Identifying predictive EEG features for cognitive overload detection in assembly workers in Industry 4.0

Jessica Morton¹, Pieter Vanneste², Charlotte Larmuseau², Bram B. Van Acker¹, Annelies Raes², Klaas Bombeke¹, Frederik Cornillie², Jelle Saldien¹, Lieven De Marez¹

¹imec-mict-UGent, Platteberg 11, 9000 Gent, Belgium
²imec-itec-KU Leuven, Etienne Sabbelaan 51, 8500 Kortrijk, Belgium

Abstract. Industry 4.0 will be characterized by far-reaching production automation because of recent advancements in robotics and artificial intelligence. As a consequence, a lot of simple, repetitive assembly tasks will no longer be performed by factory workers, but by machines. However, at the same time, consumers demand more and more personalized products, increasing the need for human assembly workers who can adapt quickly to new and more complex assembly procedures. This need for adaptation is most likely to increase the cognitive workload and potentially overload assembly workers that are used to traditional assembly work tasks. Several studies have tried to identify this cognitive overload in the EEG signal, but many failed because of poor experimental measurement procedures, bad data quality and low sample sizes. In this paper, we therefore designed a highly controlled lab experiment to collect EEG data of a large number of participants (N=46) performing an assembly task under various levels of cognitive load (low, high, overload). This systematic approach allowed us to study which EEG features are particularly useful and valid for cognitive overload assessment in the context of assembly work.

Keywords: Industry 4.0, assembly work, cognitive overload, EEG

1 Introduction

1.1 Assembly work in Industry 4.0 and cognitive load

Industry 4.0 or “smart factories” of the future will be characterized by wide-scale automatization, connectivity and AI-driven technology, resulting in a manufacturing process that will become more and more efficient [1,2,3]. It is beyond any doubt that many jobs involving simple, repetitive tasks will disappear in favor of robots or at least cobots (i.e., machines that physically interact with human workers). However, at the same time, it is expected that customer demand will push the industry towards
increasing product variety to allow for broad product personalization [4,5,6]. For example, in car manufacturing, it is more common that customers have the ability to decide on design specifications compared to the past.

Hence, amidst this evolution stands the human assembly worker who will need to operate more and more in a flexible way and will be required to constantly adjust his or her skills to changing job demands and technology [7,8,9]. Since it is not unlikely that this increasing complexity and need for flexibility will make it harder for this human worker to do the job in a proper way, it is highly important to accurately measure cognitive load and explore ways to avoid or reduce this load from a cognitive ergonomic point of view [10,11]. In this paper, we therefore focus on cognitive load detection in the context of personalized assembly work.

Throughout the history of (cognitive) ergonomics, the construct of cognitive load has been playing a substantive role in the prevention of occupational error, safety hazard, and negative (physical) stress caused by overload [4,12]. Cognitive load is a multi-dimensional, rather than a unitary construct and covers working memory processes ranging from attention and perception to memory and decision making [13]. Originally, the concept of cognitive load evolved from early work in the instructional and educational research field, eventually coming together in a widely-applied theory called cognitive load theory (CLT) [14,15,16]. Resonating with the multidimensional nature of the cognitive load concept, cognitive load measures are equally various in nature. In general, the literature converges towards assessing cognitive load based on subjective self-reporting and psychophysiological measurements [13,17,18,19,20,21].

Whereas a lot of research has been done on how to accurately question people about their cognitive load using questionnaires and in-depth interviews, there is still a lot of work to be done with respect to using psychophysiological data to assess cognitive load. Interestingly enough, recent innovations and advancements in wearable technology have led to low-cost, easy-to-wear, energy-efficient devices to measure electrical activity at the human scalp. Therefore, it is expected that cognitive load measurement based on psychophysiological EEG data will become very prominent in the future [11,22,23,24,25]. Being able to rather noninvasively measure brain activity in a real-world context in a relatively cheap way has triggered the interest of both the industry and academic cognitive load community. As a consequence, there are already many studies available in which researchers looked at the relationship between cognitive load and changes in the EEG signal [for a review, 26,27,28]. Unfortunately, many of these studies do not succeed in obtaining valid and reliable conclusions because of methodological flaws in design, issues with poor experimental measurement procedures or settings, bad data quality, and low sample sizes. This is not surprising, since inter-individual differences in EEG recordings can be very high and signals are prone to artefacts caused by technical malfunctioning, facial muscle activity and static noise coming from other electrical sources in the assessment setting. For this reason and because replication is an important characteristic of scientific research, we chose to take one step back and study predictive EEG features for cognitive overload in a highly controlled lab setting instead of at the factory floor right away (although it is beyond any doubt that the latter should be the end goal). By choosing this approach, we hope to overcome the aforementioned problems.
1.2 EEG and cognitive load

There are basically two approaches to analyze EEG data. First, spectral analysis of oscillatory activity can be used to convert time series data (electrical current fluctuating over time) to frequency domain data (the frequencies that represent these fluctuations). By separating the signal into different frequency "bands" (i.e., delta, theta, alpha, beta, and gamma, representing slower to faster signals), different cognitive and affective processes can be monitored [29]. The most interesting finding with respect to cognitive load measurements is that alpha activity suppression (decrease in power of frequencies oscillating between 8 and 12 Hz at parietal regions) has found to be associated with increasing task difficulty and load across a wide variety of tasks [30,31].

The second approach to analyze EEG data is to look at the event-related potentials, representing the changes in mean voltage preceding or following a stimulus or action of interest (hearing a sound or pressing a button). By averaging over several trial repetitions, this analysis is focuses on the specific stimulus-related activity and decreases the impact of any activity that is not related or within the time window of the occurrence of the event (which increases the signal-to-noise (SNR) ratio) [32]. Interestingly enough, some ERP components can reflect the extent to which cognitive function and sensory processing are affected by mental workload. The amplitudes of ERP components such as N1, N2, P2 and P3 are expected to be reduced when a primary task becomes more demanding and workload increases [31,33]. An auditory oddball paradigm, in which sounds of different frequencies are presented, allows us to study the high demand on general processing resources reflected in the ERP components. With an irrelevant-probe technique this can be done in a non-intrusive way, without interference on the task flow [34,35,36]. Contrary to the standard ERP design, in this technique the ERP-eliciting stimuli (sounds) are presented without requiring participants to actively attend or respond to them, thus not co-varying with task demands.

1.3 The current study

As mentioned before, we wanted to focus on cognitive load detection in the context of personalized assembly work using the EEG method. More precisely, the main goal is the collection and in-depth analysis of both performance measures, subjective measures and psychophysiological measures, in a highly controlled lab context that overcomes some of the issues that previous studies had to deal with. The main difference with previous work is that next to the low and high load condition, we also introduced a condition in which cognitive overload was induced. In the current study, load was induced in an experimental setting by manipulating complexity levels of a set of Tangram tasks combined with working memory load (i.e., remembering visual stimuli). The cognitive overload condition included the most difficult Tangram puzzles and the greatest amount of stimuli to remember. We expected that the majority of the participants would not be able to succeed in these tasks and that this would be accompanied by feelings of despair, giving up and being discouraged in completing the task. Also, the manipulation with Tangram puzzles was used in order to have a representative task for assembly performance, which requires similar spatial
Another additive value in this study is that all three load conditions have an equal length of duration, keeping the data balanced for statistical comparisons. All conditions lasted for 10 minutes, which is a substantial amount of time to be able to measure cognitive load. Additionally, baselining was carefully conducted with 4 minute measures in rest state before and after the experimental block with load conditions. Finally, this study had a multimodal approach with additional sensors in order to explore other potential and less studied cognitive load markers (i.e., heart rate, skin temperature, galvanic skin response, electro-ocular activity, motion analysis, facial video analysis). Also, additional EEG features such as band power activity of other band frequencies, other event-related components, the time-frequency spectrogram, spectral entropy, individual alpha peak frequency, and auto-correlation can be explored in this dataset. All these features are beyond the scope of this manuscript and will be analyzed in the future.

With this optimized research design, we aimed to investigate the following hypotheses:

H1: The induced cognitive load by manipulation of complexity levels of the Tangram task is also reflected in subjective ratings of mental investment. The more complex the task, the more mental investment will be reported.

H2: Task performance on the Tangram task will be reduced in the cognitive overload condition, compared to the high and low load condition.

H3: Alpha power is decreased at parietal electrode sites when performing Tangram tasks that induce a cognitive overload compared to Tangram tasks that induce low or high load.

H4: The auditory processing of sounds presented during the performance of the Tangram task that induces cognitive overload can be reflected in a decreased N2 amplitude, when compared to the Tangram tasks that induce low or high load.

2 Method

2.1 Participants

This research got the approval of the ethics committee of the Faculty of Political and Social Sciences at Ghent University. In addition, all participants read and agreed to sign an informed consent with information about the procedure, purpose, voluntary participation, right to decline, access and storage of data.

In this study, 46 participants aged between 19 and 40 years old (M = 25.8, SD = 4.19) were recruited based on a questionnaire inquiring education, hair type, and other requirements via different social media channels (Facebook, the channel of the public library and the channel of the University).

Each session had a duration of approximately 90 minutes. We strived for a more or less equal number of male (N=21) and female (N=25) participants. The participants differed somewhat regarding their background in education: 11 participants had secondary education as highest degree, 6 participants completed a professional
bachelor, 4 participants completed an academic bachelor, 24 participants completed an academic master, and 1 participant owned a PhD as highest degree.

To control for prior experience and knowledge, participants were first asked about their experience with Tangram puzzles. Most participants (67 percent) indicated to be rather inexperienced regarding the Tangram task to be conducted in the experiment, while 11 percent were neutral and 22 percent indicated to have had some amount of experience with Tangram puzzles. Additionally, a spatial ability test was conducted with an adapted version of the Revised Minnesota Task Load Index [37]. Only 20 of the total of 64 questions were included, still covering the entire difficulty range. The histogram in Figure 1 shows the results on the spatial ability test, indicating a desired variance.

Fig. 1. Histogram of participants’ average score on the spatial intelligence test (max score 20).

2.2 Research design & procedure

Design. In this experiment, a within-subjects design was used in which each participant was exposed to all experimental conditions (i.e., three levels of cognitive load: respectively a low, high, and overload level). As mentioned before, the length of the Tangram task was kept equal for all experimental conditions (i.e., 10 minutes). Thus, the experiment consisted of three phases for which a counterbalanced design, with 6 possible orders, was used to exclude possible learning effects and order effects. Baselines were measured before the first and after the last experimental phase. The independent variable was the induced cognitive load.

Procedure. At first, participants filled in the informed consent and the pretest measuring their spatial ability. Next, the testing equipment was prepared (i.e., external electrodes on mastoids and the EEG set). After the set-up, each participant got detailed instructions about the experimental procedure in a systematic way and could try two puzzles in a test trial. Before starting the first experimental phase, a resting state measurement was conducted for the baseline. Participants subsequently opened and closed their eyes, each for 2 minutes. Next, the main experimental phase started in which participants spent 10 minutes in each condition. After each condition, participants completed a one-page questionnaire gauging perceived load, perceived affective states, and memory of visual stimuli. Finally, after completing all the experimental conditions, a post baseline measurement was conducted. The final step
consisted of a participant briefing and the clean-up. Figure 2 shows the experimental setting when performing the Tangram task.

![Fig. 2. Experimental setting when performing the Tangram task.](image)

### 2.3 Materials & questionnaires

**Materials.** The aim of the experimental design was to induce different levels of cognitive load, which should allow us to identify physiological parameters that are explanatory for cognitive load. Because each method of inducing cognitive load may have different shortcomings, a combination of different methods was employed (see Table 1).

The first method to induce cognitive load was by manipulating the complexity of the task. In the low load phase, participants assembled a Tangram puzzle of which the contours of each of the seven pieces were individually visible. In the high load phase, three pairs of two pieces touched each other, so only the surrounding contour of the pair was visible. In the overload phase all seven pieces touched each other, which created only one surrounding contour, and making it even more challenging to find the correct assembly. Two different versions were created for each load phase in which the order of the Tangram puzzles was randomly shuffled.

In order to induce additional cognitive load, the participants’ working memory was addressed by asking them to remember visual stimuli simultaneously while performing the assembly task. Participants were asked to write down the stimuli they remembered after each phase. Two different kinds of stimuli were alternately presented: pictures representing a tool that is typically used in industry (such as a safety helmet, a conveyor belt or a drilling machine) or a two-digit number. The number of stimuli that had to be remembered differed for each phase. During the low load phase, two pictures and two numbers were presented. During the high load phase, three pictures and three numbers were presented. And finally, five pictures and five numbers had to be remembered in the overload phase.

The third way to experimentally vary the level of cognitive load was to include background sounds and noise. For generalization to real-life assembly work, ambient factory floor sounds were played in the background. Additionally, two different sounds that differed in frequency were presented for the ERP analysis. About every 5 seconds (with some jitter to avoid predictability and rhythmic effects) a beep tone was played. 80 percent of these beeps were standard sounds with a low tone, while 20 percent were
deviant oddballs with a higher pitch. This manipulation allowed us to study sensory processing of the sounds under different levels of cognitive load.

Table 1. Methods for inducing three different levels of cognitive load.

<table>
<thead>
<tr>
<th>Low cognitive load</th>
<th>High cognitive load</th>
<th>Cognitive overload</th>
</tr>
</thead>
<tbody>
<tr>
<td>For each condition participants perform a series of Tangram tasks for 10 minutes, built up analogously as the examples below.</td>
<td>Three pairs of pieces have touching sides. The contour of the seventh piece is visible.</td>
<td>All pieces can have multiple touching sides.</td>
</tr>
<tr>
<td>The contours of all seven pieces are each individually visible (no touching sides).</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Visual stimuli that are alternately shown on a computer screen in front of the participant have to be remembered while performing the Tangram task.

Performance. Tangram task performance was measured by the number and percentage of correctly assembled puzzles and the percentage of remembered visual stimuli.

Questionnaires. The subjective experience of cognitive load was measured by a continuous scale (ranging from 0 to 100) and a Likert scale (ranging from 1 to 7). Based on an adapted version of the NASA TLX questionnaire task complexity and mental investment were inquired [38].
2.4 Apparatus & analysis

The EEG was acquired with a Biosemi ActiveTwo measurement system (BioSemi, Amsterdam, Netherlands), using 64 Ag-AgCl scalp electrodes attached to a standard international 10–20 system cap. Two additional external electrodes were attached to the left and right mastoids, which were used for offline re-referencing. Signals were amplified and digitized with a sampling rate of 1024 Hz. Triggers were sent through a serial port via Psychopy, an open-source application for a wide range of neuroscience, psychology and psychophysics experiments, written in Python language [39]. The recording computer received these triggers for the start of both baselines, all experimental conditions and every trial a standard or oddball sound was presented.

EEG analysis was performed in Python with MNE, an open-source Python software for exploring, visualizing, and analyzing human neurophysiological data, and custom-made code [40,41]. The raw EEG data preprocessing included re-referencing to the mastoid channels, interpolation of bad channels, and a bandpass filter with a high cut-off frequency of 1 Hz and a low cut-off frequency of 45 Hz to eliminate movements and electric noise. Regarding the future of real-time measurement, no manual artefact removal was done. The preprocessed data was also normalized by subtracting the average baseline activity, measured when participants relaxed with their eyes open for 2 minutes. Finally, for ease of analyzing purposes and processing speed, data was downsampled to 100 Hz.

Next, the pre-processed signal was transformed to the frequency domain with Fourier Transform for power analysis (focus on alpha oscillations). The power spectral density (PSD) was computed using Welch’s method [42]. The Python function scipy.signal.welch computed an estimate of the PSD by averaging consecutive Fourier transform of small windows of the signal (segments of 2 seconds) without overlapping, resulting in a frequency resolution of 0.50 Hz. Absolute alpha bandpower was calculated by taking the absolute mean of the power for the frequency band within its range of 8 to 12 Hz.

Finally, the preprocessed signal was kept in the time domain for the analysis of the event-related component (i.e., the amplitude of the N2 component, negative peak). All standard sound and oddball sound trials were epoched with a time window of [-200,500], and combined in an overall value at electrodes C3 and C4 at the central region.

3 Results

3.1 Self-reported cognitive load & task performance

Task complexity. The experimental manipulation in terms of complexity was as desired. Participants indicated the low load condition as the least complex and the overload condition as the most complex, with the high load condition in between, F(2,84) = 172.04, p < .001, ηp² = .79 (see Table 2, Figure 3A). No significant correlation was established between spatial intelligence of the participants and how complex they perceived the task, r = -.11, p = .19.
Table 2. Comparison of all conditions with pairwise t-tests for task complexity ratings (p-adjusted Holm).

<table>
<thead>
<tr>
<th>Cond A</th>
<th>Cond B</th>
<th>T</th>
<th>p</th>
<th>Hedges g</th>
</tr>
</thead>
<tbody>
<tr>
<td>high</td>
<td>low</td>
<td>10.836</td>
<td>&lt;.001</td>
<td>2.173</td>
</tr>
<tr>
<td>high</td>
<td>overload</td>
<td>-6.153</td>
<td>&lt;.001</td>
<td>-1.318</td>
</tr>
<tr>
<td>low</td>
<td>overload</td>
<td>-21.514</td>
<td>&lt;.001</td>
<td>-4.075</td>
</tr>
</tbody>
</table>

**Mental investment.** The results indicate that there is a significant main effect for the different conditions on mental investment measured on the continuous scale, F(2,90) = 196.62, p < .001, η² = .81 (see Table 3, Figure 3B). The more load that was induced in the Tangram task, the more mental investment participants reported. One remark, the first three participants didn’t respond to this question. Also, there is no significant correlation between participants’ spatial intelligence and their experienced mental investment during the task, r = -.13, p = .13.

Table 3. Comparison of all conditions with pairwise t-tests for mental investment ratings (p-adjusted Holm).

<table>
<thead>
<tr>
<th>Cond A</th>
<th>Cond B</th>
<th>T</th>
<th>p</th>
<th>Hedges g</th>
</tr>
</thead>
<tbody>
<tr>
<td>high</td>
<td>low</td>
<td>9.218</td>
<td>&lt;.001</td>
<td>1.761</td>
</tr>
<tr>
<td>high</td>
<td>overload</td>
<td>-6.641</td>
<td>&lt;.001</td>
<td>-1.340</td>
</tr>
<tr>
<td>low</td>
<td>overload</td>
<td>-13.539</td>
<td>&lt;.001</td>
<td>-3.078</td>
</tr>
</tbody>
</table>

**Fig. 3.** A) Task complexity score (max score 7) and B) Mental investment score (max score 100) rated by the participants after each experimental condition.
**Task performance.** As expected, results show that the amount of correctly assembled Tangram puzzles significantly differs across conditions, $F(2,90) = 539, p < .001, \eta^2 = .92$ (see Table 4, Figure 4A). A smaller amount of Tangram puzzles were correctly assembled with the increasing task complexity. There is a significant correlation between the spatial intelligence and the number of correctly assembled Tangram puzzles ($r = .40, p < .001$). In a similar way, the proportion of remembered stimuli decreased with increasing complexity, $F(2,90) = 51.68, p < .001, \eta^2 = .54$ (see Table 5, Figure 4B).

**Table 4.** Comparison of all conditions with pairwise t-tests for performance Tangram task results ($p$-adjusted Holm).

<table>
<thead>
<tr>
<th>Cond A</th>
<th>Cond B</th>
<th>T</th>
<th>p</th>
<th>Hedges g</th>
</tr>
</thead>
<tbody>
<tr>
<td>low</td>
<td>high</td>
<td>20.680</td>
<td>&lt;.001</td>
<td>2.849</td>
</tr>
<tr>
<td>low</td>
<td>overload</td>
<td>34.215</td>
<td>&lt;.001</td>
<td>7.797</td>
</tr>
<tr>
<td>high</td>
<td>overload</td>
<td>10.858</td>
<td>&lt;.001</td>
<td>2.416</td>
</tr>
</tbody>
</table>

**Table 5.** Comparison of all conditions with pairwise t-tests for performance Memory task results ($p$-adjusted Holm).

<table>
<thead>
<tr>
<th>Cond A</th>
<th>Cond B</th>
<th>T</th>
<th>p</th>
<th>Hedges g</th>
</tr>
</thead>
<tbody>
<tr>
<td>high</td>
<td>low</td>
<td>-3.812</td>
<td>&lt;.001</td>
<td>-0.827</td>
</tr>
<tr>
<td>high</td>
<td>overload</td>
<td>4.596</td>
<td>&lt;.001</td>
<td>0.894</td>
</tr>
<tr>
<td>low</td>
<td>overload</td>
<td>10.652</td>
<td>&lt;.001</td>
<td>1.881</td>
</tr>
</tbody>
</table>

**Fig. 4.** A) Amount of correctly assembled Tangram puzzles and B) percentage of remembered visual stimuli in experimental load conditions.
3.2 Alpha power

**Absolute mean power.** As expected, alpha activity in the region of interest (i.e., the four selected electrodes in the parietal region: Pz, POz, P1 and P2) differed between conditions in a one-way ANOVA repeated measures test, $F(4,176) = 45.12, p < .001, \eta^2_p = .51$. Table 6 summarizes the pairwise t-tests and Holm adjusted p-values. Both baseline conditions showed more alpha power activity compared to the cognitive load conditions. More importantly, alpha power differed between load conditions, $F(2,88) = 4.70, p = 0.01, \eta^2_p = .07$. A lower amount of alpha power in parietal electrodes was found for the overload condition when compared to the high load condition, and similarly when compared to the low load condition. High load condition also showed a lower amount of alpha power when compared to the low load condition (see Figure 5).

**Table 6.** Comparison of all conditions with pairwise t-tests for alpha activity results ($p$-adjusted Holm).

<table>
<thead>
<tr>
<th>Cond A</th>
<th>Cond B</th>
<th>$T$</th>
<th>$p$</th>
<th>Hedges g</th>
</tr>
</thead>
<tbody>
<tr>
<td>base post</td>
<td>base pre</td>
<td>10.997</td>
<td>&lt;.001</td>
<td>0.217</td>
</tr>
<tr>
<td>base post</td>
<td>overload</td>
<td>16.359</td>
<td>&lt;.001</td>
<td>1.289</td>
</tr>
<tr>
<td>base post</td>
<td>low</td>
<td>14.386</td>
<td>&lt;.001</td>
<td>1.176</td>
</tr>
<tr>
<td>base post</td>
<td>high</td>
<td>16.246</td>
<td>&lt;.001</td>
<td>1.220</td>
</tr>
<tr>
<td>base pre</td>
<td>overload</td>
<td>11.962</td>
<td>&lt;.001</td>
<td>0.995</td>
</tr>
<tr>
<td>base pre</td>
<td>low</td>
<td>10.227</td>
<td>&lt;.001</td>
<td>0.886</td>
</tr>
<tr>
<td>base pre</td>
<td>high</td>
<td>11.724</td>
<td>&lt;.001</td>
<td>0.933</td>
</tr>
<tr>
<td>overload</td>
<td>low</td>
<td>-5.473</td>
<td>&lt;.001</td>
<td>-0.140</td>
</tr>
<tr>
<td>overload</td>
<td>high</td>
<td>-2.695</td>
<td>0.008</td>
<td>-0.057</td>
</tr>
<tr>
<td>low</td>
<td>high</td>
<td>3.499</td>
<td>0.001</td>
<td>0.080</td>
</tr>
</tbody>
</table>
Fig. 5. A) Alpha log power activity results for all conditions and B) experimental load conditions only.

Differences in alpha activity power were also observed when looking closer at the separate electrodes, $F(3,132) = 38.59, p = 0.00, \eta_p^2 = .47$. Especially the POz electrode showed to have greater alpha power overall, without taking load conditions into account (see Table 7 and Figure 6). There were no differences between electrodes in predicting the decreased alpha power effect in the different conditions, $F(6,264) = 1.25, p = 0.28$.

Table 7. Comparison of all electrodes with pairwise t-tests for alpha activity results ($p$-adjusted Holm).

<table>
<thead>
<tr>
<th>Electr A</th>
<th>Electr B</th>
<th>T</th>
<th>p</th>
<th>Hedges g</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1</td>
<td>P2</td>
<td>-1.715</td>
<td>0.089</td>
<td>-0.071</td>
</tr>
<tr>
<td>P1</td>
<td>POz</td>
<td>-24.151</td>
<td>&lt;.001</td>
<td>-0.587</td>
</tr>
<tr>
<td>P1</td>
<td>Pz</td>
<td>-11.141</td>
<td>&lt;.001</td>
<td>-0.135</td>
</tr>
<tr>
<td>P2</td>
<td>POz</td>
<td>-10.936</td>
<td>&lt;.001</td>
<td>-0.488</td>
</tr>
<tr>
<td>P2</td>
<td>Pz</td>
<td>-1.315</td>
<td>0.191</td>
<td>-0.057</td>
</tr>
<tr>
<td>POz</td>
<td>Pz</td>
<td>17.729</td>
<td>&lt;.001</td>
<td>0.462</td>
</tr>
</tbody>
</table>
3.3 Auditory event-related potentials

The auditory processing of standard and oddball sounds is measured by the N2 component, averaged at central electrode sites C3 and C4. There were no significant differences found between the three load conditions, $F(2,84) = 0.50$, $p = 0.61$ (see Figure 7 and Figure 8).

Fig. 6. Alpha log power activity results for experimental load conditions at four selected electrode sites in the parietal region (Pz, POz, P1, P2).

Fig. 7. ERP plot for experimental load conditions averaged for two electrodes at the central region (C3+C4).
3.4 Additional correlation analysis

Correlation analysis indicated that no correlation was found between Tangram performance and alpha log power activity in load conditions. But percentage of memory performance and alpha log power activity were significantly correlated, only in the high load condition, \( r = -0.37, p = .01 \). This indicated that only in the high load condition, reduced alpha power activity is correlated with a greater amount of visual stimuli participants could remember (see Table 8, Figure 9).

Table 8. Correlation analysis results for two performance features (amount of correctly assembled Tangram puzzles and percentage of visual stimuli remembered) with alpha log power activity in all experimental load conditions. Values in brackets indicate \( p \) value for each correlation.

<table>
<thead>
<tr>
<th>Performance feature</th>
<th>Low</th>
<th>High</th>
<th>Overload</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tangram</td>
<td>-0.13 (.41)</td>
<td>-0.16 (.28)</td>
<td>-0.11 (.49)</td>
</tr>
<tr>
<td>Memory (%)</td>
<td>-0.17 (.25)</td>
<td>-0.37 (.01)</td>
<td>0.01 (.93)</td>
</tr>
</tbody>
</table>
4 Discussion

The current study focused on the detection of cognitive load and overload in the context of personalized assembly work with psychophysiological sensors, performance measures, and subjective measures. We manipulated cognitive load in three conditions (i.e., low, high, overload) by creating different complexity levels of a dual task, which included a Tangram puzzle task and a working memory load (i.e., remembering visual stimuli). This task was performed with a large sample size and in a highly controlled lab context to overcome some of the methodological issues that previous studies had to deal with. The experimental design also allowed us to conduct considerable baseline measures and compare load conditions that lasted for an equal and substantial length of time. Finally, while this manuscript only focuses on EEG features, other sensors (i.e., heart rate, skin temperature, galvanic skin response, electro-ocular activity, motion analysis, facial video analysis) were implemented in the current study and will also be explored as potential cognitive load markers in the future.

Our successful manipulation of cognitive load was reflected in task performance and subjective rating results. First, the majority of the participants were not able to succeed on the dual task in the overload condition, assembling almost none of the puzzles and remembering only a small amount of the presented visual stimuli. Second, on subjective ratings they indicated that the task in the overload condition was the most complex and required the greatest amount of mental investment.

More importantly, results for the EEG features alpha power activity and auditory response (i.e., N2 amplitude) are partially in line with expectations. A greater amount of cognitive load was indicated by reduced alpha power activity at parietal electrodes, especially at the POz electrode. On the contrary, no significant effect was found on the auditory response. The reduced alpha power effect found in the current study validates this EEG feature as a marker for estimating cognitive load, in line with previous research [10, 11, 31, 43, 44]. However, we expected there would be a greater effect on
alpha power in the overload condition when compared to the other load conditions. This could be due to the task being too complex and overwhelming, making participants give up and not staying motivated to invest mental effort and resources anymore. Participants confirmed that these puzzles were too difficult and some believed they were actually unsolvable, which was reflected in the nervous laughs and freeze reactions. Regarding the assembly work context, this could be reflected in dropout, bad quality and errors because operators are becoming apathetic to the task performance [19,45]. Motion analysis of the videos or additional EEG features that study the EEG signal over time (i.e., time frequency spectrogram, auto-correlation) could provide more insights.

Also, the small effect size in these findings indicates that the alpha power may be not sensitive enough for differentiating between different levels of cognitive load in real-time. The differentiation between conditions in resting state (i.e., pre and post baselines) and conditions that require mental effort (i.e., experimental load conditions) was more pronounced than the comparison amidst only load conditions. The baselines had a distinctly lower amount of alpha power when compared to the load conditions. Consequently, the real-time differentiation between cognitive load versus overload in an assembly work context is challenging, especially when using alpha power activity as a deciding marker.

The results regarding the sensory processing with ERP analysis could not validate the N2 amplitude as a marker for estimating cognitive load. The sensory processing of the presented sounds was similar in all load conditions. First of all, factory noise and the ERP-eliciting sounds were presented in order to create additional cognitive load and reflect the ambience of assembly work for all conditions. Because a dual task was already created for manipulation of cognitive load, also attending to these sounds would have been too difficult. That way we would not have been able to create a low load condition. Consequently, in this ERP paradigm participants did not have to actively attend to the standard and oddball sounds. The N2 amplitude was possibly not sensitive enough as a marker of cognitive load because it was not part of the task flow. Another possible feature for measuring the sensory processing without overt action or attention to the presented sensory stimuli is the mismatch negativity (MMN) component and could be explored in the future [34]. The MMN indicates the event-related response to sudden changes in auditory stimuli. Finally, the low amount of trial repetitions is another possible confound in our ERP paradigm. Even though conditions lasted for 10 minutes, presenting a sound every second would have been too interfering with the primary task. Additionally, because an ERP design requires enough repetitive trials for filtering out noise and obtaining reliable conclusions, it may even be unsuitable in real-time assessment of cognitive load [11].

With regard to the finding that alpha power activity did not correlate with performance measures, we can remark that this spectral power feature may not be sensitive enough to discriminate on an aggregated level. A lot of information is lost because values for alpha power activity are averaged for the whole duration of the condition. The investigation of lower (8-10Hz) and upper (10-12Hz) alpha bands could give more detailed insights on specific frequency effects that are not distinct when only looking at the broad alpha range [44, 46].

As previously mentioned, we would like to explore other potential markers for cognitive load in our EEG dataset. Exploring the power activity in time and auto-
correlation analysis of the raw data could possibly unravel more in-depth insights about fluctuations or recurring patterns in the signal, especially when synchronized with video motion analysis. Future research will also focus on other aggregated EEG features such as the alpha peak power frequency, the frequency bin where maximum power (i.e., local peak) is found within the 8-12 Hz range [44, 47]. We expect to find lower peak frequency values in the overload condition, indicating “less integrated and interconnected feedback loops among brain areas” [47, p.419] compared to lower load conditions. This is also reflected in the deteriorated task performance results. Another approach for a more nuanced analysis of our results is the use of advanced machine learning techniques in order to classify ‘overload’ within conditions based on the data from multiple sensors.

We can conclude that our results encourage to measure and evaluate EEG features for estimating cognitive load in a highly controlled lab setting with experimental design. The results from the current study validated alpha power activity as a potential marker for estimating cognitive load, while the auditory N2 response failed to differentiate between load conditions. Our future research focusing on the real-time measurement of cognitive load will aim to validate these findings with a wearable EEG headset that is applicable in the assembly work context. Especially alpha power activity at the POz electrode will be considered as a potential marker of load. However, researchers should be aware that aggregated EEG features (i.e., alpha power and N2 amplitude) at group level are not sensitive enough for detecting cognitive overload. Other features analyzed over time, the use of a longitudinal design and training a statistical load model with data from several individual sessions may be preferred for a more nuanced approach in the exploration of other potential markers for estimating cognitive (over)load.
References


43. Jensen, O., Gelfand, J., Kounios, J., Lisman, J.E.: Oscillations in the Alpha Band (9-12 Hz) Increase with Memory Load during Retention in a Short-term Memory Task. Cerebral Cortex 12, 8 (2002). https://doi.org/10.1093/cercor/12.8.877


