**Methods**

**Study Design and Participants**

The current data were collected as part of a larger research project EEG Post-Covid (EPOC), which investigates neurophysiological parameters identified from neuropsychological paradigms using a high-resolution stationary laboratory EEG to reflect cognitive impairments and fatigue. The primary goal of the EPOC study is to find EEG parameters that can serve as neurophysiological markers for progression- and therapy-evaluation concerning cognitive functions in PCS.

Participants for the EPOC study were recruited from COVIDOM, a population-based, prospective multi-centre study to investigate PCS within the German National Pandemic Cohort Network (NAPKON). COVIDOM participants had been recruited through public health authorities in Kiel, Berlin, and Würzburg. Patients were assessed between November 15, 2020, and September 19, 2021, at University Medical Center Schleswig-Holstein, Campus Kiel, and University Hospital Würzburg in Germany (Bahmer et al., 2022; Horn et al., 2021; Schons et al., 2022).

The participants were included based on the following criteria: A polymerase chain reaction (PCR) confirmed SARS-CoV-2 infection at least 6 months before study visit, a primary residence in one of the three study regions, age ≥ 18 years at the time of recruitment (Berlin) or infection (Würzburg, Kiel). Participants with acute reinfection of SARS-CoV-2 at the time of the scheduled study visit were excluded (Horn et al., 2021).

In the EPOC study, a subset of individuals from Schleswig-Holstein who participated in COVIDOM was selected, constituting of those with PCS and a control group without PCS.

As EPOC is still ongoing at the time of writing, the analysis was conducted based on a preliminary subset of 79 participants (mean age 48.52, range 22–78, female 48, male 31, diverse 0, years of education mean 15.27 years min 9 to 24 years) with PCS (49 participants, age mean 50.29 years min 22–78, F=32 M=17, education: mean 15.04 9–23) and without PCS (30 participants, age mean 45.63 years min 22–77, f=16,m=14 d 0, education mean 15.63 min 10–24). The study was conducted at the University Medical Center Schleswig-Holstein (UKSH), Campus Kiel. Participants did not receive payment/financial compensation for their participation. Transportation and parking costs were reimbursed.

**Ethics statement**

The study was approved by the Ethics Committee of the medical faculty of the Christian-Albrechts-University of Kiel, Germany (record identification: D 446/23). In accordance with the Declaration of Helsinki, informed written consent was obtained from all participants.

**Procedure/Study Design**

In brief, the experiment consisted of neuropsychological tests, assessing cognitive domains such as working memory, attention, preprocessing speed, cognitive flexibility, executive functions, and multisensory integration, EEG recordings, and questionnaires assessing fatigue, sleep quality, depression, and anxiety.

Participants first filled out a questionnaire on demographic data (e.g., age, education) and psychological and neurological conditions, followed by neuropsychological testing to assess cognitive domains such as working memory, attention, preprocessing speed, cognitive flexibility, executive functions, and multisensory integration. The first test administered was the Trial Marking Test (TMT). Following this, the EEG cap was placed, and participants completed a series of other neuropsychological tests, starting with the redundant target effect (RTE), followed by an oddball paradigm, an n-back task, and lastly the psychomotor vigilance task (PVT). Electroencephalographic activity was recorded continuously throughout these tests. Finally, resting state was measured, 5 minutes with eyes open and 5 minutes with eyes closed. During EEG recordings, participants were seated comfortably and instructed to minimize movement, and to focus on a fixation cross displayed on the screen in front of them to reduce eye movements, while the light was turned off. After completing the resting state measurement, the EEG cap was removed, and participants filled out three questionnaires assessing fatigue (FACIT-F), sleep quality (PSQI), depression (HADS-D), and anxiety (HADS-A). The participants got then the chance to wash their scalp/hair. In all, the experiment took up to 3 hours.

Since this thesis focuses on behavioral data obtained from the TMT, n-back and PVT, as well as EEG resting state data, and data from the questionnaires, the RTE, and the oddball task will not be further explained. In addition, the MoCA score was measured in the previous COVIDOM study.

**Cognitive tasks**

**MoCA**

The Montreal Cognitive Assessment (MoCA; Nasreddine et al., 2005) is a widely used, validated screening tool originally designed to detect mild cognitive impairment (MCI) (Hauffe, 2024). It assesses several cognitive domains, including visuospatial skills/ability, executive function, naming, memory (short-term and delayed recall), working memory, attention and concentration, language, abstraction, and orientation (Freitas, Simões, Alves & Santana, 2013; Hobson, 2015; Kang et al., 2018; Nasreddine et al., 2005). The MoCA has a total possible score of 30 points, with a score of ≥ 26 considered normal (Nasreddine et al., 2005). Cognitive performance on the MoCA is influenced by sociodemographic factors such as age and education (Kang et al., 2018; Larouche et al., 2016). To account for educational background the MoCA test manual specifies that one additional point is added for individual with ≤ 12 years of formal education, allowing for a maximum score of 30 points (Nasreddine et al., 2005). Additionally, to address variations in performance related to age, normative data for the MoCA are available for precise interpretation of scores (Larouche et al., 2016).

**TMT Part A and B**

Originally, developed as part of the Army Individual Test Battery (AITB) in 1944, the Trail Marking Test (TMT) was later integrated into the Halstead-Reitan Battery (Reitan & Wolfson, 1985; Tombaugh, 2004). It is now one of the most popular and widely used neuropsychological assessments, included in most test batteries (Tombaugh, 2004). Its widespread use is supported by strong evidence of its validity (Arbuthnott & Frank, 2000; Sánchez-Cubillo et al., 2009). The TMT assesses cognitive processing speed and executive functioning (Lezak, 1995; Mitrushina et al., 2005; Reitan, 1992; Sánchez-Cubillo et al., 2009; Strauss et al., 2016; Tombaugh, 2004), as well as visual search/scanning, and mental flexibility (Sánchez-Cubillo et al., 2009; Tombaugh, 2004).

The TMT consists of two parts: Part A (TMT-A), a number-connection task, and Part B (TMT-B), a number-letter alternation task. Both parts were administered (in this study) according to the guidelines provided by Strauss et al. (2006).

In TMT-A, participants were instructed to connect consecutively numbered circles from 1 to 25 on an A4 page by drawing lines between them, aiming to complete the task as quickly and accurately as possible. In TMT-B, the task becomes more complex (Gaudino, Geisler & Squires, 1995). Participants were instructed to draw lines alternating between numbered circles from 1 to 13 and lettered circles from A to L in sequential order (e.g, 1 to A, to 2, to B, etc.) on an A4 page. The aim, again, was to complete the task as quickly and accurately as possible.

The administration of the TMT began with TMT-A, followed by TMT-B. For each part, participants were first given an example to familiarize themselves with the task. After completing the example, they proceeded to the actual test. If participants made a mistake, the experimenter immediately pointed it out, and the participant was required to correct it before continuing. The experimenter timed each part, with the time of completion for each part representing its direct score. In addition to the direct scores, the difference between TMT-B and TMT-A (TMT-B – TMT-A) was calculated.

**N-Back Part A and B**

The n-back task (Kirchner, 1958) has become a widely used tool in neuroscience for assessing working memory (Jaeggi, Buschkuehl, Perrig &Meier, 2010; Pelegrina et al., 2015). N-back tasks are continuous-recognition measures, that present sequences of stimuli (Kane, Conway, Miura & Colflesh, 2007). In these tasks, participants must determine whether a given stimulus matches one that was presented “n” trials before. Reliability of the n-back task varies across studies, with more complex levels (e.g., 2-back, 3-back) generally yielding higher reliability coefficients (Jaeggi et al., 2010; Pelegrina et al., 2015).

In this study, participants completed two blocks of the n-back task: A 1-back task followed by a 2-back task, with a pause between blocks during which the instructor provided additional instructions before participants proceeded to the second block. The task was programmed using … and presented on a 27-inch computer screen.

In both conditions, participants were shown a series of 60 linguistic stimuli, consisting of 16 different consonants (B, C, D, F, G, H, J, K, M, Q, R, S, T, V, X, Z) presented individually in the center of the screen. Each block contained 20 target trials and 40 non-target trials. A trial began with a 250 ms fixation period (a red dot was shown on screen, for the participant to fixate), followed by a 150 ms black screen. The stimulus letter then appeared for 500 ms, succeeded by a variable inter-trial interval of 180 to 220 ms (black screen). Total trial duration ranged from 1080 to 1120 ms.

For the 1-back task, participants were instructed to press the spacebar when the current letter matched the previous one. For example, in the sequence “B, C, C, D,” participants were supposed to respond to the second “C” as it matches the previous letter. In the 2-back task, they were instructed to press the spacebar when the current letter matched the letter presented two trials prior. For instance, in the sequence “B, B, D, F, D,” participants should press the spacebar when the second “D” occurred, as it matches the letter presented two trials before. The response window was limited to the 500 ms stimulus presentation period. Reaction time, hits, misses, and false alarms were recorded. In total, the experiment took around 5 minutes. The light was turned off during the experiment.

**PVT**

The Psychomotor Vigilance Task (PVT) is a widely used reaction time test developed in 1985 to assess sustained attention, particularly in contexts involving fatigue and sleep deprivation (Drummond et al., 2005). It has been shown to be sensitive to sleepiness in clinical and experimental settings (Molina, Sanabria, Jung & Correa, 2019).

The key feature of the PVT is its monotonous and unpredictable target presentation which makes participants highly prone to lapses of attention. This unpredictability minimizes learning effects, ensuring that performance remains largely independent of prior abilities and experience (Basner and Dinges, 2011) (stolen from Molina et al., 2019). Reaction time measured by the PVT has been linked to cognitive function in both healthy subjects and patients, supporting its validity as an assessment tool (Jakobsen, Sorensen, Rask, Jensen & Kondrup, 2011).

This study employed a 5-minute version of the PVT, which has been established as a valid alternative to the traditional 10-minute PVT-192 (for assessing fatigue) (Lamond et al., 2008).

The dynamic stimulus appeared as a red number, counting up in milliseconds, representing the participant’s reaction time. Participants were instructed to respond immediately, when the stimulus/red number occurred, by pressing the spacebar. Between trials, a white fixation cross was displayed on a black screen for a variable interval ranging from 2 to 10 seconds. Participants were required to maintain their gaze on this fixation cross. Each trial concluded when a response was made. Following each response, the participant’s reaction time was displayed on the screen for 50 ms as feedback before the next trial began.

After receiving instructions, participants underwent a training block of 8 trials to familiarize themselves with the task. Following the training, participants proceeded to the main experiment, which consisted of 50 stimulus presentations. The light was turned off during the experiment.

**Resting state**

Do I even need to write something here?

Maybe to investigate delta and beta frequency eyes closed condition and why?

**Questionnaires**

**PSQI**

The Pittsburgh Sleep Quality Index (PSQI), developed by Buysse, Reynolds, Monk, Berman, and Kupfer in 1988 is a self-rated questionnaire that assesses sleep quality and disturbance over the past month. The PSQI is the most commonly used sleep health assessment tool in both clinical and research settings. Its reliability and validity have been consistently demonstrated in multiple studies (e.g. Carpenter & Andrykowski, 1998; Manzar et al., 2018; Mollayeva et al., 2016). The questionnaire consists of 24 items in total, 19 of which are self-reported by the patient and 5 of which require input from a room or bed partner. Only the 19 self-reported items are used for the quantitative evaluation of sleep quality, as perceived by the patient (Buysse et al., 1989; Manzar et al., 2018). The response formats across the items vary, including the recording of usual bed and wake times, number of hours slept, minutes taken to fall asleep, as well as forced-choice Likert-type responses (Buysse et al., 1989). The items are categorized into seven components, which are sleep quality, sleep latency, sleep duration, habitual sleep efficiency, sleep disturbance, use of sleeping medications, and daytime dysfunction, for each component given a score. Together, these component scores generate a global sleep quality score ranging from 0 to 21, with scores >5 indicating poor sleep quality (Buysse et al., 1989; Hinz et al., 2017).

**HADS**

The self-assessment Hospital Anxiety and Depression Scale (HADS) was originally developed by Zigmond & Snaith in 1983 to identify the presence of anxiety and depression states among patients in non-psychiatric hospital clinics. HADS is an extensively used, reliable, and valid instrument to measure anxiety and depression, not only in psychiatric, and clinical patients (Herrmann, 1997) but in general populations (Bjelland, Dahl, Haug & Neckelmann, 2002; Herrero et al., 2003; Spinhoven et al., 1997). The questionnaire assesses anxiety and depression symptoms during the past week, excluding symptoms also related to physical disorders, e.g., headache, dizziness, or insomnia (Bjelland et al., 2002; Hinz & Braehler, 2011; Zigmond & Snaith, 1983). The scale consists of 14 items, divided into a 7-item anxiety (HADS-A), and a 7-item depression subscale (HADS-D). Both subscales are rated on a four-point Likert scale, giving subscale scores ranging from 0 to 21 (Zigmond & Snaith, 1983). There is no universally accepted cut-off score for the HADS (Herrero et al., 2003; Spinhoven et al., 1997). In this study, the cut-off point was set to eight, indicating elevated/caseness anxiety and depression for scores ≥8, following the recommendations by Zigmond and Snaith (1983), as well as Bjelland et al. (2002) and Herrero et al. (2003).

In this study the German Version (HADS-D) of the scale was used.

**FACIT-F**

The 13-item Functional Assessment of Chronic Illness Therapy (FACIT)-Fatigue Scale (FACIT-F; Version 4) was used to assess self-reported fatigue and its impact on daily activities and functions (Cella, Lai, Chang, Peterman & Slavin, 2002; Yellen et al., 1997) during the last 7 days. While it was originally developed for cancer-related fatigue (Cella et al., 2002; Yellen et al., 1997), it has been shown, that the FACIT-F is a reliable and valid measure of fatigue across various health conditions (Cella et al., 2002), making it a widely used tool for both clinical practice and research (Cella et al., 2022; Butt et al., 2013; Montan, Löwe, Cella, Mehnert & Hinz, 2018; Tinsley, Macklin, Korzenik & Sands, 2011). This questionnaire utilizes a five-point Likert scale, with total scores ranging from 0 (severe fatigue) to 52 (no fatigue). Based on general population data, scores ≤30 indicate clinically significant fatigue (Piper & Cella, 2010).

**Cluster Analysis**

The cluster analysis is an explorative statistical method used to organize objects, data points, or observations into homogeneous groups, known as clusters, based on similarities (Ketchen & Shook, 1996). The goal is to achieve high homogeneity within groups (intragroup homogeneity) and high heterogeneity between groups (intergroup heterogeneity) (Bacher, Pöge & Wenzig, 2010; Backhaus, Erichson, Gensler, Weiber & Weiber, 2011). In this study, the behavioral cognitive data from the PVT, TMT, MoCA, and n-back task will be utilized as cluster variables, aiming to identify two clusters that differ in their cognitive performance levels suggesting, that one group may perform better or worse than the other.

**Data Preprocessing.** All participant data was imported from a TSV file (participants.tsv) into R (using read.delim()), where preprocessing and analysis of the data was conducted. Specific variables of interest were selected and stored as subset including demographic information, cognitive test scores and clinical/questionnaire measures. The MoCA variable was converted to a binary variable: Scores ≤25 indicate cognitive impairment, while scores >25 indicate no impairment. Missing values were added for n-back task, if participant mentioned that they did not understand the task. Rows with missing values in cone of the relevant cognitive test variables (pvt….) were removed. Missing values in the relevant cognitive test variables (PVT reaction time, n-back miss 1, n-back miss 2, TMT a time, TMT b time (need to mention earlier)) were detected/checked is.na() and rows (participants) with missing values in key cognitive variables were removed using drop\_na(), to ensure complete cases for analysis. 9 rows have been deleted because of missing values, leaving the dataset with 70 participants. Winzorising was used to replace outliers by capping extreme values beyond 1.5 times the interquartile range (IQR). Outliers are identified and winsorized for the before mentioned cognitive variables. A function winsorize\_variable is defined to perform winsorization, replacing values beyond 1.5 \* IQR with the values at Q1 – 1.5 \* IQR or Q3 + 1.5 \* IQR. Winsorized versions of variables were created with \_w suffix. 4 outliers were detected for PVT reaction time and winsorized. 2 outliers for TMT\_a\_time. 4 outliers TMT\_b\_time. A custom function winsorize\_variable() was implemented to replace extreme values beyond 1.5 times the interquartile range (IQR) with the nearest non-outlier values (Q1 - 1.5 \* IQR or Q3 + 1.5 \* IQR). New variable TMT\_diff was calculated as the difference between TMT B and A (B-A).

**Variable transformation**

To account for the influence of age on cognitive performance, participants were divided into age groups, and z-scores were calculated within each group to adjust the data.

Two functions were defined to categorize participants into age groups. Participants were divided into four distinct age groups, 18-34 years (12 participants), 35-49 years (19 participants), 50-64 years (33 participants), and 65-80 years (6 participants). Age groups orientated from TMT norms. Why decided for this age groups? A separate categorization was used specifically for the TMT difference score. Also 4 age groups but 18-24 years (2 participants), 25-54 years (34 participants), 55-64 years (28 participants), and 65-80 years (6 participants). Why different age groups for TMT difference and all other variables?

After age groups were created, mean and standard deviation for each cognitive variable (PVT and TMTa, TMTb and TMT difference) were calculated within each age group. The function calculate\_z\_scores\_individual() was used to compute z-scores for each participant based on age group norms, adjusting for age-related differences in cognitive performance. Creating new variables with \_z suffix. Additionally n-back miss scores (miss 1 and miss 2) are standardized using the scale() function, creating new variables with \_s suffix. When looking at the n-back means in each age groups, no age related trend could be detected, therefore no z-score was calculated for the n-back values, instead was standardized. N-back miss scores were standardized using the scale() function to ensure comparability across variables.

The final cleaned and processed dataset was saved for further use as a file named “clean\_data.Rdata” in the cluster analysis. Saved as an R data file named clean\_data.Rdata.

**Cluster Analysis/hierarchical clustering.** A hierarchical cluster analysis was performed on the preprocessed test data to identify clusters among participants. Hierarchical cluster analysis was performed on the preprocessed cognitive test data to identify (potential) subgroups/clusters within the participant pool/among participants. Which packages were loaded? Tidyverse, dplyr, dendextend, ggplot2, gridExtra, purrr, vroom. This approach was employed to uncover/reveal patterns in cognitive performance of the participants. Cognitive variables were compared as winsorized, standardized, as well as the original scores were compared.

Only include if good reason to think they will define the clusters. First all cognitive variables, after reduced, why?

At first PVT reaction time, TMT A, TMT B, n-back miss 1 and n-back miss 2, MoCA were used as variables within the clustering. Explorative approach. Using n-back let to bad clustering results. Because everyone performed quite bad. Therefore n-back was excluded from the cluster analysis. TMT difference was not used, since it would double information (not independent). Too high correlation. Why did I choose the variables I chose? Where do I need to write that? Why not used MoCA?

As the hierarchical

The first step in the hierarchical cluster analysis process was to compute a distance matrix. In this approach Euclidean distance method was used Proximitätsmaß/Distanzmaß bevorzugt bei WARD. All values continuous numerical values that is why used euclidean. Tried different kind of linkage methods and than decided which one performed best, based on. Why? Ward methods was chosen as the algorithm. Specified linkage method via method argument. Dendrogram was build by plotting hierarchical cluster object with hclust. Created desired number of clusters. Cut\_mean <- cutree (hclust\_median, k = 2) In consideration of the research question and the two groups, withPCS and withoutPCS, a two-cluster solution was chosen. K-means/elbow measure suggested 4 clusters. Therefore a 4-cluster solution was also looked at. As validation, the 2-cluster solution was compared to their self repoRTEd group assignments (withPCS or withoutPCS. Do I here need to mention, how many people are in which cluster? Or is that already result? To visualize cluster on dendrogram abline function used. Stability tested, different proximity measures have been used. But with euclidean best result. Also different algorithms have been tested. Also non-hierarchical clustering was compared to the hierarchical clustering (k-means). Why did I do the analysis with ward and not with k-means? What was my decision there? For the selected number of cluster, three additional analyses were performed using the complete, single, and weighted-average linkage methods. The agreement(Übereinstimmung) was assessed using the adjusted Rand index (Hubert and Arabie, 1985). Change of algorithm and alteration of number (which numbers where tested?) of clusters was variated. For adjusted rand index: library(fossil) adjusted rand index calculated with rand.index function.

PVT reaction time, TMT a, TMT b, TMT b-a, MoCA, and n-back scores were included. Euclidean distance was used as a distance matrix, and Ward’s method was selected for clustering, as it is widely used in practice and known for its effectiveness in identifying distinct clusters (Backhaus et al., 2011). It is considered a reliable algorithm, provided that the variables are on a metric scale, are uncorrelated, and do not contain outliers (Wentura & Pospeschill, 2015). As stated earlier, cluster analysis is an explorative method used to identify patterns in data. However, in this study, the approach is only semi-exploratory, as the number of clusters to be generated was predetermined based on prior knowledge. A two-cluster solution was explored, as that aligns with the self-reported groups (with PCS, and without PCS), thereby allowing for good comparisons between the cluster solution and the self-reported groups. One could also consider this as a confirmatory cluster analysis (Bacher et al., 2010). The stability of the clusters was tested by comparing different proximity measures and algorithms using the adjusted Rand index (Hubert & Arabie, 1985). **Verwendete Kriterien zur Bestimmung der Clusteranzahl**

Inhaltlich, da zwei verschiedene Gruppen. Later 4 because of k-means

**DurchgefühRTE Stabilitätsprüfung**

Für ausgewählze Clusteranzahl noch drei weitere Analysen mit dem Complete-, Single- und Weighted-Average-Linkage gerechnet. Die Übereinstimmung wurde mittels des adjustieRTEn Randindex (Hubert und Arabie 1985) beuRTEilt.

Auch: Wechsel des Algorithmus und Veränderung Gruppenzahl (Ein Cluster bleibt gleich)

**DurchgefühRTE Validitätsprüfung**

Zur Validitätsprüfung wurde auf Variablen Z1, Z2 usw. zurückgeriffen

**Statistical analysis.** The two clusters were compared in several aspects. First, the two clusters were compared in their cognitive performance levels to validate whether significant differences exist between clusters. Clusters were then compared across demographic variables and results in questionnaires. Of particular interest was to examine how those two clusters differ from or align with the self-reported perception of cognitive performance level. To investigate differences between objective and subjective cognitive performance levels, comparisons occurred not only between two clusters but also within the clusters between the subjective groups with PCS and without PCS. Additionally, to maximize the insights from the cluster analysis, the with PCS groups in cluster 1 was compared to the with PCS group in cluster 2, and similarly for the without PCS groups. A t-test was used for these comparisons. Effect size and cohens d were also compared (need to check why)

The clusters were compared in several expects with each other. In demographical variables (sex, age, and years of education), in the used variables for cluster analysis. But also in their other cognitive variables (PVT, TMT, n-back, MoCa). Also results in the scores from questionnaires were compared. Not only were the two groups compared between each other, but also within comparison took place. WithPCS and withoutPCS within one cluster were compared. Also withPCS and withoutPCS were compared between clusters (that means, withPCS in Cluster 1 was compared to withPCS in Cluster 2 to clarify). All comparisons were tested by t-test. T-test robust to….. Data is not normal distributed. That was tested by… cat function was used.

The two clusters where compared

Alongside the comparisons of demographic, cognitive data, and questionnaire results, the clusters were also examined for their EEG resting state patterns.

**EEG Recording and Analysis**

For each group (withPCS and withoutPCS), 5 minutes of resting state with eyes open and 5 minutes of resting state with eyes closed were recorded using high-density EEG. Since the eyes-closed condition represents a simple, standardized procedure (Babiloni et al., 2016), it is the most commonly used (Babiloni et al., 2022) and will therefore be analyzed in this study to ensure comparability. EEG signals were recorded using a 128-channel EEG cap (128Ch Standard Brain Cap for actiCHamp Plus, Easycap GmbH, Wörthsee, Germany) with electrodes positioned in an equidistant layout, connected to an actiCHamp Plus Amplifier (Brain Products GmbH, Gilching, Germany).

The sampling rate was 1000 Hz with an amplitude resolution of 0.1 µV. Electrolyte gel was applied to improve conductivity between skin and electrodes, ensuring impedances remained below 20 kΩ. Eye movements and changes in the resting potential of the retina (EOG activity) were monitored using two EOG electrodes placed below each eye, with impedances also maintained below 20 kΩ. In addition, a ground electrode was positioned on the forehead, and a reference electrode was positioned on the tip of the nose. Impedances for both the reference and ground electrode were kept below 5 kΩ.

**Preprocessing**

Data preprocessing/analysis was performed using the FieldTrip toolbox (Fieldtrip-20240504; Oostenveld, Fries, Maris & Schoffelen, 2011) and the EEGLab toolbox (v2024.0; Delorme & Makeig, 2004) in Matlab (v24.1.0.2578822 (R2024a) Mathworks Inc., 2024, MathWorks® https://de.mathworks.com) on Windows.

The participants' EEG data were organized in BIDS (Brain Imaging Data Structure) format (Gorgolewski et al., 2016; Pernet et al., 2019). BIDS is a community standard that ensures homogeneity in the organization and description of raw neurocognitive/brain-derived/neuroscientific data, enabling efficient data sharing, minimizing errors, and supporting completely automated analysis workflows (Gorgolewski et al., 2016; Pernet et al., 2019; Truong, Robbins, Delmore & Makeig, 2023). The rsEEG data, organized according to this standard, were identified and imported into MATLAB using the FieldTrip Toolbox. A trial defining function was built to select the data from the eyes-open condition for subsequent processing. This resulted in approximately 300 s per participant.

**Filtering and Resampling**

A finite impulse response (FIR) windowed-sinc (firws) filter, designed with a hamming windowed sinc function and implemented in the FieldTrip toolbox, was used for both high-pass and low-pass filtering of the continuous data. For high-pass filtering, a cut-off frequency of 0.1 Hz was applied to eliminate very low frequencies (drift) (Keil et al., 2013). This cut-off was based on the findings of Delorme (2023) and Winkler, Debener, Müller and Tangermann (2015), where filtering at 0.1 Hz or higher significantly improved data quality compared to no filtering. Filters above 0.1 were not used due to….

Prior to applying low-pass filtering, the data was downsampled from 1000 Hz to 250 Hz, to reduce computational load while preserving sufficient temporal resolution for subsequent analysis. A cut-off frequency of 45 Hz was then used to eliminate high-frequency noise and mitigate potential 50 Hz line noise (Delorme, 2023). Finally, the data underwent re-referencing using the Common Average Reference (CAR) technique to remove the influence of the reference and improve signal quality (Ludwig et al., 2009). As the name implies, an average of the recordings from all electrode sites was computed and used as the reference (Ludwig et al., 2009; Offner, 1950). The function reref() was used (to perform this step). The data was converted into the EEGLAB data structure for further processing.

Due to empty dataset from one participant, the participant was excluded, leaving the dataset with 69 participants (something like that. But where should I write that?)

**Artifact removal**

After the initial filtering and resampling, the preprocessing pipeline continued with detecting and removing artifacts. First, large artifacts, including the removal of flat-line channels, noisy channels, and short-time bursts of noise, were removed from the data using the EEGLAB pop\_clean\_rawdata() function with specific parameters. Channels with flat lines for more than 5 seconds were removed (FlatlineCriterion = 5), based on the default recommendation (for this parameter) by Pernet et al. (2021). This ensured the exclusion of “dead” or disconnected channels, thereby improving data quality. Channels were further excluded if their signal could not be predicted from a randomly selected subset of the remaining channels for at least 85% of the recording time (ChannelCriterion = 0.85), to remove those that were highly dissimilar from the rest of the channels (Gil Ávila et al., 2023; Pernet et al., 2021). The euclidean distance metric was used to calculate the similarity between channels. Data segments with abnormally high amplitude bursts, exceeding 100 SD compared to neighboring segments, were eliminated (BurstCriterion = 100), as such extreme bursts are considered unlikely to reflect brain signals (Chang et al., 2018). The default BurstCriterion is set to 20, but it may be adjusted if the default setting results in rejecting too many data segments. Some scientist recommend setting the threshold to 100 (EEGLAB, "Automated Pipeline Tutorial", 2024), which aligns with the optimal cut-off range of 10 to 100 suggested by Chang et al. (2018). Therefore, a mild threshold of 100 was chosen here, as it still effectively removes large-amplitude artifacts while retaining valuable data (Chang et al., 2018). Time windows where more than 40% of the channels were marked as noisy were removed (WindowCriterion = 0.4), to ensure the quality of the remaining data. A more lenient threshold of 0.4 was chosen over the default of 0.25 to retain more data (even if it is potentially noisier). How many “bad” (excessively noisy) channels were detected or removed in this process?

Again the data is re-referenced to the average reference (CAR), this time using the EEGLAB function pop\_reref() (Gil Ávila et al., 2023).

Secondly, Independent Component Analysis (ICA) (Bell and Sejnowski, 1995; Hyvärinen, 2013, Jung et al., 2000; Lee et al., 1999; Palmer et al., 2008) was performed on the data, to detect and reject further artifacts, such as eye or muscle movements (Makeig et al., 1995). ICA was performed with the “runica” algorithm and function pop\_runica() with the extended InfoMax method. The runica algorithm was employed with the extended InfoMax method. using the pop\_runica function implemented in EEGLAB. To avoid rank deficiency, the number of components was set to one less than the total number of channels (Kim, Luo, Chu, Cannard, Hoffman & Miyakoshi, 2023). This approach decomposes the EEG signal into independent components, potentially separating artifacts from neural activity. Due to the non-deterministic nature of the ICA algorithm, its results vary across repetitions. That is, every repetition of the ICA algorithm leads to small differences in the reconstructed time series after removing artifactual components (Gil Ávila et al., 2023). The resulting ICA weights, which represent the transformation matrix for this decomposition, were saved in a separate file.

Automatic component rejection was implemented using ICALabel (Pion-Tonachini, Kreutz-Delgado & Makeig, 2019), as automatic artifact rejection is preferred over the manual one to ensure standardization (Miljevic et al., 2022). Artifactual components are automatically classified by the ICLabel classifier (Pion-Tonachini et al., 2019). Thresholds were set at probabilities of 0.8 (80%) for muscle-related components (Pernet et al., 2021) and 0.5 (50%) for eye-related components. Components exceeding these thresholds were flagged and automatically removed using the EEGLAB function pop\_subcomp(). By default, only components whose probability of being “muscle” is higher than 80% were subtracted from the data (Pernet et al., 2021). The two EOG channels (31 and 32) were removed from the dataset. The cleaned dataset was then checked for consistency using eeg\_checkset().

After this steps, an average of 110.3 good channels remained in Cluster 1 (*SD* = 11.4, Range = 75 - 125) and 109 good channels in Cluster 2 (*SD* = 11, Range = 78 - 124) in the two-cluster solution. This corresponds to approximately 85-86% good channels in both groups.

In the four-cluster solution an average of **110.8 good channels remained** in **Cluster 1** (SD = 11.84, range = 75–125), **109 good channels** in **Cluster 2** (SD = 11.03, range = 78–124), **108.6 good channels** in **Cluster 3** (SD = 12.19, range = 77–122), and **113 good channels** in **Cluster 4** (SD = 6.64, range = 105–123) after ICA. This corresponds to approximately 85-88% good channels.

Thirdly and finally, an additional artifacts removal step was implemented to address any remaining problematic channels. This process involved a statistical approach to identify outlier channels based on their signal characteristics. The standard deviation and mean were calculated for each channel across all time points. Then, overall mean values for these standard deviations and means were computed across all channels. Thresholds were established at 2.5 standard deviations above and below the overall mean, creating an acceptable range for channel activity. Channels with standard deviations falling outside this range were identified as outliers. These outlier channels were then removed from the dataset using the EEGLAB function pop\_select(), further refining the EEG data quality. This step ensures that channels with unusually high or low variability, which might represent persistent artifacts or malfunctioning electrodes, are excluded from subsequent analyses. As a result, the dataset retained an average 108 good channels in Cluster 1 (*SD* = 11.2, Range = 73 - 121) and 106.6 good channels in Cluster 2 (*SD* = 10.9, Range = 77 - 121) in the two-cluster solution. This corresponds to approximately 83-84% good channels in both groups.

In the four-cluster solution an average of **108.4 good channels remained** in **Cluster 1** (SD = 11.46, range = 73–121), **106.6 good channels** in **Cluster 2** (SD = 10.92, range = 77–121), **106.3 good channels** in **Cluster 3** (SD = 12.17, range = 74–120), and **111.1 good channels** in **Cluster 4** (SD = 6.20, range = 104–121). This corresponds to approximately 83-87% good channels.

**Interpolate bad channels**

Channels removed in the previous step were interpolated using the EEGLAB function pop\_interp() with the default spherical splines method (Perrin, Pernier, Bertrand & Echallier, 1989), ensuring a consistent number of channels across participants (Gil Ávila et al., 2023). Interpolated channels were inserted into the original channel order. On average 16-17% of the channels in each group were interpolated.

**Epoch length and number**

Lastly, the continuous data were segmented into epochs using the eeg\_regepochs() function implemented in the EEGLAB toolbox, which defaults to 2-second epochs (Gil Ávila et al., 2023). However, longer epochs improve frequency resolution, making them preferable for slower frequencies like delta. To achieve higher resolution while maintaining an adequate trial count, EEG data for each participant were segmented into 5-second nonoverlapping epochs. This function then outputs the new epoch EEG as a dataset on EEGLAB (Bonello, Garg, Garg & Audu, 2018).

The preprocessing resulted in an average of 37.7 good epochs (SD = 14.7, Range = 4 – 60) in Cluster 1, 37.9 good epochs (SD = 16.6, Range = 3 – 59) in Cluster 2 in the two-cluster solution. A two-sided *t-*test did not indicate a significant differences in epoch number between groups, t = -0,0585, p = .95. In the four-cluster solution, the dataset retained an average of **37.6 good epochs** (SD = 16.2, range = 4–58) in **Cluster 1, 37.9 good epochs** (SD = 16.6, range = 3–59) in **Cluster 2, 35.9 good epochs** (SD = 14.5, range = 9–60) in **Cluster 3**, and **42.0 good epochs** (SD = 12.5, range = 24–60) in **Cluster 4**.

**Power Analysis**

For relative power, the values were expressed as a percentage of power in a frequency band divided by the total power across all seven frequency bands.

**Statistical Analysis**

T-test for two group comparing. For four groups different test.