International Journal of Human–Computer Interaction

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/hihc20

A Multimodal Approach Exploiting EEG to Investigate the Effects of VR Environment on Mental Workload

Marta Mondellini, Ileana Pirovano, Vera Colombo, Sara Arlati, Marco Sacco, Giovanna Rizzo & Alfonso Mastropietro

To cite this article: Marta Mondellini, Ileana Pirovano, Vera Colombo, Sara Arlati, Marco Sacco, Giovanna Rizzo & Alfonso Mastropietro (19 Sep 2023): A Multimodal Approach Exploiting EEG to Investigate the Effects of VR Environment on Mental Workload, International Journal of Human–Computer Interaction, DOI: 10.1080/10447318.2023.2258017

To link to this article: https://doi.org/10.1080/10447318.2023.2258017

INTERNATIONAL JOURNAL OF HUMAN–COMPUTER INTERACTION
https://doi.org/10.1080/10447318.2023.2258017

A Multimodal Approach Exploiting EEG to Investigate the Effects of VR Environment on Mental Workload

Marta Mondellini[a], Ileana Pirovano[b], Vera Colombo[a], Sara Arlati[a], Marco Sacco[a], Giovanna Rizzo[b], and Alfonso Mastropietro[b]

[a]Istituto di Sistemi e Tecnologie Industriali Intelligenti per il Manifatturiero Avanzato, Consiglio Nazionale delle Ricerche, Lecco, Italy;
[b]Istituto di Tecnologie Biomediche, Consiglio Nazionale delle Ricerche, Milan, Italy

ABSTRACT
Virtual reality (VR) is a technology that allows users to experience multisensory and interactive environments that simulate real or imaginary scenarios. The effect of different VR immersive technology on mental workload (MWL), i.e., the amount of resources required to perform a task, is still debated; however the potential role of EEG in this context was never exploited. This paper aims to investigate the effects on MWL of performing a cognitive task in a VR environment in two conditions characterized by different degrees of immersion using a multimodal approach which combines well-assessed subjective evaluations of MWL with physiological EEG measures. A cognitive task based on the n-back test was proposed to compare the performance and MWL of participants who used either a head-mounted display (HMD) or a desktop computer to present the stimuli. The task had four different complexity levels ($n = 1$ or 2 with either visual or visual and audio stimuli). Twenty-seven healthy participants were enrolled in this study and performed the tasks in both conditions. EEG data and NASA Task Load indeX (NASA-TLX) were used to assess changes in objective and subjective MWL, respectively. Error rates (ERs) and reaction times (RTs) were also collected for each condition and task level. Task levels had significant effects on MWL, increasing subjective measures and decreasing performance, in both conditions. EEG MWL index have shown a significant increase especially if compared to rest. Different degrees of immersion did not show significant differences neither in individual's performance nor in MWL as estimated by subjective ratings. However, HMD reduced the EEG-derived MWL in most conditions indicating a lower cognitive load. In conclusion, HMD may reduce the cognitive load of some tasks. The reduced level of MWL, as depicted by the EEG MWL index, may have implications for the design and future evaluation of VR-based applications.

KEYWORDS
Mental workload; EEG; virtual reality; immersive environments; N-back tasks

1. Introduction

Mental workload—(MWL) is a multidimensional construct that reflects the amount of mental resources required to perform a task or a set of tasks. While there's still no consensus on MWL definition, a novel, inclusive and operational definition of human MWL can be found in a recent paper by Longo and colleagues (Longo et al., 2022). MWL depends on various factors such as the complexity (Bläsing & Bornewasser, 2021) and duration (Khaksari et al., 2019) of the task, environmental and situational conditions, and individual characteristics and skills (Brookhuis et al., 2009; Verwey, 2000). High MWL levels, as well as too low levels, can lead to decreased performance and increased errors (Fan & Smith, 2017; Kantowitz, 2000; Lysaght et al., 1989). This can affect human performance, well-being, and safety in complex human-machine systems (Midha et al., 2022).

Nowadays, people are required to multitask and maintain prolonged vigilance while still performing well (Haavisto et al., 2010; Holm et al., 2009). This can happen during daily activities or in working environments, or even more specifically in rehabilitation and training settings (Falkenstein & Gajewski, 2021; Karbach & Strobach, 2022; Smithers et al., 2018). Furthermore, it is worth noticing that some individuals may have additional cognitive demands due to aging (Tucker-Drob et al., 2019) or some psychophysiological conditions, such as after an accident or a stroke affecting brain regions (Riese, 1999). In this perspective, assessing the level of MWL is essential and can help optimize task design, rehabilitation strategies, training, and in general human-machine interactions (Asgher et al., 2021; Wickens, 2017). This can potentially lead to the informed development of new technologies, information-based procedures, and user interfaces that maximize human performance.

The complexity of the MWL paradigm led to the use of a multidomain evaluation of MWL involving a combination methods, including subjective ratings, task performance

CONTACT Giovanna Rizzo giovanna.rizzo@cnr.it Istituto di Tecnologie Biomediche, Consiglio Nazionale delle Ricerche, Segrate, Milan, Italy
Supplemental data for this article is available online at https://doi.org/10.1080/10447318.2023.2258017

outcomes, and psychophysiological measures (Arana-De las Casas et al., 2021; Charles & Nixon, 2019; Tao et al., 2019), providing a more comprehensive assessment of an individual's mental workload (Albuquerque et al., 2020; Borgheai et al., 2019).

Over the last 25 years, VR-based technologies have played a growing role in mediating human-machine interactions across different areas spanning from training and entertainment to education and rehabilitation (Slater & Sanchez-Vives, 2016). VR is a technology that allows users to experience multisensory and interactive environments that simulate real or imaginary scenarios. More recently, the advent of good-quality and affordable devices, such as HTC Vive and Oculus Quest, has greatly increased the widespread use of immersive virtual reality (IVR). Such devices, namely Head Mounted Displays (HMDs), can completely immerse the users in a virtual scenario, thus increasing their Sense of Presence and involvement (Krokos et al., 2019; Mondellini et al., 2018; North & North & North, 2016; Pallavicini et al., 2019; Schwind et al., 2019). On the other hand, IVR also poses some challenges and risks for the users, as it may induce cybersickness, anxiety, or mental fatigue, which can strongly impact the perceived quality of the whole experience (Weech et al., 2019).

In this context, the evaluation of the effects of VR on MWL is a hot topic. Previous studies have shown that VR might increase MWL (Jost et al., 2020; Wu et al., 2020) because it can require additional information processing from the user, such as the processing of visual and auditory stimuli, which can increase the user's cognitive load. Conversely, other papers have shown that VR can reduce mental workload. For example, Chang et al. (2019) have shown that VR reduces the individual's MWL during the process of programming an industrial robot, and Chao et al. (2017) reported that VR training can reduce mental workload compared to traditional training methods. In other papers, VR had no significant effect on any of the dimensions of workload (Luong et al., 2019; Xi et al., 2023). Specifically, Chen and colleagues have shown that VR had no significant effect on any workload sub-dimensions in a shopping-related task, whereas the results described by Luong and colleagues support the fact that there is no specific additional mental effort related to the immersion in a VE using a VR HMD. In conclusion, there is still a lack of consensus on whether the degree of immersion provided by a VR device could affect the MWL of the users. This might be due to the different methodological approaches that have been used for the MWL evaluation, as well as different experimental protocols and stimulation. However, the use of electroencephalography (EEG) has not been explored yet, although EEG is a widely used technique, in other contexts, for the estimation of MWL. Indeed, EEG allows obtaining a direct non-invasive measurement of brain activity in different conditions (Borghini et al., 2014; Chikhi et al., 2022; Holm et al., 2009).

Therefore, the main goal of this paper was to explore the effects of different levels of immersion on MWL using a multimodal approach that combines subjective evaluations and physiological EEG measures. We also aimed to compare different task levels (n-back tests) in terms of their impact on MWL. We hypothesize that the VR environment, as well as the complexity of the tasks, has a significant effect on MWL.

2. Materials and methods

2.1. Participants

27 participants (16 males and 11 females) aged between 24 and 41 years (mean = 31.37, $SD = 4.32$) were enrolled. The study was conducted according to the principles expressed in the Declaration of Helsinki and was approved by the ethics committee of the National Research Council of Italy. All participants had normal or corrected-to-normal vision and no history of neurological, cognitive, or psychiatric disorders. The participants signed a written informed consent before participating in the study.

2.2. Experimental protocol

The experimental sessions were carried out in a controlled laboratory environment to avoid any external noise sources that could interfere with the participants' cognitive performance. The same room was used for all sessions and the time range was fixed to avoid variations in circadian rhythms.

Participants were required to have a sufficient night's sleep before each session and to refrain from consuming any caffeinated beverages within 4 h before the experiment. They were randomly assigned to one of two groups: one group used the desktop computer (Desktop) in the first session and then switched to immersive virtual reality (HMD) after a week, while the other group did the opposite. The proposed cognitive task, based on the n-back test (see section 2.3 for further details), was identical in both sessions and was performed within a natural environment (see section 2.4 for further details). The experimenters explained to the participants how to perform the task and to keep their heads still and restrain from sudden movements as far as possible during the entire EEG acquisition. They also gave them instructions on how to fill out the NASA Task Load indeX (NASA-TLX) (see section 2.6 for further details) that was integrated into the digital application and administered at the end of each level of the task. The participants performed a 2-min familiarization session with the cognitive tasks; this activity was performed at least 10 min before the beginning of the experimental test. During this phase, doubts about the execution of the test and the meaning of the questions were clarified. Furthermore, participants could see the original NASA-TLX on paper.

The session proceeded with the dressing of the EEG helmet followed by the HMD in case of immersive VR sessions. We carefully checked that any EEG channels would not detach during the HMD wearing, that the electrodes' impedance would remain in the accepted values range (below 20 kΩ) and that the channels' bridge would not occur due to the spreading of gel. The HMD used in this study is an

HTC Vive Pro (2011-2023 HTC corporation©) with a wireless adapter, characterized by a resolution of 1440 × 1600 pixels per eye, a field of view of 110°, and a 90 Hz refresh rate. The HTC Vive Pro is a Hi-res certified headset, with Hi-res certified headphones integrated with 3D spatial audio. Resting-state EEG signals were then recorded for 3 min before each session; subsequently, the experimental phase started. The whole task was divided into four blocks of 4 min each in which the user performed the cognitive task with incremental difficulty (see section 2.3). At the end of each level of cognitive task, the NASA-TLX was administered to evaluate subjective MWL. A schematic representation of the whole protocol is displayed in Figure 1.

2.3. N-back test

N-back test is a tool that measures a part of working memory and working memory capacity (Jaeggi et al., 2008). It involves presenting a sequence of stimuli (such as images or letters) one by one and asking the participants to indicate if the current stimulus is the same as the one that appeared n steps before (one or two steps in our protocol). In the dual-task paradigm (Jaeggi et al., 2003), two independent sequences are presented simultaneously, in our case an auditory and a visual stimulus. In this way, the attentive resources from the participants should be distributed in two cognitive functions. The load factor n was set to 1 in the first (Level 1) and third levels (Level 3), and to 2 in the second (Level 2) and fourth levels (Level 4). In the first two levels (Level 1 and Level 2), the participant responded only to visual stimuli. In levels 3 and 4 (Level 3 and Level 4), the auditory stimuli were added to the visual stimuli. At each level, the participants were required to react differently depending if the repeated stimulus was visual or auditory. They were also instructed to respond as quickly and accurately as possible.

2.4. Digital environment

The digital environment—developed in Unity (version 2020.3.29f1)—consisted of two scenes. The first one—loaded when the application is launched—was used for the REST assessment (Figure 2(a)). The participant was asked to relax and focus his/her gaze on a white cross placed over a black background. The participant was asked to press the Space bar to start the rest phase, which then lasts three minutes.

After this period, the second scene was automatically loaded. The virtual scenario reproduced a park where the user moved automatically along a predefined path at a constant speed of 60 revolutions per minute. The participant observed the virtual environment from a first-person perspective. The virtual park scenario derives from previous works carried out by the STIIMA research group (Colombo et al., 2023; Mrakic-Sposta et al., 2018). Such an environmental setting was considered suitable as relaxing and not containing too many elements that could distract the participants' attention from the main cognitive task (e.g., no sudden events). The original setup, designed for dual-task training, for rehabilitation purposes, foresees that the user controls the movement in the virtual environment by cycling on a stationary bike (Arlati et al., 2019; Pedroli et al., 2018). In the present study, aimed at specifically evaluating how increasing the immersion of such VR training could impact the mental workload, we maintained the original setup with the exclusion of the cycling part to focus the evaluation on the cognitive aspect and avoid other influencing factors. Letting the user the possibility to explore the

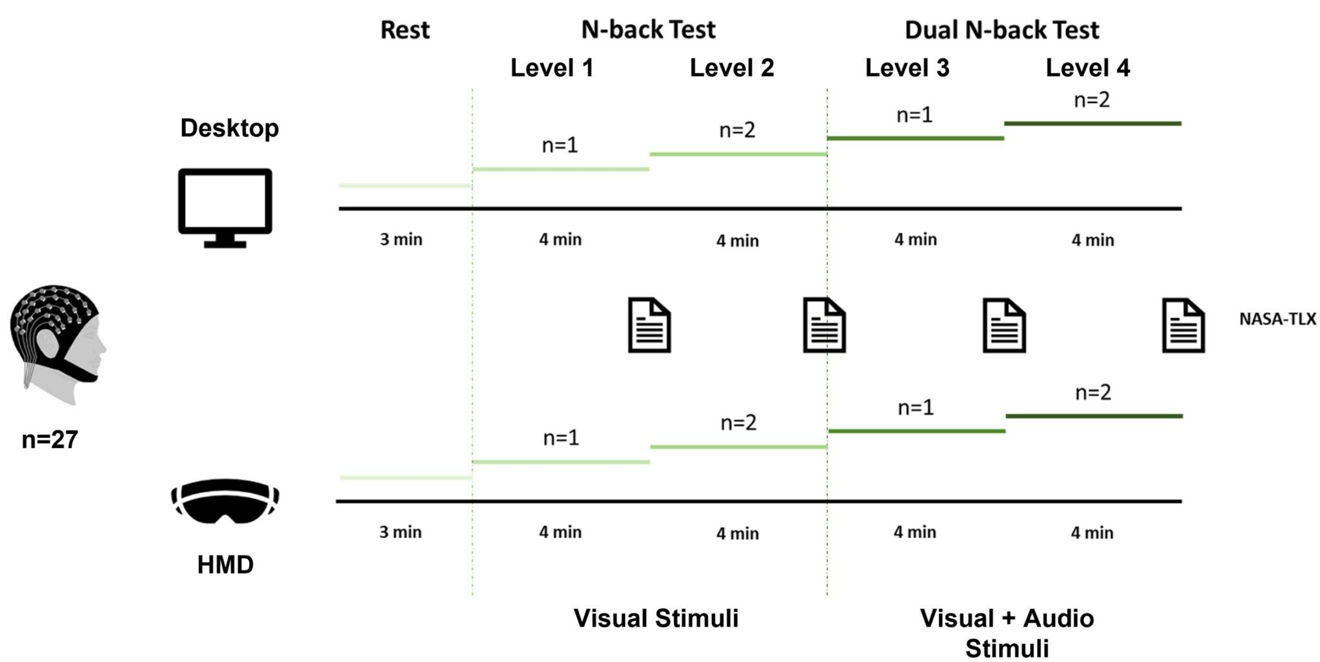

Figure 1. Experimental protocol. The figure shows the timeline of events in each session. The participants performed a rest phase followed by four levels of cognitive task based on the n-back test. The task was presented in either a desktop computer or a head-mounted display (HMD) condition. The EEG signals were recorded throughout the session. The NASA-TLX was administered after each level of the task to assess the subjective MWL.

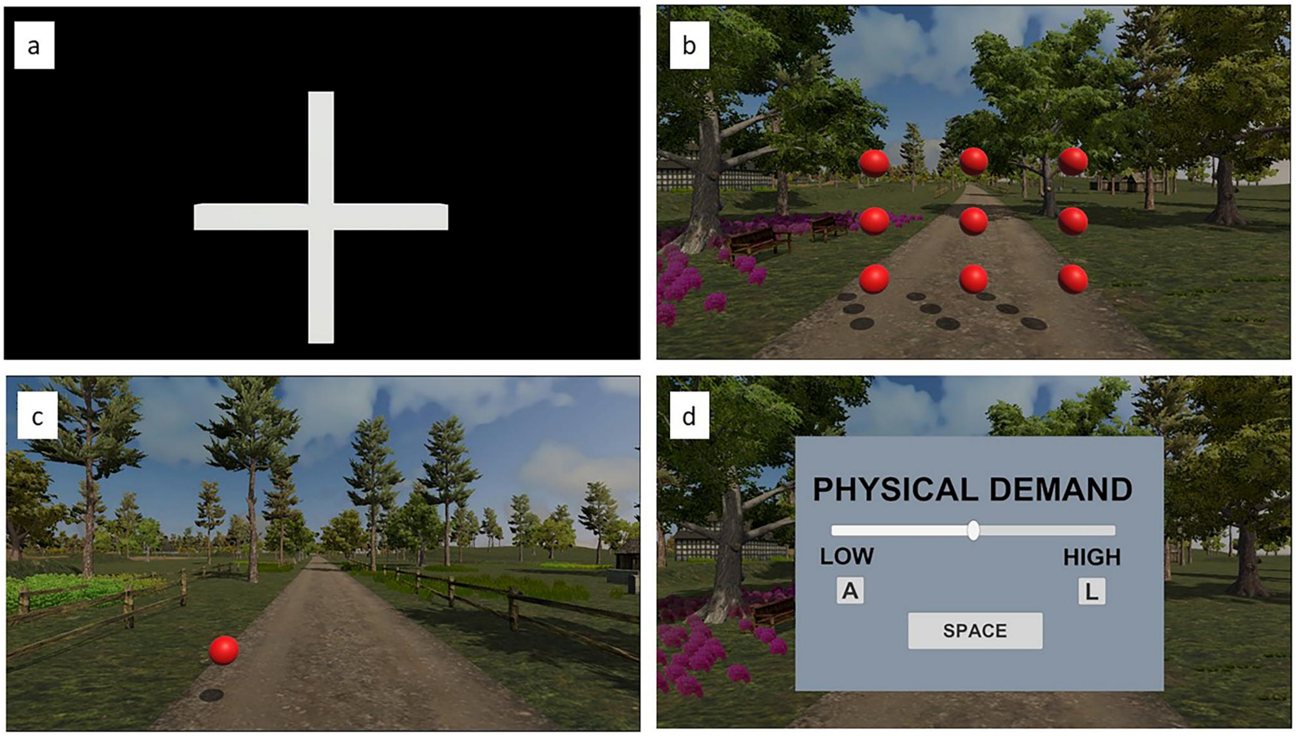

Figure 2. Digital environment. The figure shows the screenshots of the digital environment developed in Unity. (a) The environment proposed for rest phase; (b) The nine possible positions for visual stimuli; (c) A visual stimulus appears in position 7; (d) The NASA-TLX questionnaire.

virtual park scenario, even without actively controlling the movement, is important to facilitate the transfer of results to future rehabilitation settings. The following features characterizing the n-back cognitive task were added for the purpose of this study.

The LSL4Unity plugin was imported into the Unity project to allow EEG acquisition synchronization. The package enables the integration of the labstreaminglayer—an open-source networked middleware ecosystem to stream, receive, synchronize, and record neural, physiological, and behavioral data streams acquired from diverse sensor hardware (Kothe et al., 2014). Before starting the task, the participant was shown all the visual stimuli, i.e., nine red spheres on a 3×3 invisible grid, to familiarize with the nine positions (Figure 2b). Since the purpose of the present study was not to evaluate a specific cognitive ability—e.g., the discrimination of shape or position in space—one of the visual features of the stimulus already used with the n-back task, namely the spatial position of the stimulus (Jaiswal et al., 2019) was chosen. For the audio stimulus, a pre-recorded voice spelled a letter among nine possible ones (a, e, i, o, u, h, g, m, j). The choice of this kind of verbal auditory stimulus was based on the fact that, in the Italian language, the letters proposed are easily distinguishable sounds. The participant had to respond by pressing the "A" key on the keyboard when the visual target appeared and the "L" key if the auditory target occurred. In our experiment, the time between the disappearance of one stimulus and the appearance of the next was 2500 ms, and each visual stimulus remained on the screen for 500 ms. Reaction times (RTs) and the correctness of the answers were automatically collected by the application. The four tasks—as described in section 2.3—were presented in sequence; an example in which the sphere appears in position 7 is shown in Figure 2(c). At the end of each task level, the NASA-TLX was presented on a UI panel at the center of the user's point of view (Figure 2(d)).

2.5. Performance metrics

The objective cognitive task performances were evaluated using the error rate (ER) and the reaction time (RT). The ER was calculated by dividing the number of incorrect answers (False Positive, FP and False Negative, FN) by the total number of stimuli requiring an answer (True Positive, TP; FP and FN) from the participants.

The ER is defined as:

$$ER = \frac{FP + FN}{TP + FP + FN} \tag{1}$$

which is equivalent to 1-Jaccard Index (Jaccard, 1912; Taha & Hanbury, 2015)

2.6. Subjective MWL assessment

Subjective MWL was evaluated using NASA-TLX (Hart, 2006; Hart & Staveland, 1988), a widely used multidimensional assessment tool that rates perceived workload across six dimensions: mental demand, physical demand, temporal demand, performance, effort, and frustration. Each dimension of NASA-TLX is rated on a scale from 0 (very low) to 20 (very high). The participants completed the questionnaire after each task's level using the digital version of the questionnaire embedded within the software developed to present stimuli. Both the scores in the six dimensions and the sum of the scores (Overall NASA-TLX Score) were considered in this study.

The NASA-TLX proposed to participants slightly differs from the original version. First of all, in the original, the six scales are all reported on the same page, one below the other; in the present study, this was not possible as this choice would have made the six scales impossible to read, especially in HMD mode. This is why the six scales were proposed in six successive steps/pages, one after the other. However, the original scale was shown to all participants before the test was performed. Additionally, each NASA-TLX scale in this study was proposed with a default value in the center, and the scale was not graduated, unlike the original NASA-TLX. However, in this study, the goal was to evaluate the variations related to the difficulty of the tasks and the degree of immersion; therefore any underestimation or overestimation effects should not affect the results.

2.7. Objective MWL assessment

2.7.1. EEG acquisitions

Continuous EEG data were collected using a compact 32-channel system (eego^{TM}sports 32, ANT Neuro®, Enschede, The Netherlands). A gel-based electrode cap with sintered Ag/AgCl electrodes was used (Waveguard, ANT Neuro®, 10–20 system). The online reference was placed at the CPz electrode. The signal was acquired with eego sports acquisition software connected to a 24 bits amplifier at a sampling rate of 500 Hz. Impedances for all electrodes were kept below 20 kΩ. EEG signals were recorded across 30 channels: Fp1, Fpz, Fp2, F7, F3, Fz, F4, F8, FC5, FC1, FC2, FC6, T7, C3, Cz, C4, T8, CP5, CP1, CP2, CP6, P7, P3, Pz, P4, P8, POz, O1, Oz, and O2 excluding the mastoids electrodes (M1 and M2). The starting and ending points of each trial composing the whole session (e.g., Rest, Task 1, Task 2, Task 3, Task 4) were automatically labeled using the lab stream layer (LSL).

2.7.2. EEG pre-processing

EEG signals were processed as previously suggested by Mastropietro et al. (Mastropietro et al., 2023). In particular, EEG signals were band-pass filtered in the range 1–45 Hz using a Hamming windowed sinc Finite Impulse Response (FIR) filter. Bad channels were removed by evaluating the normed joint probability of the average log power across the channels (Gabard-Durnam et al., 2018). Channels whose probability falls more than three standard deviations from the mean are removed as bad channels. Subsequently, the Artifact Subspace Reconstruction (ASR) algorithm (Mullen et al., 2013) was used to interpolate artifact "bursts" with a variance higher than fifteen standard deviations different from the automatedly detected reference signal, as previously suggested (Chang et al., 2018). Independent Component Analysis (ICA) was then used to detect and remove artifacts (such as eye movements and electrocardiographic signals) that usually overlay with brain activity in EEG recordings. The extended Infomax (Bell & Sejnowski, 1995) ICA algorithm was used in this work. ICLabel (Pion-Tonachini et al., 2019) was used to automatically reject independent components having a probability to be plausible brain sources of less than 40%. Channels that were removed as "bad channels" were replaced by data interpolated from nearby "artifact-free" channels using a spherical function, and EEG signals were re-referenced to the average of the channels. All the pre-processing steps were implemented in MATLAB (R2021b, The MathWorks) using the EEGLAB toolbox (Delorme & Makeig, 2004).

2.7.3. MWL index calculation

Welch's method was used to analyze the pre-processed EEG signals in the frequency domain and obtain the power spectra in the 1–45 Hz range. We divided the EEG signal into non-overlapping segments of 1 s length (500 samples) and multiplied each segment by a Hamming window. We then computed the discrete Fourier transform of each segment and squared its magnitude to get the periodogram. We averaged the periodograms across segments for each of the three experimental blocks to get the power spectra for each task. Next, we computed the absolute band power for each channel by integrating the power spectrum over theta (4–8 Hz) and alpha (8–13 Hz) frequency bands. Finally, according to the formulation reported in Holm et al. (2009), we calculated the MWL index (MWLI) of each block by dividing the absolute power θ at Fz, corresponding the a frontal midline electrode, by the absolute power α at Pz, corresponding to a parietal midline electrode.

$$MWLI = \frac{\theta_{Fz}}{\alpha_{Pz}} \tag{2}$$

2.8. Statistical Analysis

Descriptive analyses of performance metrics, subjective and EEG-based objective MWL were run. Due to the small sample size, asymmetry and kurtosis of some variables, a transformation of the data has been performed in their respective ranks. The reliability of NASA-TLX was evaluated by Alpha and Omega coefficients (Dunn et al., 2014; Sijtsma, 2009). A frequently cited acceptable range of Alpha coefficient is a value of 0.70 or above (Hair et al., 2010) and the same cut-off is set for Omega coefficient (McNeish, 2018). The repeated measure two-way ANOVA test on the rank-transformed data (Iman & Conover, 1979) was used to examine whether the interaction of the two within-subjects factors (Immersion modality and task levels) had a significant effect on the MWL-related metrics measured by different tools (e.g., EEG, Error Rates, Reaction Times, Overall NASA-TLX Score, etc.). A pairwise two-sided Wilcoxon test was then conducted for multiple comparisons between groups and *p*-values were adjusted using the false discovery rate method. We considered p-values $\leq$ 0.05 as significant. Finally, Spearman's correlation coefficient was calculated to assess the intra- and inter-domain correlations between MWL metrics, considering total experience scores and not individual task levels. Statistical analysis and tests were performed in R (version 4.2.1) (R Core Team, 2022) embedded in RStudio (2022.12.0 Build 353) and IBM SPSS v.28 (IBM Corp., 2021).

3. Results

To provide a comprehensive overview of the findings and describe all the relevant results, this paragraph is divided into four sections. The first section describes changes in objective measures of performance (i.e., ER and RT). The second section describes changes in subjective measures of MWL determined through the NASA-TLX. The third section describes quantitative changes in MWL metrics derived from EEG. Finally, the fourth section shows significant correlations among parameters.

3.1. Performance assessment

As shown in Figure 3(a,b), in both Desktop and HMD conditions, ERs to visual stimuli showed a tendency to increase over task levels. Similarly, ERs to auditory stimuli increased over task levels in both conditions. Two-way ANOVA revealed that task levels had a significant effect on the ERs, as indicated by a p-value < 0.001 for both visual and auditory stimuli. All of the differences in ERs among task levels were statistically significant, except those in response to visual stimuli between levels 2 and 3 for Desktop and levels 1 and 3 for HMD as listed in Table 1.

As shown in Figure 3(c,d), in both Desktop and HMD conditions, RTs to visual stimuli constantly increased over task levels. Similarly, RTs to auditory stimuli increased over task levels in both conditions. Two-way ANOVA revealed a significant effect of task levels on the RTs with p-value $<$ 0.001 in both visual and auditory stimuli. Even in this case, the differences in RTs among task levels were statistically significant, except those in response to visual stimuli between levels 2 and 3 for both Desktop and VR as shown in Table 1.

Considering both the visual and audio tasks, the degree of immersion had not a significant effect on ERs ($p > 0.05$) whereas the two-way ANOVA revealed a significant effect of HMD environment on RTs in response to auditory stimuli ($p = 0.038$). Regardless of the task levels, neither ERs nor the RTs showed any significant difference between Desktop and HMD after the pairwise comparison.

3.2. Subjective MWL assessment

3.2.1. NASA-TLX Overall Score

The overall scale obtained good reliability in both Desktop ($\alpha = 0.75$ and $\omega = 0.74$ level 1; $\alpha = 0.82$ and $\omega = 0.82$ level 2;

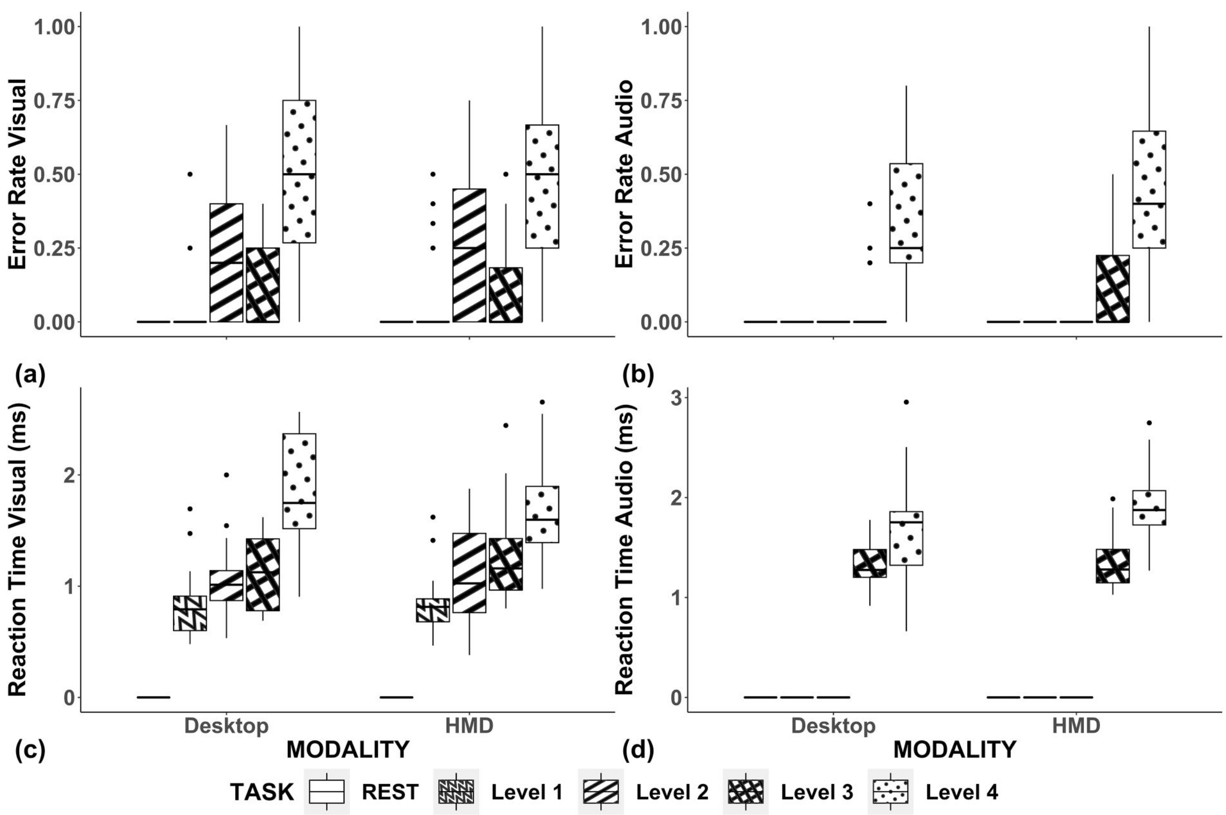

Figure 3. Performance metrics. The figure shows the boxplots of error rates (ERs) (a) and (b) and reaction times (RTs) (c) and (d) for visual and auditory stimuli in desktop and HMD conditions. For Desktop, ERs for visual stimuli had median values of 0, 0.2, 0 and 0.5 in levels 1, 2, 3 and 4 respectively. As to HMD, ERs had median values of 0 in level 1, 0.25 in level 2, 0 in level 3 and 0.5 in level 4. ERs for auditory stimuli had median values of 0 and 0.25 in levels 3 and 4 respectively for Desktop whereas the median values for HMD were 0 and 0.4 respectively. For Desktop, RTs to visual stimuli had median values of 0.79 s in level 1, 1.01 s in level 2, 1.12s in level 3, and 1.75 s in level 4. For HMD, RTs had median values of 0.81 s in level 1, 1.02 s in level 2, 1.16 s in level 3, and 1.60 s in level 4. In the case of auditory stimuli, RTs had median values of 1.27 s and 1.75 s in levels 3 and 4 respectively for Desktop whereas the median values for HMD were 1.28 s and 1.88 s respectively. There were no significant differences between desktop and HMD conditions in terms of ERs and RTs.

Table 1. Pairwise comparisons between task levels for both ERs and RTs, using the Wilcoxon test.

			ER visual		RT visual		ER audio		RT audio	
MODALITY	Group 1	Group 2	*p* adj	*p* adj signif	*p* adj	*p* adj signif	*p* adj	*p* adj signif	*p* adj	*p* adj signif
DESKTOP	Level 1	Level 2	0.002	**	0.005	**				
DESKTOP	Level 1	Level 3	0.034	*	0.01	*				
DESKTOP	Level 1	Level 4	0.000166	***	0.0000519	****				
DESKTOP	Level 2	Level 3	0.179	ns	0.264	ns				
DESKTOP	Level 2	Level 4	0.002	**	0.0000522	****				
DESKTOP	Level 3	Level 4	0.000166	***	0.000106	***	0.001	**	0.001	**
HMD	Level 1	Level 2	0.011	*	0.006	**				
HMD	Level 1	Level 3	0.592	ns	0.000165	***				
HMD	Level 1	Level 4	0.000454	***	8.94E-07	****				
HMD	Level 2	Level 3	0.041	*	0.243	ns				
HMD	Level 2	Level 4	0.015	*	0.000189	***				
HMD	Level 3	Level 4	0.000516	***	0.000788	***	0.0000432	****	0.0000432	****

The adjusted *p*-values and statistical significance for both visual and auditory stimuli are listed.

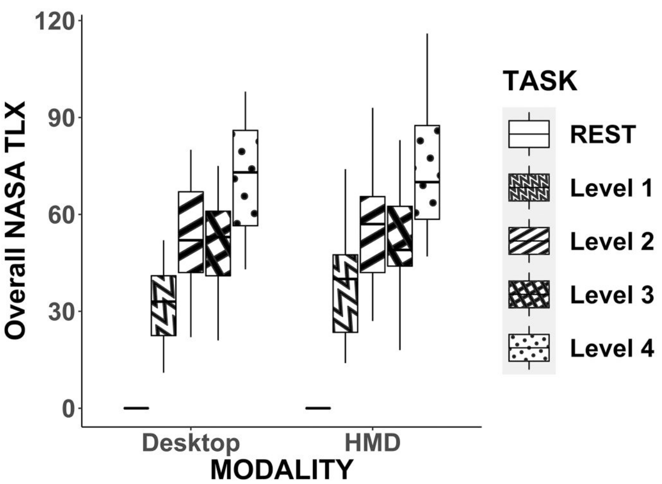

Figure 4. Subjective MWL assessment. The figure shows the boxplots of the overall NASA-TLX score in desktop and HMD conditions. In the case of Desktop, the Overall NASA-TLX had median values of 33, 52, 53 and 73 in levels 1, 2, 3 and 4 respectively. In the case of VR, Overall NASA-TLX showed median values of 40 in level 1, 57 in level 2, 49 in level 3 and 70 in level 4. There were no significant differences between desktop and HMD conditions in terms of the overall NASA-TLX score.

Table 2. Pairwise comparisons between task levels for Overall NASA-TLX, using the Wilcoxon test.

			Overall NASA-TLX	
MODALITY	Group 1	Group 2	*p* adj	*p* adj signif
DESKTOP	Level 1	Level 2	0.0000111	****
DESKTOP	Level 1	Level 3	0.0000111	****
DESKTOP	Level 1	Level 4	0.0000111	****
DESKTOP	Level 2	Level 3	0.736	ns
DESKTOP	Level 2	Level 4	0.0000384	****
DESKTOP	Level 3	Level 4	0.0000111	****
HMD	Level 1	Level 2	0.0000118	****
HMD	Level 1	Level 3	0.0000192	****
HMD	Level 1	Level 4	0.0000118	****
HMD	Level 2	Level 3	0.258	ns
HMD	Level 2	Level 4	0.0000124	****
HMD	Level 3	Level 4	0.0000118	****

The adjusted *p*-values and statistical significance are listed.

$\alpha = 0.81$ and $\omega = 0.82$ level 3; $\alpha = 0.75$ and $\omega = 0.75$ level 4) and HMD ($\alpha = 0.78$ and $\omega = 0.77$ level 1; $\alpha = 0.86$ and $\omega = 0.87$ level 2; $\alpha = 0.81$ and $\omega = 0.82$ level 3; $\alpha = 0.83$ and $\omega = 0.82$ level 4) conditions.

As displayed in Figure 4, in both Desktop and HMD conditions, the Overall NASA-TLX Score showed a tendency to increase over task levels, analogously to what was previously described for ERs and RTs.

Most of the differences in Overall NASA-TLX Score among task levels were statistically significant, except those in response to visual stimuli between levels 2 and 3 for both Desktop and HMD as listed in Table 2.

Even for the Overall NASA-TLX Score, regardless of the task levels, there was no significant difference between Desktop and HMD and no effect was revealed by two-way ANOVA.

3.2.2. NASA-TLX subscales

Median and interquartile differences for each subscale of the NASA-TLX are reported in Table 3.

Regarding Mental Demand, Temporal Demand, Performance, Effort, and Frustration no differences between the two conditions were found, at any level. Physical Demand differed between conditions in the first level ($p = 0.002$) and the third level ($p = 0.022$).

The differences in scores between levels, separated by experimental condition, are reported in supplementary files Table 1.

3.3. Objective MWL assessment by EEG

As shown in Figure 5, MWLI assessed using EEG signals did not reveal any significant trends with respect to task levels. However, there was a consistent increase in MWLI during tasks compared to rest in both Desktop and HMD conditions, as shown in Table 4. Additionally, when using a desktop, a slight but significant difference was observed between Levels 3 and 4 whereas in the case of HMD, Levels 1, 2, and 4 were all significantly different from Level 3.

As regards the effect of HMD environment on MWLI values, the two-way ANOVA test has shown a significant result ($p = 0.024$). In particular, the pairwise comparisons showed significant lower MWLI in HMD if compared to Desktop between levels 2, 3 and 4 as shown in Table 5.

3.4. Correlations among metrics

As to the performance metrics, both ERs and RTs were significantly moderately correlated with each other (desktop condition: $\rho = 0.46$, $p < 0.001$ for visual and $\rho = 0.33$,

Table 3. Median and interquartile difference (in the brackets) for Mental Demand (MD), Physical Demand (PD), Temporal Demand (TD), Performance (P), Effort (E), and Frustration (F) in both conditions.

		MD		PD		TD		P		E		F	
Level	Condition	DESK	HMD	DESK	HMD	DESK	HMD	DESK	HMD	DESK	HMD	DESK	HMD
Lev.1		6 (6)	8 (6)	3 (4)	5 (7)	7 (5)	7 (4)	5 (6)	6 (6)	5 (4)	7 (6)	6 (5)	6 (6)
Lev.2		11 (6)	12 (5)	5 (6)	8 (7)	10 (6)	10 (5)	8 (6)	8 (5)	11 (5)	11 (5)	9 (6)	8 (6)
Lev.3		11 (5)	11 (4)	6 (6)	8 (6)	10 (5)	10 (5)	7 (4)	7 (5)	11 (5)	10 (6)	9 (6)	8 (6)
Lev.4		16 (6)	15 (5)	7 (8)	10 (9)	10 (7)	12 (7)	12 (6)	12 (6)	13 (6)	14 (6)	11 (8)	11 (8)

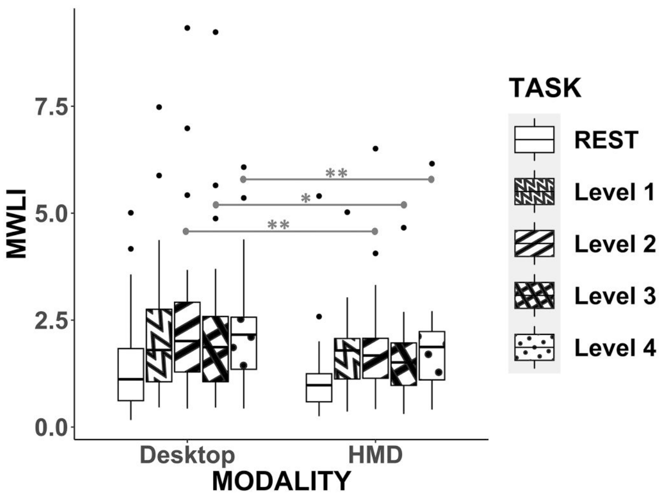

Figure 5. Objective MWL assessment by EEG. The figure shows the boxplots of the mental workload index (MWLI) derived from EEG signals in desktop and HMD conditions. MWLI, in the case of Desktop, had median values of 1.12 at rest, 1.80 in level 1, 2.01 in level 2, 1.87 in level 3 and 2.16 in level 4. Considering HMD, MWLI showed median values of 0.978 at rest, 1.79 at level 1, 1.67 at level 2, 1.51 at level 3 and 1.87 at level 4. The MWLI was significatly lower in HMD than in desktop condition in levels 2, 3 and 4.

Table 4. Pairwise comparisons between task levels for MWLI, using the Wilcoxon test.

			MWLI	
MODALITY	Group 1	Group 2	*p* adj	*p* adj signif
DESKTOP	REST	Level 1	9.93E-08	****
DESKTOP	REST	Level 2	1.12E-07	****
DESKTOP	REST	Level 3	9.93E-08	****
DESKTOP	REST	Level 4	9.93E-08	****
DESKTOP	Level 1	Level 2	0.159	ns
DESKTOP	Level 1	Level 3	0.732	ns
DESKTOP	Level 1	Level 4	0.169	ns
DESKTOP	Level 2	Level 3	0.06	ns
DESKTOP	Level 2	Level 4	0.732	ns
DESKTOP	Level 3	Level 4	0.046	*
HMD	REST	Level 1	0.0000084	****
HMD	REST	Level 2	1.42E-06	****
HMD	REST	Level 3	0.000209	***
HMD	REST	Level 4	1.42E-06	****
HMD	Level 1	Level 2	0.698	ns
HMD	Level 1	Level 3	0.029	*
HMD	Level 1	Level 4	0.534	ns
HMD	Level 2	Level 3	0.006	**
HMD	Level 2	Level 4	0.953	ns
HMD	Level 3	Level 4	0.005	**

The adjusted *p*-values and statistical significance are listed.

$p = 0.02$ for audio stimuli; HMD: $\rho = 0.33$, $p < 0.001$ for visual and $\rho = 0.34$, $p = 0.012$ for audio stimuli). When considering the relationship between performance metrics and subjective MWL, ERs exhibited stronger significant correlations with the Overall NASA-TLX Score. The correlation's indexes between RT, ER and NASA-TLX total score in both

Table 5. Pairwise comparisons between different VR environments for each task level for MWLI using the Wilcoxon test.

			MWLI	
TASK	Group 1	Group 2	*p* adj	*p* adj signif
REST	DESKTOP	HMD	0.455	ns
Level 1	DESKTOP	HMD	0.106	ns
Level 2	DESKTOP	HMD	0.006	**
Level 3	DESKTOP	HMD	0.03	*
Level 4	DESKTOP	HMD	0.006	**

The adjusted *p*-values and statistical significance are listed.

conditions are reported in Figure 6 for both visual and auditory stimuli.

ERs and RTs for visual and auditory stimuli were found to be correlated with several domains composing the NASA-TLX, in both conditions. All correlation indexes between NASA-TLX subscales and performance metrics are shown in Supplementary Figures 1–4. The strongest correlation coefficients were found among NASA-TLX domains with values up to 0.87 as in the case of the correlation between Mental Demand and NASA-TLX Total Score in desktop condition, and up to 0.90 in the case of Effort and NASA-TLX total score using HMD, when visual stimuli were presented. No correlations were found between MWLI and either the objective performance metrics or the subjective metrics.

4. Discussion

This study investigated the effects of different levels of immersion (Desktop vs HMD) in a VR environment on MWL using a multimodal approach that combines subjective and physiological measures along with the assessment of individual's performance.

4.1. Effect of task levels on MWL

The results showed that ERs and RTs increased over task levels in both desktop and HMD conditions; this is consistent with what was reported by the Norman and Bobrow model (Norman & Bobrow, 1975), whereby performance in a task can be affected when the task overloads the amount of available resources, in our case the working memory span (Navon & Gopher, 2014). Similarly, the Overall NASA-TLX Score, namely the subjective MWL, increased over task levels in both desktop and HMD conditions, as reported in other studies (e.g., Aksu et al., 2023). Furthermore, the correlation that emerged between objective performance and perceived workload is consistent with what is reported in the literature (Moray, 1982). This proves that N-back tasks effectively elicit varying levels of MWL, as demonstrated by

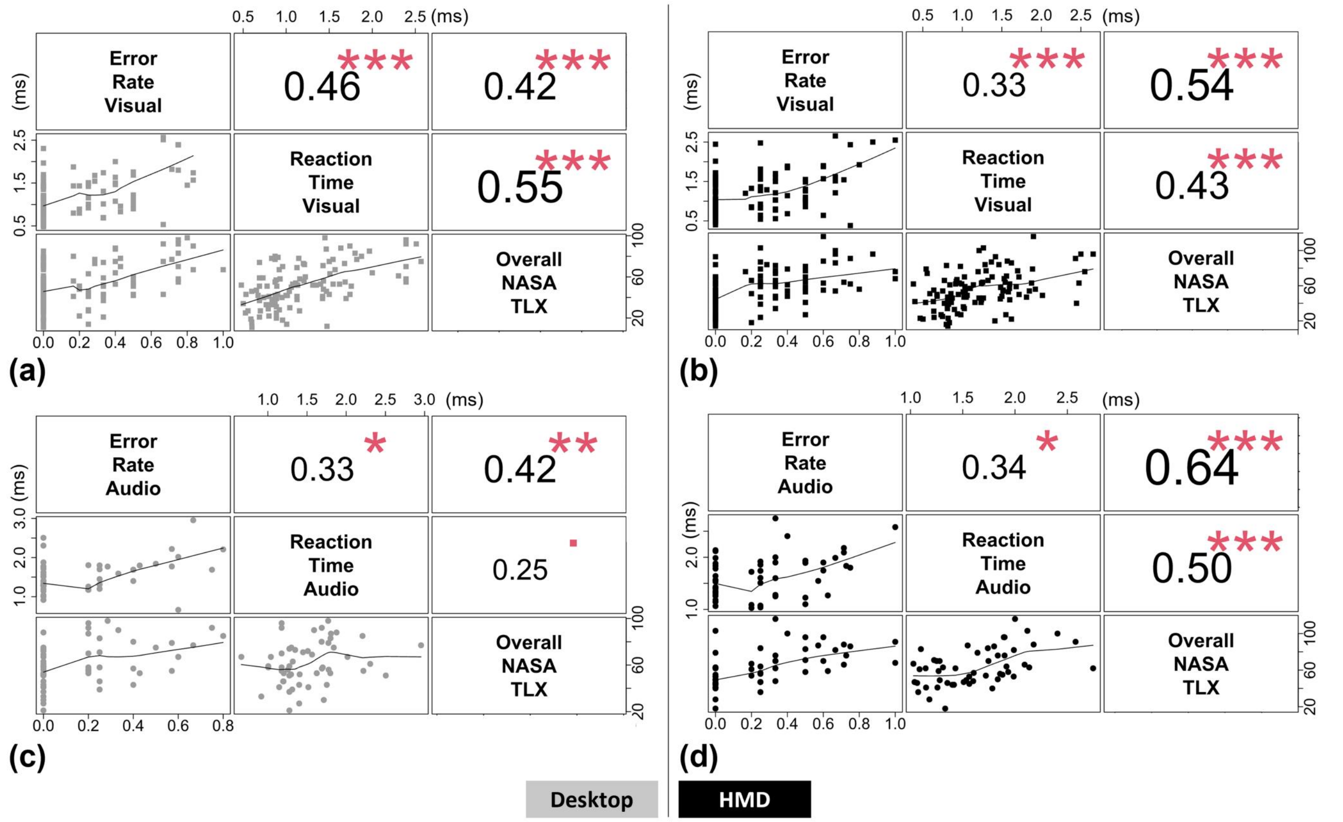

Figure 6. Upper Panels show Spearman's correlation indexes for visual ERs and TRs, as well as the total NASA-TLX score in both desktop (a) and HMD (b) conditions. Lower Panels display the same indexes for auditory ERs (c) and TRs (d). In each panel, on the bottom of the diagonal the bivariate scatter plots with a fitted line are displayed. On the top of the diagonal the value of the correlation plus the significance level as stars. Each significance level is associated to a symbol: p-values (0, 0.001, 0.01, 0.05, 0.1, 1) $<=>$ symbols ("***", "**", "*", "." ").

both subjective and objective metrics, and it is consistent with previous research (Herff et al., 2013; Luong et al., 2019).

However, our study found consistent significant differences in the EEG MWL index mainly between rest and each task level while only incidentally among task levels. Additionally, no significant correlations were found between the EEG-derived MWL index and other subjective measures or performance metrics. To explain this partial inconsistency, it should be considered that our definition of the MWL index considers the contributions of theta band power, which is associated with working memory processes, and alpha band power, which has long been considered an indicator of the brain's "wakefulness" state due to its desynchronization during cognitive tasks (Chikhi et al., 2022). In our study, the cognitive workload increased compared to the resting condition but remained relatively stable regardless of task complexity. This may be due to the low frequency of events requiring active user intervention (a maximum of five events in each task) during the proposed sessions. As a result, the average MWL index derived from EEG signals over the 4-min task duration may not be sensitive enough to the low effect on working memory and attention elicited by such a small number of events to be detected. Recently, Tremmel and colleagues (Tremmel et al., 2019) found that the spectral power of EEG signals can effectively discriminate workload levels during an N-back task presented in a VR environment. Specifically, they found a correlation between workload levels and β and γ power for frontal electrodes. However, they did not find the typical frontoparietal θ/α variations consistently in all subjects, which is in agreement with our results.

4.2. Effect of immersive VR environments on MWL

As to the effect of different immersive VR environments on MWL, no significant differences between desktop and HMD condition was revealed neither by performance measures nor by subjective MWL metrics. However, interestingly, the study revealed that when a HMD was used, the EEG-derived MWLI decreased in most conditions. To control for possible biases due to the setup variations in desktop and HMD tests, we performed a signals sanity check, both during the acquisitions and the off-line pre-processing. Moreover, we did not find significant differences between the baseline EEG-MWLI acquired in the two conditions, suggesting that significant changes observed during the tasks could be ascribed to the VR modality. We may hypothesize that the reduced MWL recorded in the HMD condition can be due to the higher sense of presence provided by the device (Gorini et al., 2011; Paes et al., 2021; Pallavicini et al., 2019), which have contributed in distracting the participants from the demands of the n-back cognitive task (as it happens, e.g., for pain (Shahrbanian et al., 2012) or perceived effort

(Colombo et al., 2022; Gomez et al., 2022)). However, more specific studies investigating this relationship are needed to confirm this hypothesis. The effects of VR on MWL are still debated and controversial results have been reported in the recent literature, as described in the introduction. Our results, considering just the subjective and objective performance metrics, agree with those described in some previous papers (Xi et al., 2023) that did not show any significant effect of an immersive environment on MWL levels. Specifically, Luong and colleagues conducted an experiment to assess the effect of being immersed in a virtual environment using an HMD on the user's mental effort while performing a standardized cognitive task (the well-known N-back task, with three difficulty levels). Their results support the fact that there is no specific additional mental effort related to immersion in a virtual environment using a VR HMD. However, the decrease in the EEG MWL index in HMD condition observed in this study is interesting and unprecedented, as no prior research has utilized EEG measurements in this particular context. Other physiological measurements (e.g., Blood Volume Pulse; Photo Plethysmography; Electrodermal Activity, Heart Rate Variability) have failed to show any difference (Luong et al., 2019) or seemed to be more sensitive in the comparison of a non-immersive VR-based industrial training versus a traditional method (Chao et al., 2017). The methodology proposed in this study may be replicated in future studies to extend the evaluation of MWL to other factors that characterize VR, e.g., the ways of interaction, the types of instructions and/or feedback provided. In addition, other cognitive tests to elicit a stress response, e.g. the Stroop Test, specifically adapted to VR applications, may be employed as proposed and validated by Gradl and colleagues (Gradl et al., 2019) but not yet tested with EEG-derived indexes.

The reduced level of MWL, as depicted by the EEG MWL index, may have implications for the design and evaluation of VR-based applications. In the rehabilitation field, measuring MWL can be essential to understand better which technology can favor the improvement of patients' clinical outcomes. For instance, for administering cognitive training to patients or older adults with cognitive decline, it is key that only the task itself requires MWL and not the interaction with the technological means. Indeed, if acting in the virtual environment becomes too demanding, it may happen that: first, the cognitive training may not address the correct cognitive domains, and second, the task may become too challenging, jeopardizing any benefit for the patients. Again, in an assistive scenario, measuring the MWL can be essential to choose the right assistive technologies, beyond the rehabilitation tasks themselves, that can favor patient mobility and autonomy. In this context, recently, a lightweight wearable robotic exoskeleton was proposed to assist potential stroke patients with an integrated portable brain interface using MWL signals acquired with a portable functional near-infrared spectroscopy (fNIRS) system (Asgher et al., 2021). The system may generate command signals for operating a wearable robotic exoskeleton hand using two-state MWL signals, thus adapting to the user's status.

4.3. Limitations and future works

The main limitation of this study is the low number of volunteers involved in the experimentation. In the future, to generalize our findings, a larger sample size could be used to confirm the results pointed out in this study. If this happens, the fact that the MWL is lower using immersive VR environments will open up new perspectives; for example, in the field of training and rehabilitation, in which the subjects have to perform the same exercise numerous times, minimize the cognitive load in patients already suffering from cognitive and/or motor deficits could be a support to increase not only performance but the motivation itself for the activity. Another limitation of our study is that the target stimuli in the n-back task were few and infrequent. This probably did not stimulate the users' attention and concentration and could have affected some parameters related to the performance, as reported in other works (Young et al., 2015; Young & Stanton, 2002). However, the minimum number of stimuli to be offered to the participants is not indicated in the literature and there are currently no guidelines on the development of an n-back task with respect to various variables. Furthermore, it can be noted that, despite the small number of target stimuli, the parameters relating to reaction times and the number of errors changed as the difficulty of the tasks progressed thus indicating the effectiveness of the proposed stimulation protocol. In future studies, proposing a task with more numerous target stimuli will therefore be more appropriate.

5. Conclusions

This study suggests that physiological EEG measures can be a valuable approach to investigate the impact of different levels of immersive VR on the complex MWL construct since it directly takes into account the underlying brain activity. Moreover, in our experimental protocol, it emerges that the use of HMD may reduce the cognitive load in most of conditions. However, further research is needed to understand the relationship between immersion and MWL better and to develop more effective methods for measuring MWL in VR settings. Such information will be helpful in the future to make an informed design of novel VR-based applications in several fields, such as motor rehabilitation, cognitive training, and industrial training.

Acknowledgement(s)

The authors wish to thank Dr. Emanuele Gruppioni and Dr. Angelo Davalli for their useful discussions and suggestions.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Data availability statement

Raw EEG data used in this paper can be found at the following link: https://doi.org/10.17605/OSF.IO/C7M6Z

Funding

This work was supported by INAIL (Istituto Nazionale per l'Assicurazione contro gli Infortuni sul Lavoro) within the project PR19-SV-P1 "Rip@rto" - Simulatore di guida per assistere operatori nella valutazione delle capacità di guida dell'utente e nella scelta degli ausili di cui dotare l'automobile.

ORCID

Marta Mondellini http://orcid.org/0000-0002-9578-4586
Ileana Pirovano http://orcid.org/0000-0003-4905-3990
Vera Colombo http://orcid.org/0000-0001-6648-1917
Sara Arlati http://orcid.org/0000-0001-8092-1499
Marco Sacco http://orcid.org/0000-0002-5055-4031
Giovanna Rizzo http://orcid.org/0000-0002-6341-1304
Alfonso Mastropietro http://orcid.org/0000-0003-3358-7111

References

Aksu, Ş. H., Çak It, E., & Dagdeviren, M. (2023). Investigating the relationship between EEG features and n-back task difficulty levels with NASA-TLX scores among undergraduate students. *Intelligent Human Systems Integration (IHSI 2023): Integrating People and Intelligent Systems*, *69*(69). https://doi.org/10.54941/ahfe1002828

Albuquerque, I., Tiwari, A., Parent, M., Cassani, R., Gagnon, J.-F., Lafond, D., Tremblay, S., & Falk, T. H. (2020). Wauc: A multimodal database for mental workload assessment under physical activity. *Frontiers in Neuroscience*, *14*, 549524. https://doi.org/10.3389/fnins.2020.549524

Arana-De las Casas, N. I., Maldonado-Macias, A. A., De La Riva-Rodríguez, J., Sáenz-Zamarrón, D., Alatorre-Ávila, J. F., & García-Grajeda, E. (2021). Mental workload management and evaluation: A literature review for sustainable processes and organizations. In A. Realyvásquez Vargas, J. L. García-Alcaraz, E. Z-Flores (Eds.), *New perspectives on applied industrial ergonomics* (pp. 55–78). Springer. https://doi.org/10.1007/978-3-030-73468-8_3

Arlati, S., Colombo, V., Spoladore, D., Greci, L., Pedroli, E., Serino, S., Cipresso, P., Goulene, K., Stramba-Badiale, M., Riva, G., Gaggioli, A., Fserrigno, G., & Sacco, M. (2019). A social virtual reality-based application for the physical and cognitive training of the elderly at home. *Sensors*, *19*(2), 261. https://doi.org/10.3390/s19020261

Asgher, U., Khan, M. J., Asif Nizami, M. H., Khalil, K., Ahmad, R., Ayaz, Y., & Naseer, N. (2021). Motor training using mental workload (MWL) with an assistive soft exoskeleton system: A functional near-infrared spectroscopy (fNIRS) study for brain–machine interface (BMI). *Frontiers in Neurorobotics*, *15*, 605751. https://doi.org/10.3389/fnbot.2021.605751

Bell, A. J., & Sejnowski, T. J. (1995). An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, *7*(6), 1129–1159. https://doi.org/10.1162/neco.1995.7.6.1129

Bläsing, D., & Bornewasser, M. (2021). Influence of increasing task complexity and use of informational assistance systems on mental workload. *Brain Sciences*, *11*(1), 102. https://doi.org/10.3390/brainsci11010102

Borgheai, S. B., Deligani, R., McLinden, J., Abtahi, M., Ostadabbas, S., Mankodiya, K., & Shahriari, Y. (2019). Multimodal evaluation of mental workload using a hybrid EEG-fNIRS brain-computer interface system. In *2019 9th international IEEE/EMBS conference on neural engineering (NER)* (pp. 973–976). IEEE. https://doi.org/10.1109/NER.2019.8717118

Borghini, G., Astolfi, L., Vecchiato, G., Mattia, D., & Babiloni, F. (2014). Measuring neurophysiological signals in aircraft pilots and car drivers for the assessment of mental workload, fatigue and drowsiness. *Neuroscience and Biobehavioral Reviews*, *44*, 58–75. https://doi.org/10.1016/j.neubiorev.2012.10.003

Brookhuis, K. A., van Driel, C. J., Hof, T., van Arem, B., & Hoedemaeker, M. (2009). Driving with a congestion assistant; mental workload and acceptance. *Applied Ergonomics*, *40*(6), 1019–1025. https://doi.org/10.1016/j.apergo.2008.06.010

Chang, C.-Y., Hsu, S.-H., Pion-Tonachini, L., & Jung, T.-P. (2018). Evaluation of artifact subspace reconstruction for automatic EEG artifact removal. In *2018 40th annual international conference of the IEEE engineering in medicine and biology society (EMBC)* (pp. 1242–1245). IEEE. https://doi.org/10.1109/EMBC.2018.8512547

Chang, Y., Devine, K., & Klitzing, G. (2019). Can virtual reality enhance user performance and experience by reducing an individual's cognitive workload? In *ASEE EDGD 74th midyear conference proceedings*. ERIC.

Chao, C.-J., Wu, S.-Y., Yau, Y.-J., Feng, W.-Y., & Tseng, F.-Y. (2017). Effects of three-dimensional virtual reality and traditional training methods on mental workload and training performance. *Human Factors and Ergonomics in Manufacturing & Service Industries*, *27*(4), 187–196. https://doi.org/10.1002/hfm.20702

Charles, R. L., & Nixon, J. (2019). Measuring mental workload using physiological measures: A systematic review. *Applied Ergonomics*, *74*, 221–232. https://doi.org/10.1016/j.apergo.2018.08.028

Chikhi, S., Matton, N., & Blanchet, S. (2022). EEG power spectral measures of cognitive workload: A meta-analysis. *Psychophysiology*, *59*(6), e14009. https://doi.org/10.1111/psyp.14009

Colombo, V., Bocca, G., Mondellini, M., Sacco, M., & Aliverti, A. (2022). Evaluating the effects of virtual reality on perceived effort during cycling: Preliminary results on healthy young adults. In *2022 IEEE international symposium on medical measurements and applications (MEMEA)* (pp. 1–6). IEEE. https://doi.org/10.1109/MeMeA54994.2022.9856467

Colombo, V., Mondellini, M., Fumagalli, A., Aliverti, A., & Sacco, M. (2023). A virtual reality-based endurance training program for copd patients: Acceptability and user experience. *Disability and Rehabilitation. Assistive Technology*, 1–10. Advance online publication. https://doi.org/10.1080/17483107.2023.2219699

Delorme, A., & Makeig, S. (2004). Eeglab: An open source toolbox for analysis of single-trial eeg dynamics including independent component analysis. *Journal of Neuroscience Methods*, *134*(1), 9–21. https://doi.org/10.1016/j.jneumeth.2003.10.009

Dunn, T. J., Baguley, T., & Brunsden, V. (2014). From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. *British Journal of Psychology (London, England: 1953)*, *105*(3), 399–412. https://doi.org/10.1111/bjop.12046

Falkenstein, M., & Gajewski, P. D. (2021). Changes of electrical brain activity due to cognitive training in old adults and older industrial workers. In T. Strobach, J. Karbach (Eds.), *Cognitive training*. Springer. https://doi.org/10.1007/978-3-030-39292-5_25

Fan, J., Smith, A. P. (2017). The impact of workload and fatigue on performance. In L. Longo & M. Leva (Eds.), *Human mental workload: Models and applications. H-WORKLOAD 2017. Communications in Computer and Information Science* (Vol. 726). Springer. https://doi.org/10.1007/978-3-319-61061-0

Gabard-Durnam, L. J., Mendez Leal, A. S., Wilkinson, C. L., & Levin, A. R. (2018). The harvard automated processing pipeline for electroencephalography (HAPPE): Standardized processing software for developmental and high-artifact data. *Frontiers in Neuroscience*, *12*, 97. https://doi.org/10.3389/fnins.2018.00097

Gomez, D., Browne, J. D., Almalouhi, A., Abundex, M., Hu, J., Nason, S., Kull, N., Mills, C., Harris, Q., Ku, R., Neufeld, E. V., & Dolezal, B. A. (2022). Muscle activity and user-perceived exertion during immersive virtual reality exergaming incorporating an adaptive cable resistance system. *International Journal of Exercise Science*, *15*(7), 261–275. https://digitalcommons.wku.edu/ijes/vol15/iss7/2

Gorini, A., Capideville, C. S., De Leo, G., Mantovani, F., & Riva, G. (2011). The role of immersion and narrative in mediated presence: The virtual hospital experience. *Cyberpsychology, Behavior and Social Networking*, *14*(3), 99–105. https://doi.org/10.1089/cyber.2010.0100

Gradl, S., Wirth, M., Mächtlinger, N., Poguntke, R., Wonner, A., Rohleder, N., & Eskofier, B. M. (2019). The stroop room: A virtual reality-enhanced stroop test. In *Proceedings of the 25th acm symposium on virtual reality software and technology*. Association for Computing Machinery. Retrieved from https://doi.org/10.1145/3359996.3364247

Haavisto, M.-L., Porkka-Heiskanen, T., Hublin, C., Härmä, M., Mutanen, P., Müller, K., Virkkala, J., & Sallinen, M. (2010). Sleep restriction for the duration of a work week impairs multitasking performance. *Journal of Sleep Research*, *19*(3), 444–454. https://doi.org/10.1111/j.1365-2869.2010.00823.x

Hair, J., Black, W., Babin, B., & Anderson, R. (2010). *Multivariate data analysis: Pearson college division*. Pearson.

Hart, S. G. (2006). Nasa-task load index (NASA-TLX); 20 years later. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *50*(9), 904–908. https://doi.org/10.1177/154193120605000909

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (task load index): Results of empirical and theoretical research. In *Advances in psychology* (Vol. 52, pp. 139–183). Elsevier.

Herff, C., Heger, D., Fortmann, O., Hennrich, J., Putze, F., & Schultz, T. (2013). Mental workload during n-back task—quantified in the prefrontal cortex using fnirs. *Frontiers in Human Neuroscience*, *7*, 935. https://doi.org/10.3389/fnhum.2013.00935

Holm, A., Lukander, K., Korpela, J., Sallinen, M., & Müller, K. M. (2009). Estimating brain load from the EEG. *TheScientificWorldJournal*, *9*, 639–651. https://doi.org/10.1100/tsw.2009.83

IBM Corp (2021). *IBM SPSS statistics for windows*. IBM Corp. https://www.ibm.com/it-it/products/spss-statistics

Iman, R. L., & Conover, W. J. (1979). The use of the rank transform in regression. *Technometrics*, *21*(4), 499–509. https://doi.org/10.1080/00401706.1979.10489820

Jaccard, P. (1912). The distribution of the flora in the alpine zone. *New Phytologist*, *11*(2), 37–50. https://doi.org/10.1111/j.1469-8137.1912.tb05611.x

Jaeggi, S. M., Buschkuehl, M., Jonides, J., & Perrig, W. J. (2008). Improving fluid intelligence with training on working memory. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(19), 6829–6833. https://doi.org/10.1073/pnas.0801268105

Jaeggi, S. M., Seewer, R., Nirkko, A. C., Eckstein, D., Schroth, G., Groner, R., & Gutbrod, K. (2003). Does excessive memory load attenuate activation in the prefrontal cortex? Load-dependent processing in single and dual tasks: Functional magnetic resonance imaging study. *NeuroImage*, *19*(2 Pt 1), 210–225. https://doi.org/10.1016/s1053-8119(03)00098-3

Jaiswal, D., Chowdhury, A., Banerjee, T., & Chatterjee, D. (2019). Effect of mental workload on breathing pattern and heart rate for a working memory task: A pilot study. In *2019 41st annual international conference of the IEEE engineering in medicine and biology society (EMBC)* (pp. 2202–2206). IEEE. https://doi.org/10.1109/EMBC.2019.8856458

Jost, P., Cobb, S., & Hämmerle, I. (2020). Reality-based interaction affecting mental workload in virtual reality mental arithmetic training. *Behaviour & Information Technology*, *39*(10), 1062–1078. https://doi.org/10.1080/0144929X.2019.1641228

Kantowitz, B. H. (2000). Attention and mental workload. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *44*, 3–456. https://doi.org/10.1177/154193120004402121

Karbach, J., & Strobach, T. (2022). Multitasking training. In *Handbook of human multitasking* (pp. 305–345). Springer.

Khaksari, K., Condy, E., Millerhagen, J. B., Anderson, A. A., Dashtestani, H., & Gandjbakhche, A. H. (2019). Effects of performance and task duration on mental workload during working memory task. In *Photonics*, 6, 94. https://doi.org/10.3390/photonics6030094

Kothe, C., Medine, D., Boulay, C., Grivich, M., & Stenner, T. (2014). *Lab streaming layer*. https://github.com/sccn/labstreaminglayer.

Krokos, E., Plaisant, C., & Varshney, A. (2019). Virtual memory palaces: Immersion aids recall. *Virtual Reality*, *23*(1), 1–15. https://doi.org/10.1007/s10055-018-0346-3

Longo, L., Wickens, C. D., Hancock, G., & Hancock, P. A. (2022). Human mental workload: A survey and a novel inclusive definition. *Frontiers in Psychology*, *13*, 883321. https://doi.org/10.3389/fpsyg.2022.883321

Luong, T., Martin, N., Argelaguet, F., & Lécuyer, A. (2019). Studying the mental effort in virtual versus real environments. In *2019 IEEE conference on virtual reality and 3D user interfaces (VR)* (pp. 809–816). IEEE. https://doi.org/10.1109/VR.2019.8798029

Lysaght, R. J., Hill, S. G., Dick, A., Plamondon, B. D., & Linton, P. M. (1989). *Operator workload: Comprehensive review and evaluation of operator workload methodologies*. Defense Technical Information Center.

Mastropietro, A., Pirovano, I., Marciano, A., Porcelli, S., & Rizzo, G. (2023). Reliability of mental workload index assessed by eeg with different electrode configurations and signal pre-processing pipelines. *Sensors*, *23*(3), 1367. https://doi.org/10.3390/s23031367

McNeish, D. (2018). Thanks coefficient alpha, we'll take it from here. *Psychological Methods*, *23*(3), 412–433. https://doi.org/10.1037/met0000144

Midha, S., Wilson, M. L., & Sharples, S. (2022). Lived experiences of mental workload in everyday life. In *Proceedings of the 2022 chi conference on human factors in computing systems* (pp. 1–16). ACM. https://doi.org/10.1145/3491102.3517690

Mondellini, M., Arlati, S., Greci, L., Ferrigno, G., & Sacco, M. (2018). Sense of presence and cybersickness while cycling in virtual environments: Their contribution to subjective experience. In L. De Paolis, P. Bourdot (Eds.), *Augmented reality, virtual reality, and computer graphics. AVR 2018. Lecture Notes in Computer Science* (Vol. 10850). Springer. https://doi.org/10.1007/978-3-319-95270-3_1

Moray, N. (1982). Subjective mental workload. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *24*(1), 25–40. https://doi.org/10.1177/001872088202400104

Mrakic-Sposta, S., Di Santo, S. G., Franchini, F., Arlati, S., Zangiacomi, A., Greci, L., Moretti, S., Jesuthasan, N., Marzorati, M., Rizzo, G., Sacco, M., & Vezzoli, A. (2018). Effects of combined physical and cognitive virtual reality-based training on cognitive impairment and oxidative stress in mci patients: A pilot study. *Frontiers in Aging Neuroscience*, *10*, 282. https://doi.org/10.3389/fnagi.2018.00282

Mullen, T., Kothe, C., Chi, Y. M., Ojeda, A., Kerth, T., Makeig, S., … Jung, T.-P. (2013). Real-time modeling and 3D visualization of source dynamics and connectivity using wearable EEG. In *2013 35th annual international conference of the IEEE engineering in medicine and biology society (EMBC)* (pp. 2184–2187). IEEE. https://doi.org/10.1109/EMBC.2013.6609968

Navon, D., & Gopher, D. (2014). Task difficulty, resources, and dual-task performance. In *Attention and performance VIII* (pp. 297–315). Psychology Press.

Norman, D. A., & Bobrow, D. G. (1975). On data-limited and resource-limited processes. *Cognitive Psychology*, *7*(1), 44–64. https://doi.org/10.1016/0010-0285(75)90004-3

North, M. M., & North, S. M. (2016). A comparative study of sense of presence of traditional virtual reality and immersive environments. *Australasian Journal of Information Systems*, *20*. https://doi.org/10.3127/ajis.v20i0.1168

Paes, D., Irizarry, J., & Pujoni, D. (2021). An evidence of cognitive benefits from immersive design review: Comparing three-dimensional perception and presence between immersive and non-immersive virtual environments. *Automation in Construction*, *130*, 103849. https://doi.org/10.1016/j.autcon.2021.103849

Pallavicini, F., Pepe, A., & Minissi, M. E. (2019). Gaming in virtual reality: What changes in terms of usability, emotional response and sense of presence compared to non-immersive video games? *Simulation & Gaming*, *50*(2), 136–159. https://doi.org/10.1177/1046878119831420

Pedroli, E., Greci, L., Colombo, D., Serino, S., Cipresso, P., Arlati, S., Mondellini, M., Boilini, L., Giussani, V., Goulene, K., Agostoni, M., Sacco, M., Stramba-Badiale, M., Riva, G., & Gaggioli, A. (2018). Characteristics, usability, and users experience of a system combining cognitive and physical therapy in a virtual environment: Positive bike. *Sensors*, *18*(7), 2343. https://doi.org/10.3390/s18072343

Pion-Tonachini, L., Kreutz-Delgado, K., & Makeig, S. (2019). Iclabel: An automated electroencephalographic independent component classifier, dataset, and website. *NeuroImage*, *198*, 181–197. https://doi.org/10.1016/j.neuroimage.2019.05.026

R Core Team (2022). *R: A language and environment for statistical computing [Computer software manual]*. R Core Team. https://www.R-project.org/

Riese, H. (1999). Mental fatigue after very severe closed head injury: Sustained performance, mental effort, and distress at two levels of workload in a driving simulator. *Neuropsychological Rehabilitation*, *9*(2), 189–205. https://doi.org/10.1080/713755600

Schwind, V., Knierim, P., Haas, N., & Henze, N. (2019). Using presence questionnaires in virtual reality. In *Proceedings of the 2019 chi conference on human factors in computing systems* (pp. 1–12). ACM. https://doi.org/10.1145/3290605.3300590

Shahrbanian, S., Ma, X., Aghaei, N., Korner-Bitensky, N., Moshiri, K., & Simmonds, M. J. (2012). Use of virtual reality (immersive vs. non immersive) for pain management in children and adults: A systematic review of evidence from randomized controlled trials. *European Journal of Experimental Biology*, *2*(5), 1408–1422.

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika*, *74*(1), 107–120. https://doi.org/10.1007/s11336-008-9101-0

Slater, M., & Sanchez-Vives, M. V. (2016). Enhancing our lives with immersive virtual reality. *Frontiers in Robotics and AI*, *3*, 74. https://doi.org/10.3389/frobt.2016.00074

Smithers, L. G., Sawyer, A. C., Chittleborough, C. R., Davies, N. M., Davey Smith, G., & Lynch, J. W. (2018). A systematic review and meta-analysis of effects of early life non-cognitive skills on academic, psychosocial, cognitive and health outcomes. *Nature Human Behaviour*, *2*(11), 867–880. https://doi.org/10.1038/s41562-018-0461-x

Taha, A. A., & Hanbury, A. (2015). Metrics for evaluating 3d medical image segmentation: Analysis, selection, and tool. *BMC Medical Imaging*, *15*(1), 29. https://doi.org/10.1186/s12880-015-0068-x

Tao, D., Tan, H., Wang, H., Zhang, X., Qu, X., & Zhang, T. (2019). A systematic review of physiological measures of mental workload. *International Journal of Environmental Research and Public Health*, *16*(15), 2716. https://doi.org/10.3390/ijerph16152716

Tremmel, C., Herff, C., Sato, T., Rechowicz, K., Yamani, Y., & Krusienski, D. J. (2019). Estimating cognitive workload in an interactive virtual reality environment using EEG. *Frontiers in Human Neuroscience*, *13*, 401. https://doi.org/10.3389/fnhum.2019.00401

Tucker-Drob, E. M., Brandmaier, A. M., & Lindenberger, U. (2019). Coupled cognitive changes in adulthood: A meta-analysis. *Psychological Bulletin*, *145*(3), 273–301. https://doi.org/10.1037/bul0000179

Verwey, W. B. (2000). On-line driver workload estimation. Effects of road situation and age on secondary task measures. *Ergonomics*, *43*(2), 187–209. https://doi.org/10.1080/001401300184558

Weech, S., Kenny, S., & Barnett-Cowan, M. (2019). Presence and cybersickness in virtual reality are negatively related: A review. *Frontiers in Psychology*, *10*, 158. https://doi.org/10.3389/fpsyg.2019.00158

Wickens, C. D. (2017). Mental workload: Assessment, prediction and consequences. In L. Longo & M. Leva (Eds.), *Human mental workload: Models and applications. H-WORKLOAD 2017. Communications in Computer and Information Science* (Vol. 726). Springer. https://doi.org/10.1007/978-3-319-61061-0_2

Wu, B., Hu, Y., & Wang, M. (2020). How do head-mounted displays and planning strategy influence problem-solving-based learning in introductory electrical circuit design? *Educational Technology & Society*, *23*(3), 40–52. https://www.jstor.org/stable/26926425

Xi, N., Chen, J., Gama, F., Riar, M., & Hamari, J. (2023). The challenges of entering the metaverse: An experiment on the effect of extended reality on workload. *Information Systems Frontiers*, *25*, 659–680. https://doi.org/10.1007/s10796-022-10244-x

Young, M. S., Brookhuis, K. A., Wickens, C. D., & Hancock, P. A. (2015). State of science: Mental workload in ergonomics. *Ergonomics*, *58*(1), 1–17. https://doi.org/10.1080/00140139.2014.956151

Young, M. S., & Stanton, N. A. (2002). Attention and automation: New perspectives on mental underload and performance. *Theoretical Issues in Ergonomics Science*, *3*(2), 178–194. https://doi.org/10.1080/14639220210123789

About the authors

Marta Mondellini graduated in psychology in 2014 and is a research fellow at CNR-STIIMA since 2017. She is currently attending the Ph.D. course in Psychology. Her work focuses on the Human Factor in interaction with technologies for rehabilitation and training, and on motivation and emotions in particular.

Ileana Pirovano is a research fellow at the Institute for Biomedical Technologies of Italian CNR since November 2020. She obtained bachelor's and master's degrees in biomedical engineering and a PhD in Physics at Politecnico di Milano in 2020. She works on EEG signals acquisition, processing, and methods' development.

Vera Colombo received the M.Sc. and Ph.D. in Bioengineering from the Politecnico di Milano, in 2015 and 2022, respectively. Since 2016, she is a research fellow at CNR-STIIMA. Her research activities focus on the design, development and validation of eXtended Reality applications for rehabilitation and training.

Sara Arlati received her MSc (2013) and her Ph.D. (2020) degrees in Bioengineering from Politecnico di Milano. She is currently a full-time researcher at CNR-STIIMA. Her research deals with the use of eXtended Reality (XR) for health; she mainly designs, develops and validates applications for neuromotor rehabilitation and cognitive training.

Marco Sacco is the Head of CNR-STIIMA in Lecco. Vice-President of EuroXR with proxy in EU and collaboration issues. His experiences (since 1994) are in XR technologies applied to both product/process/factory life cycle as well as to health and wellbeing sectors. Since March 2021 member of the CNR Center of competence I4.0 for digital technologies.

Giovanna Rizzo is an electronic engineer and a senior researcher at ITB-CNR in Italy. Her main research interests are in the field of biomedical signals and image processing and analysis, mainly in oncology and neurology. She has published more than 100 articles and contributed to numerous national and international research projects.

Alfonso Mastropietro is a bioengineer and a researcher at the Institute of Biomedical Technologies of the National Research Council in Italy. He holds a PhD in bioengineering from the Politecnico di Milano. His research interests are in the field of biomedical signals and image processing and analysis, mainly in neurological applications.