valuable information about momentary cognitive and emotional states, showing consistent increases with higher cognitive load [26].

2.3 Cognitive Load in VR

Most studies have explored CL measurement in traditional, screen-based devices [6, 21, 41, 42]. However, virtual reality (VR) experiences introduce additional challenges, such as multisensory environments, users freely moving around–which may affect physiological signals–, and a generally higher cognitive load due to increased realism and immersion. Assessing such levels of CL is paramount for enhancing user experience. For example, Wen et al. [57] used guidance techniques in VR learning setups that were reported (NASA-TLX) to reduce CL and improve learning outcomes. Chiossi et al. [19] leveraged real-time EDA to dynamically adjust induced CL by manipulating a VR game difficulty, and Marucci et al. [39] showed that using multisensory stimuli (vibrotactile, auditory, and visual) reduced CL and improved performance.

Motivated by the potential of physiological signals for assessing CL in VR environments, some works have gathered datasets encompassing physiological responses and CL. Some of these studies have collected datasets to train machine-learning based models across the past years [5] for distinguishing between different levels of CL [14, 46, 56], with a strong focus on classification rather than on in-depth analysis of the underlying mechanisms and rationale underneath. Other studies have investigated the effects of CL on specific physiological markers. For instance, Lee et al. [33] examined pupillometry in educational VR settings and found that increased task difficulty, associated with higher CL, led to larger pupil diameters, a trend also reported by Ahmadi et al. [1]. However, both studies acknowledge that pupil-based measurements are sensitive to lighting conditions and individual light reflexes, which may limit their reliability in practical, real-world settings. Ahmadi et al. [1] also evaluated HR, electroencephalogram (EEG), and EDA for CL during tasks involving moderate arm movement. They observed increased EDA responses with higher task difficulty but found no clear relationship with HR, likely due to a short measurement window. Marucci et al. [39] similarly reported an increase in EDA with higher CL and further examined EEG and electrooculography to assess experienced workload.

Despite these efforts, a robust and reliable physiological marker for CL assessment has yet to be identified, mainly due to the high complexity of virtual environments and the large number of confounding factors. We aim to exhaustively explore the influence of CL on different physiological signals. We have captured a dataset that includes ECG, EDA, pupillometry, eye-tracking, and PPG signals, and is designed for a systematic and comparative analysis across the different physiological signals. In this work, and motivated by previous literature, we focus on markers derived from the ECG and EDA signals, since they are reported to be the most promising ones. Beyond prior work, we also introduce two different levels of participant movement, enabling us to examine how movement affects physiological measurements and their reliability for CL assessment.

3 Experimental Setup

Our study is designed to elicit different degrees of cognitive load in the participants, and measure their physiological signals as they conduct one or more tasks while immersed in a virtual environment. The tasks involve multisensory inputs: the main task is a visual search one, while the secondary task (which may or may not be present) is based on recognition of specific words within auditory input. The search is conducted either within a 90° area in front of the participant, or in the full 360° around them.

Four different conditions were thus tested in the experiment, resulting from the combination of two binary independent variables in a full factorial design. Those two variables are the *level of cognitive load* and the *search area*. The two levels of cognitive load differed in whether the participant had to conduct *only* the main task (low cognitive load, LCL) or the main task *plus* the secondary task (high cognitive load, HCL). The two search areas tested were 90°, in which the visual content was only in front of the participant, and 360°, in which the content fully surrounded the participant, who had to turn around as they explored the scene to conduct the search task.

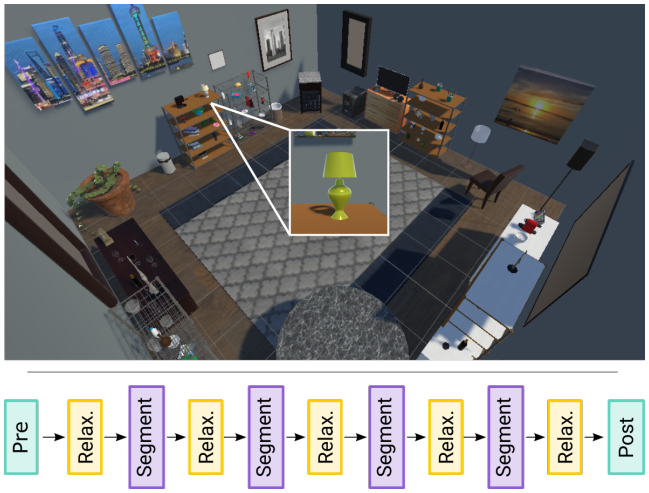

Figure 2: **Top:** Virtual scene used during our experiments. Participants sat in the center of a furnished room, in which they were instructed to find an object that was highlighted as golden (**inset**). **Bottom:** Overview of our experimental procedure, which included an initial relaxation segment to acquire a reliable baseline for the physiological measures, and four test segments, one per experimental condition (combinations of search area levels (90° and 360°) and cognitive load (CL) levels (high and low), see Section 3), followed by another relaxation segment to recover from the previous one.

In addition to physiological signals, we also measure participant performance when conducting the tasks, and gather subjective responses to questionnaires on sickness and workload (NASA-TXL).

Hardware We used the Varjo XR-4 head-mounted display (HMD), which incorporates eye-tracking technology capable of following participants' eye movements with sub-degree accuracy. The HMD, one of the leading headsets in VR in terms of image quality, has a spatial resolution of 3840×3744 per eye, a field of view (FOV) of 120°×150°, a temporal resolution of 90 Hz, and 51 pixels per degree (PPD). We performed five-point eye-tracker calibration and recorded gaze and eye measurements at a rate of 100 Hz. Although the Varjo HMD has integrated speakers, the audio in the experiment was reproduced through the ASUS TUF Gaming H3 headphones for better audio quality. The ECG signals were recorded using the Shimmer3 ECG unit, and the EDA and PPG signals were recorded using the Shimmer3 GSR+ unit. Both units were connected to a separate computer via Bluetooth, and recorded the data in their respective SD cards. The signals from the shimmer ECG unit and the shimmer GSR+ unit were synchronized using the UNIX TimeStamps provided by the devices. More details (e.g., sampling rate) can be found in Sec. 3.1 and in the supplementary material.

Participants We performed a power analysis for a repeated-measures design with two within-subjects factors (2x2 design). The estimated sample size for detecting a medium effect size is 36 participants. Therefore, we recruited 36 participants for the study, 12 of which were female and none of them identified themselves as *non-binary*, *not listed* or *prefer not to disclose*. Ages of the participants ranged from 22 to 62 (mean age of 31.94 and STD of 11.16). All participants provided written consent. We tested for vision and audition (as described below, in the Procedure); one participant was excluded following these tests and did not proceed with the experiment. None of the participants reported heart conditions. The

whole experimental procedure was approved by Comité de Ética de la Investigación de la Comunidad Autónoma de Aragón: CEICA.

Virtual environment and tasks The virtual environment was a room in which the wall (90° search area) or walls (360° search area) were lined up with shelves filled with different objects (see Fig. 2, top). The participant sat in the middle of the room, in a swivel chair, and could rotate but was instructed not to translate.

The main task involved searching for objects highlighted in golden color among the numerous objects on the shelves. At each point, only one object was highlighted, and when the participant found it (by signaling it with the controller), a new object was highlighted from the set of objects present. In the case of a 90° search area, the next object to highlight is picked randomly among those in the only non-empty wall (the north wall); in the case of the 360° search area, a displacement angle with respect to the previous object was randomly chosen from the set {0°, 90°, 180°, 270°}, a ray was traced adding this displacement angle to the yaw angle of the participant's forward vector, and the closest object to where this ray landed was chosen to highlight. In this way, we fostered a broader search while reducing the variability of the so-called displacement angles (angle between a target object and the next).

The secondary task consisted of recognizing certain words within an auditory input: the participant had to press the "A" button on the controller every time they heard an odd number. The audio track used narrates a coherent story, which was manually crafted to include a significant portion of odd and even numbers. The audio task was present for both HCL segments (90° and 360°). To avoid learning effects, the story heard during the second HCL segment encountered was a continuation of the first and not a repetition. The distribution of odd and even numbers in both halves was equal.

Procedure At the beginning of the study, participants were informed of which data would be captured and what the experiment consisted of. They were asked to provide written consent and were tested for visual and auditory capabilities. Vision acuity was tested with the Snellen test [49], color blindness with the Ishihara test [15], and depth perception with the Titmus test [23]. Audio perception was tested by playing different spatialized audio tracks and asking the participant where they perceived the sound was coming from. One of the participants was discarded as a result of this initial assessment. Next, the instructor put the ECG and GSR+ sensors on the participant, calibration for the HMD eye-tracking was performed, and the actual experiment began.

The experiment itself consisted of four *test* segments (one per condition) with a duration of 4 minutes each, as shown in Fig. 2, bottom. The order of the four conditions (LCL-90, HCL-90, LCL-360, HCL-360) was randomized. Before the start of each test segment, the participant was informed of the next segment's search area and whether there would be a secondary task or not; they were told both with textual cues shown before the segment and by the instructor of the experiment. Additionally, the participant was always reoriented towards the north wall at the start of each segment to avoid spurious delays, particularly in the 90º search area case. Participants were also instructed to avoid raising their arms above their shoulder height or moving their non-dominant hand (which had the GSR+ sensor attached) to minimize signal noise.

Prior to the first test segment, participants underwent a 4-minute *relaxation* segment. In it, they encountered a calm night scene next to a river and a bonfire with soothing music. This lasted for 3 minutes and 40 seconds, followed by a 20-second segment where the participant was transported (fade-in transition) to the room of the experiment, still hearing the soothing music. The purpose of this relaxation segment was to help participants relax and create a reliable baseline for the physiological signals. The final 20 seconds in the experiment room were intended to create a baseline for pupil data in similar lighting conditions. Once the relaxation was over, participants did a short tutorial prior to the start of the first segment to make sure they understood the tasks and controls. Additionally, between test segments there were short relaxation segments lasting two minutes each. These intermediate segments had the purpose of separating test segments, under the reasonable assumption that the ANS returns to baseline state within such time.

Once the last test segment was completed, the participant was asked to fill out questionnaires regarding sickness and load per task (NASA-TLX). Finally, they were asked to answer verbal questions about the story they had listened to during the HCL segments (they had not been told that they would be asked about the story). It is worth mentioning that the audio used for the secondary task was one complete track narrating a coherent story which was separated in two segments: The first half of the audio was played during the first HCL segment that appeared in the experiment, and the second half of the audio resumed for the second HCL segment (the audio track not only included an equal number of odd and even numbers, but also a similar number-to-word ratio in both halves).

3.1 Collected dataset

We have three sources of information recorded from 35 participants during the experiment: physiological signals, subjective responses to questionnaires, and objective measures of task performance.

Physiological signals include: Electrocardiogram (ECG) and Impedance Pneumography (RESP) from the Shimmer3 ECG unit; Inertial Measurement Unit (IMU), Photoplethysmography (PPG) and Electrodermal Activity (EDA) from the Shimmer GSR+ unit; gaze data and Pupil Dilation (PD) from the Varjo XR-4's integrated eye-tracker. ECG signals were recorded at 512Hz, GSR+ signals at 256Hz –since a lower temporal resolution is sufficient in this case–, and eye tracking at 100Hz.

Subjective data include responses to: the NASA-TLX questionnaire [28], measuring workload along the six usual dimensions (Mental Demand, Physical Demand, Temporal Demand, Performance, Effort and Frustration) on a scale from 0 to 20, for both the high and the low cognitive load conditions; and to a sickness questionnaire including responses to General Discomfort, Headache, Eye Strain and Nausea on a 4-point scale of None, Slight, Moderate and Severe.

Finally, in terms of performance measures, we report mean search time per object and total number of objects found for the main (visual search) task, and accuracy for the secondary (auditory recognition) task. More details on these measures can be found in Sec. 4.3.

Table 1 compiles an overview of the dataset. This wealth of data from different sources and for different experimental conditions (two levels of cognitive load and two search areas) enables an in-depth study of cognitive load and its relationship with different measurements. In the next section, we analyze the data along key dimensions. Besides, our dataset is publicly available[1] for researchers to study and build upon.

4 Results

With the overarching goal of analyzing the effect of cognitive load on different physiological signals and their temporal evolution, we conduct here several analyses. First, we assert that our experimental setup does indeed induce the intended low and high levels of cognitive load, as reported by participants through the NASA-TLX questionnaire (Sec. 4.2), and look into the effects of this cognitive load on user performance on the different tasks (Sec. 4.3). Then, we study the effect of *cognitive load*, as well as our other factor *search area*, on the physiological signals (Sec. 4.4). Finally, we extract insights on the temporal aspect of these signals, comparing the shifts in the different markers for different conditions (Sec. 4.5).

4.1 Physiological Data Processing

From the physiological data gathered per participant (Sec. 3.1), we have processed the ECG and EDA raw signals to obtain meaningful

[1] https://graphics.unizar.es/projects/CL_Biosignals/

Table 1: Overview of the data available in our dataset, encompassing physiological signals, subjective measurements, and objective performance metrics. We provide such data for a total of 35 participants across the different experimental conditions, including two levels of cognitive load (LCL and HCL) and two search areas (90° and 360°). Please refer to Section 3 for details on the experimental and data collection procedures.

Source	Data	Description
Shimmer ECG unit	Electrocardiogram (ECG)	Electrical activity of the heart (mV)
	Impedance Pneumography (RESP)	Respiratory effort measured through thoracic impedance variations (kΩ)
Shimmer GSR+ unit	Inertial Measurement Unit (IMU)	Acceleration (accelerometer, g), angular velocity (gyroscope, deg/s), magnetic field (magnetometer, gauss)
	Photoplethysmography (PPG)	Peripheral blood volume variations measured optically (a.u.)
	Electrodermal Activity (EDA)	Skin conductance (μS) reflecting sweat gland activity
Varjo XR-4	Gaze data	Gaze fixation coordinates inside the participant's field of view
	Pupil Dilation (PD)	Estimation of the diameter of the pupil and the iris-pupil ratio
Questionnaires	NASA TLX scores	Self-reported evaluation of CL level in six categories for LCL and HCL
	Sickness questionnaires	Self-reported evaluation of sickness levels before and after the experiment
Performance	Search time (visual)	Performance temporally annotated (i.e., time to find a highlighted object, ms)
	Nº. of objects per segment	Total number of objects found during each segment of the visual search task
	Accuracy (auditory)	Ratio of odd numbers correctly identified by the participant (from 0 to 1)

physiological markers to be used in subsequent analyses. The collected data underwent thorough processing to ensure its quality; we briefly outline this processing here, while full technical details can be found in the supplementary material.

Heart rate ECG signals were first resampled to 1000 Hz, and R-wave detection performed using a wavelet-based algorithm [38]. Using the R-wave occurrence time, the time intervals between consecutive normal heartbeats (the so-called RR series) were obtained, after ectopic removal and outlier correction through the Integral Pulse Frequency Modulation (IPFM) model [8]. This allowed us to extract derived markers of participants' heart rate. From them, we have focused on the mean heart rate (HR) in the analysis, since it has been found to correlate with cognitive load before [7].

Respiration An estimation of the respiratory signal during the experiment was obtained via ECG-derived respiration (EDR) using a QRS slope-based approach [32], later combined with the impedance pneumography recorded by the sensor to obtain the final markers, from which we extract mean respiratory frequency (RF).

Electrodermal Activity EDA signals were analyzed using the convex optimization approach cvxEDA [26], robustly decomposing the signals into their tonic and phasic components. From these components, a number of markers can be extracted, among which we analyze the mean of the tonic (EDA_{tonic}) and phasic (EDA_{phasic}) components, which have been studied as indicators of CL variations [7]. Data for three participants was corrupted, and their EDA signals are not included in the dataset or the analysis.

Outlier Rejection We performed outlier rejection on the data across all the metrics analyzed, including performance and the four physiological markers extracted. Outliers were defined based on interquartile range (IQR), and values considered as outliers were removed for each metric and condition. All participants completed the full experiment, and none of them reported sickness symptoms.

4.2 Subjective Assessment of Cognitive Load

We collected participant responses to NASA-TLX questionnaires to assess their self-perceived workload for each of the two cognitive load conditions: performing the visual search task alone (low cognitive load, LCL) or together with the auditory task (high cognitive load, HCL). This well-known tool asks participants to evaluate their experience according to six categories: mental demand, physical demand, temporal demand, performance, effort, and frustration [28]. Following previous work [56], we decided to use the mental demand score as representative of the self-perceived cognitive load of the participant. The scores for all the categories of the questionnaire can be found in the supplementary material.

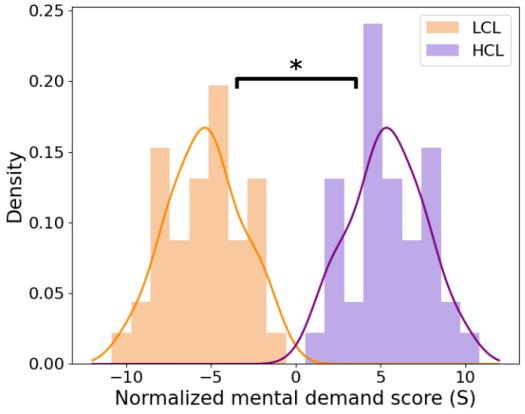

Figure 3: Distribution of the normalized mental demand scores S for the low (LCL, orange) and high cognitive load (HCL, purple) conditions. Histograms have been normalized to represent probability density. The horizontal shift between plots indicates that conditions labeled as high cognitive load indeed entail a larger mental demand.

Given the high variability between subjects, we perform within-subject normalization of the mental demand score per participant by subtracting the mean score of that participant across cognitive load levels. As a result, the mental demand score S for participant i and cognitive load level c is computed as:

$$S_{i,c} = S'_{i,c} - \frac{S'_{i,L} + S'_{i,H}}{2}, \tag{1}$$

where S' is used to denote the raw, reported scores and $c \in \{H, L\}$.

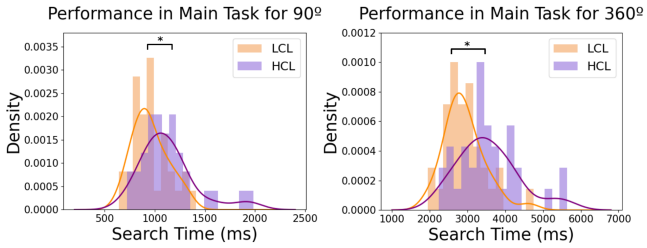

Figure 4: Distributions of main task (visual search) performance for all participants, for the 90° (left) and 360° (right) conditions, colored according to the cognitive load level (orange is LCL, purple is HCL). Statistical differences are marked with an asterisk ($p < 0.0001$). Histograms have been normalized to represent probability density. Note that in both search areas, performance is significantly different between CL levels, with search times being lower for the LCL, suggesting that higher levels of CL lead to longer search times.

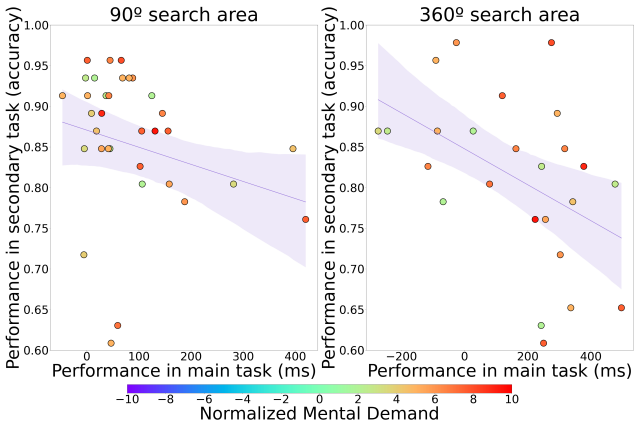

Figure 5: Correlation between the performance in the main (visual search) and secondary (auditory recognition) tasks for both the 90° (left) and 360° (right) search areas in the high cognitive load (HCL) case. The x-axis shows the mean search time in the visual task for each participant, normalized within-subject, while the y-axis shows their accuracy in the auditory task. Statistical analyses show that, for the 360° search area, there is a mild, yet significant, negative correlation between both performance metrics: Generally, an increase in search time correlated to lower accuracies in the auditory one, i.e., performance was consistent across tasks. However, no direct correlation was found between levels of performance and CL scores.

Fig. 3 shows the distribution of the normalized scores S from all participants, for the LCL (orange) and HCL (purple) conditions (see Sec. 6.1 in the supplementary material for the raw scores). It shows how participants clearly perceived the two conditions as eliciting different degrees of mental demand. Specifically, the peaks of both distributions are circa 10 points apart, with the raw scores $S \in [0, 20]$. Additionally, we used Wilcoxon Signed-Rank tests and found a significant difference between the scores of LCL and HCL conditions, both for the normalized scores ($W = 0$, $p < 0.0001$) and the raw scores ($W = 0$, $p < 0.0001$). This validates our experimental setup with regard to the levels of cognitive load. At the same time, it suggests that self-reported scores may not be robust indicators, due to the high variability in the responses within each condition.

4.3 Task Performance and Cognitive Load

To investigate the effect of cognitive load (CL) on task performance, we record objective measures of performance for each participant in both the main (visual search) and secondary (auditory recognition) tasks. For the *visual search task*, we record the time (ms) it takes the participant to select each object since the instant the object turns golden. We then average these times across all searches in a segment to obtain our main task performance measure, *mean search time*. For the *auditory recognition task*, we record the instant when the participants press a controller button and compare it with the instant they hear an odd number. We established a three-second window for a button press to be eligible as correct. If more than one odd number appeared within that window, valid button presses were associated with the numbers following a first-in-first-out strategy. We then calculate the *accuracy* as the percentage of odd numbers correctly identified and use it as our secondary task performance measure.

We first look into the effect of cognitive load on the main task performance, given by the mean search time. Results are shown in Fig. 4, separated by search area (90° and 360°). Given the within-subject design and the presence of both fixed-effects factors (cognitive load and search area) and random effects (participant and trial order), we analyzed task performance using a generalized linear mixed-effects model (GLMM) fitted with restricted maximum likelihood. To stabilize variance, we applied a logarithmic transformation to the response variable. Post hoc pairwise comparisons were conducted using estimated marginal means with false discovery rate (FDR) correction [11]. Tables with the full results can be found in the supplementary material. We found a significant main effect of search area ($p < 0.0001$, $t = 48.13$), with longer search times in the 360° condition, as expected given the larger spatial extent of the search area. We also observed a significant main effect of cognitive load, with longer search times in the high load (HCL) condition ($p < 0.0001$, $t = 7.42$). However, the interaction between cognitive load and search area was not significant ($p = 0.54$, $t = 0.62$), suggesting that the effect of cognitive load was consistent across both search areas. Post hocs show significant differences between the low (LCL) and high (HCL) cognitive load conditions for both the 90° ($p < 0.0001$, $t = 5.48$) and 360° ($p < 0.0001$, $t = 6.44$) search areas. The effect was moderate at 90° ($\eta_p^2 = 0.42$) and very large at 360° ($\eta_p^2 = 0.97$). As expected, participants took longer to complete their visual search task when being simultaneously prompted to pay attention to the auditory task. Accuracy rates in the secondary task were in the range [0.59,0.96], indicating that participants did not ignore this secondary task in favor of the first one. All this suggests that the auditory task was absorbing a significant part of the cognitive resources that would otherwise be focused on the visual search, and thus aligns with the hypothesis that having additional tasks involving multisensory processing can hinder task performance.

We then analyze, for the high cognitive load (HCL) conditions, the correlation between performance in both tasks (main and secondary), as well as with the self-reported mental demand. Fig. 5 shows scatter plots (one per search area) depicting the relationship between both performance measures; additionally, points are color-coded according to mental demand. To make it comparable between subjects, mean search time is normalized within each subject (akin to S in Eq. 1). In the case of the 360° search area, statistical analyses show a mild negative correlation between performance in both tasks (Pearson's $r = -0.4628$, $p = 0.0198$ and Spearman's $\rho = -0.4533$, $p = 0.0229$). This suggests that factors like expertise might have a significant influence. Interestingly, this correlation is not found in the 90° search area. A discussion on this can be found in Sec. 5.

With regard to the correlation between mental demand and performance measures within the HCL conditions, we find no significant correlation of mental demand with either performance measure, as also hinted by Fig. 5. In the case of the LCL conditions, there is also no significant correlation between performance and mental demand. This suggests that, while both mental demand and main task perfor-

mance were affected by the level of cognitive load (Figs. 3 and 4), within a certain cognitive load level we find no clear relationship between perceived mental demand and performance.

4.4 Effect of Cognitive Load on Physiological Signals

Here we look into the effect of cognitive load on various physiological signals collected in our study. We focus on electrocardiogram (ECG) and electrodermal (EDA) signals, which are known to be affected by high-level cognitive processes like changes in cognitive load [1, 39]. In particular, we extract the following markers from those signals: mean heart rate (HR, measured in beats/minute), mean respiration frequency (RF, breaths/minute), median EDA tonic (EDA_{tonic}, μs), median EDA phasic (EDA_{phasic}, μs). In our analyses, each of them was a dependent variable, with a value per test segment. We normalized every marker's value by computing the difference between the value in the test segment and the participant baseline recorded in the first relaxation segment (see Sec. 3).

Given that we have a repeated measures experimental design, with both fixed-effects factors (cognitive load and search area) and random factors (participant, gender and experimental order in which the test segments were presented), we fitted a generalized linear mixed-effects model (GLMM) using a restricted maximum likelihood approach with a Gaussian distribution and identity link for each response variable using R. Each variable thus depends on two fixed-effects factors with two levels each (cognitive load (HCL vs. LCL) and search area (90° vs. 360°)), a first-level interaction between said factors, and two random intercepts (participant and order). For all physiological markers analyzed, the search area had a significant main effect, while the interaction between factors were not significant. In the following, we focus on the effects of cognitive load. Tables with the full results from the GLMMs, along with the post-hoc pairwise comparisons, can be found in the supplementary material, while we report here the main findings.

Fig. 6 depicts the distributions of the four markers for both cognitive load levels, grouped by search area (first row) and then separated (90° in the second row and 360° in the third row). Details from the global analysis are presented in Tab. 1 of the supplementary material. When looking at the global results (first row), both for HR and RF we find a significant influence of CL of medium effect size ($t = -2.577, p = 0.012, \eta_p^2 = 0.07$ and $t = -2.602, p = 0.011, \eta_p^2 = 0.07$, respectively), and a large-sized effect of search area ($t = -9.943, p < 0.0001, \eta_p^2 = 0.51$ and $t = 5.220, p < 0.0001, \eta_p^2 = 0.22$, respectively). In both cases, the interaction is not significant. EDA signals behave differently: In the case of EDA_{tonic}, we find a large-sized, significant effect of both CL ($t = -4.516, p < 0.0001, \eta_p^2 = 0.22$) and search area ($t = -6.005, p < 0.0001, \eta_p^2 = 0.33$), yet no interaction between both. Finally, EDA_{phasic} is only significantly affected by search area ($t = -7.288, p < 0.0001, \eta_p^2 = 0.39$). This difference between the tonic and the phasic component of EDA could reflect different mechanisms for instantaneous and sustained responses to CL. Further analysis can be found in Sec. 4.5.

We then analyze the effect of CL within each search area through post-hoc pairwise comparisons, results in Tab. 2 of the supplementary material. In the case of the 90º search area (Fig. 6, second row), we find no significant difference between LCL and HCL for all markers, except for EDA_{tonic} ($t = -3.370, p = 0.0017$). In the case of the 360° search area (Fig. 6, third row), both HR ($t = -2.081, p = 0.0481$) and EDA_{tonic} ($t = -3.009, p = 0.0041$) exhibit significant differences between LCL and HCL. These results suggest that the influence of different search areas and levels of movement should not be neglected when measuring shifts in CL, especially when using signals that are sensitive to movement like HR or RF. At the same time, the results also hint at the robustness of EDA_{tonic} as a CL indicator across search areas.

Regarding random effects, we found that participant-level variance was strongest in EDA_{tonic} ($SD = 0.51$), RF ($SD = 2.08$), and HR ($SD = 2.79$), suggesting that their baseline states varied independent of experimental conditions. Unlike them, EDA_{phasic} ($SD = 0.05$) showed minimal participant-level variance. This suggests that individual differences should be considered when designing virtual experiences, highlighting the importance of adaptive experiences. On the other hand, trial order effects remained smaller ($SD = 0.57$ being the largest, belonging to RF), suggesting that our counterbalanced design was effective and there was no significant habituation effect. Our analysis did not show significant interactions between gender and the way CL levels affected any of the markers: HR ($p = 0.335$), EDA_{tonic} ($p = 0.502$), EDA_{phasic} ($p = 0.441$) and RF ($p = 0.177$). However, a more diverse population could show more robust results regarding gender-specific comparisons.

4.5 Temporal Evolution of Physiological Markers

A common approach in CL analysis using physiological markers is to compute average measures over the entire task duration. However, different physiological markers may exhibit distinct temporal dynamics. For instance, the tonic component of EDA changes gradually, requiring a longer time to reflect meaningful variations, whereas the phasic component responds more rapidly, capturing changes over shorter time windows [7]. In this section, we explore how the different physiological markers we study evolve over time, and the effect that the different CL levels have on them.

Specifically, we study changes in each marker between a temporal window at the beginning and a temporal window at the end of each test segment, named w_i and w_f, respectively. For each test segment (i.e., for each experimental condition), we compute the participant-averaged mean HR and RF, and the median EDA_{tonic} and EDA_{phasic} over the duration of w_i and w_f. An analysis of its influence (see supplementary material) led us to establish a window length of 40 seconds. We term this additional factor (whether we measure the marker in the first or last window of the segment) the observation *period*. As shown in Fig. 7, markers exhibit diverse temporal patterns: some vary significantly between the two periods, while others remain constant.

To assess these differences, we fitted a GLMM for each marker, including period, search area, and cognitive load as fixed effects, and participant and trial order as random effects for within-subject variability and repeated measures order. Full results are provided in the supplementary material. All markers except EDA_{tonic} showed significant effects of observation period, with large effect sizes for HR ($p < 0.0001, t = 7.74, \eta_p^2 = 0.21$) and EDA_{phasic} ($p < 0.0001$, $t = 6.91, \eta_p^2 = 0.20$), and a medium effect for RF ($p < 0.0001, t = -5.28, \eta_p^2 = 0.11$). This highlights EDA_{tonic} as the most temporally stable marker, yielding consistent results across periods.

We further conducted post hoc pairwise comparisons of the estimated marginal means, with p-values corrected using the False Discovery Rate (FDR) method [11] (see supplementary material). Results show a consistent decrease in HR between the first and last periods across all conditions: HCL–90° ($p = 0.007$, $t = 2.943$), LCL–90° ($p = 0.039, t = 2.299$), HCL–360° ($p < 0.0001$, $t = 4.917$), and LCL–360° ($p < 0.0001$, $t = 5.331$). Similarly, EDA_{phasic} shows significant changes in HCL–90° ($p = 0.042$, $t = 2.204$), HCL–360° ($p < 0.0001$, $t = 5.634$), and LCL–360° ($p = 0.0001, t = 4.350$). In contrast, EDA_{tonic} remains stable (i.e., with no statistically-significant difference between periods) for most conditions, with a single significant change observed for LCL–90° ($p = 0.004, t = 3.097$). RF also shows minimal change, except in HCL–360° ($p < 0.0001, t = -4.598$). Again, EDA_{tonic} stands out for its stability, while HR and EDA_{phasic} vary more strongly over time, possibly because they are influenced by factors beyond CL. These results underscore the importance of temporal considerations when analyzing physiological markers for CL, as marker sensitivity

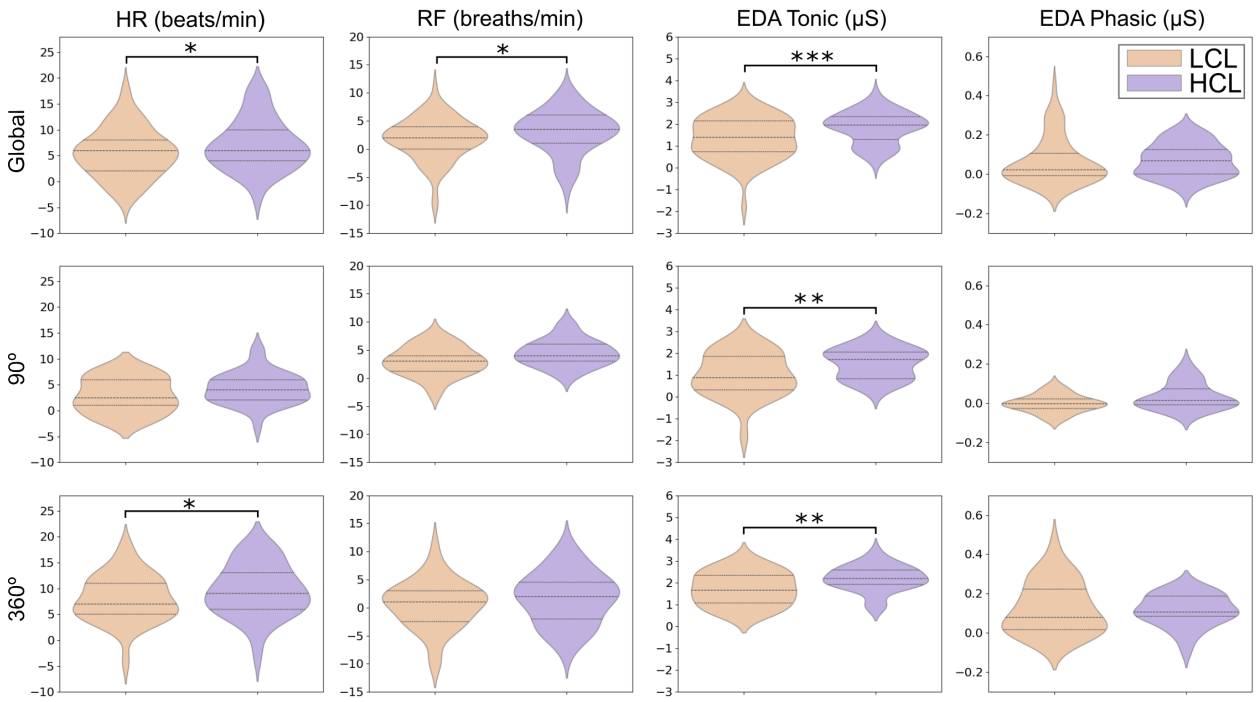

Figure 6: Distribution across participants of physiological markers –one per column, from left to right: HR, RF, EDA$_{tonic}$ and EDA$_{phasic}$– for the two cognitive load levels (LCL in orange, HCL in purple). Different rows show the results grouped by search area (first row) or separated (90° in the second row and 360° in the third row). While HR and RF are differently affected by different experimental conditions, EDA$_{phasic}$ remains unaffected by cognitive load and EDA$_{tonic}$ shows consistent significant differences with CL regardless of search area.

can depend strongly on the selected observation window.

5 Discussion

Our study shows that cognitive load has a measurable effect on both quantitative and qualitative responses. In particular, we find a significant effect between HCL and LCL levels for subjective assessment of cognitive load, task performance, physiological signals, and their temporal evolution. These robust trends highlight the transversal effect of CL in physiological processes. As expected, the analysis of the NASA-TLX questionnaires shows a clear difference between LCL and HCL levels for mental demand despite the high inter-user variability. This first correlation confirms that participants perceive different levels of cognitive load in our user study. Additionally, the high inter-user variability also highlights the importance of identifying reliable physiological indicators of cognitive load.

Regarding performance, the main task was significantly affected by CL (with worse results under HCL levels), regardless of the search area. Participants consistently showed slower response times for the segments with high cognitive load when compared to the segments with low cognitive load. The secondary task had a relatively high accuracy rate (over 0.59), suggesting that our participants were indeed paying attention to both tasks. In this context, multitasking may lead to a division in the allocation of limited cognitive resources. However, we did not find significant correlations between performance and perceived mental demand. This could be due to several reasons: not enough precision when self-reporting subjective data, a difference between perceived and measured performance, personal strategies to fulfill the assigned tasks, or individual skill levels. Further research should be carried out in order to investigate if there is indeed a lack of correlation between self-perceived CL and task performance. Performance for both tasks shows a weak correlation, suggesting that individual skill plays an important role.

Moving on to the effect of cognitive load on physiological markers, EDA$_{tonic}$ seems to be a more reliable marker for CL from the four studied signals. Unlike RF and HR, EDA$_{tonic}$ shows consistent differences regardless of the search area. EDA$_{phasic}$, on the other hand, does not yield significant differences for complete segments in any search area, nor globally. This suggests that some signals, such as EDA$_{tonic}$, are more robust to variations in context or movement, while others may be more affected by user mobility. EDA$_{tonic}$ poses as a stable CL indicator, reflecting sustained sympathetic arousal. This same stability, however, would most likely hinder its ability to account for fast variations. Discerning the optimal indicator for such scenarios would require a user study specifically designed for sudden stimuli. Moreover, the individual variation found in the random intercept of the statistical models reinforces the need to consider adaptive solutions when studying these responses. Our statistical analyses did not reveal significant interactions between gender and cognitive load levels for any of the physiological markers or performance measures. With 24 male and 12 female participants, the generalizability of our results is limited, and although the sample size was adequate for the main analyses, some effects might not be visible, and thus a more balanced population may be necessary to draw specific conclusions regarding gender-related effects. Additionally, there are interesting, significant differences in markers when considering the temporal evolution of the physiological signals. HR and EDA$_{phasic}$ show significant changes between the first and last seconds of each segment. In the case of EDA$_{phasic}$, differences between experimental conditions were more pronounced at the beginning of each segment. On the other hand, RF and EDA$_{tonic}$ seem quite consistent from start to finish, especially the latter, while changes in RF could be influenced by movement. Interestingly, RF appears to highlight cognitive load differences more clearly toward the end of the 360°

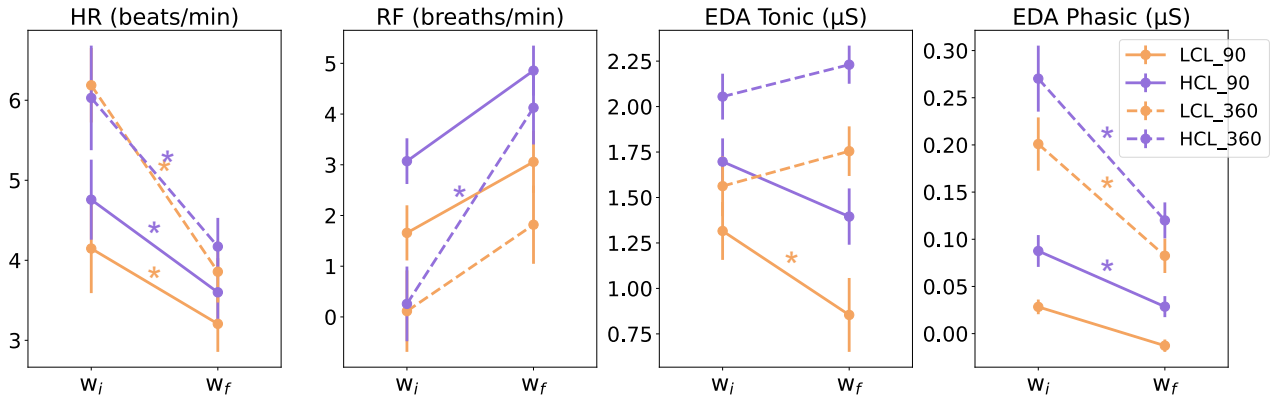

Figure 7: Evolution of physiological markers with time within a test segment. Plots show mean value across participants for HR, RF, EDA_{tonic}, and EDA_{phasic} for the first (w_i) and last (w_f) 40 seconds of each segment. The different colors indicate the level of cognitive load (LCL in orange, HCL in purple), while both search areas are represented with continuous (90°) and dashed lines (360°). Error bars represent the standard error. Significant differences between the beginning and the end of the segment are marked with an asterisk.

segments, suggesting that temporal positioning and task scope may modulate its sensitivity. Notably, we found a link between the temporal stability of markers and their overall reliability. Markers that presented more drastic changes in time, like HR or EDA_{phasic}, offered weaker overall differences, while more constant markers like EDA_{tonic} seemed to be robust for all the segments. These differences in temporal evolution indicate that researchers should direct their focus to different aspects of the response variables depending on the signal and the temporal window that are being considered. Further investigating the temporal changes in physiological signals, together with additional markers from our dataset, remains a future avenue.

All in all, our results highlight the transversal effect of cognitive load across a variety of responses from task performance to physiological activity. Although we have found some seemingly robust markers (like EDA_{tonic}), more sensitive signals could be useful in different contexts. A central takeaway from our results is that the sensitivity of physiological markers to cognitive load is shaped by user variability, experimental context, and temporal dynamics. Choosing the most appropriate marker should therefore be a deliberate decision guided by the specific goals of each study. In the absence of clear guidelines, we advocate for recording multiple signals and comparing their behavior to ensure robust interpretation. We believe our inclusion of multisensory stimuli, different levels of movement and the study of time progression enhance ecological validity and bring our insights closer to real-world applications.

5.1 Limitations and Future Work

Our work provides an in-depth analysis of the relationship between cognitive load, performance, and physiological signals in VR, yet plenty of avenues for future exploration and improvement remain.

First, a strength of our work is the fact that we measure a large number of physiological signals per participant, and analyze and compare several key physiological markers. However, many more markers could be derived from the raw signals we measure, and added to the analysis. According to our findings, markers derived from the EDA signal seem to be among the most promising. We suggest possible markers in the supplementary material, and have made our dataset publicly available for future research to build on.

Our experimental setup is designed to cover different use case scenarios, including ones in which the participant needs to rotate and move when doing search tasks in a 360° area. Measurements of respiration obtained from impedance pneumography are sensitive to movement and can produce artifacts, even after mitigating the effect with ECG-derived respiration. As a result, the signal may not only be influenced by cognitive load variations, but also by movement. Isolating those confounding factors, even in controlled environments like this, remains a challenge, but our aim is to find informative signals in ecologically-valid scenarios, thus our design.

While our focus here is on analyzing the impact of cognitive load on measurable variables, searching for an objective and reliable indicator of the degree of cognitive load is a long-standing goal in the field. Our results suggest that inferring coarse levels of cognitive load from physiological markers is feasible. Finer inference would require reliable, and ideally continuous, quantification of cognitive load that can be used as ground truth, which is still a challenge [56].

A valuable future direction would be to design studies that disentangle the sources of cognitive load. While this work focuses on overall CL effects, distinguishing between intrinsic, germane, and extraneous load could inform better VR experience design.

Finally, our findings also reflect a high inter-subject variability in certain response variables, especially in subjective measurements and performance, but also in some of the physiological markers studied. This speaks to the importance of working towards adaptive VR experiences that factor in this variability, and further exploration of it can contribute to the development of more engaging VR systems.

6 Conclusion

In this work, we have provided a complete and comprehensive dataset encompassing physiological signals, subjective measurements of CL, and performance metrics for two distinct CL conditions and for two search areas with different movement requirements. Using this data, we have performed a detailed analysis of the impact of CL on task performance and several physiological markers in VR. We identified EDA_{tonic} as a robust indicator of CL, and acknowledged the limitations of other markers like HR or RF, sensitive to different levels of movement. Our temporal analysis also revealed the consistency of EDA_{tonic} across time, while HR and EDA_{phasic} presented noticeable shifts from start to finish. Ultimately, our work offers a rich set of publicly available data, along with valuable insights that can foster future research on incorporating cognitive load into user-centric VR applications.

Acknowledgments

This research was supported by grant PID2022-141539NB-I00, funded by MICIU/AEI/10.13039/501100011033 and by ERDF, EU; by the Aragon Institute for Engineering Research (I3A) through the Impulso program; and by the Gobierno de Aragon through project "Human-VR" (Proy_T25_24). Additionally, J. Pina was supported by an FPI predoctoral grant (PRE2023-UZ-16) and E. Bernal-Berdun by a Gobierno de Aragon predoctoral grant (2021-2025).

REFERENCES

[1] M. Ahmadi, S. W. Michalka, S. Lenzoni, M. Ahmadi Najafabadi, H. Bai, A. Sumich, B. Wuensche, and M. Billinghurst. Cognitive load measurement with physiological sensors in virtual reality during physical activity. In *Proceedings of the 29th ACM Symposium on Virtual Reality Software and Technology*, pp. 1–11, 2023.

[2] M. A. AlGerafi, Y. Zhou, M. Oubibi, and T. T. Wijaya. Unlocking the potential: A comprehensive evaluation of augmented reality and virtual reality in education. *Electronics*, 12(18):3953, 2023.

[3] E. W. Anderson, K. C. Potter, L. E. Matzen, J. F. Shepherd, G. A. Preston, and C. T. Silva. A user study of visualization effectiveness using eeg and cognitive load. In *Computer Graphics Forum*, vol. 30, pp. 791–800. Wiley Online Library, 2011.

[4] A. Armougum, E. Orriols, A. Gaston-Bellegarde, C. Joie-La Marle, and P. Piolino. Virtual reality: A new method to investigate cognitive load during navigation. *Journal of Environmental Psychology*, 65:101338, 2019.

[5] S. M. Asish, B. B. Karki, B. KC, N. Kolahchi, and S. Sutradhar. Synthesizing Six Years of AR/VR Research: A Systematic Review of Machine and Deep Learning Applications . In *2025 IEEE Conference Virtual Reality and 3D User Interfaces (VR)*, pp. 175–185. IEEE Computer Society, 2025. doi: 10.1109/VR59515.2025.00042

[6] N. Ayala, C. A. Martin Calderon, N. Sharma, E. Irving, S. Cao, S. Kearns, and E. Niechwiej-Szwedo. Examining the utility of blink rate as a proxy for cognitive load in flight simulation. In *Proceedings of the 2024 Symposium on Eye Tracking Research and Applications*, pp. 1–6, 2024.

[7] P. Ayres, J. Y. Lee, F. Paas, and J. J. Van Merrienboer. The validity of physiological measures to identify differences in intrinsic cognitive load. *Frontiers in psychology*, 12:702538, 2021.

[8] R. Bailón, G. Laouini, C. Grao, M. Orini, P. Laguna, and O. Meste. The integral pulse frequency modulation model with time-varying threshold: application to heart rate variability analysis during exercise stress testing. *IEEE transactions on biomedical engineering*, 58(3):642–652, 2010.

[9] M. Bannert. Managing cognitive load—recent trends in cognitive load theory. *Learning and instruction*, 12(1):139–146, 2002.

[10] P. Barrouillet, S. Bernardin, S. Portrat, E. Vergauwe, and V. Camos. Time and cognitive load in working memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 33(3):570, 2007.

[11] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.

[12] E. Bernal-Berdun, J. Pina, M. Vallejo, A. Serrano, D. Martin, and B. Masia. Avisal360: Audiovisual saliency prediction for 360° video. In *IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 1246–1255, 2024.

[13] E. Bernal-Berdun, M. Vallejo, Q. Sun, A. Serrano, and D. Gutierrez. Modeling the impact of head-body rotations on audio-visual spatial perception for virtual reality applications. *IEEE Trans. on Visualization and Computer Graphics*, 2024.

[14] S. S. Bhat, C. Dobbins, A. Dey, and O. Sharma. Multi-modal classification of cognitive load in a vr-based training system. In *IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 503–512, 2023.

[15] J. Birch. Efficiency of the ishihara test for identifying red-green colour deficiency. *Ophthalmic and Physiological Optics*, 17(5):403–408, 1997.

[16] F.-Y. Chao, C. Ozcinar, C. Wang, E. Zerman, L. Zhang, W. Hamidouche, O. Deforges, and A. Smolic. Audio-visual perception of omnidirectional video for virtual reality applications. In *2020 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pp. 1–6, 2020. doi: 10.1109/ICMEW46912.2020.9105956

[17] F. Chen, J. Zhou, Y. Wang, K. Yu, S. Z. Arshad, A. Khawaji, and D. Conway. *Robust multimodal cognitive load measurement*. Springer, 2016.

[18] O. Chen, F. Paas, and J. Sweller. A cognitive load theory approach to defining and measuring task complexity through element interactivity. *Educational Psychology Review*, 2023.

[19] F. Chiossi, R. Welsch, S. Villa, L. Chuang, and S. Mayer. Virtual reality adaptation using electrodermal activity to support the user experience. *Big Data and Cognitive Computing*, 6(2):55, 2022.

[20] M. Csikszentmihalyi and I. Csikszentmihalyi. *Optimal Experience: Psychological Studies of Flow in Consciousness*. Donald F. Koch American Philosophy Collection. Cambridge University Press, 1992.

[21] A. Das, Z. Wu, I. Skrjanec, and A. M. Feit. Shifting focus with hceye: Exploring the dynamics of visual highlighting and cognitive load on user attention and saliency prediction. *Proceedings of the ACM on Human-Computer Interaction*, 8:1–18, 2024.

[22] J. Engström, G. Markkula, T. Victor, and N. Merat. Effects of cognitive load on driving performance: The cognitive control hypothesis. *Human factors*, 59(5):734–764, 2017.

[23] S. L. Fawcett. An evaluation of the agreement between contour-based circles and random dot-based near stereoacuity tests. *Journal of American Association for Pediatric Ophthalmology and Strabismus*, 9(6):572–578, 2005.

[24] E. Gil, M. Orini, R. Bailón, J. M. Vergara, L. Mainardi, and P. Laguna. Photoplethysmography pulse rate variability as a surrogate measurement of heart rate variability during non-stationary conditions. *Physiological Measurement*, 31(9):1271, aug 2010. doi: 10.1088/0967-3334/31/9/015

[25] M. Grassmann, E. Vlemincx, A. von Leupoldt, J. M. Mittelstädt, and O. Van den Bergh. Respiratory changes in response to cognitive load: A systematic review. *Neural Plasticity*, 2016(1):8146809, 2016. doi: 10.1155/2016/8146809

[26] A. Greco, G. Valenza, A. Lanata, E. P. Scilingo, and L. Citi. cvxeda: A convex optimization approach to electrodermal activity processing. *IEEE transactions on biomedical engineering*, 63(4):797–804, 2015.

[27] A. Greco, G. Valenza, J. Lazaro, J. M. Garzon-Rey, J. Aguilo, C. de la Cámara, R. Bailón, and E. P. Scilingo. Acute stress state classification based on electrodermal activity modeling. *IEEE Transactions on Affective Computing*, 14(1):788–799, 2021.

[28] S. G. Hart and L. E. Staveland. Development of nasa-tlx (task load index): Results of empirical and theoretical research. In *Advances in psychology*, vol. 52, pp. 139–183. Elsevier, 1988.

[29] N. Hollender, C. Hofmann, M. Deneke, and B. Schmitz. Integrating cognitive load theory and concepts of human–computer interaction. *Computers in human behavior*, 26(6):1278–1288, 2010.

[30] H. Kim and I.-K. Lee. Studying the effects of congruence of auditory and visual stimuli on virtual reality experiences. *IEEE Transactions on Visualization and Computer Graphics*, 28(5):2080–2090, 2022. doi: 10.1109/TVCG.2022.3150514

[31] M. M. Kouijzer, H. Kip, Y. H. Bouman, and S. M. Kelders. Implementation of virtual reality in healthcare: a scoping review on the implementation process of virtual reality in various healthcare settings. *Implementation science communications*, 4(1):67, 2023.

[32] J. Lázaro, A. Alcaine, D. Romero, E. Gil, P. Laguna, E. Pueyo, and R. Bailón. Electrocardiogram derived respiratory rate from qrs slopes and r-wave angle. *Annals of biomedical engineering*, 42:2072–2083, 2014.

[33] J. Y. Lee, N. de Jong, J. Donkers, H. Jarodzka, and J. J. van Merriënboer. Measuring cognitive load in virtual reality training via pupillometry. *IEEE Transactions on Learning Technologies*, 17:704–710, 2023.

[34] K. A. Longstaffe, B. M. Hood, and I. D. Gilchrist. The influence of cognitive load on spatial search performance. *Attention, Perception, & Psychophysics*, 76:49–63, 2014.

[35] J. Lázaro, E. Gil, M. Orini, P. Laguna, and R. Bailón. Baroreflex sensitivity measured by pulse photoplethysmography. *Frontiers in Neuroscience*, 13, 2019. doi: 10.3389/fnins.2019.00339

[36] S. Malpica, D. Martin, A. Serrano, D. Gutierrez, and B. Masia. Task-dependent visual behavior in immersive environments: A comparative study of free exploration, memory and visual search. *IEEE Trans. on Visualization and Computer Graphics*, 29(11):4417–4425, 2023.

[37] D. Martin, S. Malpica, D. Gutierrez, B. Masia, and A. Serrano. Multimodality in vr: A survey. *ACM Computing Surveys (CSUR)*, 54(10s):1–36, 2022.

[38] J. P. Martínez, R. Almeida, S. Olmos, A. P. Rocha, and P. Laguna. A wavelet-based ecg delineator: evaluation on standard databases. *IEEE*

Transactions on biomedical engineering, 51(4):570–581, 2004.

[39] M. Marucci, G. Di Flumeri, G. Borghini, N. Sciaraffa, M. Scandola, E. F. Pavone, F. Babiloni, V. Betti, and P. Aricò. The impact of multisensory integration and perceptual load in virtual reality settings on performance, workload and presence. *Scientific Reports*, 11(1):4831, 2021.

[40] M. Melo, G. Gonçalves, P. Monteiro, H. Coelho, J. Vasconcelos-Raposo, and M. Bessa. Do multisensory stimuli benefit the virtual reality experience? a systematic review. *IEEE Transactions on Visualization and Computer Graphics*, 28(2):1428–1442, 2022. doi: 10.1109/TVCG.2020.3010088

[41] H. Mitre-Hernandez, R. C. Carrillo, C. Lara-Alvarez, et al. Pupillary responses for cognitive load measurement to classify difficulty levels in an educational video game: Empirical study. *JMIR Serious Games*, 9(1):e21620, 2021.

[42] N. Nourbakhsh, Y. Wang, and F. Chen. Gsr and blink features for cognitive load classification. In *Human-Computer Interaction–INTERACT 2013: 14th IFIP TC 13 International Conference*, pp. 159–166, 2013.

[43] S. Oviatt. Human-centered design meets cognitive load theory: designing interfaces that help people think. In *Proceedings of the 14th ACM international conference on Multimedia*, pp. 871–880, 2006.

[44] J. L. Plass, R. Moreno, and R. Brünken. Cognitive load theory. 2010.

[45] W. Schnotz and C. Kürschner. A reconsideration of cognitive load theory. *Educational psychology review*, 19:469–508, 2007.

[46] J. N. Setu, J. M. Le, R. K. Kundu, B. Giesbrecht, T. Höllerer, K. A. Hoque, K. Desai, and J. Quarles. Predicting and explaining cognitive load, attention, and working memory in virtual multitasking. *IEEE Trans. on Visualization and Computer Graphics*, 2025.

[47] A. Skulmowski and K. M. Xu. Understanding cognitive load in digital and online learning: A new perspective on extraneous cognitive load. *Educational psychology review*, 34(1):171–196, 2022.

[48] K. Smith and M. Apter. *A Theory of Psychological Reversals*. Picton Publishing, 1975.

[49] H. Snellen. *Probebuchstaben zur bestimmung der sehschärfe*. H. Peters, 1873.

[50] J. Sweller. Cognitive load during problem solving: Effects on learning. *Cognitive Science*, 1988.

[51] J. Sweller. Cognitive load theory. In *Psychology of learning and motivation*, vol. 55, pp. 37–76. 2011.

[52] J. Sweller. The development of cognitive load theory: Replication crises and incorporation of other theories can lead to theory expansion. *Educational Psychology Review*, 2023.

[53] T. Van Gog, F. Paas, and J. Sweller. Cognitive load theory: Advances in research on worked examples, animations, and cognitive load measurement. *Educational Psychology Review*, 2010.

[54] P. Vanneste, A. Raes, J. Morton, K. Bombeke, B. B. Van Acker, C. Larmuseau, F. Depaepe, and W. Van den Noortgate. Towards measuring cognitive load through multimodal physiological data. *Cognition, Technology & Work*, 23:567–585, 2021.

[55] C. Varon, J. Morales, J. Lázaro, M. Orini, M. Deviaene, S. Kontaxis, D. Testelmans, B. Buyse, P. Borzée, L. Sörnmo, et al. A comparative study of ecg-derived respiration in ambulatory monitoring using the single-lead ecg. *Scientific reports*, 10(1):5704, 2020.

[56] J. Wei, E. Siegel, P. Sundaramoorthy, A. Gomes, S. Zhang, M. Vankipuram, K. Smathers, S. Ghosh, H. Horii, and J. Bailenson. Cognitive load inference using physiological markers in virtual reality. *IEEE Conference on Virtual Reality and 3D User Interfaces (IEEE VR)*, 2025.

[57] P. Wen, F. Lu, and A. Z. Mohamad Ali. Using attentional guidance methods in virtual reality laboratories reduces students' cognitive load and improves their academic performance. *Virtual Reality*, 28(2):110, 2024.

[58] R. M. Yerkes and J. D. Dodson. The relation of strength of stimulus to rapidity of habit-formation. *Journal of Comparative Neurology and Psychology*, 18(5):459–482, 1908.