

## Article

# Wearable Technologies for Mental Workload, Stress, and Emotional State Assessment during Working-Like Tasks: A Comparison with Laboratory Technologies

Andrea Giorgi <sup>1,\*</sup>, Vincenzo Ronca <sup>1,2</sup>, Alessia Vozzi <sup>1,2</sup>, Nicolina Sciaraffa <sup>1,3</sup>, Antonello di Florio <sup>1</sup>, Luca Tamborra <sup>2,4</sup>, Ilaria Simonetti <sup>2,4</sup>, Pietro Aricò <sup>1,5,6</sup>, Gianluca Di Flumeri <sup>1,5,6</sup>, Dario Rossi <sup>1,6</sup> and Gianluca Borghini <sup>1,5,6</sup>

<sup>1</sup> BrainSigns, SRL, 00185 Rome, Italy; vincenzo.ronca@uniroma1.it (V.R.); alessia.vozzi@uniroma1.it (A.V.); nicolina.sciaraffa@uniroma1.it (N.S.); antonello.diflorio@brainsigns.com (A.d.F.); pietro.arico@uniroma1.it (P.A.); gianluca.diflumeri@uniroma1.it (G.D.F.); dario.rossi@brainsigns.com (D.R.); gianluca.borghini@uniroma1.it (G.B.)

<sup>2</sup> Department of Anatomical, Histological, Forensic and Orthopaedic Sciences, Sapienza University, 00185 Rome, Italy; luca.tamborra@it.ey.com (L.T.); ilaria.simonetti@it.ey.com (I.S.)

<sup>3</sup> Department of Molecular Medicine, Sapienza University of Rome, 00185 Rome, Italy

<sup>4</sup> Ernst & Young, Department People Advisory Services, 00187 Rome, Italy

<sup>5</sup> IRCCS Fondazione Santa Lucia, 00179 Rome, Italy

<sup>6</sup> Department of Business and Management, LUISS University, 00197 Rome, Italy

\* Correspondence: andrea.giorgi@brainsigns.com



**Citation:** Giorgi, A.; Ronca, V.; Vozzi, A.; Sciaraffa, N.; di Florio, A.; Tamborra, L.; Simonetti, I.; Aricò, P.; Di Flumeri, G.; Rossi, D.; et al. Wearable Technologies for Mental Workload, Stress, and Emotional State Assessment during Working-Like Tasks: A Comparison with Laboratory Technologies. *Sensors* **2021**, *21*, 2332. <https://doi.org/10.3390/s21072332>

Academic Editor: Alan Jović

Received: 24 February 2021

Accepted: 24 March 2021

Published: 26 March 2021

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Abstract:** The capability of monitoring user's performance represents a crucial aspect to improve safety and efficiency of several human-related activities. Human errors are indeed among the major causes of work-related accidents. Assessing human factors (HFs) could prevent these accidents through specific neurophysiological signals' evaluation but laboratory sensors require highly-specialized operators and imply a certain grade of invasiveness which could negatively interfere with the worker's activity. On the contrary, consumer wearables are characterized by their ease of use and their comfortability, other than being cheaper compared to laboratory technologies. Therefore, wearable sensors could represent an ideal substitute for laboratory technologies for a real-time assessment of human performances in ecological settings. The present study aimed at assessing the reliability and capability of consumer wearable devices (i.e., Empatica E4 and Muse 2) in discriminating specific mental states compared to laboratory equipment. The electrooculographic (EOG), electrodermal activity (EDA) and photoplethysmographic (PPG) signals were acquired from a group of 17 volunteers who took part to the experimental protocol in which different working scenarios were simulated to induce different levels of mental workload, stress, and emotional state. The results demonstrated that the parameters computed by the consumer wearable and laboratory sensors were positively and significantly correlated and exhibited the same evidences in terms of mental states discrimination.

**Keywords:** wearable device; emotional state; mental workload; stress; heart rate; eye blinks rate; skin conductance level

## 1. Introduction

This paper aims to investigate the capability of two consumer wearable devices (i.e., Empatica 4 and Muse 2) in assessing different levels of mental and emotional states. The consumer devices were compared to laboratory ones (i.e., BeMicro and Shimmer) in order to validate their reliability in scientific research.

### 1.1. Monitoring Mental States

In recent years there was an increasing interest toward wearable monitoring devices to assess physiological and mental activity, both in research and industry [1,2].



**Copyright:** © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

These devices are particularly important to the world's increasingly aging population since this aspect constitutes a relevant risk factor for work-related accidents [3]. Both in research and industry domains the mental states' monitoring is becoming really important. Starting from few decades ago, there was a shift in the focus from operators' physical demands to their cognitive demands. This shift is particularly evident for some complex and safety-critical human activities such as air traffic control, and car and rail train driving [4–7]. In these contexts, it is evident that most of the fatal and non-fatal accidents occur because of Human Factors (HFs) concerns [8–11]. Among all the HFs, stress, mental overload, and lack of vigilance could cause tragic human errors in several working environments [12–14]. Giving the limitations imposed by subjective evaluation of mental states [15–17] and due to the fact that in some specific activities it is not possible to interrupt operators while working, researchers started to acquire biosignals to monitor and assess operators' mental states. Biomarkers such as skin conductance level (SCL), heart rate (HR), and eye blink rate (EBR) are investigated as correlates of users' mental states to develop a monitoring system to diminish and prevent fatal and non-fatal accidents [4,6,16,18–20]. For this reason, it is important to reduce at minimum the invasiveness of the monitoring equipment. Furthermore, the interest in consumer wearable devices was supported by the increasing advances in microelectronics which allowed to overcome the limitations imposed by the size of the electronic components and of the measuring sensor itself [21]. The size reduction, other than costs reduction and easiness to use, enhanced the application of such wearable devices to areas of research which were usually investigated using laboratory technologies, considered in scientific literature as the gold-standard [22–24]. Indeed, despite the improvements of the technology behind laboratory equipment such devices are often uncomfortable and obtrusive for the participants leading to a non-optimal condition to ecologically assess mental states [25].

### 1.2. Consumer Wearables in Scientific Research

Consumer wearable devices are ideal candidates to record operators' biosignals without negatively interfere with their activities and tasks. Given the emergence of an incredible amount of commercial and user-friendly wearable devices [23,24] and given the fact that they seem to better adapt to daily-life activities, their accuracy has to be investigated deeply. The reliability of wearable devices in measuring biomarkers such as HR and SCL was demonstrated. In fact, compared to gold-standard equipment, consumer wearable devices showed a similar accuracy in measuring different biomarkers such as HR, HRV and SCL in different conditions [26–28]. Ragot and colleagues successfully adopted the Empatica E4 wrist-band to measure physiological response in an emotion recognition task [29]. Based on these evidence wearable devices were also used to assess different mental states. Setz and colleagues [30] compared the reliability of a consumer wearable device (Empatica E4) in detecting drowsiness during a driving simulation task using HRV. The authors found that E4 wristband showed similar results compared to a medical-grade device and argued that the latter device could be substituted with the E4 in order to detect drowsiness. The possibility to discriminate between different levels of the same mental states was also explored. A study on simulated train traffic controlling [25] demonstrated that it is possible to differentiate between different level of mental workload (WL) using HRV acquired via wearable device. Compared to an FDA-approved medical device authors showed that a consumer wearable sensor (EmWave Pro, Boulder Creek, California, USA) had similar results in estimating changes in HRV, while the Empatica E3, a different consumer wearable device included in the same study, did not show the same reliability. The potentiality of consumer wearable devices in acquiring biosignals in an unobtrusive way brought to the development of devices to collect electroencephalographic (EEG) and electrooculographic (EOG) signals. Krigolson and colleagues [31] validated Muse 2 wearable device for ERP research demonstrating an adequate level of accuracy in measuring N200 and P300 components compared to standard 10–20 electrode configuration. Other researchers investigated

the possibility to use Muse 2 to discriminate between different levels of enjoyment while playing videogames [32]. The authors reported no significant difference in cortical activity while subjective reports did but the absence of a gold standard reference did not allow to objectively assess the accuracy of the consumer wearable EEG device considered.

### 1.3. Aim of the Present Study

Summarizing, there are contrasting evidence in literature about the reliability of consumer wearable devices. The possibility of these devices to estimate different biosignals is well accepted [26–29,32]. Additionally, some authors successfully differentiated between several mental states using the neurometrics collected with consumer wearables devices [25,30] but in other cases a failure was reported [25,31]. This paper fits into this contest by comparing the Empatica E4 and Muse 2 with laboratory equipment. The reliability and capability of the two consumer wearable devices were investigated for stress, mental workload (WL), and emotional state (EmS) evaluation while participants were performing three working-like tasks, comparing them with laboratory equipment. To summarize, this paper aimed at responding to the following research questions (RQ):

- RQ1: Are the above-mentioned neurophysiological parameters (EBR, SCL and HR) gathered through consumer wearable devices comparable with those acquired with laboratory equipment?
- RQ2: Are consumer wearable devices reliable in discriminating different levels of the mental states considered (WL, Stress and EmS)?

## 2. Materials and Methods

### 2.1. Participants

Seventeen (17) participants were recruited from the Sapienza University of Rome (ten males and seven females,  $31.1 \pm 3.7$  years old) with normal or corrected-to-normal vision. Due to artifacts and missing data caused by technical issues after signals processing twelve (12) participants were considered valid for the analysis. Informed consent was obtained from each participant after explanation of the study. The experiment was conducted following the principles outlined in the Declaration of Helsinki of 1975, as revised in 2000 and was approved by the Sapienza University of Rome Ethical Committee in Charge for the Department of Molecular Medicine (protocol number: 2507/2020, approved on 4 August 2020). To respect the privacy of participants, only aggregate results were reported.

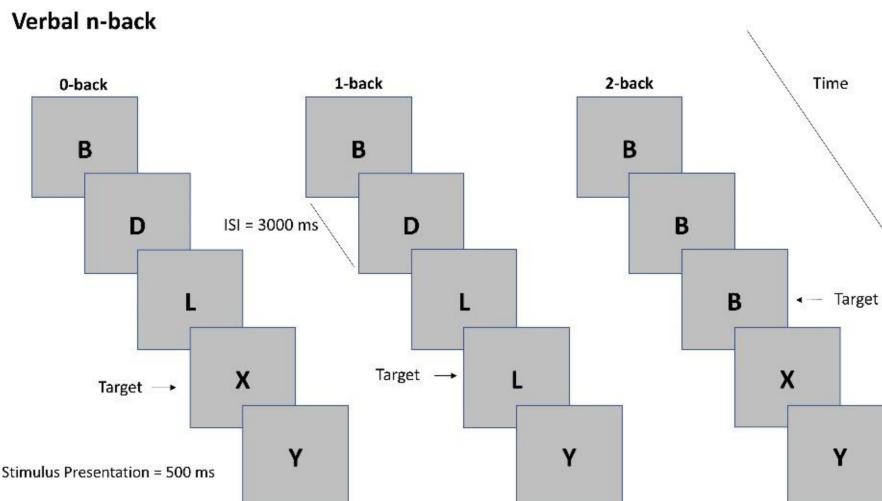
### 2.2. Procedures

In order to test the reliability of consumer wearable devices in WL, stress, and EmS evaluation, an experimental protocol was designed including three tasks: N-back task, Doctor Game task, and Webcall task. These tasks were selected to respectively simulate an office-like environment, an assembly-line and a teleworking activity. N-back task was used to simulate an office related activity which usually do not demand a pronounced physical effort whilst keeping high the mental one. Doctor Game (i.e., “Operation”) represents a fine motor skill task requiring participant to use a pair of tweezers to extract several items from their slots. This task was adopted because of its analogy with the assembly line activities. Finally, Webcall task was used to reproduce a teleworking case, in which people are often asked to communicate and coordinate with someone who is not physically present. The order of tasks completion was balanced and randomized among participants.

#### 2.2.1. N-Back Task

The N-back task (NB) (Figure 1) is a robust psychological test to manipulate working memory load [33], one of the major components and a reasonable approximation of WL [34]. Participants are presented with a sequence of letters on a screen. The goal is to press a button when the letters appearing on the screen is the same that occurred in the series  $n$  steps before. The difficulty of the task can be manipulated increasing the value of  $n$ , thus

forcing participants to retain more items in their mind. In this study, the task was composed of a baseline and three conditions: Low WL, high WL, and stress. Under all conditions, 21 uppercase letters were used, which were displayed for 500 ms and an inter-stimulus interval randomized between 500 to 3000 ms; 33% of the displayed letters were targets.



**Figure 1.** Example of N-back task under the 0-back, 1-back, and 2-back conditions.

- Baseline: Participants were instructed to watch the sequence of letters without giving any response.
- Low WL: 0-back. The task consisted in indicating when the stimulus on the screen matches a predetermined letter.
- High WL: 2-back. The task consisted in indicating when the stimulus occurred in the series 2 steps before. When investigating stress assessment, we referred to this condition as 'No Stress' condition (i.e., in the comparison 'No Stress vs. Stress') as it differed from the Stress one only in the presence of the stressors whilst the difficulty level was the same.
- Stress: The task was practically equivalent to the High WL one (indicating when the stimulus occurred in the series two steps before) but simultaneously high intensity noise was played (85 db) and the white-coat effect was used to stress the participant [35]. Four-minute relaxing music and video was played at the end of this phase for letting the participants recover from the stressful event before continuing with the remaining experimental conditions [36].

In all conditions, behavioral data like reaction time and number of errors were collected. The low WL and the high WL conditions were performed randomly while the baseline and the stress conditions were performed respectively at the beginning and at the end of the experimental task. Before the 0-back and the 2-back task, the participant performed a training session containing 21 stimuli, 33% of which were targets.

## 2.2.2. Doctor Game Task

This task is a fine motor skill task. We adopted the "Doctor Game" (DG) (i.e., "Operation") board game (Figure 2). Its goal consisted in removing small objects from the board without touching the metal edges. In this task a baseline, two difficulty levels and one stressful condition were performed as well.



**Figure 2.** The Doctor Game task consisted in extracting as many objects as possible from the “patient” without touching the metal border. If an error occurred, the nose will emit a red light and the board will vibrate.

- Baseline: Participants were instructed to watch the board game without touching the board itself nor the objects.
- Low WL: Participants were asked to remove five predefined objects (the easiest ones). They had three minutes to complete the task.
- High WL: Participants were asked to remove all 12 objects. They had three minutes to complete the task. When investigating stress assessment, we referred to this condition as ‘No Stress’ condition (i.e., in the comparison ‘No Stress vs. Stress’) as it differed from the Stress one only in the presence of the stressors whilst the difficulty level was the same.
- Stress: Participants were asked to remove all 12 objects. They had one minute to complete the task. Additionally, high intensity noise was played (85 db) and the white-coat effect was used to stress the participant [35]. Then, a four-minute relaxing music and video was played at the end of this phase. This was done to let participants recover from the stressful event before continuing with the experiment.

In all conditions, behavioral data like number of objects removed and accomplishment time were collected. The Low WL and the High WL conditions were performed randomly while the baseline and the stress conditions were performed respectively at the beginning and at the end of the experimental task. Before the baseline the participant performed a training session by extracting a couple of objects from the board.

### 2.2.3. Webcall Task

This task consisted in an interactive Webcall to simulate a teleconference in smart-working condition. This task comprised a baseline, a positive, and a negative condition of two minutes each. The positive and negative conditions were achieved by asking the participant to respectively recall the happiest and the saddest memory of their past, while during the baseline condition the participant was asked to watch the teleconference platform interface without reacting. The positive condition was always performed at the beginning to avoid transients due to strong negative memories. One experimenter was sitting in another room interacting with the participant. The hypothesis was that asking the participant to talk about saddest/happiest memories will naturally induce these emotions and thereby enable them to feel and display the relevant expressions of emotions via multiple modalities, including physiological reactions [37,38].

### 2.3. Performance Assessment

Participants' performance was assessed for NB and DG tasks. Webcall task did not imply a right or wrong response therefore no performance was computed. Performance in NB was assessed using the Inverse Efficiency Score (IES) [34] computed as reported in Equation (1):

$$\text{IES} = \frac{\text{RT}}{1 - \text{PE}} \quad (1)$$

where RT is the participant's average (correct) reaction time within the condition considered, and PE is the participant's proportion of errors in the same condition. IES can be considered as the RT corrected for the amount of errors committed [34].

For the DG task we combined the number of errors, number of extracted objects, and the time spent to complete the task, in order to have an overall value representing the performance. Since no standard Performance Index (PI) are reported in the literature, we proposed the following one:

$$PI = \frac{\frac{\text{OBJ}}{\text{OBJ}_{\max}} + \left(1 - \frac{\text{ERR}}{\text{TIME}}\right)}{2} \quad (2)$$

where OBJ is the number of extracted objects,  $\text{OBJ}_{\max}$  is the total number of objects in the condition (5 in the low WL condition and 12 in the high WL and stress ones), ERR is the maximum number of errors a participant could make in the condition (one error per second, 180 in Low WL and High WL conditions and 60 in Stress one) and TIME is the time the participant spent to complete the task in the condition.

### 2.4. Subjective Reports

After each experimental condition, including the baseline, two questionnaires were administered to the participants:

- NASA Task-Load Index (NASA-TLX): It consists of six sub-scales representing independent groups of variables: mental, physical and temporal demands, frustration, effort and performance [39]. The participants were initially asked to rate on a scale from "low" to "high" (from 0 to 100) each of the six dimensions during the task. Afterwards, they had to choose the most important factor along pairwise comparisons. The NASA-TLX was selected for subjectively quantify the mental demand perceived by the participants with respect to the experimental condition of DG and NB tasks.
- GENEVA Emotion Wheel (GEW): It is a validated instrument to measure emotional reactions to several stimuli [40]. The participants were asked to indicate the emotion he/she experienced by choosing intensities for a single emotion or a blend of several emotions out of 20 distinct emotion families. Given the nature of the task, in this analysis we decided to use only the type of emotions selected by participants, without considering their intensities.

The reason why we selected these questionnaires is because they have been adopted in several studies. In particular, the NASA-TLX has been used for WL [41,42] subjective reports and GENEVA has been used for emotion categorization [40,43]. For the stress self-report, we utilized only the temporal demand and frustration parameters because they are the main components of the stressor used in this study.

## 2.5. EOG Recording and Analysis for Mental Workload Assessment

The vertical EOG pattern was estimated by acquiring simultaneously the EEG Fpz channel of the BeMicro (EB Neuro, Florence, Italy) and the EEG TP9 channel of the Muse 2 (Interaxon Inc, Toronto, OH, USA), with a sampling frequency of 256 Hz and 64 Hz respectively. Details are summarized in Table 1. The aim of the EOG analysis was to detect the eye blinks in order to estimate the eye blink rate (EBR) and finally correlate it with the WL variations) [7,44]. The same algorithm was adopted for the analysis of both datasets. Firstly, the EOG signal was band-pass filtered using a 5th-order Butterworth filter within the frequency range of 2–10 Hz, since in this range the main frequency contribute of eye blinks is contained [45,46].

**Table 1.** A summary of the devices and signals used in the presented work.

Signal	Laboratory Device	Consumer Wearable Device	Extracted Feature	Filter Frequency Range	Time Window
EOG	BeMicro	Muse 2	EBR	2–10 Hz	-
EDA	Shimmer	Empatica 4	SCL	1 Hz	60 s
PPG	-	Empatica 4	HR	1–4 Hz	60 s
ECG	BeMicro	-	HR	1–15 Hz	60 s

Secondly, the eyes open condition was used to identify a threshold for each participant that, when exceeded, identified a potential blink. The threshold was calculated as follows:

$$\text{Threshold} = \text{mean}(EOG \text{ Eyes Open}) + 3 * \text{robustStdDev} \quad (3)$$

where robustStdDev is the mean absolute deviation of the corresponding EOG channel.

Finally, every time the EOG signal exceeded the computed threshold, the Pearson correlation between a common blink template (the template was built averaging the blinks estimated from five random participants during the eyes open condition) and the EOG signal was computed within each experimental condition (i.e., pattern-matching phase). If this value was higher than 0.9, a potential blink would be classified as “real blink”, similarly to what performed by the BLINKER algorithm [47].

The EBR estimated for each participant in each condition were calculated as the total number of blinks in every condition divided by the condition duration. EBR was evaluated under the different WL conditions to assess if it could differentiate user’s mental workload. Previous studies demonstrated the capability of this parameter in estimating WL demand [16,44,48].

## 2.6. EDA Recording and Analysis for Stress Assessment

The EDA was recorded by both laboratory and consumer wearable devices. The sampling frequency of the Shimmer3 GSR+ unit (Shimmer Sensing, Dublin, Ireland) laboratory device was 64 Hz while the sampling frequency of the Empatica E4 was 4 Hz. Shimmer sensors were placed on the participant’s no-dominant hand on the second and third fingers. In Empatica E4 the two electrodes are placed on the bottom part of the wrist. The EDA was firstly low-pass filtered with a cut-off frequency of 1 Hz and then processed by using the Ledalab suite [49], a specific open-source toolbox implemented within the MATLAB (MathWorks, Natik, Massachusetts) environment for EDA processing (details in Table 1). The continuous decomposition analysis [50] was applied in order to estimate

the tonic (SCL) and the phasic (SCR) components [51]. The SCL is the slow-changing component of the EDA signal, mostly related to the global arousal of the participant. On the contrary, the SCR is the fast-changing component of the EDA signal usually related to single stimuli reactions. The EDA components, as well as the other neurophysiological parameters, were estimated both using a 60 s time resolution and averaging within each experimental condition. Finally, only the SCL was analyzed accordingly with the objectives of the present study as demonstrated by Borghini et al. [7]. This parameter was chosen for stress estimation since previous studies demonstrated its relation with this mental state [7,52].

### 2.7. ECG Signal Recording and Analysis for Emotional State Assessment

Additionally, the HR estimation was performed using laboratory and consumer wearable technologies. ECG signal was collected using an electrode fixed on the participant's chest (laboratory device BeMicro) and referred to the potential recorded at both the earlobes with a sampling frequency of 256 Hz. At the same time, photoplethysmographic signal (PPG) was collected by means of Empatica E4 (Empatica, Milan, Italy). First, the ECG and PPG signal were filtered using a 5th-order Butterworth band-pass filter (1–1 Hz, and 1–4 Hz, respectively) in order to reject the continuous component and the high-frequency interferences, such as that related to the mains power source (details in Table 1). Another purpose of this filtering was to emphasize the QRS process of the ECG signal [53]. The following step consisted in computing the ECG (PPG) signal to the power of 3 to emphasize the heartbeat peaks, as they generally have the highest amplitude, and at the same time reduce spurious artifact peaks. Finally, the distance between consecutive peaks (i.e., each R peak corresponds to a heartbeat) was measured to estimate the HR values every 60 s. The Pan-Tompkins algorithm [54] was used for the HR estimation. A combination of HR and SCL measurements was adopted in order to estimate EmS [55,56]. In this regard, an Emotional Index (EI) was defined as:

$$EI = |SCL| * HR \quad (4)$$

where SCL and HR were normalized by subtracting the corresponding baseline and dividing by the corresponding standard deviation. The resulting values were then averaged within the considered experimental condition. The combination of these two parameters was adopted because the sensitivity of this emotional index was already described in previous works [56].

### 2.8. Statistical Analysis

Statistical analyses were performed after normalizing each data condition with the corresponding task Baseline. For each participant, EBR, SCL, and HR data collected during baseline were subtracted from data collected during experimental conditions. The new EBR, SCL and HR values were named respectively EBR', SCL' and HR'. The Shapiro–Wilk test was used to assess the normality of the distribution related to each of the considered parameters. If normality was confirmed, Student's t-test would have been performed to pairwise compare the conditions (e.g., 'Low WL vs. High WL', or 'laboratory device vs. wearable device'). In case of non-normal distribution, the Wilcoxon signed-rank test was performed. In case of comparisons between three or more distributions, the analysis of variance (ANOVA) or its non-parametric equivalent (Friedman ANOVA) was performed. For all tests, statistical significance was set at  $\alpha = 0.05$ .

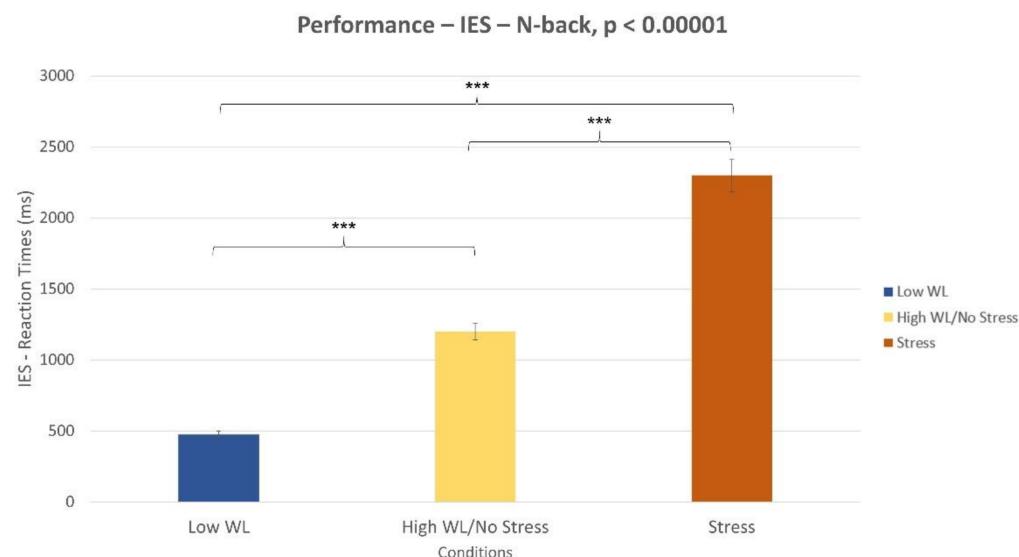
Pearson's repeated measure correlation (rmcorr) analysis [57] was then used to assess the reliability of the parameters estimated by the wearable device with respect to the laboratory one both at single- participant level and on the entire group. The rmcorr was performed on the average values of each parameter of wearable and laboratory devices gathered during the entire experimental session.

### 3. Results

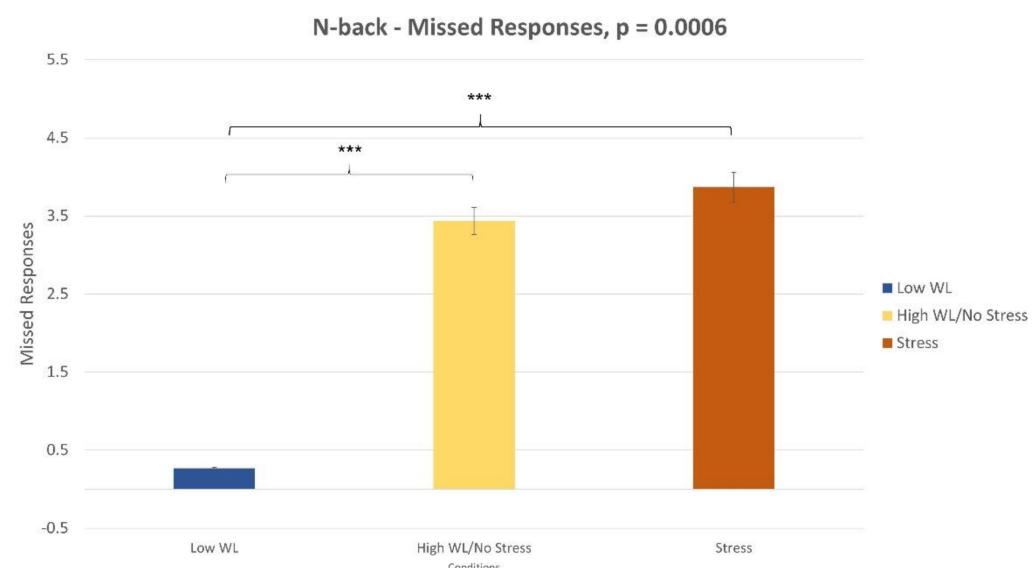
#### 3.1. Performance

##### 3.1.1. N-back task

The Wilcoxon signed-rank test on the IES (Figure 3) revealed a significant difference between the low WL and high WL conditions ( $p < 0.001$ ) and between the “no stress” (i.e., high WL) and stress conditions ( $p < 0.001$ ). Furthermore, the three parameters involved in the IES computation (i.e., reaction times, wrong response, missed response) were analyzed to determine the one was most affecting the decreasing performance while executing the task. The Wilcoxon signed-rank test showed that both in high WL and Stress conditions (Figure 4) the number of missed responses increased significantly compared to the low WL condition ( $p < 0.001$ ).



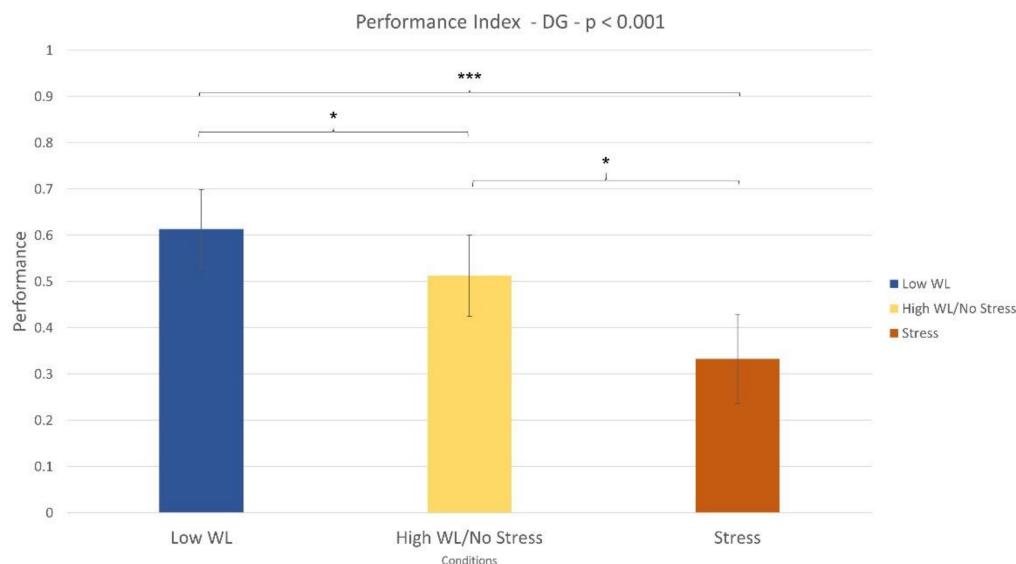
**Figure 3.** Difference in subjective performance during N-back task. Low vs. high Workload (WL) conditions ( $p < 0.001$ ). No stress vs. stress conditions ( $p < 0.001$ ).



**Figure 4.** The number of missed responses was higher in high WL and stress conditions compared to the low WL condition ( $p < 0.001$ ).

### 3.1.2. Doctor Game Task

The Wilcoxon signed-rank test revealed that the performance index significantly decreased ( $p = 0.03$ ) during the high WL condition compared to the low WL one (Figure 5). The same was observed during the Stress condition when compared with the no stress one ( $p = 0.02$ ).



**Figure 5.** The performance index significantly decreased during the high WL condition compared to Low WL condition ( $p = 0.03$ ). The same result was found in the stress vs. no stress comparison ( $p = 0.001$ ).

## 3.2. Subjective Reports

### 3.2.1. N-back task

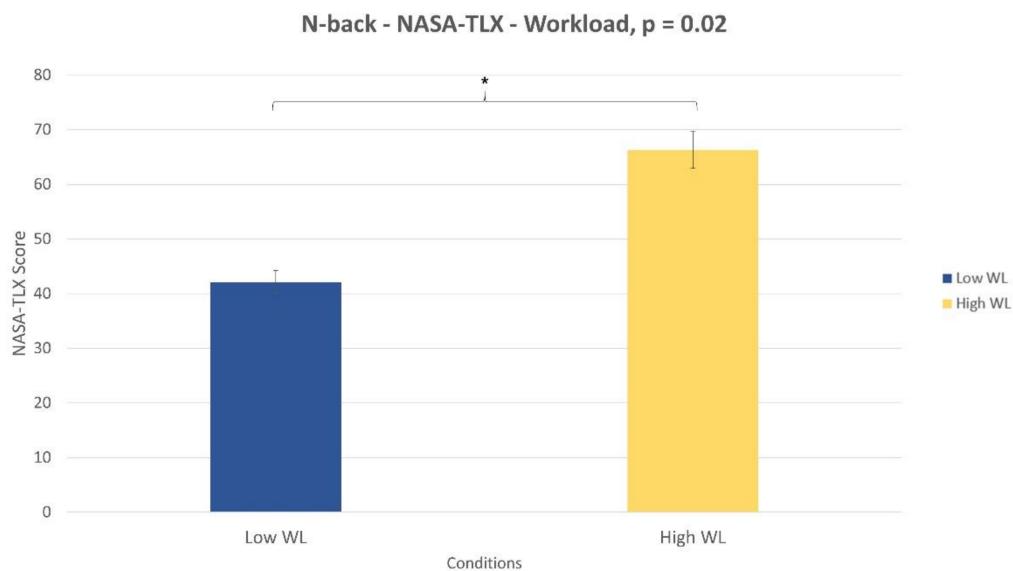
The Wilcoxon signed-rank test performed on the NASA-TLX demonstrated that participants perceived the High WL condition significantly more demanding ( $p = 0.02$ ) than Low WL one (Figure 6). Additionally, at the end of the experiments they reported that the High WL condition resulted too difficult to be performed and that for this reason they did not or could not attend the task properly. Regarding the subjective stress evaluation, the combination of frustration and temporal demand parameters of the NASA-TLX was considered. These two parameters were selected accordingly with the relevant audio noise and the white-coat effect induced within the stress condition. The statistical analysis showed no significant difference ( $p = 0.4$ ) in terms of perceived stress between no-stress and stress conditions.

### 3.2.2. Doctor Game Task

Looking at NASA-TLX total score, participants did not perceive the High WL condition to be significantly harder than Low WL condition ( $p = 0.9$ ). Additionally, in this task we considered the frustration and temporal demand parameters of the NASA-TLX to assess the perceived stress, and no significant difference was found between the no-stress and stress conditions ( $p = 0.8$ ).

### 3.2.3. Webcall Task

As showed in Table 2, during the positive condition participants rated mostly positive emotions than the negative ones. Instead, during negative conditions the rated emotions were mostly negative. However, some participants selected negative emotions during the positive calls while others positive emotions during the negative one.



**Figure 6.** NASA-TLX total score during the low WL and high WL conditions ( $p = 0.02$ ).

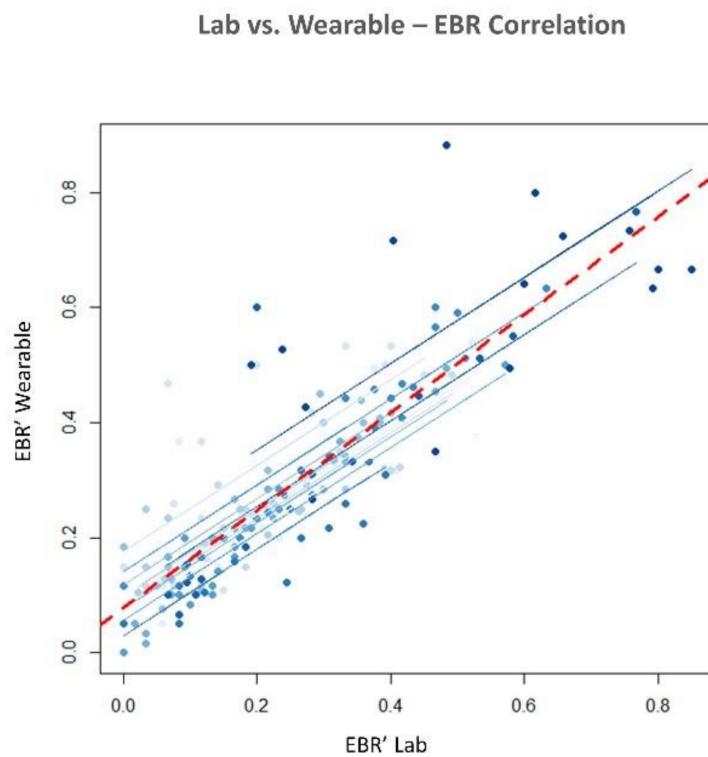
**Table 2.** Frequency of the emotions selected after positive and negative conditions of the Webcall.

Emotions (Geneva Emotion Wheel)	Positive Webcall	Negative Webcall
Admiration	1	
Contentment	1	1
Joy	12	
Love	3	2
Pleasure	6	
Pride	3	1
Relief		1
Interest	6	2
Embarrassment	1	
Compassion		1
Anger	1	2
Disappointment		4
Disgust		1
Fear		3
Guilt		3
Regret	1	1
Sadness	2	11
Shame	1	3

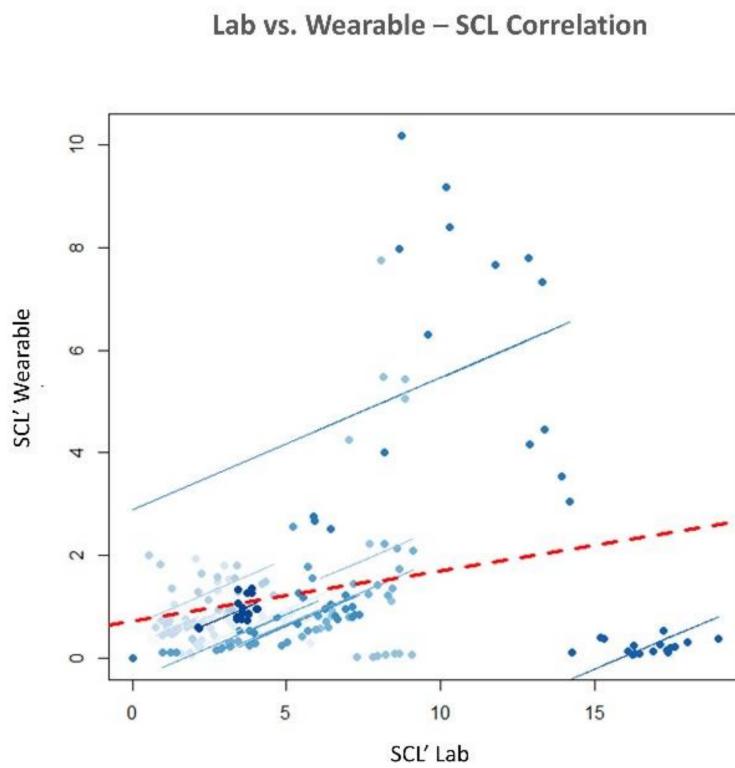
### 3.3. Neurophysiological Results

#### 3.3.1. Methods comparisons

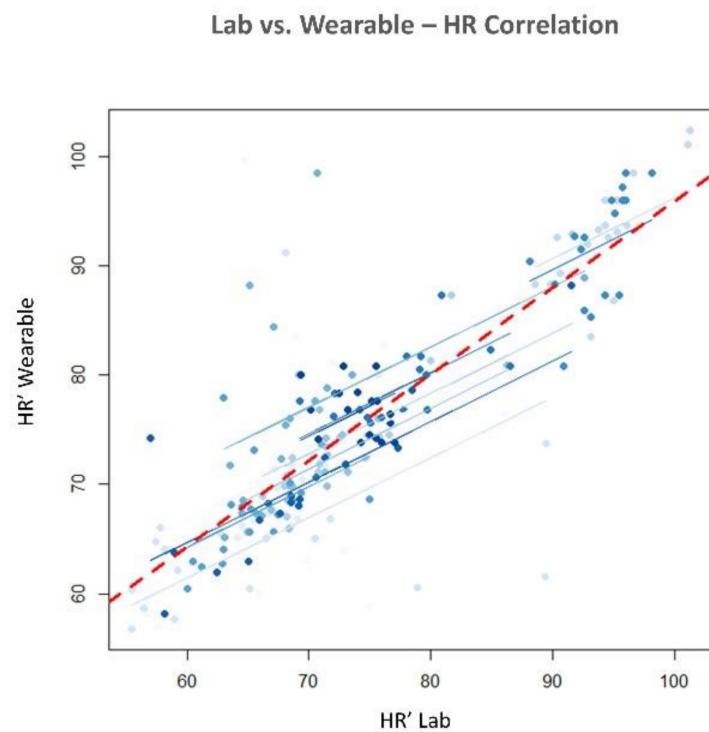
The statistical analysis revealed no significant difference in terms of EBR' between the consumer wearable and laboratory equipment during both NB ( $p = 0.65$ ) and DG ( $p = 0.69$ ). Similarly, the Wilcoxon signed-rank tests on the SCL' and HR' showed no significant differences in terms of SCL' (NB:  $p = 0.09$ ; DG:  $p = 0.4$ ) and HR' (NB:  $p = 0.18$ ; DG:  $p = 0.69$ ) estimation. Correlation analysis between the neurophysiological parameter estimated with wearable and laboratory devices was performed. All the parameters were significantly correlated ( $p < 0.05$ ). EBR estimated with laboratory and wearable devices resulted highly and positively correlated ( $R = 0.83$ ,  $p < 10^{-47}$ ) (Figure 7). Correlation for SCL and HR resulted less strong but however significant. SCL correlation analysis (Figure 8) reported and R of 0.4 ( $p < 10^{-6}$ ). Finally, R value for HR correlation (Figure 9) was 0.51 ( $p < 10^{-14}$ ). To support correlation results, time dynamics of the investigated parameters acquired in a representative participant are depicted in Figures 10–12.



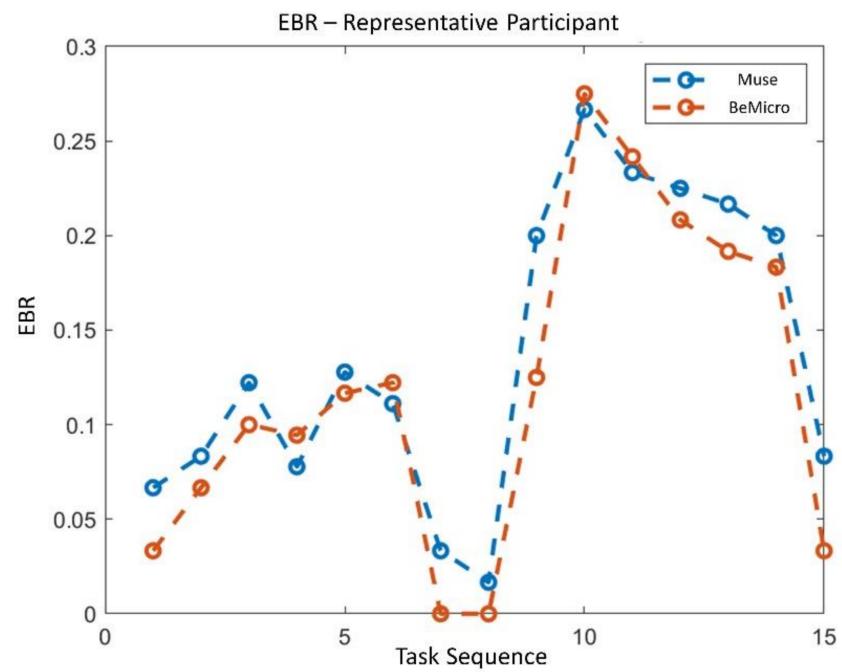
**Figure 7.** Pearson's repeated measure correlation for the Eyeblink Rate (EBR) estimated with laboratory and wearable devices.  $R = 0.83, p < 10^{-47}$ .



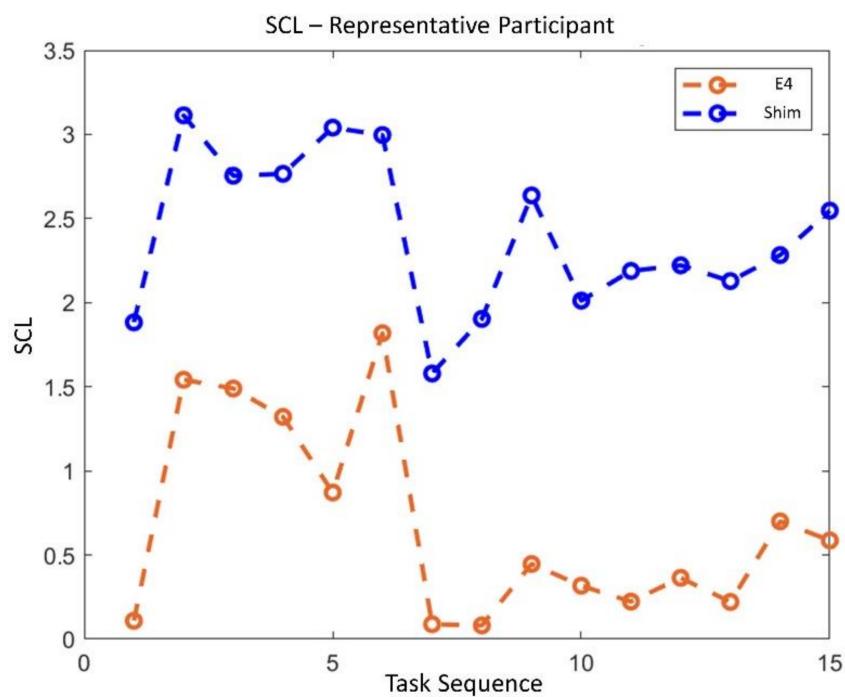
**Figure 8.** Pearson's repeated measure correlation for the Skin Conductance Level (SCL) estimated with laboratory and wearable devices.  $R = 0.4, p < 10^{-6}$ .



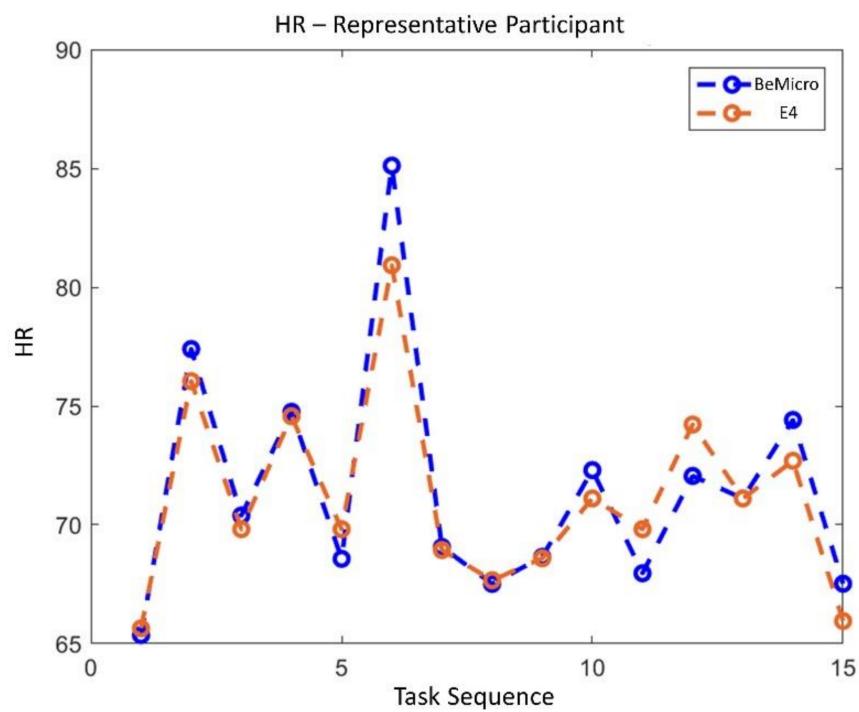
**Figure 9.** Pearson's repeated measure correlation for the Heart Rate (HR) estimated with laboratory and wearable devices.  $R = 0.51, p < 10^{-14}$ .



**Figure 10.** Time dynamics of EBR across all experimental task and conditions for both consumer wearable (blue) and laboratory device (red).



**Figure 11.** Time dynamics of SCL across all experimental task and conditions for both consumer wearable (red) and laboratory device (blue).



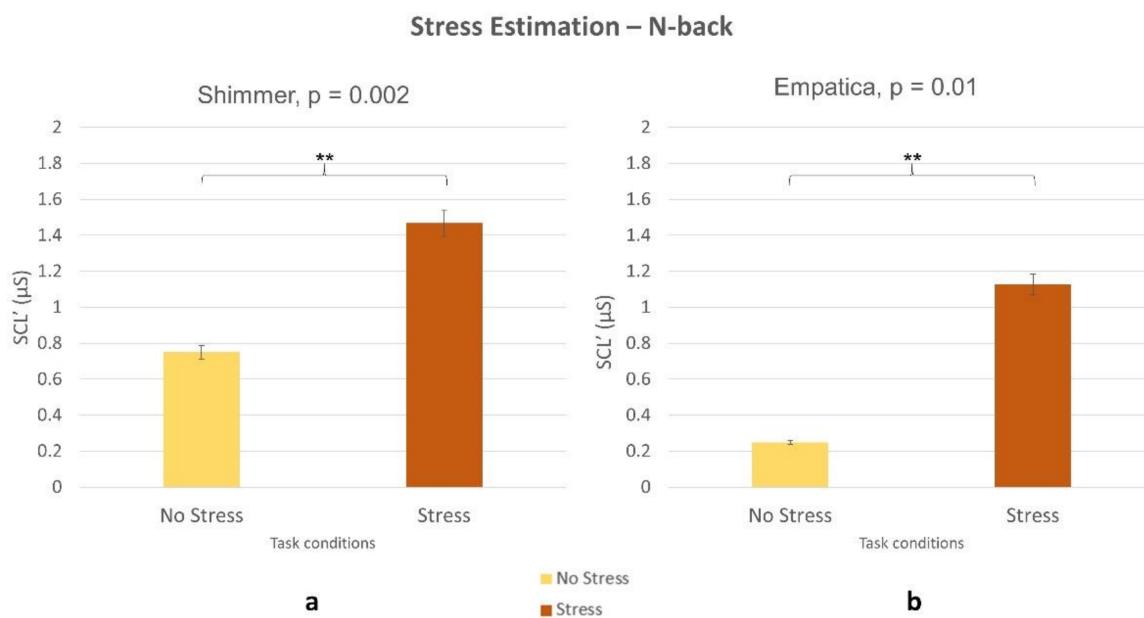
**Figure 12.** Time dynamics of EBR across all experimental task and conditions for both consumer wearable (red) and laboratory device (blue).

### 3.3.2. Mental workload

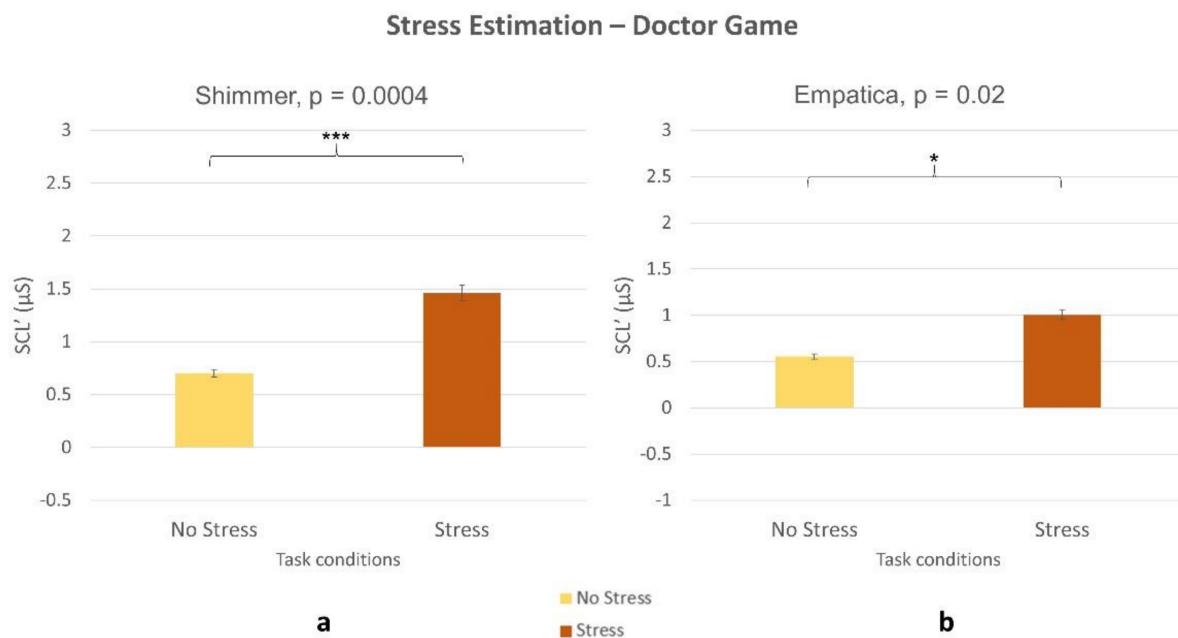
For both wearable and laboratory device the Wilcoxon signed-rank tests did not reveal significant differences (consumer wearable:  $p = 0.64$ ; laboratory:  $p = 0.96$ ) in terms of EBR' when comparing high WL vs. low WL conditions.

### 3.3.3. Stress

The Wilcoxon signed-rank tests on SCL' parameter estimated by the laboratory device and the wearable one returned significant difference showing higher values during the stress condition (all  $p < 0.05$ ) both for the NB (Figure 13) and DG (Figure 14) task.



**Figure 13.** Increased SCL' in stress vs. no stress condition during NB task. Statistical analysis revealed significant difference between the conditions for both (a) laboratory equipment ( $p = 0.002$ ) and (b) wearable device ( $p = 0.1$ ).



**Figure 14.** Increased SCL' in stress vs. no stress condition during DG task. Statistical analysis revealed significant difference between the conditions for both (a) laboratory equipment ( $p = 0.0004$ ) and (b) wearable device ( $p = 0.02$ ).

### 3.3.4. Emotional State

The Wilcoxon signed-rank test demonstrated no statistical differences (wearable:  $p = 0.1$ ; laboratory:  $p = 0.4$ ) in terms of EI between the positive and negative conditions.

## 4. Discussion

The objectives of the study consisted in assessing the reliability and capability of commercial wearable devices with respect to laboratory devices in estimating EBR, SCL and HR parameters and discriminating different levels of mental workload, stress, and emotional state.

### 4.1. Research Questions

Regarding the RQ1 (i.e., “Are the above-mentioned neurophysiological parameters (EBR, SCL and HR) gathered through consumer wearable devices comparable with those acquired with laboratory equipment?”), our results confirmed the feasibility to measure EBR, HR and SCL using consumer wearable devices. The parameters estimated with wearable and laboratory devices showed significant positive correlations as a demonstration that the two devices provided similar neurophysiological results (Figures 7–9). Additionally, no statistical differences were observed in terms of EBR, HR, and SCL estimation between the two technologies considered (i.e., consumer wearable and laboratory). In fact, for each of the parameters considered the statistical analysis showed no significant difference in the averaged. These results support the adoption of consumer wearable devices and the relative collected metrics to disentangle complex mental and emotional events in real-life environments. This aspect leads to the RQ2 (i.e., “Are consumer wearable devices reliable in assessing different levels of several mental states?”). EBR was used as a neurophysiological correlate of WL, and the Muse 2 (wearable) and BeMicro (laboratory) devices were compared. No difference was found in terms of mental workload variation during the NB and DG task.

### 4.2. Workload Assessment

Regarding the DG task, the absence of WL changes was probably due to the fact that the High WL condition was not so hard as expected. Indeed, even if performance decreased in the high WL condition compared to the low WL one, participants did not perceive the high WL condition to be harder. It is arguable that adding more items resulted in a similar WL demand between low and high WL conditions with no difference when comparing EBR’ correlates.

Similarly, for the NB task, combining together performance and subjective reports, it could be argued that the absence of WL correlates was due to the difficulty of the task itself. In fact, NASA-TLX showed participants perceiving high WL condition to be harder than low WL one (Figure 6). However, at the end of the experimental session they reported that the High WL condition was too hard to be performed and for this reason they did not or could not attend the task properly. This finding is supported by performance analysis, where it was found number of missed responses significantly increased in high WL condition compared to low WL one (Figure 4). In this view, the absence of WL correlates could be a result of participants’ abandoning the task. Alternatively, the lack of EBR’ variations in both tasks could be motivated by EBR sensitivity. EBR parameter could be less sensitive to slight changes in task WL demand than other parameters (HR, HRV, PSD, ERP, etc.). This means that other parameters than EBR could have detected WL correlates in the same conditions. This points out directions for future works. The same paradigm could be tested using different neurophysiological correlates of WL to test their sensitivity and to support their adoption in different environments.

### 4.3. Stress Assessment

In terms of stress assessment, the SCL parameter was used as a neurophysiological correlate. The Empatica E4 (consumer wearable) and Shimmer (laboratory) evaluated an increased stress level during Stress condition compared to no stress one, both within NB (Figure 13) and DG (Figure 14) tasks. Even if stress correlates are accompanied with a decreased performance in both experimental tasks (Figure 3, Figure 5), participants were not able to perceive stress variations. In accordance with this, previous studies highlighted

the limit in assessing perceived stress using subjective reports [34]. This study, therefore, confirmed the utility of using neurometrics to assess perceived stress [7]. It was also demonstrated that consumer wearable devices could substitute laboratory equipment to acquire such neurometrics. The possibility to detect stress in an obtrusive way is one of the most promising aspects of wearable devices.

#### 4.4. Emotional State Assessment

Finally, regarding the possibility to discriminate between a positive EmS and a negative one using a combination of SCL and HR [55], both technologies were not able to differentiate these two conditions. Even if after positive condition participants selected mostly positive emotions (and negative ones after negative conditions), we found that after positive condition participants selected also some negative emotions and vice versa. It is arguable what arose from the two conditions was a blend of emotions, with no pure positive or negative connotations. Additionally, there is the possibility that two minutes interaction with a stranger in a simulated webcall was not enough to elicit a measurable neurophysiological change in the participants' emotional states. As exposed, considering performance and subjective evaluations, the reason for the absence of WL and EmS correlates could be the experimental design itself, which did not elicit the desired mental states. This limit points out direction for next works. Future studies should design an experiment to more accurately define WL and EmS conditions.

#### 4.5. Limits and Future Directions

Although both the reliability of consumer wearable devices in estimating neurophysiological signals and their capability in discriminating different levels of stress is promising, some limitations must be discussed. An experimental design and tasks capable of eliciting the desired levels of the mental states must be implemented to better investigate the usability of wearable devices. For NB and DG, an improved design should elicit the proper level of workload while for the emotional state evaluations a longer duration of the task should be considered in order to elicit a stronger and measurable emotional, and therefore autonomic, response in the participants. Consumer wearable devices are user-friendly and non-invasive technologies, allowing their usage in dynamics condition in which laboratory equipment would not be adequate. The possibility to use these devices in dynamics environments must be supported by a good quality of the gathered signals. This is a challenging aspect for consumer wearable devices and their utilization must be carefully evaluated considering the recording settings and protocol in order to acquire a valid signal. In particular, after this preliminary evaluation of wearables reliability, their capability in differentiating between different mental states should will be tested in real-working conditions with attention to the processing and analysis of the data gathered with these devices and the results will be considered for the next study. Additionally, it should be underlined that one of the considered consumer wearable devices, the Empatica E4, can be classified as a high-level wearable device. The elevated cost of high-quality wearables could represent a limit in their adoption. For this reason, the possibility to estimate the considered signals and the related mental states using commercial and low-cost wearable devices should be also explored in order to broad the mental state monitoring in the consumer world, without limiting their adoption to the scientific research.

Furthermore, future works should investigate these and other mental states in a larger group of participants and investigate the impact of participants' movements on the quality of collected data with a particular attention to the devices/parameters affected by the movements and the intensity of the considered signals. Specifically, an important aspect that will be investigated in the next study is the comparison of the number of artifacts and the percentage of data loss found in consumer wearable devices with those of laboratory equipment. Additionally, reliability of investigated parameters in estimating mental states correlates in working-like tasks should be compared to other physiological signals (such as EEG and HRV) in order to detect the one that better fits to the recording conditions.

Consequently, the adoption of other physiological signals must be accompanied by an adequate task duration to provide reliable data. Once reliability of wearable devices has been confirmed, the possibility to discriminate mental states in real-time must be investigated. Finally, consumer wearable devices are optimal candidate for health and well-being monitoring [58,59]. When appropriate algorithms are applied it is possible to monitor patients' health by remote in real-time and prevent fatal and non-fatal occurrences. For this reason, it is important to investigate the acceptance of this wearable devices and their easiness to use [60]. This will be especially important for monitoring elderly population [61].

## 5. Conclusions

The study demonstrated that signal recorded with consumer wearable and laboratory devices showed a statistically positive correlation and no significant difference (RQ1). Additionally, it was demonstrated the capability in differentiating stress levels (RQ2). Within this experimental design it was impossible to differentiate between different levels of WL and EmS (RQ2).

The possibility to measure neurophysiological parameters at the same level laboratory devices do but with a limited invasiveness is one of the greatest points of strength of consumer wearable devices. On the other side, unobtrusiveness is achieved with reduced size which comports a limited duration of the battery, limiting these devices to short periods of testing. Furthermore, it is reported that the contact between wearable devices and the body is not always optimal, leading to missing or altered data [25]. This limits the use of consumer wearables to those cases in which movement is compatible with data collecting.

Taken together, these findings support the adoption of low-cost wearable device to monitor operators' mental states in laboratory and real-life environments. The possibility to unobtrusively assess mental states has broad applications. It could be possible to monitor air-traffic controllers, medical operators, surgeons, while working without interfering with the performance. Hopefully, the ability to better differentiate between mental states will reduce the effect of tragic occurrences.

**Author Contributions:** Conceptualization: G.B. and V.R.; methodology: G.B. and V.R.; software: G.B., A.d.F., D.R., G.D.F., N.S., and V.R.; formal analysis: G.B., V.R., L.T., I.S., and A.G.; investigation: G.B. and V.R.; resources: G.B., P.A., A.G., A.V., and V.R.; data curation: G.B., V.R., and A.G.; writing—original draft preparation: A.G.; writing—review and editing: G.B., P.A., G.D.F., N.S., and A.V.; visualization: G.B. and A.G.; supervision: G.B.; funding acquisition: G.B. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was European Commission by Horizon2020 projects “WORKINGAGE: Smart Working environments for all Ages” (GA no. 826232); “SIMUSAFFE: Simulator Of Behavioral Aspects For Safer Transport” (GA n. 723386); “SAFEMODE: Strengthening Synergies between Aviation and Maritime in the area of Human Factors towards Achieving more Efficient and Resilient MODE of Transportation” (GA no. 814961), “BRAINSAFEDRIVE: A Technology to Detect Mental States during Driving for Improving the Safety of the Road” (Italy-Sweden collaboration) with a grant of Ministero dell’Istruzione dell’Università e della Ricerca della Repubblica Italiana, “MINDTOOTH: Wearable device to decode human mind by neurometrics for a new concept of smart interaction with the surrounding environment” (GA no. 950998) and H2020-SESAR-2019-2 project: Transparent Artificial Intelligence and Automation to Air Traffic Management Systems, “ARTIMATION,” GA no. 894238.

**Institutional Review Board Statement:** The study was conducted according to the guidelines of the Declaration of Helsinki and approved by the Sapienza University of Rome Ethical Committee in Charge for the Department of Molecular Medicine (protocol number: 2507/2020, approved on 4 August 2020).

**Informed Consent Statement:** Informed consent was obtained from all subjects involved in the study.

**Data Availability Statement:** The aggregated data presented in this study might be available on request from the corresponding author. The data are not publicly available because they were collected within the EU Project “WORKINGAGE: Smart Working environments for all Ages” (GA n.826232) and they are property of the Consortium.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Hagh, M.; Thurow, K.; Stoll, R. Wearable devices in medical internet of things: Scientific research and commercially available devices. *Healthc. Inform. Res.* **2017**, *23*, 4–15. [[CrossRef](#)]
2. Ranavolo, A.; Draicchio, F.; Varrecchia, T.; Silvetti, A.; Iavicoli, S. Wearable monitoring devices for biomechanical risk assessment at work: Current status and future challenges—A systematic review. *Int. J. Environ. Res. Public Health* **2018**, *15*, 2001. [[CrossRef](#)] [[PubMed](#)]
3. White, M.S.; Burns, C.; Conlon, H.A. The impact of an aging population in the workplace. *Work. Health Saf.* **2018**, *66*, 493–498. [[CrossRef](#)] [[PubMed](#)]
4. Borghini, G.; Astolfi, L.; Vecchiato, G.; Mattia, D.; Babiloni, F. Measuring neurophysiological signals in aircraft pilots and car drivers for the assessment of mental workload, fatigue and drowsiness. *Neurosci. Biobehav. Rev.* **2014**, *44*, 58–75. [[CrossRef](#)] [[PubMed](#)]
5. Young, M.S.; Brookhuis, K.A.; Wickens, C.D.; Hancock, P.A. State of science: Mental workload in ergonomics. *Ergonomics* **2015**, *58*, 1–17. [[CrossRef](#)]
6. Borghini, G.; Aricò, P.; Di Flumeri, G.; Cartocci, G.; Colosimo, A.; Bonelli, S.; Golfetti, A.; Imbert, J.P.; Granger, G.; Benhacene, R.; et al. EEG-based cognitive control behaviour assessment: An ecological study with professional air traffic controllers. *Sci. Rep.* **2017**, *7*, 1–16. [[CrossRef](#)]
7. Borghini, G.; Di Flumeri, G.; Aricò, P.; Sciaraffa, N.; Bonelli, S.; Ragosta, M.; Tomasello, P.; Drogoul, F.; Turhan, U.; Acikel, B.; et al. A multimodal and signals fusion approach for assessing the impact of stressful events on Air Traffic Controllers. *Sci. Rep.* **2020**, *10*, 1–18. [[CrossRef](#)]
8. Hansen, F. Human error: A concept analysis. *J. Air Transp.* **2006**, *11*, 61–77. Available online: <https://ntrs.nasa.gov/search.jsp?R=20070022530> (accessed on 24 March 2021).
9. Arico, P.; Borghini, G.; Di Flumeri, G.; Colosimo, A.; Graziani, I.; Imbert, J.-P.; Granger, G.; Benhacene, R.; Terenzi, M.; Pozzi, S.; et al. Reliability over time of EEG-based mental workload evaluation during Air Traffic Management (ATM) tasks. In Proceedings of the 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Milano, Italy, 25–29 August 2015; pp. 7242–7245. [[CrossRef](#)]
10. Aricò, P.; Reynal, M.; Di Flumeri, G.; Borghini, G.; Sciaraffa, N.; Imbert, J.-P.; Hurter, C.; Terenzi, M.; Ferreira, A.; Pozzi, S.; et al. How neurophysiological measures can be used to enhance the evaluation of remote tower solutions. *Front. Hum. Neurosci.* **2019**, *13*, 303. [[CrossRef](#)]
11. Borghini, G.; Aricò, P.; Astolfi, L.; Toppi, J.; Cincotti, F.; Mattia, D.; Cherubino, P.; Vecchiato, G.; Maglione, A.G.; Graziani, I.; et al. Frontal EEG theta changes assess the training improvements of novices in flight simulation tasks. In Proceedings of the 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC); Institute of Electrical and Electronics Engineers, Osaka, Japan, 3–7 July 2013; Volume 2013, pp. 6619–6622.
12. Jahangiri, M.; Hoboubi, N.; Rostamabadi, A.; Keshavarzi, S.; Hosseini, A.A. Human error analysis in a permit to work system: A case study in a chemical plant. *Saf. Health Work* **2016**, *7*, 6–11. [[CrossRef](#)] [[PubMed](#)]
13. Kondrateva, O.; Kravchenko, M.; Loktionov, O. Development of the methods for assessing the risk of damage to health of the employees of the electric power industry. *Bezop. Promyshlennosti* **2019**, *2019*, 63–68. [[CrossRef](#)]
14. Bevilacqua, M.; Ciarapica, F.E. Human factor risk management in the process industry: A case study. *Reliab. Eng. Syst. Saf.* **2018**, *169*, 149–159. [[CrossRef](#)]
15. Babiloni, F. Mental workload monitoring: New perspectives from neuroscience. In *Communications in Computer and Information Science*; Springer: Cham, Switzerland, 2019; Volume 1107, pp. 3–19.
16. Arico, P.; Borghini, G.; Di Flumeri, G.; Bonelli, S.; Golfetti, A.; Graziani, I.; Pozzi, S.; Imbert, J.-P.; Granger, G.; Benhacene, R.; et al. Human factors and neurophysiological metrics in air traffic control: A critical review. *IEEE Rev. Biomed. Eng.* **2017**, *10*, 250–263. [[CrossRef](#)]
17. Wall, T.D.; Michie, J.; Patterson, M.; Wood, S.J.; Sheehan, M.; Clegg, C.W.; West, M. On the validity of subjective measures of company performance. *Pers. Psychol.* **2004**, *57*, 95–118. [[CrossRef](#)]
18. Sciaraffa, N.; Borghini, G.; Aricò, P.; Di Flumeri, G.; Colosimo, A.; Bezerianos, A.; Thakor, N.V.; Babiloni, F. Brain interaction during cooperation: Evaluating local properties of multiple-brain network. *Brain Sci.* **2017**, *7*, 90. [[CrossRef](#)]
19. Fairclough, S.H. Fundamentals of physiological computing. *Interact. Comput.* **2009**, *21*, 133–145. [[CrossRef](#)]
20. Cartocci, G.; Maglione, A.G.; Vecchiato, G.; Di Flumeri, G.; Colosimo, A.; Scorpecci, A.; Marsella, P.; Giannantonio, S.; Malerba, P.; Borghini, G.; et al. Mental workload estimations in unilateral deafened children. In Proceedings of the 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Milano, Italy, 29–29 August 2015; Volume 2015, pp. 1654–1657. [[CrossRef](#)]

21. Patel, S.; Park, H.; Bonato, P.; Chan, L.; Rodgers, M. A review of wearable sensors and systems with application in rehabilitation. *J. Neuroeng. Rehabil.* **2012**, *9*, 21. [CrossRef]
22. Di Flumeri, G.; Aricò, P.; Borghini, G.; Sciaraffa, N.; Di Florio, A.; Babiloni, F. The dry revolution: Evaluation of three different EEG dry electrode types in terms of signal spectral features, mental states classification and usability. *Sensors* **2019**, *19*, 1365. [CrossRef] [PubMed]
23. Gradl, S.; Wirth, M.; Richer, R.; Rohleder, N.; Eskofier, B.M. An overview of the feasibility of permanent, real-time, unobtrusive stress measurement with current wearables. In Proceedings of the 13th EAI International Conference on Pervasive Computing Technologies for Healthcare, Trento, Italy, 20–23 May 2019; Volume 19, pp. 360–365.
24. E Dooley, E.; Golaszewski, N.M.; Bartholomew, J.B. Estimating accuracy at exercise intensities: A comparative study of self-monitoring heart rate and physical activity wearable devices. *JMIR mHealth uHealth* **2017**, *5*, e34. [CrossRef]
25. Lo, J.C.; Sehic, E.; Meijer, S.A. Measuring mental workload with low-cost and wearable sensors: Insights into the accuracy, obtrusiveness, and research usability of three instruments. *J. Cogn. Eng. Decis. Mak.* **2017**, *11*, 323–336. [CrossRef]
26. Menghini, L.; Gianfranchi, E.; Cellini, N.; Patron, E.; Tagliabue, M.; Sarlo, M. Stressing the accuracy: Wrist-worn wearable sensor validation over different conditions. *Psychophysiology* **2019**, *56*, e13441. [CrossRef]
27. Shcherbina, A.; Mattsson, C.M.; Waggett, D.; Salisbury, H.; Christle, J.W.; Hastie, T.; Wheeler, M.T.; Ashley, E.A. Accuracy in wrist-worn, sensor-based measurements of heart rate and energy expenditure in a diverse cohort. *J. Pers. Med.* **2017**, *7*, 3. [CrossRef]
28. McCarthy, C.; Pradhan, N.; Redpath, C.; Adler, A. Validation of the empatica E4 wristband. In Proceedings of the IEEE EMBS International Student Conference (ISC), Ottawa, ON, Canada, 29–31 May 2016; pp. 1–4.
29. Ragot, M.; Martin, N.; Em, S.; Pallamin, N.; Diverrez, J.-M. Emotion recognition using physiological signals: Laboratory vs. wearable sensors. *Adv. Intell. Syst. Comput.* **2018**, *608*, 15–22.
30. Setz, C.; Arnrich, B.; Schumm, J.; La Marca, R.; Tröster, G.; Ehrlert, U. Discriminating stress from cognitive load using a wearable EDA device. *IEEE Trans. Inf. Technol. Biomed.* **2010**, *14*, 410–417. [CrossRef] [PubMed]
31. Krigolson, O.E.; Williams, C.C.; Norton, A.; Hassall, C.D.; Colino, F.L. Choosing MUSE: Validation of a low-cost, portable EEG system for ERP research. *Front. Neurosci.* **2017**, *11*, 109. [CrossRef] [PubMed]
32. Abujelala, M.; Abellanoza, C.; Sharma, A.; Makedon, F. Brain-EE: Brain enjoyment evaluation using commercial EEG headband. In Proceedings of the PETRA 2016, Corfu, Greece, 29 June–1 July 2016. [CrossRef]
33. Kirchner, W.K. Age differences in short-term retention of rapidly changing information. *J. Exp. Psychol.* **1958**, *55*, 352–358. [CrossRef] [PubMed]
34. Berka, C.; Levendowski, D.J.; Lumicao, M.N.; Yau, A.; Davis, G.; Zivkovic, T.; Olmstead, R.E.; Tremoulet, P.; Craven, P.L. EEG correlates of task engagement and mental workload in vigilance, learning, and memory tasks. *Aviat. Space Environ. Med.* **2007**, *78*, B231–B244. Available online: [https://www.researchgate.net/publication/6289900\\_EEG\\_correlates\\_of\\_task\\_engagement\\_and\\_mental\\_workload\\_in\\_vigilance\\_learning\\_and\\_memory\\_tasks](https://www.researchgate.net/publication/6289900_EEG_correlates_of_task_engagement_and_mental_workload_in_vigilance_learning_and_memory_tasks) (accessed on 24 January 2021). [PubMed]
35. Skoluda, N.; Strahler, J.; Schlotz, W.; Niederberger, L.; Marques, S.; Fischer, S.; Thoma, M.V.; Spoerri, C.; Ehrlert, U.; Nater, U.M. Intra-individual psychological and physiological responses to acute laboratory stressors of different intensity. *Psychoneuroendocrinology* **2015**, *51*, 227–236. [CrossRef] [PubMed]
36. De La Torre-Luque, A.; Caparros-Gonzalez, R.A.; Bastard, T.; Vico, F.J.; Buela-Casal, G. Acute stress recovery through listening to Melomics relaxing music: A randomized controlled trial. *Nord. J. Music Ther.* **2016**, *26*, 124–141. [CrossRef]
37. Ceccarelli, L.A.; Giuliano, R.J.; Glazebrook, C.M.; Strachan, S.M. Self-compassion and psycho-physiological recovery from recalled sport failure. *Front. Psychol.* **2019**, *10*, 1564. [CrossRef] [PubMed]
38. Konečni, V.J.; Brown, A.; Wanic, R.A. Comparative effects of music and recalled life-events on emotional state. *Psychol. Music* **2008**, *36*, 289–308. [CrossRef]
39. Hart, S.; Staveland, L. Development of NASA-TLX (Task Load Index) results of empirical and theoretical research. *Adv. Psychol.* **1988**, *52*, 139–183.
40. Coyne, A.K.; Murtagh, A.; McGinn, C. Using the Geneva Emotion Wheel to measure perceived affect in human-robot interaction. In Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction, Cambridge, UK, 23–26 March 2020; pp. 491–498.
41. Zheng, B.; Jiang, X.; Tien, G.; Meneghetti, A.; Panton, O.N.M.; Atkins, M.S. Workload assessment of surgeons: Correlation between NASA TLX and blinks. *Surg. Endosc.* **2012**, *26*, 2746–2750. [CrossRef] [PubMed]
42. Grier, R.A. How High is High? A Meta-Analysis of NASA-TLX Global Workload Scores. *Proc. Hum. Factors Ergon. Soc. Annu. Meet.* **2015**, *59*, 1727–1731. [CrossRef]
43. Shuman, V.; Schlegel, K.; Scherer, K. Geneva Emotion Wheel Rating Study PROPERMO View Project a Developmental Perspective of Emotion Regulation View Project. 2015. Available online: <https://www.researchgate.net/publication/280880848> (accessed on 26 January 2021).
44. Faure, V.; Lobjois, R.; Benguigui, N. The effects of driving environment complexity and dual tasking on drivers' mental workload and eye blink behavior. *Transp. Res. Part F Traffic Psychol. Behav.* **2016**, *40*, 78–90. [CrossRef]
45. Di Flumeri, G.; Arico, P.; Borghini, G.; Colosimo, A.; Babiloni, F. A new regression-based method for the eye blinks artifacts correction in the EEG signal, without using any EOG channel. In Proceedings of the 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Orlando, FL, USA, 16–20 August 2016; Volume 2016, pp. 3187–3190.

46. Abbas, S.N.; Abo-Zahhad, M. *Eye Blinking EOG Signals as Biometrics*; Springer: Cham, Switzerland, 2017; pp. 121–140.
47. Kleifges, K.; Bigdely-Shamlo, N.; Kerick, S.E.; Robbins, K.A. BLINKER: Automated extraction of ocular indices from eeg enabling large-scale analysis. *Front. Neurosci.* **2017**, *11*, 12. [[CrossRef](#)]
48. Borghini, G.; Ronca, V.; Vozzi, A.; Aricò, P.; Di Flumeri, G.; Babiloni, F. Monitoring performance of professional and occupational operators. *Handb. Clin. Neurol.* **2020**, *168*, 199–205.
49. Bach, D.R. A head-to-head comparison of SCRalyze and Ledalab, two model-based methods for skin conductance analysis. *Biol. Psychol.* **2014**, *103*, 63–68. [[CrossRef](#)] [[PubMed](#)]
50. Benedek, M.; Kaernbach, C. A continuous measure of phasic electrodermal activity. *J. Neurosci. Methods* **2010**, *190*, 80–91. [[CrossRef](#)] [[PubMed](#)]
51. Braithwaite, J.J.; Derrick, D.; Watson, G.; Jones, R.; Rowe, M. A Guide for Analysing Electrodermal Activity (EDA) and Skin Conductance Responses (SCRs) for Psychological Experiments. 2015. Available online: <https://www.birmingham.ac.uk/Documents/college-les/psych/saal/guide-electrodermal-activity.pdf> (accessed on 24 March 2021).
52. Borghini, G.; Bandini, A.; Orlandi, S.; Di Flumeri, G.; Arico, P.; Sciaraffa, N.; Ronca, V.; Bonelli, S.; Ragosta, M.; Tomasello, P.; et al. Stress assessment by combining neurophysiological signals and radio communications of air traffic controllers. In Proceedings of the 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC); Institute of Electrical and Electronics Engineers (IEEE), online event (ex-Montreal). 20–24 July 2020; Volume 2020, pp. 851–854.
53. Goovaerts, H.G.; Ros, H.H.; Akker, T.J.V.D.; Schneider, H. A digital QRS detector based on the principle of contour limiting. *IEEE Trans. Biomed. Eng.* **1976**, *23*, 154–160. [[CrossRef](#)]
54. Pan, J.; Tompkins, W.J. A real-time QRS detection algorithm. *IEEE Trans. Biomed. Eng.* **1985**, *32*, 230–236. [[CrossRef](#)] [[PubMed](#)]
55. Russell, J.A.; Barrett, L.F. Core affect, prototypical emotional episodes, and other things called emotion: Dissecting the elephant. *J. Pers. Soc. Psychol.* **1999**, *76*, 805–819. [[CrossRef](#)]
56. Vecchiato, G.; Cherubino, P.; Maglione, A.G.; Ezquierro, M.T.H.; Marinoszzi, F.; Bini, F.; Trettel, A.; Babiloni, F. How to measure cerebral correlates of emotions in marketing relevant tasks. *Cogn. Comput.* **2014**, *6*, 856–871. [[CrossRef](#)]
57. Bakdash, J.Z.; Marusich, L.R. Repeated measures correlation. *Front. Psychol.* **2017**, *8*, 456. [[CrossRef](#)] [[PubMed](#)]
58. Marakhimov, A.; Joo, J. Consumer adaptation and infusion of wearable devices for healthcare. *Comput. Hum. Behav.* **2017**, *76*, 135–148. [[CrossRef](#)]
59. Guk, K.; Han, G.; Lim, J.; Jeong, K.; Kang, T.; Lim, E.-K.; Jung, J. Evolution of wearable devices with real-time disease monitoring for personalized healthcare. *Nanomaterials* **2019**, *9*, 813. [[CrossRef](#)]
60. Tran, V.-T.; Riveros, C.; Ravaud, P. Patients' views of wearable devices and AI in healthcare: Findings from the ComPaRe e-cohort. *NPJ Digit. Med.* **2019**, *2*, 1–8. [[CrossRef](#)] [[PubMed](#)]
61. Stavropoulos, T.G.; Lazarou, I.; Strantsalis, D.; Nikolopoulos, S.; Kompatsiaris, I.; Koumanakos, G.; Frouda, M.; Tsolaki, M. Human factors and requirements of people with mild cognitive impairment, their caregivers and healthcare professionals for ehealth systems with wearable trackers. In Proceedings of the IEEE International Conference on Human-Machine Systems (ICHMS), Rome, Italy, 6–8 April 2020; pp. 1–6.