Pre-intervention test-retest reliability of EEG and ERP over four recording intervals

Ip, Cheng-Teng; Ganz, Melanie; Ozenne, Brice; Sluth, Lasse B.; Gram, Mikkel; Viardot, Geoffrey; l'Hostis, Philippe; Danjou, Philippe; Knudsen, Gitte M.; Christensen, Søren R.

Published in:
International Journal of Psychophysiology

DOI:
10.1016/j.ijpsycho.2018.09.007

Publication date:
2018

Document version
Publisher's PDF, also known as Version of record

International Journal of Psychophysiology 134 (2018) 30–43

Contents lists available at ScienceDirect

International Journal of Psychophysiology

journal homepage: www.elsevier.com/locate/ijpsycho

Pre-intervention test-retest reliability of EEG and ERP over four recording intervals

Cheng-Teng Ip[a,b], Melanie Ganz[b,c], Brice Ozenne[b,d], Lasse B. Sluth[a], Mikkel Gram[e,f], Geoffrey Viardot[g], Philippe l'Hostis[g], Philippe Danjou[g], Gitte M. Knudsen[b], Søren R. Christensen[a,*]

[a] Department of Clinical Pharmacology, H. Lundbeck A/S, Ottiliavej 9, DK-2500 Valby, Denmark
[b] Neurobiology Research Unit, University Hospital of Copenhagen, Rigshospitalet, N9201, 9 Blegdamsvej, DK-2100 Copenhagen, Denmark
[c] Department of Computer Science, University of Copenhagen, Universitetsparken 1, DK-2100 Copenhagen, Denmark
[d] Section of Biostatistics, Department of Public Health, University of Copenhagen, Østerfarimagsgade 5, 1014 Copenhagen, Denmark
[e] Mech-Sense, Department of Gastroenterology and Hepatology, Aalborg University Hospital, Hobrovej 18-22, DK-9000, Denmark
[f] Coaze IVS, Kronhjorten 129, DK-9530 Støvring, Denmark
[g] Biotrial, 35000 Cedex, Rennes, France

ARTICLE INFO

Keywords:
Test-retest reliability
EEG
ERP
Repeated measurements
Crossover design

ABSTRACT

In this study we present the test-retest reliability of pre-intervention EEG/ERP (electroencephalogram/event-related potentials) data across four recording intervals separated by a washout period (18–22 days). POz-recording-reference EEG/ERP (28 sites, average reference) were recorded from thirty-two healthy male participants. Participants were randomly allocated into different intervention sequences, each with four intervention regimens: 10 mg vortioxetine, 20 mg vortioxetine, 15 mg escitalopram and Placebo. We report classical EEG spectra: δ (1–4 Hz), θ (4–8 Hz), α (8–12 Hz), β (12–30 Hz), $\gamma 1$ (30–45 Hz) and $\gamma 2$ (45–80 Hz) of resting state and vigilance-controlled, and of auditory steady state response, as well as ERP components N100, P200 and P300 in auditory oddball task and error related negativity (ERN) and error positivity (Pe) in hybrid flanker task. Reliability was quantified using intra-class correlation coefficient (*ICC*). We found that θ, α and β of continuous EEG were highly reliable (*ICCs* $\geq$ 0.84). Evoked power of other tasks demonstrated larger variability and less reliability compared to the absolute power of continuous EEG. Furthermore, reliabilities of ERP measures were lower compared to those of the EEG spectra. We saw fair to excellent reliability of the amplitude of the components such as Pe (0.60–0.82) and P300 (0.55–0.80). Moreover, blood tests confirmed that there was no measurable drug carry-over from the previous intervention. The results support that EEG/ERP is reliable across four recording intervals, thus it can be used to assess the effect of different doses and types of drugs with CNS effects.

1. Introduction

Electroencephalography (EEG) provides a noninvasive method to measure electrical activity of the brain with high temporal resolution. The technique has shown great potential in clinical practice to monitor and access the intervention effects in diagnoses such as depression (Mulert et al., 2007; Tenke et al., 2011), Alzheimer (Brassen and Adler, 2003; Yener et al., 2007) and attention-deficit/hyperactivity disorder (ADHD) (Loo et al., 2000).

With the increased use of EEG and ERP in clinical practice, a systematic investigation of EEG and ERP reliability becomes more important, especially for commonly-used paradigms (e.g. resting state EEG and an auditory oddball task). Previous studies have investigated the reliability of EEG and ERP in various paradigms including resting state EEG with eyes-closed (Corsi-Cabrera et al., 2007) and eyes-opened (Williams et al., 2005), ERP components in an auditory oddball task (Williams et al., 2005), a working memory task (McEvoy et al., 2000) and a Sternberg task (Cassidy et al., 2012). These studies showed that a fair reliability of EEG and ERP could be obtained but that reliability could also be affected by various factors. For example, the reliability of EEG is affected by the epoch length of resting EEG (Gudmundsson et al., 2007), recording intervals (Sandman and Patterson, 2000), different reference schemes (Towers and Allen, 2009), and different aspects of the same EEG indicator (Tenke et al., 2018). In the study of Towers and

* Corresponding author.
E-mail address: SRAC@Lundbeck.com (S.R. Christensen).

https://doi.org/10.1016/j.ijpsycho.2018.09.007
Received 9 February 2018; Received in revised form 20 September 2018; Accepted 21 September 2018
Available online 22 September 2018

Allen (2009), different reference schemes including online reference, re-referencing to linked-mastoids and average were compared for the reliability of frontal α asymmetry. Their results showed that linked-mastoids demonstrated greater reliability than other reference schemes, while other reference schemes still exhibited excellent split-half reliability (> 0.9). Different spectral parameters were compared, and showed that both absolute and relative power are reliable parameters (Fernandez et al., 1993). In a recent study, researchers assessed the temporal stability of different aspects of posterior EEG α over twelve years (Tenke et al., 2018). They suggested that lower reliability of net α (eyes closed-plus-open) and α asymmetry might result from additive errors when separating the α estimates. For ERP studies, there is accumulating evidence showing that ERP amplitudes have higher reliability than ERP peak-latency measures (Cassidy et al., 2012; Walhovd and Fjell, 2002; Weinberg and Hajcak, 2011), which might be a result of the considerable variations in peak-latency detection. These variations could be due to individual differences in information processing efficiency, or induced by the appearance time of the peak amplitude, thus lowering the test-retest reliability. Since the replication of results is not always guaranteed within the field, it is essential to assess the reliability of EEG and ERP measurements.

Among all the factors that could affect reliability, the number of recording sessions bring the biggest challenge to clinical application as it is almost impossible to maintain consistency between or within subjects. So far, a number of studies have investigated the reliability of EEG and ERP over both shorter (days: (McEvoy et al., 2000); weeks: (Cassidy et al., 2012; Hämmerer et al., 2013; Huffmeijer et al., 2014)) and longer recording intervals (months: (Brunner et al., 2013; Näpflin et al., 2007); years: (Sandman and Patterson, 2000; Tenke et al., 2018)). Sandman and Patterson (2000), evaluated ERP reliability in the paradigm of a dual rare-event over a three-year period and found that ERP measurements of adjacent years (e.g. Year 1 & 2) are more similar than ERP measurements of nonadjacent years (Year 1 & 3). Meanwhile, the test-retest reliability of resting EEG was not affected by the recording intervals (Corsi-Cabrera et al., 2007). One might argue that this inconsistency could be a result of different lengths of time (3 years vs. 9 months) during which the results were evaluated. Another possibility could be that different quantifications were investigated, i.e. EEG vs ERP. It could be possible that measures of EEG are more reliable than ERP measures, thus manifesting higher reliability over time. In the study of Williams et al. (2005), they reported high to excellent reliability for EEG power while only fair to excellent reliability for ERP measures. Furthermore, it is unclear how EEG and ERP vary across multiple recording intervals since only a few studies have reported the reliability across more than two sessions (Corsi-Cabrera et al., 2007; Kinoshita et al., 1996; Sandman and Patterson, 2000). In order to address this issue, the current study included four-time points to assess the reliability of both EEG and ERP measures.

In addition to recording intervals, the age of the participants is also known to contribute to the variations in ERP reliability (Alperin et al., 2014). Older adults show higher reliability of the P3 amplitude at the fronto-central site (Cz) while young adults have higher reliability at the centro-parietal area site (Pz) (Walhovd and Fjell, 2002). Hämmerer et al. (2013) suggested that age differences might be a result of different people's signal-to-noise ratio (SNRs), with children and older adults having lower SNRs than other age groups. Despite these variations, ERP measures still exhibit moderate to high reliability when evaluated with varying recording intervals and when participants of different age groups are selected (Hämmerer et al., 2013; Walhovd and Fjell, 2002). Therefore, age was used as a covariate throughout all our analyses.

Besides the signal itself, methodological differences in the statistical analysis have also led to discrepancies in test-retest reliability in EEG/ERP studies. Different statistical methods have been adopted by studies that investigated the correlations between different recording sessions, and the test-retest reliability within sessions, such as *ICC* (Gudmundsson et al., 2007), Pearson's r (Walhovd and Fjell, 2002) and Spearman-Brown-corrected coefficients (Cassidy et al., 2012; Hämmerer et al., 2013; Walhovd and Fjell, 2002). Furthermore, there exist various types of Intra-class correlation coefficients (*ICC*) (Mcgraw and Wong, 1996) and previous studies have investigated the test-retest reliability by using different *ICC* measures. For instance, researchers have used a one-way random model of *ICC* (Gudmundsson et al., 2007), a two-way mixed model with absolute agreement (Brunner et al., 2013; Hämmerer et al., 2013) and a two-way mixed model with consistency (Rentzsch et al., 2008). When assessing EEG reliability between sessions, we define reliability as having both accuracy (i.e. no systematic bias) and precision (i.e. small variance caused by subject variability). We will therefore favor the *ICC* for absolute agreement over correlation coefficients or *ICC* for consistency, since the latter two only measure precision and will overestimate the reliability in presence of systematic biases.

Since the present study aims at assessing the reliability of EEG/ERP parameters, it can serve as a reference for investigating intervention effects. Therefore, we included spontaneous EEG, auditory steady state response, auditory oddball and hybrid flanker Go/Nogo tasks which are common measures in human cognition and executive function. In the present study, we incorporated the baseline data from four different sessions of an intervention study into one model. The carry-over drug effect from the previous session was evaluated through blood tests. The interventions included two different dosing levels of vortioxetine, one dosing level of escitalopram and placebo. The reliability of baseline data across different doses and types of antidepressants was evaluated through a linear mixed model with unstructured covariance matrix and was quantified by absolute agreement *ICC*. We hypothesized that: 1. The *ICC* of EEG and ERP measures will show at least moderate test-retest reliability across four recording intervals. 2. The power spectrum of continuous EEG will exhibit higher test-retest reliability than peak-picking ERP measures. 3. Amplitude measures will have higher test-retest reliability compared to peak latency measures.

2. Method

The study was conducted at the clinical site of Biotrial, Rennes, France. The research protocol was approved by the local ethics committee (reference No. 15835A).

2.1. Participants

Participants were recruited in this study through advertisements and were screened by a trained investigator. To minimize the variability, women were excluded to eliminate the menstrual cycle as a covariate. Thirty-two healthy male participants were enrolled in the study and were compensated for participation. Enrolled participants were aged 22 to 45 years (mean age 33.1 ± 6.8), their body mass index (BMI) ranged from 19.5 to 27.9 kg/m^2 (mean BMI 23.9 kg/m^2 ± 2.24), 94% of participants were Caucasian and 6% were African American. Exclusion criteria included use of psychoactive medication, drug or alcohol abuse, severe drug allergy or hypersensitivity and history of any medical, psychiatric, and neurological (such as immunological, cardiovascular, respiratory, metabolic neurological, or psychiatric) disease. Informed consent was obtained from all the participants before the study. All participants conducted the experiment except for one participant who has missing baseline data for three tasks (auditory steady state response (ASSR), auditory oddball and hybrid flanker task) in the 3rd session. All the collected data were included and analyzed.

2.2. Experimental protocol

This was an interventional, randomized, double-blind, placebo-controlled and four-way crossover study. The four included intervention regimens were: 10 mg vortioxetine (A), 20 mg vortioxetine (B), 15 mg escitalopram (C) and Placebo (D). Each participant was

randomly allocated into one sequence group (ABDC, BCAD, CDBA or DACB) with 8 participants in each group and was investigated under all intervention regimens separated by a washout period (20–22 days[1], median of all between sessions were 21 days) (Fig. 1). Bioanalysis was conducted before the administration of the next intervention to assess the leftover effects from the previous intervention. Within each session, an EEG battery was recorded on Day −1 (pre-intervention), Day 1 (the 1st day after intervention) and Day 3 (the 3rd day after intervention). The EEG battery included continuous EEG with resting and with vigilance-controlled, ASSR, auditory oddball and hybrid flanker tasks. Since the main purpose of this study was to assess the test-retest reliability, only the EEG recording of the four pre-interventions was considered in the subsequent analysis.

2.3. EEG battery

A previous study of antidepressants on rodents has shown a dissociation marker on different treatments, especially on the γ band (Leiser et al., 2014). Moreover, ERP components like P300 and ERN provide physiological measures associated with attentional engagement (Olbrich and Arns, 2013) and early error processing (Olvet and Hajcak, 2009a). The initiative of this study is whether the similar findings could be replicated in humans, as well as how antidepressants would affect human cognition and executive function. Therefore, we included spontaneous EEG, auditory steady state response, auditory oddball and hybrid flanker Go/Nogo tasks.

2.3.1. Continuous EEG

Continuous EEG data were acquired under two conditions: resting and vigilance-controlled. Participants were instructed to relax, keep their eyes closed and stay awake in both conditions. They were instructed to keep pressing two buttons using their thumbs of each hand under the vigilance-controlled condition. A sound would play if the participant let go of the button. Each condition was recorded at least 3 min.

2.3.2. Auditory steady state response (ASSR)

Participants were presented with a 40 Hz impulse trains sound at 89 dB binaurally through a headset (Sennheiser HD 25-1 II pro) (McFadden et al., 2014; Van Deursen et al., 2011). Each train was composed of 20 biphasic 1 ms clicks, and each click was followed by silences lasting 24 ms. There was a silent period of 700 ms after each train. These trains were repeated for 5 min.

2.3.3. Auditory oddball

The auditory oddball paradigm consisted of two acoustic stimuli with different frequencies. Participants were presented with a series of standard tones (500 Hz) and deviant tones (2000 Hz) binaurally through a headset (Sennheiser HD 25-1 II pro). They were asked to count the deviant sounds. To make sure participants performed the task, the presentations of deviant and standard tones were different in sessions. Each session consisted of on average of 35 deviants (randomized between 30 and 40) and 198 standards (randomized between 170 and 226). Deviant tones made up 15% of the presentations. The sound level for each tone was 85 dB, with duration of 100 ms and inter-stimulus-interval (ISI) of on average 1550 ms (randomized between 1200 and 1900 ms). The test lasted approximately 7 min.

2.3.4. Hybrid flanker go/Nogo

Participants performed a hybrid flanker Go/Nogo paradigm (Ruchsow et al., 2006, 2005) with a monitor approximately 100 cm from them. Stimuli consisted of one of the following letter strings (BBBBB, DDDDD, VVVVV, UUUUU, BBDBB, DDBDD, UUVUU, or VVUVV) and were presented on a computer screen for 300 ms in randomized order. Participants were required to focus on the center letter and to press a button whether it was a B or a U (Go condition), and to withhold a button press upon appearance of a D or V (NoGo condition). Each condition consisted of 420 trials. There were 840 trials overall. Strings with congruent letters made up 40% of presentations, while strings with different letters were shown in 60% of all trials. Each trial was followed by 750 ms for stimulus onset asynchrony (SOA) and 500 ms for feedback in response to the participants' performance: 'true' (i.e. correct and in time), 'faster' (i.e. correct but out of time) or 'false'. The deadline for response time was 300 ms after stimulus onset. The ISI was 800 ms (randomized between 600 and 1000 ms). Test duration was approximately 45 min.

2.4. Electrophysiological recording

All participants were seated on a comfortable armchair in a quiet room. During data acquisition, participants were instructed to keep their eyes closed during continuous EEG, auditory oddball, and ASSR recordings. Participants conducted hybrid flanker task with open eyes and were told to refrain from eyes blinking and movement. EEG was recorded from 28 scalp sites using a 10–20 electrode system, with a sample rate of 400 Hz (Comet EEG system, Grass Technologies, West Warwick, RI, USA). AFz served as the ground and POz served as the reference electrode. In order to remove ocular and muscle artifacts electrooculography (EOG) and electromyogram (EMG) were recorded at bipolar channels. Impedances across all electrodes were maintained at $< 5\,k\Omega$.

2.5. Preprocessing of all data

Eye-blink and other ocular corrections were conducted for all the collected data by the ocular artifact reduction option of NeuroScan 4.1 software. It computes a linear regression of covariance between EEG and EOG, and then performs a point-by-point proportional subtraction of the blinks (Semlitsch et al., 1986). The data were further processed in Matlab 2012a (The Matworks, Inc., Natick, MA, USA).

2.5.1. Data preprocessing for spectral analysis

A zero-phase digital IIR Butterworth bandpass filter was applied to all data. The cut-off frequencies of the filter were 1 and 80 Hz, with an order of 2. In addition, a 50 Hz notch filter with the order of 6 was applied. All data (including continuous EEG, ASSR, auditory oddball and hybrid flanker tasks) were re-referenced to the average electrode for later time-frequency analysis. Continuous EEG was cleaned by cutting sections of noisy EEG from the signal by manual inspection.

2.5.2. Data preprocessing for ERP analysis

A zero-phase digital IIR Butterworth bandpass filter was applied to auditory oddball and hybrid flanker tasks. The cut-off frequencies of the filter were 0.1 and 30 Hz, with an order of 2. ERP data were re-referenced to the averages of linked mastoid electrodes (Segalowitz et al., 2010; Weinberg and Hajcak, 2011; Williams et al., 2005).

2.6. Data analysis

2.6.1. Time-frequency analysis of all data

Since EEG data have non-stationary characteristic, all data were analyzed using a wavelet transform as this has a better time-frequency resolution than the more common Fourier transform (Akin, 2002). The continuous wavelet transform was applied using the complex Morlet wavelet as a mother wavelet function with a bandwidth of 10 Hz and a center frequency of 1 Hz. The scales for the mother wavelet were chosen to match frequencies ranging from 1 to 80 Hz with a 0.5 Hz

[1] There was one outlier (91 days) in the last washout period due to recording cancellation. This recording was rescheduled after all participants were recorded.

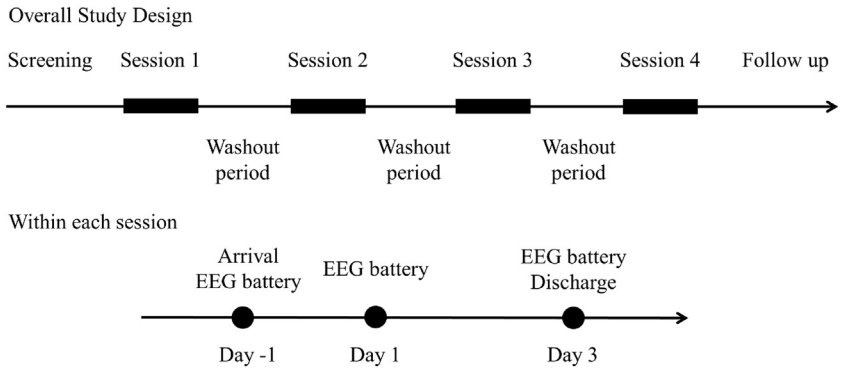

Fig. 1. Overall study design. Three interventions (A, B, C) and one placebo (D) were included in the study. Each participant was randomly allocated to one session sequence including ABDC, BCAD, CDBA and DACB. There were washout periods (median intervals were 21 days) between two sessions, and pharmacokinetic assessments were conducted to assess the carry-over drug effect from the previous intervention. Three EEG recordings were conducted within each session, including Day −1, Day 1 and Day 3. In this study, only the data from Day −1 was analyzed.

between-scale frequency interval. In the current study we worked on absolute power only, thus, the absolute values of the obtained wavelet coefficients were used for the following analysis: First, the wavelet coefficients were divided into the following standardized bands: δ (1–4 Hz), θ (4–8 Hz), α (8–12 Hz), β (12–30 Hz), $\gamma 1$ (30–45 Hz) and $\gamma 2$ (45–80 Hz). Then, the γ band was divided into two bands to deal with artifacts from muscle activity. Next, the wavelet coefficients were averaged over time and summed within each frequency band.

We applied different approaches for the continuous EEG and all the other tasks. The wavelet transform was applied on the noise-free continuous EEG data without segmentation, including resting state and vigilance-controlled. All other tasks were segmented prior to time-frequency analysis and then evoked power was calculated for each task. ASSR and auditory oddball were segmented into stimulus-locked epochs of 500 ms according to the onset of the stimulus. For the auditory oddball task, evoked power was calculated for standards and deviants separately. The hybrid flanker task was segmented from 0 to 400 ms according to the onset of error response.

The spectral analysis focused on three midline sites (Fz, Cz, Pz), therefore, the values represent the absolute values contained in each frequency band at these channels. Data were log-transformed prior to statistical analysis.

2.6.2. Grand average analysis of ERP data

In the auditory oddball task, EEG data were segmented into stimulus-locked epochs of 1000 ms (including a 200 ms pre-stimulus baseline) according to the onset of the sounds. Averaging was performed for standards and deviants separately. Epochs were rejected if the voltage in EOG channels, Fp1, Fp2 exceeded ± 75 μV. Based on prior studies investigating auditory oddball key components (Kemp et al., 2010; Poyraz et al., 2017), both peak latency and amplitude (baseline to peak) were determined on midline channels (Fz, Cz, Pz). The selected components and the corresponding latency windows for peak identification included: standard: N100 (80–140 ms), P200 (140–270 ms); deviant: N100 (80–140 ms) and P300 (270–550 ms). All epochs were manually inspected for other artifacts. A similar approach was applied to the hybrid flanker task. The main interest of the hybrid flanker task was the false positive response (Ruchsow et al., 2006, 2005), thus only responses of error commission were reported. EEG data were then segmented into response-locked epochs of 600 ms (including 200 ms pre-response baseline) according to the onset of error response. ERN (0–250) was analyzed at sites in the fronto-central area (Fz, Cz) and Pe (100–350) was analyzed at sites in the centro-parietal area (Cz, Pz) (Falkenstein et al., 2000). The number of accepted epochs is shown in Table 1.

2.7. Blood sampling

The blood samples (2 mL for each regimen) were analyzed for the plasma concentrations of vortioxetine and escitalopram. Plasma concentrations were determined by using protein precipitation followed by liquid chromatography with tandem mass spectrometric detection. The purpose of these assessments was to ensure that previous intervention was completely washed out so that it would not interfere with the current intervention administration.

2.8. Statistics

The statistics were divided into two parts and performed in SPSS version 24 (IBM Corp., Armonk, NY). First, all EEG and ERP measures were analyzed with a linear mixed model (restricted maximum likelihood estimation) using an unstructured covariance matrix with assigned sequence (ABDC, BCAD, CDBA or DACB) and pre-intervention recordings of each session (BL1, BL2, BL3, BL4) as fixed factors. This was done in order to investigate if there was an effect of session or assigned sequence on our measurements. Participant served as a random variable to account for the correlation between measurements from the same patient. An unstructured covariance matrix was employed to make minimal assumption on the covariance structure - meaning we relax the assumption of homogeneity of variance by modeling a different variance at each session and allow the correlation to vary between pairs of sessions. The structure of the covariance matrix used in the mixed models was decided upon inspection of the model fit. Using likelihood ratio tests, we found a significantly worse fit for the compound symmetry structure (i.e. assuming constant variance over time and constant correlation between any two timepoints) compared to a compound symmetry structure for some of the power measures of resting EEG and some of the ERP measures of flanker hybrid task. Therefore, an unstructured covariance matrix was employed. In all mixed models, age was included as covariate. Main effects of session and sequence were tested using *F*-tests. In post hoc analyses, regression coefficients of the different levels of the main effects were compared using Wald tests with Tukey contrasts. This was performed using the module EM Means for Linear Mixed Model in SPSS. Neither the *p*-values from the *F*-tests nor the post hoc analyses were adjusted for multiple comparisons in order to not reduce power. In this fashion we are maximizing our chance to detect any session or sequence effect despite detecting possible false positives. Second we assessed the reliability of our measurement using the intra-class correlation (*ICC*) with absolute agreement (Brunner et al., 2013; Hämmerer et al., 2013). Single measure *ICC* (A, 1) was calculated by a two-way mixed random model (Mcgraw and Wong, 1996), where participant served as random variable and session served as fixed variable. *ICC* of adjacent time points, BL1 & BL2, BL2 & BL3 and BL3 & BL4 are reported. In accordance with the classification of *ICC* levels in a previous study (Rentzsch et al., 2008), *ICC* < 0.39 would be considered poor, 0.4–0.59 fair, 0.6–0.75 good and > 0.75 would be considered excellent. Overall, time variances are reported in the supplement and were computed by the structure of compound symmetry. To provide a synthetic measure of the ICC over time, we computed "average *ICCs*" using a mixed model with a compound symmetry covariance matrix instead of an unstructured covariance matrix. This enables us to provide a graphical representation

Table 1
The number of accepted epochs for different tasks.

Task	Condition	BL1	BL2	BL3	BL4	p values
Auditory oddball	Standard	180 ± 23(117–212)[a]	169 ± 37(101–227)	167 ± 27(97–215)	174 ± 34(80–227)	F (3,92) = 1.884, p = .138
	Deviant	31 ± 5(22–38)	30 ± 7(16–40)	29 ± 6(12–38)	30 ± 7(15–40)	F (3,92) = 1.270, p = .289
Hybrid Flanker	Error	85 ± 35(28–180)	70 ± 28(8–133)	67 ± 36(8–183)	72 ± 31(3–141)	F (3,92) = 3.981, p = .01

Notes. [a] The minimum and maximum of epochs are provided in the brackets. The mean and standard deviation are reported.

of the ICC as a function of the percentage of accepted trials of across time (Fig. 8).

3. Results

Blood tests were performed to assess the carry-over drug effect of previous interventions. The blood concentration of the previous treatment, C_{max} for all participants across sessions was below 5%, which was considered as complete washout.

3.1. Behavioral results

In the hybrid flanker Nogo trials, participants demonstrated a mean false positive alarm rate of 21% (SD: 7.8) for BL1, 17% (± 7.3) for BL2, 16% (± 9.1) for BL3 and 18% (± 7.7) for BL4. A linear mixed model revealed that there were no significant effects for session and assigned sequence in error rate (p values > .05). Considering the mean reaction time, participants demonstrated a mean false positive reaction time of 283 ms (± 18) for BL1, 282 ms (± 22) for BL2, 276 ms (± 17) for BL3 and 274 ms (± 22) for BL4. There were no significant effects of session and assigned sequence in Nogo reaction time (p values > .05).

3.2. Absolute power of resting EEG

Since there was no segmentation for continuous EEG, spectra were used for presentation instead of time-frequency plots (Fig. 2). There were no significant effects of session and assigned sequence in resting condition for all frequency bands (p values > .05). In the vigilance-controlled task, γ1 at the central site exhibited a significant main effect of session (F (3, 31) = 3.41, p = .029). Post hoc analyses revealed that absolute γ1 power at the first recording session BL1 was larger than the last session BL4 (17.83 vs 16.39 μV, p = .006). No other significant effect was found.

3.3. Evoked power of ASSR, auditory oddball and hybrid flanker task

Fig. 3 shows the absolute evoked power for ASSR, auditory oddball and hybrid flanker tasks for all four recording sessions. Compared to the absolute power of continuous EEG, evoked power demonstrated more variations between sessions. Specifically, the absolute evoked power at the first recording (BL1) contributed the most to the significance.

For the ASSR task, no sequence effect was found for all frequency bands. Significant main effects of session were found for δ and γ1 at the frontal site (F (3, 31) = 3.919, p = .018; F (3, 31) = 3.567, p = .025, Fig. 3a). Post hoc analyses revealed that a smaller absolute δ power was observed at BL1 compared to BL3 and BL4 (1.35 vs. 1.51 μV, p = .02; 1.35 vs. 1.58 μV, p = .004), and larger γ1 was observed at BL1 compared to BL3 and BL4 (1.91 vs. 1.79 μV, p = .02; 1.91 vs. 1.84 μV, p = .03). Similarly, δ at the parietal site indicated a significant session effect (F (3, 31) = 3.179, p = .038), showing that the absolute δ power at BL1 was smaller than BL4 (0.61 vs. 0.83 μV, p = .014). Moreover, θ and α at the central site exhibited significant session effects (F (3, 31) = 4.352, p = .011; F (3, 31) = 3.409, p = .03). The absolute θ power of BL1 was the smallest compared to other recording sessions (p values < .05) and the absolute α of BL1 and BL2 were smaller than that of BL4 (1.13 vs. 1.30 μV, p = .03; 1.08 vs. 1.30 μV, p = .013).

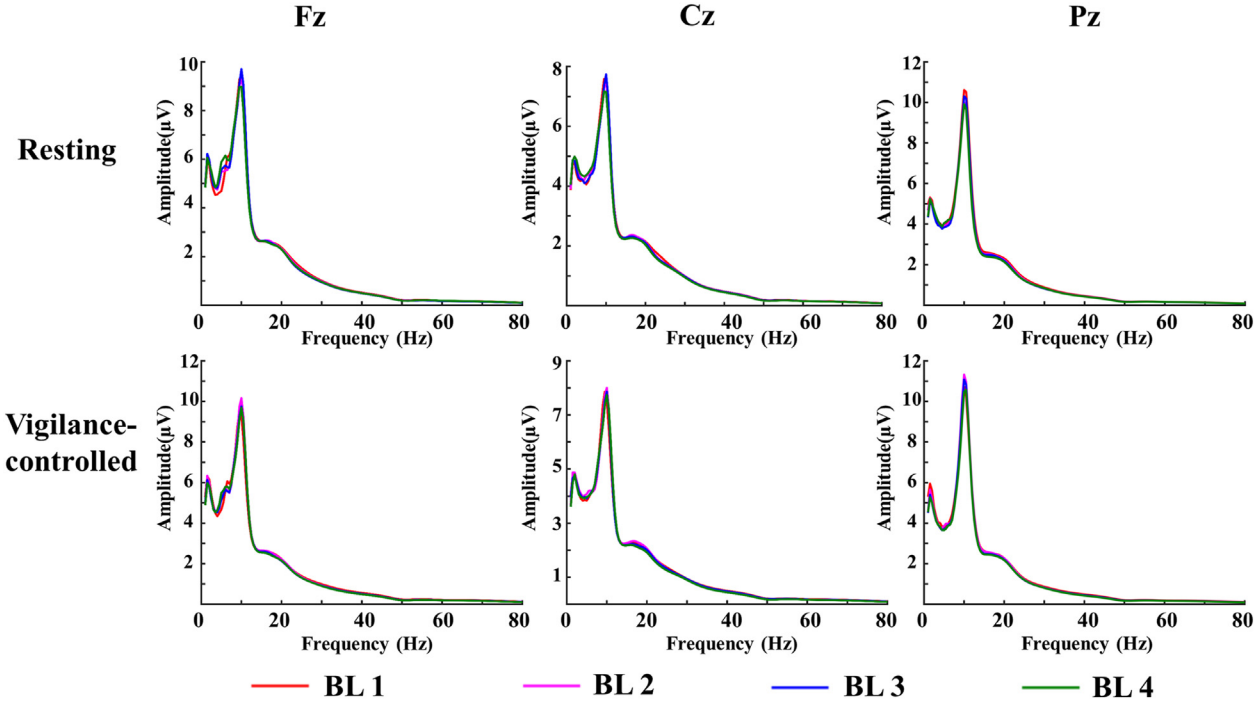

Fig. 2. Spectral results for continuous EEG including conditions of resting state and vigilance-controlled. Three midline electrodes (Fz, Cz and Pz) are shown for each condition. Four recording sessions (BL1, BL2, BL3 and BL4) are shown in different colors.

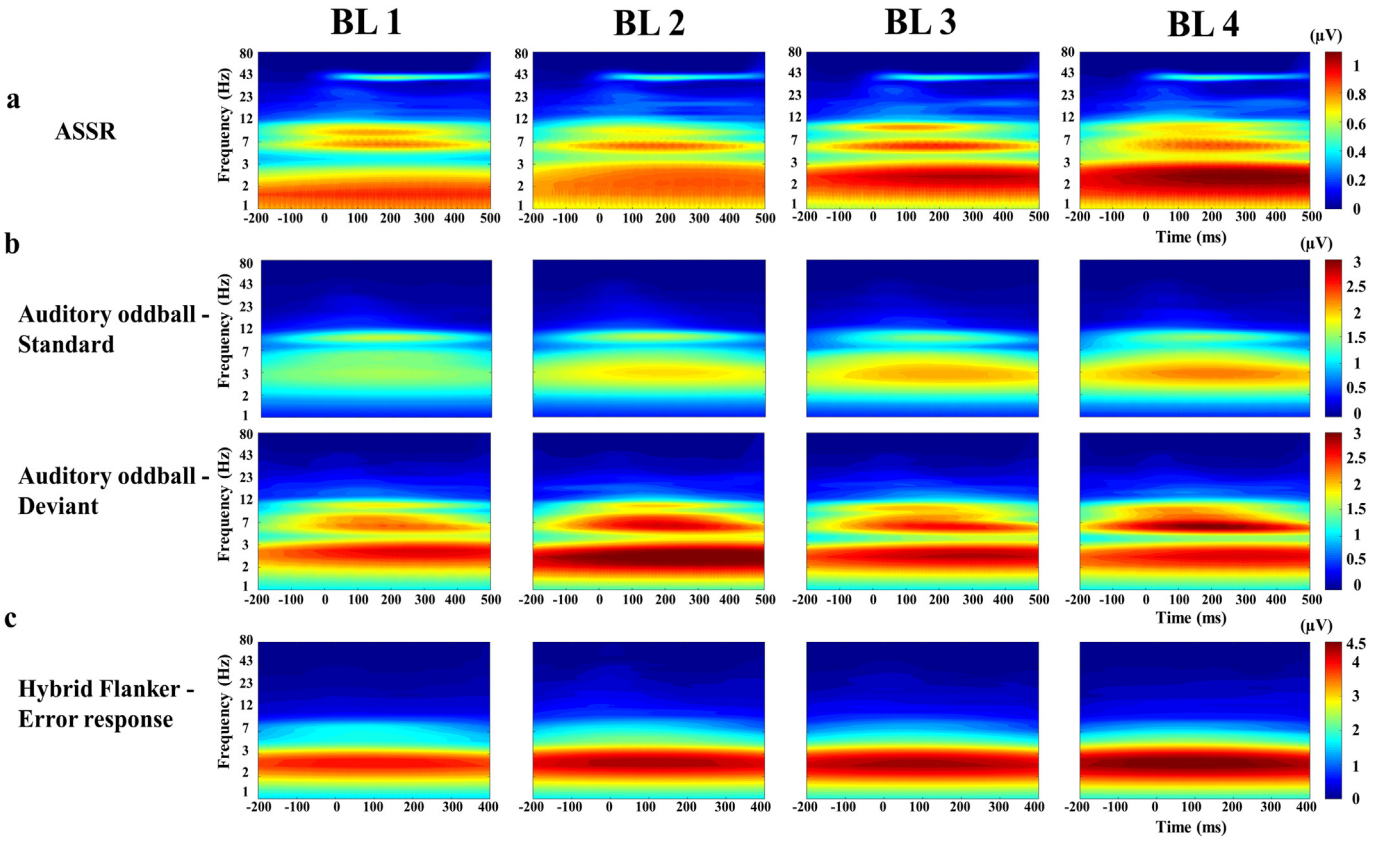

Fig. 3. Time-frequency results for ASSR, auditory oddball and hybrid flanker tasks. Only results at electrode Fz are shown here since most of the significant results were found on this electrode. Four recording sessions (BL1, BL2, BL3 and BL4) are shown in columns. Log-scale is shown for the frequency range.

There was no significant effect of session for β and $\gamma 2$ bands (p values $>$ .05).

For the standard tones in the auditory oddball task, significant session effects of the frontal and the parietal sites were observed in the δ band (F (3, 31) = 5.651, p = .003; F (3, 31) = 3.844, p = .02; Fig. 3b). Post hoc analyses revealed that the absolute frontal δ power of BL1 was the smallest (p values $<$ .05), while absolute parietal δ power was the largest among all recording sessions (p values $<$ .05). No other significant effect was found. For the deviant tones in the auditory oddball task, absolute frontal δ power showed a significant session effect (F (3, 31) = 3.111, p = .04), indicating that the absolute frontal δ of BL2 was larger than BL3 and BL4 (2.54 vs. 2.37 μV, p = .02; 2.54 vs. 2.34 μV, p = .011). Notably, absolute θ power of BL1 was significantly smaller than BL2 (2.54 vs. 2.71 μV, p = .02; 2.13 vs. 2.32 μV, p = .005), indicated by a significant main session effect at frontal and parietal sites (F (3, 31) = 3.327, p = .032; F (3, 31) = 3.185, p = .038). Moreover, absolute frontal θ power of BL1 was smaller than BL4 (2.54 vs. 2.73 μV, p = .02). No sequence effect was found (p values $<$ .05).

There were no session effects in the bands of δ, α and $\gamma 1$ for the error response of the hybrid flanker task. A significant session effect was observed for the absolute θ power at the central site (F (3, 31) = 3.52, p = .027), due to smaller absolute θ during the first two recording sessions than BL3 (Fig. 3c). Significant main effects of session were found for β and $\gamma 2$ at fronto-central sites. Post hoc analyses indicated that absolute frontal β and $\gamma 2$ powers of BL1 were the smallest (F (3, 31) = 3.679, p = .023; F (3, 31) = 3.219, p = .036) among other recording sessions. Absolute central β, $\gamma 1$ and $\gamma 2$ powers of BL1 were smaller than BL3 (F (3, 31) = 3.297, p = .033; F (3, 31) = 4.804, p = .007; F (3, 31) = 3.640, p = .023). Moreover, absolute $\gamma 2$ of BL1 at the central site was smaller than BL4. Sequence effects were observed in $\gamma 2$ at the frontal site (F (3, 24) = 4.381, p = .014), due to the greatest power observed in the sequence of ABDC among others.

3.4. Amplitude and latency analysis of auditory oddball and hybrid flanker tasks

Table 1 shows the number of accepted epochs for both auditory oddball and hybrid flanker tasks. The results of a linear mixed model indicated that there was a significant session effect (F (3,92) = 3.981, p = .01) for the number of accepted epochs in the hybrid flanker task. BL1 demonstrated a significant higher number in accepted epochs than BL2 (p = .043), and BL3 (p = .014). No session effect was observed for the auditory oddball task (p values $>$ .05).

Fig. 4 shows the mean ERP waveform for auditory oddball and hybrid flanker tasks for all four recording sessions. For the standard ERPs in the auditory oddball task, fronto-central N100 amplitude exhibited a significant session effect (F (3, 31) = 5.21, p = .005; F (3, 31) = 6.93, p = .001, Fig. 4). BL1 and BL2 showed larger fronto-central N100 amplitude than BL3 and BL4 (p values $<$ .05). No session effect was found for fronto-central N100 latency. However, parietal N100 latency indicated a significant session effect (F (3, 31) = 3.36, p = .034), showing that BL1 had longer latency than all other recording sessions (p values ≤ .052). There was no session effect on the P200 amplitude. No assigned sequence effect was found for standard ERPs. For the deviant ERPs in the auditory oddball task, there were no significant effects of session and assigned sequence on the N100 amplitude. The central N100 latency was shortest for the last recording (BL4, p values $<$ .05), as suggested by a significant session effect (F (3, 31) = 3.26, p = .034). The fronto-central P300 amplitude seemed not to be affected by session or assigned sequence (Fig. 4). However, a significant session effect was also found for the parietal P300 latency (F (3, 31) = 4.13, p = .014), showing that a shorter P300 latency was observed at BL1 compared to BL4 (310 vs. 325 ms, p = .009).

For the error ERPs in the hybrid flanker task, there were no significant effects of session and assigned sequence for ERN and Pe

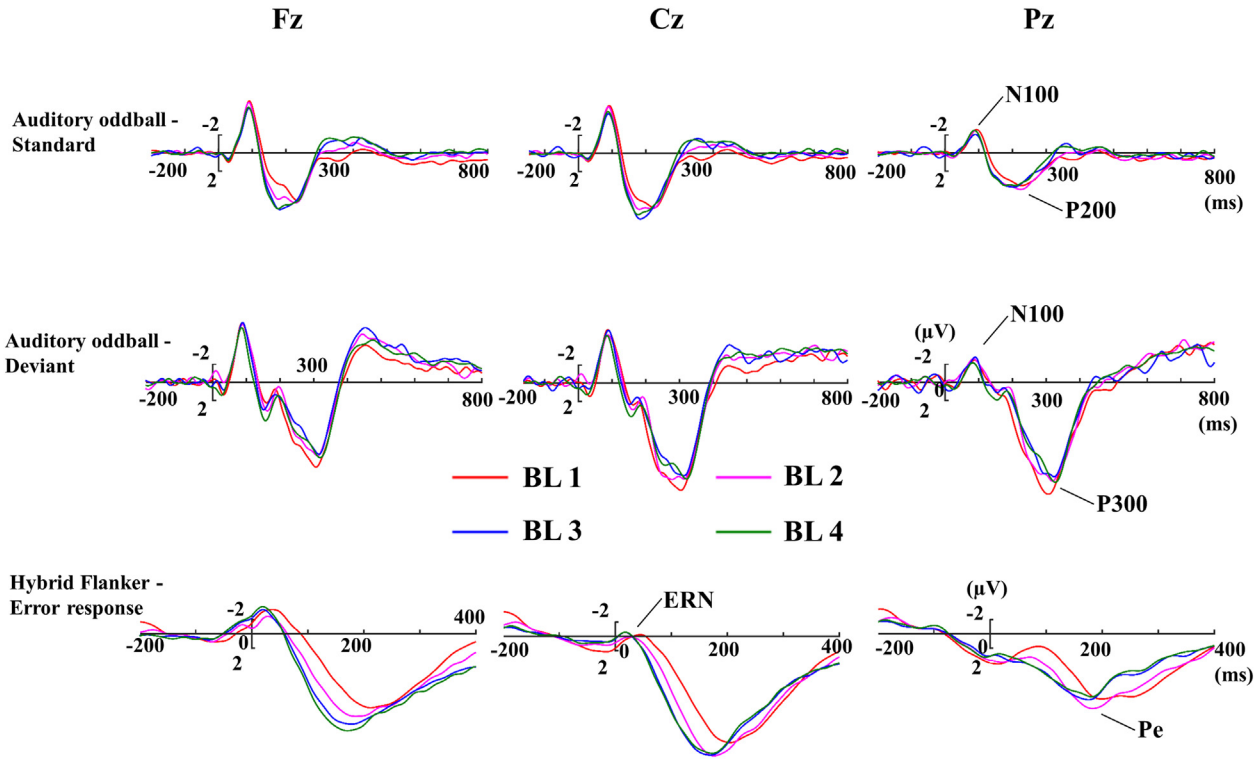

Fig. 4. The grand-averaged ERP waveforms for the auditory oddball (epoched by the stimuli) and hybrid flanker task (epoched by the error response). Three midline electrodes (Fz, Cz and Pz) are shown for each task/component. Four recording sessions (BL1, BL2, BL3 and BL4) are shown in different colors.

amplitudes (see Fig. 4). However, session effects were observed for the latency measures. The fronto-central ERN latency exhibited a significant session effect (F (3, 31) = 3.78, p = .02; F (3, 31) = 6.91, p = .001), showing that a longer ERN latency was observed at BL1 than at BL3 and BL4 (p values < .05). The centro-parietal Pe latency was longer at BL1 and BL2 than BL3 and BL4 (p values < .05), as indicated by a significant session effect (F (3, 31) = 29.34, p < .001; F (3, 31) = 22.66, p < .001).

3.5. Test-retest reliability

3.5.1. Absolute power of resting EEG

Between session *ICCs* for continuous EEG (both resting state and vigilance-controlled) are presented in Fig. 5. The test-retest reliabilities were similar in both conditions. The *ICCs* of adjacent sessions showed excellent test-retest reliability (0.84–0.97) in the frequency bands of θ, α and β. Midline δ and $\gamma 1$ bands were less robust but still indicated good to excellent levels of reliability (0.62–0.87). *ICCs* for midline $\gamma 2$ exhibited the least reliability among all other bands (0.30–0.66). Across time, *ICC* showed similar results to adjacent time points. Midline θ, α and β had excellent reliability (0.86–0.93) while δ and $\gamma 1$ bands showed good to excellent reliability (0.66–0.82). Compared to adjacent time points, $\gamma 2$ ICC across time performed worse with poor to fair levels of reliability (0.37–0.52).

3.5.2. Evoked power of ASSR, auditory oddball and hybrid flanker task

Between session *ICCs* for evoked power of ASSR, auditory oddball and hybrid flanker tasks are presented in Fig. 6. Across time *ICC* showed similar results to adjacent time points.

For the ASSR task, midline $\gamma 1$ –which contains the stimulation frequency- exhibited good to excellent reliability for both adjacent sessions and across time (0.66–0.86), except for the fair *ICC* measured between the last two sessions at the frontal site (0.57). The *ICCs* of δ, β and $\gamma 1$ were less robust but still indicative of fair to good levels of reliability (0.44–0.76). Midline θ exhibited larger variations in different recording sessions, where the reliability varied from poor to excellent (0.37–0.83). The *ICCs* of the α band demonstrated poor to fair levels of reliability in the ASSR task (0.19–0.56).

For the standard tones of the auditory oddball task, midline β and $\gamma 1$ revealed fair to excellent levels of reliability for both adjacent sessions and across time (0.44–0.85). The *ICCs* of the δ, θ and α bands exhibited larger variation between sessions compared to the β and $\gamma 1$ bands, in the range of poor to excellent (0.29–0.84). Compared to other frequency bands, midline $\gamma 2$ of standard tones showed less robust reliability with poor to good levels of *ICC* (0.36–0.62). In general, deviant tones were less robust compared to standard tones. The *ICCs* of δ were in the range of good to excellent (0.63–0.83). Midline θ had poor to excellent reliability (0.34–0.82) while the ICCs of other bands were in the range of poor to good (0.08–0.75).

For the error response of the hybrid flanker task, midline θ tended to exhibit the best reliability among other bands for both adjacent sessions and across time (0.50–0.85). The *ICCs* of α were fair to good (0.47–73) while the *ICCs* of δ showed more variability, in the range of poor to excellent (0.24–0.80). Midline β, $\gamma 1$ and $\gamma 2$ bands demonstrated similar reliability, in the range of poor to good levels of reliability (0.25–74).

3.5.3. Amplitude and latency analysis of ERP task

Between session *ICCs* for peak amplitude and latency measures of the auditory oddball and hybrid flanker tasks are presented in Fig. 7. Generally, amplitude and latency analysis of ERP showed lower reliability compared to the power spectrum analysis of ERP data. Furthermore, latency measures were less stable than amplitude measures.

For standard ERPs in the auditory oddball task, the fronto-central N100 amplitude showed good to excellent reliability for adjacent sessions and across time (*ICC* > 0.70), and the parietal N100 amplitude demonstrated poor to fair levels of reliability (0.39–0.53). The P200 amplitude exhibited similar reliability, with good to excellent levels of reliability at fronto-central sites (0.65–0.83) and less stable performance at parietal site (0.49–0.68). Latency measures exhibited more variations from session to session. The *ICCs* of the N100 latency were in

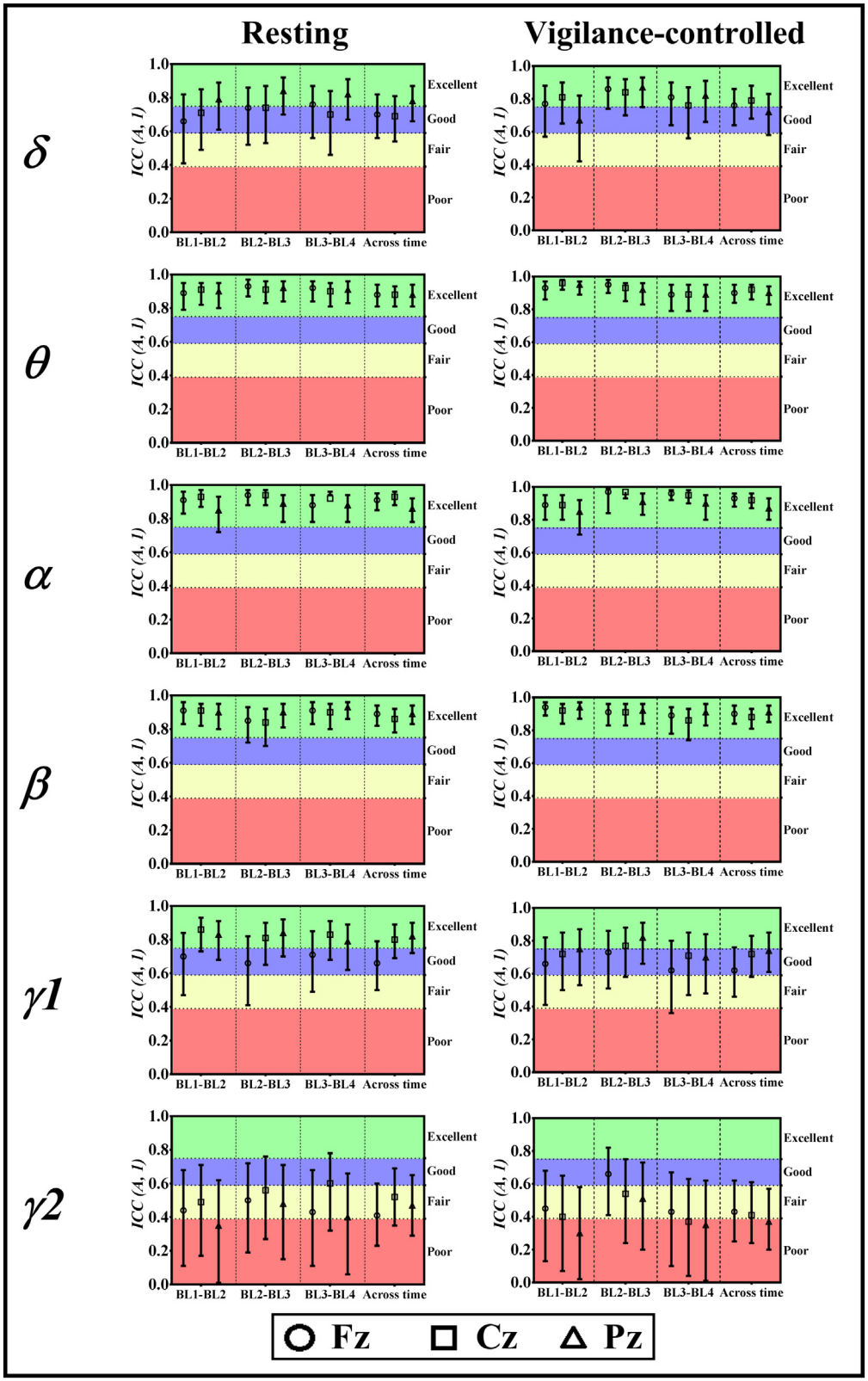

Fig. 5. Intra-class correlation coefficient (*ICC*) for continuous EEG across four sessions. *ICCs* of adjacent time and across four-time points are reported. Test-retest reliability is estimated by the single measure *ICC (A, 1)*. The mean and confident intervals for *ICCs* are shown in the figure.

the range of fair to excellent (0.43–0.89), except for the *ICC* of first two sessions, which showed only poor reliability (0.28). The *ICCs* of the P200 latency were in the range of poor and fair (−0.48–0.49). Compared to standard tones, the fronto-central N100 amplitude of deviant tones exhibited lower reliability, with the *ICC* range of poor to fair (0.18–0.45). The parietal N100 amplitude showed poor reliability (−0.05–0.19). The midline P300 amplitude yielded fair to excellent reliability for adjacent sessions and across time (0.55–0.80). Compared

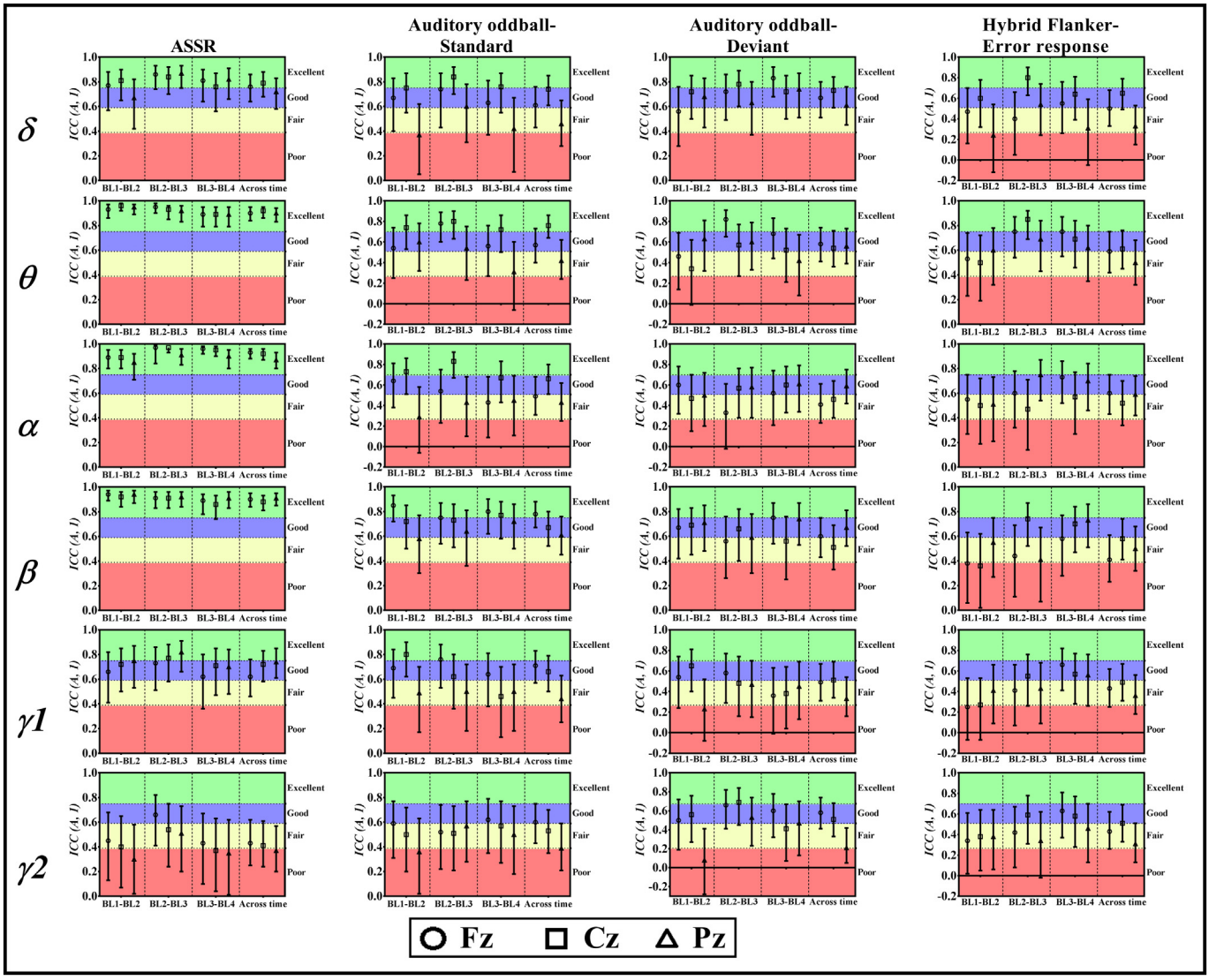

Fig. 6. Intra-class correlation coefficient (*ICC*) for ASSR, auditory oddball and hybrid flanker tasks across four sessions. *ICCs* of adjacent time and across four-time points are reported. Test-retest reliability is estimated by the single measure *ICC (A, 1)*. The mean and confident intervals for *ICCs* are shown in the figure.

to the standard ERPs, the latency measures for deviant ERPs demonstrated less variations between sessions but were still indicative of poor to good levels of reliability (N100: −0.10–0.57; P300: 0.19–0.63).

For the error ERPs in the hybrid flanker task, the fronto-central ERN amplitude demonstrated poor to good levels of reliability for adjacent sessions (0.12–0.61) and poor reliability across time (0.32–0.38). The centro-parietal Pe tented to exhibited higher reliability compared to ERN, with the *ICC* ranging of good to excellent for both adjacent time points and across time (0.60–0.82). Latency measures showed less reliability compared to amplitude measures. The *ICCs* of ERN latency were poor (0.12–0.35) while Pe latencies were poor to good (0.19–0.71).

3.5.4. Exploratory analysis: test-retest reliability with increasing percentage of accepted trials

Across time *ICCs* for the auditory oddball and hybrid flanker tasks for an increasing percentage of accepted trials are presented in Fig. 8 (see Supplementary materials for adjacent time points). Four percentages were assessed with 25% as an increment: 25%, 50%, 75% and 100%. Percentages were calculated relative to the total amount of accepted trials individually. Then the corresponding number of trials would be successively selected from the total amount of accepted trials, i.e. the first 25% (or 50% and 75%) of the total accepted trials. The number of accepted epochs for different percentages is shown in Table 2. Mean amplitude, which was calculated using the same window as peak amplitude, was included here for comparison to peak amplitude.

As expected, reliability increased with increasing percentage of accepted trials. Peak amplitude demonstrated comparable results with mean amplitude for all components. Latency measures were more susceptible to changes of percentage compared to peak amplitude and mean amplitude measures.

For the ERPs in the auditory oddball task, the *ICCs* increased with increasing percentage of accepted trials. Hence, it could be possible that increasing the number of accepted trials could increase the test-retest reliability. The grand average (100%) exhibited the highest reliability for almost all components, except for the N1 latency evoked by deviant tones, where the *ICCs* for grand average were lower than that of the first 75% of accepted trials.

For the error ERPs in the hybrid flanker task, the results for the ERN and Pe measures are similar. They were less affected by the increasing percentage of accepted trials. The *ICCs* increased slightly with increasing percentage of accepted trials up to the first 50%, but then remained at the same level of reliability as the grand average.

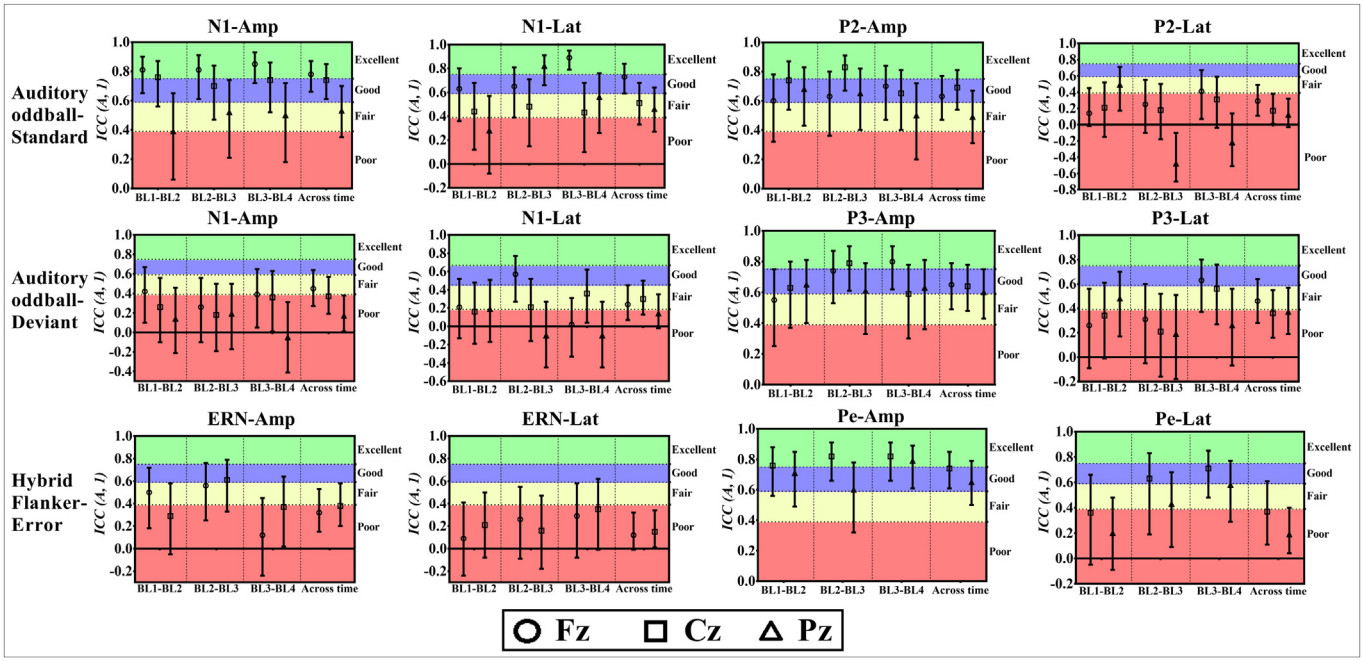

Fig. 7. Intra-class correlation coefficient (*ICC*) for peak amplitude and latency measures of auditory oddball and hybrid flanker tasks across four sessions. *ICCs* of adjacent time and across four-time points are reported. Test-retest reliability is estimated by the single measure *ICC (A, 1)*. The mean and confident intervals for *ICCs* are shown in the figure.

4. Discussion

In this study, we examined the test-retest reliability of an EEG battery over four recording intervals. The EEG battery was comprised of continuous EEG (resting state and vigilance-controlled), ASSR as well as an auditory oddball paradigm and a hybrid flanker task. A linear mixed model with unstructured covariance matrix was used to identify any significant effect of recording session or the assigned intervention sequence. The test-retest reliability was quantified by an absolute agreement type of *ICC*. For healthy participants, the results demonstrated that the EEG battery was found to be reliable over four sessions. The absolute power of continuous EEG showed excellent reliability in θ, α and β ($ICC > 0.84$). Evoked power for ERP tasks demonstrated itself to be less stable compared to the absolute power of continuous EEG. The absolute evoked power of ASSR showed fair reliability in δ, β, $\gamma 1$ and $\gamma 2$ bands. For the auditory oddball task, the β band exhibited fair reliability ($ICC > 0.51$) in both standard and deviant conditions. The *ICCs* of θ in the hybrid flanker task were the most stable among all the frequency bands. While the ERP components showed lower reliability than the power spectral analysis, they still showed good test-retest reliability at their maximal sites. The P300 amplitude obtained from the auditory oddball paradigm had consistently fair to excellent reliability at the central sites ($ICC = 0.55–0.80$) as well as the amplitude of the midline P2 ($ICC = 0.49–0.83$). The centro-parietal Pe amplitude obtained from the hybrid flanker task also exhibited good to excellent reliability ($ICC = 0.60–0.82$). Compared to amplitude measures, peak latency measures showed poor to good reliability with greater variability, thus they are less reliable compared to other measures. A washout period and pharmacokinetic assessment were included to avoid a carry-over drug effect from the previous intervention.

4.1. The absolute power analysis of continuous EEG is highly reliable

The observed excellent test-retest reliability of continuous EEG in the power analysis is consistent with previous studies (Corsi-Cabrera et al., 2007; Gudmundsson et al., 2007; McEvoy et al., 2000; Williams et al., 2005). In the study of Gudmundsson et al. (2007), researchers compared different qEEG features such as power spectral parameters, entropy, complexity and coherence measures and suggested that power spectral analysis exhibits higher reliability than others types of analysis. This was confirmed by our results, and indicates that power spectral analysis of continuous EEG is reliable over time and is sufficient for clinical use.

4.2. The reliability of ERP measures was affected by various factors

Our results showed that ERP measures exhibited more variation and are less stable compared to continuous EEG. According to previous results, there are many factors that could cause the variability of ERP measures. They include the number of averaged trials (Larson et al., 2010) and the scoring methods (Brunner et al., 2013). Larson et al. (2010) investigated the influence of the number of averaged trials on error-related ERP components, and showed that adding trials increases the test-retest reliability for both the amplitude and latency measures. This was confirmed by our exploratory analysis. The results highlighted that increasing the percentage of accepted trials improved the test-retest reliability. For the ERPs in the auditory oddball task, the first 75% of accepted trials produced a comparable reliability to the grand average. Particularly for the latency measures, the increasing trend indicated that an increasing number of accepted trials improved the reliability. Except for the N1 latency of deviant tones, the *ICCs* of the grand average had lower reliabilities than the corresponding *ICCs* based on the first 75% of accepted trials. This result was mainly due to the high reliability of the first 75% of accepted trials at BL1-BL2 (Fig. S1). We surmised that fatigue/impatience due to the unfamiliarity of the task might be a possible reason. Participants could be tired or lose motivation during the last 25% of the accepted trials. Since this phenomenon didn't extend to the subsequent sessions, the result reiterated the importance of guiding the participants in a proper way so that they can be more comfortable with the experiments during the first recording session. For the error ERPs in the hybrid flanker task, the reliability of different percentages was similar, especially for percentages above the first 25%. This finding supports previous findings by Olvet and Hajcak (2009b), in which they reported that stable ERN and Pe can

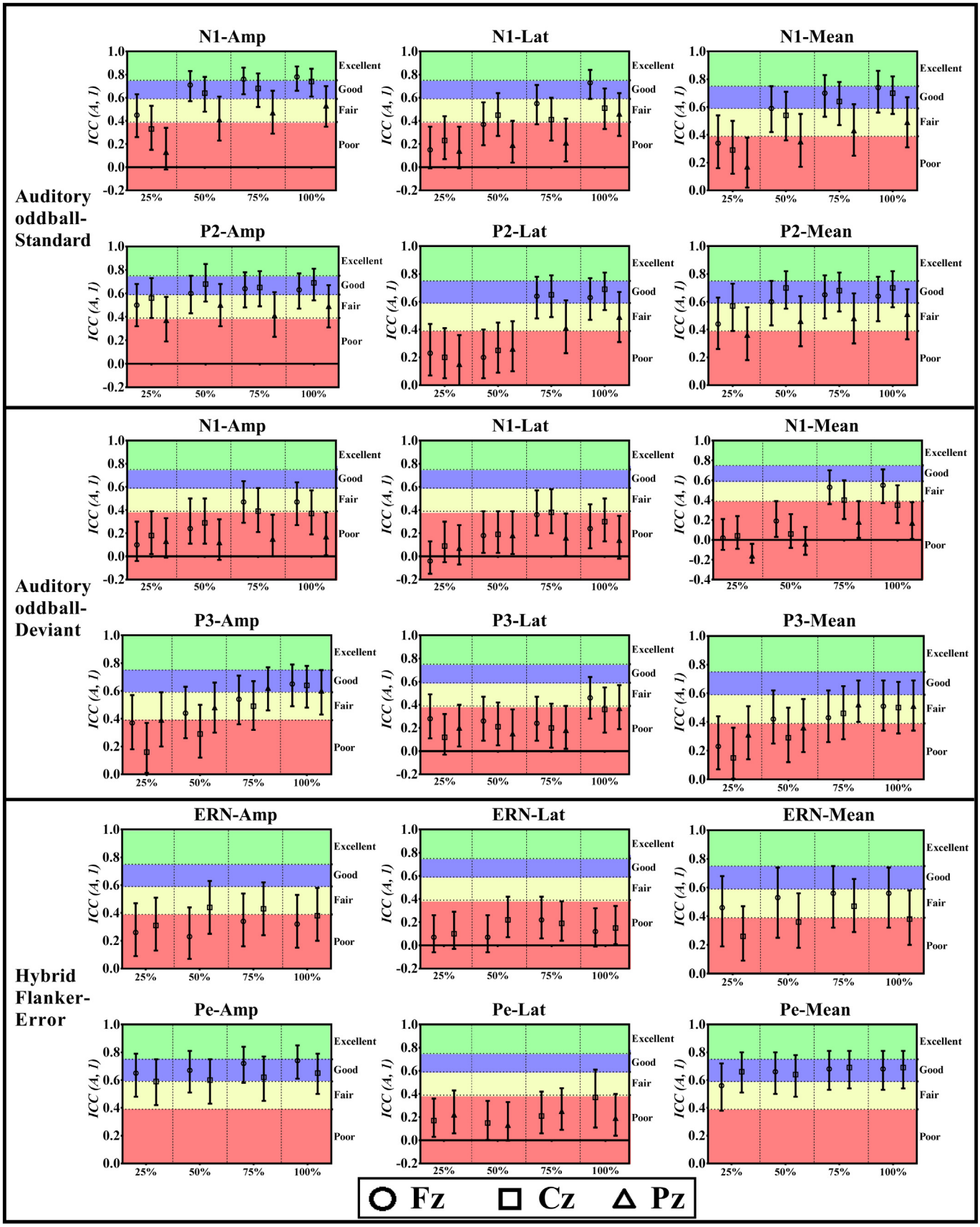

Fig. 8. Intra-class correlation coefficient (*ICC*) with increasing percentage of accepted trials of across time for ERP tasks (please refer to the Supplementary materials for adjacent time points). Four percentages of accepted epochs are assessed: 25%, 50%, 75% and 100%. Test-retest reliability is estimated by the single measure *ICC (A, 1)*. The mean and confident intervals for *ICC*s are shown in the figure.

Table 2
The number of accepted epochs for different percentages.

Task	Percentage	Condition	BL1	BL2	BL3	BL4
Auditory oddball	25%	Standard	44 ± 7(30–53)[a]	39 ± 9(25–57)	40 ± 8(24–54)	41 ± 8(20–57)
		Deviant	8 ± 1(6–10)	7 ± 2(4–10)	7 ± 2(3 − 10)	7 ± 2(4–10)
	50%	Standard	87 ± 13(59–106)	78 ± 19(51–114)	79 ± 16(49–108)	81 ± 17(40–114)
		Deviant	15 ± 2(11–19)	14 ± 4(8–20)	14 ± 3(6–19)	14 ± 3(8–20)
	75%	Standard	130 ± 20(88–159)	116 ± 28(76–170)	119 ± 24(73–161)	122 ± 25(60–170)
		Deviant	23 ± 3(17–29)	21 ± 5(12 − 30)	21 ± 5(9–29)	21 ± 5(11 − 30)
	100%	Standard	180 ± 23(117–212)	169 ± 37(101–227)	167 ± 27(97–215)	174 ± 34(80–227)
		Deviant	31 ± 5(22–38)	30 ± 7(16–40)	29 ± 6(12–38)	30 ± 7(15–40)
Hybrid Flanker	25%	Error	22 ± 9(7–45)	18 ± 7(2–34)	17 ± 9(2–46)	18 ± 8(1–36)
	50%		43 ± 18(14–90)	35 ± 14(4–67)	34 ± 18(4–92)	36 ± 16(2–71)
	75%		64 ± 26(21–135)	53 ± 21(6–100)	51 ± 28(6–138)	54 ± 23(3–106)
	100%		85 ± 35(28–180)	70 ± 28(8–133)	67 ± 36(8–183)	72 ± 31(3–141)

Notes. [a] The minimum and maximum of epochs are provided in the brackets. The mean and standard deviation are reported.

be obtained with 6 and 2 error trials, correspondingly. Our data extended the results by presenting similar reliability when the number of accepted trials was increased. Moreover, we found that the reliability of ERP measures is affected by the size of the components. Smaller-sized components such as N100 and ERN exhibit lower reliability relative to larger-sized components, such as P300 and Pe (Fig. 7). This discrepancy could be caused by the difference in SNRs existing in different sizes of ERP components (Luck, 2005). Increasing the number of averaged trials and a better control of artifacts could increase the SNR for ERP components, thereby leading to a higher reliability.

Furthermore, we found that amplitude measures are more stable than latency measures, which is consistent with previous findings (Cassidy et al., 2012). The scoring method could play an important role in causing the discrepancies between ERP parameters, in which amplitude seems to be less susceptible to different scoring methods (Olvet and Hajcak, 2009a; Weinberg and Hajcak, 2011) than latency (Brunner et al., 2013). Brunner et al. (2013) compared the reliability of conventional peak measures and of the fractional area approach (FA) for the measures of independent component analysis (ICA). Their results suggested that the FA approach leads to an increase in the reliability of latency measures between two recording sessions, especially for the late components. On the other hand, Olvet and Hajcak (2009a) found similar reliabilities using both the area and peak measures, which indicate that the reliability of amplitude measurements is affected by different scoring methods to a lesser degree. In the present study, only peak-picking analysis was implemented, and thus it was difficult to capture the best method for the reliability of latency measures. Further investigation is needed to improve the reliability of latency measures in general.

4.3. Statistical methodology

To date, EEG/ERP reliability studies have mainly been conducted over two recording sessions, while more than two sessions are involved in most pharmacological studies. It is important to evaluate how EEG changes across longer periods of time and across multiple sessions. The reason for choosing a linear mixed model for the present analysis is that it can evaluate different recording sessions through an unstructured covariance matrix, an approach which is assumption-free on the covariance matrix, given the fact that we cannot be sure whether EEG/ERP parameters decay, increase or remain stable between the different recordings. This approach is adequate for assessing multiple recordings. We used the *ICC* coefficients to quantify the reliability between subsequent recording sessions instead of the Pearson correlation coefficient since the latter is not a proper measure of reliability (see Chapter 1, (Lin et al., 2012)).

4.4. Limitations

There are some considerations that need to be taken into account before an interpretation and further generalization of our results can be made. First, to reduce the variability of our data, we excluded women in our recruited population. Even though the effect of menstrual cycle was not the main interest in the current study, it could result in lower reliabilities based on previous findings (O'Reilly et al., 2004; Walpurger et al., 2004). Moreover, Bazanova et al. (2017) demonstrated how the α amplitude suppression could change in different phases of the menstrual cycle, but the effects of different phases and the relation between phases and the reliability of EEG (or ERP) remain unclear. Hence, the presented results should be carefully interpreted since menstrual cycle could influence the test-retest reliability. Although there was previous evidence showing that the test-retest reliability is highly comparable for both genders (Tenke et al., 2018), future studies must address whether test-retest reliability changes across genders, e.g. with the menstrual phases.

In addition, the mixed model showed a significant effect of assigned sequence. We believe that this might be caused by the spurious age differences within the different sequence groups (F (3, 31) = 4.057, $p < .05$), since unfortunately age was not taken into account when the participants were randomized. This could cause low EEG/ERP reliability since age does have an impact on EEG/ERP (Hämmerer et al., 2013). However, we cannot be certain due to the relatively small sample size of the present study (eight participants per intervention sequence).

Another related issue is that of the low number of accepted trials of error ERPs at BL4 which was observed for one participant (Table 1, the minimum is 3). This participant contributed the lowest number of accepted trials (the second lowest is 12) in all sessions. This was the case due to low committed errors instead of a noisy signal (i.e. non-physiological signal). We didn't exclude the participant for two main reasons: first, this study is a clinical trial and thus it is important to report the actual data. Second, in the study of Olvet and Hajcak (2009b), they demonstrated that stable error ERPs could be measured with a minimum of six error trials, even two trials for the component Pe. Olvet and Hajcak (2009a) extended the results by comparing the reliability of high versus low number of error trials, and demonstrated similar test-retest reliability between groups. Therefore, it could be possible that the reported reliability for EPRs is underestimated but we don't believe the exclusion of this single participant's data would improve the reliability significantly.

Finally, carry-over drug effects might have existed in our dataset even though blood tests confirmed that there was a complete washout. This is because while during later baseline measurements, blood tests can eliminate the presence of previous interventions in the bloodstream (BL2, BL3, BL4), it cannot completely rule out the indirect influence a previous treatment had on a participant's subsequent test performance.

5. Conclusions

In conclusion, we find excellent test-retest reliability for power spectra measures of continuous EEG in the frequency bands of θ, α and β and an acceptable reliability for the evoked power of ERP tasks. Amplitude measures are more consistent across sessions relative to latency measures, and the amplitudes of P300 and Pe are more reliable than the amplitudes of ERN and Pe. Our results support that these EEG/ERP parameters are reliable across three-week intervals and thus are sufficiently reliable for future investigations examining pharmacological effects.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ijpsycho.2018.09.007.

Declaration of interest

All authors declare no conflict of interest.

Acknowledgement

Collection of the data included in the study was supported by H. Lundbeck A/S. C.I. was supported by the Innovationsfonden, Denmark (5189-00087A). M.G. was supported by the Lundbeck Foundation (R181-2014-3586). B.O. was supported by the Lundbeck Foundation (R231-2016-3236) and Marie-Curie-NEUROMODEL (746850).

References

Akin, M., 2002. Comparison of wavelet transform and FFT methods in the analysis of EEG signals. J. Med. Syst. 26, 241–247. https://doi.org/10.1023/A:1015075101937.

Alperin, B.R., Mott, K.K., Rentz, D.M., Holcomb, P.J., Daffner, K.R., 2014. Investigating the age-related "anterior shift" in the scalp distribution of the P3b component using principal component analysis. 51, 620–633. https://doi.org/10.1111/psyp.12206.

Bazanova, O.M., Nikolenko, E.D., Barry, R.J., 2017. Reactivity of alpha rhythms to eyes opening (the Berger effect) during menstrual cycle phases. Int. J. Psychophysiol. 122, 56–64. https://doi.org/10.1016/j.ijpsycho.2017.05.001.

Brassen, S., Adler, G., 2003. Short-term effects of acetylcholinesterase inhibitor treatment on EEG and memory performance in Alzheimer patients: an open, controlled trial. Pharmacopsychiatry 36, 304–308. https://doi.org/10.1055/s-2003-45118.

Brunner, J.F., Hansen, T.I., Olsen, A., Skandsen, T., Håberg, A., Kropotov, J., 2013. Long-term test-retest reliability of the P3 NoGo wave and two independent components decomposed from the P3 NoGo wave in a visual Go/NoGo task. Int. J. Psychophysiol. 89, 106–114. https://doi.org/10.1016/j.ijpsycho.2013.06.005.

Cassidy, S.M., Robertson, I.H., O'Connell, R.G., 2012. Retest reliability of event-related potentials: evidence from a variety of paradigms. Psychophysiology 49, 659–664. https://doi.org/10.1111/j.1469-8986.2011.01349.x.

Corsi-Cabrera, M., Calindo-Vilchis, L., Del-Río-Portilla, Y., Arce, C., Ramos-Loyo, J., 2007. Within-subject reliability and inter-session stability of EEG power and coherent activity in women evaluated monthly over nine months. Clin. Neurophysiol. 118, 9–21. https://doi.org/10.1016/j.clinph.2006.08.013.

Falkenstein, M., Hoormann, J., Christ, S., Hohnsbein, J., 2000. ERP components on reaction errors and their functional significance: a tutorial. Biol. Psychol. 51, 87–107. https://doi.org/10.1016/S0301-0511(99)00031-9.

Fernandez, T., Harmony, T., Rodriguez, M., Reyes, A., Marosi, E., Bernal, J., 1993. Test-retest reliability of EEG spectral parameters during cognitive tasks: I. Absolute and relative power. Int. J. Neurosci. 68, 255–261. https://doi.org/10.3109/00207459308994280.

Gudmundsson, S., Runarsson, T.P., Sigurdsson, S., Eiriksdottir, G., Johnsen, K., 2007. Reliability of quantitative EEG features. Clin. Neurophysiol. 118, 2162–2171. https://doi.org/10.1016/j.clinph.2007.06.018.

Hämmerer, D., Li, S.-C., Völkle, M., Müller, V., Lindenberger, U., 2013. A lifespan comparison of the reliability, test-retest stability, and signal-to-noise ratio of event-related potentials assessed during performance monitoring. Psychophysiology 50, 111–123. https://doi.org/10.1111/j.1469-8986.2012.01476.x.

Huffmeijer, R., Bakermans-Kranenburg, M.J., Alink, L.R.A., Van IJzendoorn, M.H., 2014. Reliability of event-related potentials: the influence of number of trials and electrodes. Physiol. Behav. 130, 13–22. https://doi.org/10.1016/j.physbeh.2014.03.008.

Kemp, A.H., Pe Benito, L., Quintana, D.S., Clark, C.R., McFarlane, A., Mayur, P., Harris, A., Boyce, P., Williams, L.M., 2010. Impact of depression heterogeneity on attention: an auditory oddball event related potential study. J. Affect. Disord. 123, 202–207. https://doi.org/10.1016/j.jad.2009.08.010.

Kinoshita, S., Inoue, M., Maeda, H., Nakamura, J., Morita, K., 1996. Long-term patterns of change in ERPs across repeated measurements. Physiol. Behav. 60, 1087–1092. https://doi.org/10.1016/0031-9384(96)00130-8.

Larson, M.J., Baldwin, S.A., Good, D.A., Fair, J.E., 2010. Temporal stability of the error-related negativity (ERN) and post-error positivity (Pe): the role of number of trials. Psychophysiology 47, 1167–1171. https://doi.org/10.1111/j.1469-8986.2010.01022.x.

Leiser, S.C., Pehrson, A.L., Robichaud, P.J., Sanchez, C., 2014. Multimodal antidepressant vortioxetine increases frontal cortical oscillations unlike escitalopram and duloxetine - a quantitative EEG study in rats. Br. J. Pharmacol. 171, 4255–4272. https://doi.org/10.1111/bph.12782.

Lin, L., Hedayat, A.S., Wu, W., 2012. Statistical Tools for Measuring Agreement. Springer, New York. https://doi.org/10.1007/978-1-4614-0562-7.

Loo, S.K., Specter, E., Smolen, A., Hopfer, C., Teale, P.D., Reite, M.L., 2000. Functional effects of the DAT1 polymorphism on EEG measures in ADHD. J. Am. Acad. Child Adolesc. Psychiatry 42, 986–993. https://doi.org/10.1097/01.CHI.0000046890.27264.88.

Luck, S.J., 2005. Ten Simple Rules for Designing ERP Experiments. In: Handy, T.C. (Ed.), Event-Related Potentials: A Methods Handbook. MIT Press, Cambridge, MA. https://doi.org/10.1371/journal.pcbi.0020012.

McEvoy, L.K., Smith, M.K., Gevins, A., 2000. Test–retest reliability of cognitive EEG. Clin. Neurophysiol. 111, 457–463. https://doi.org/10.1016/S1388-2457(99)00258-8.

McFadden, K.L., Steinmetz, S.E., Carroll, A.M., Simon, S.T., Wallace, A., Rojas, D.C., 2014. Test-retest reliability of the 40 Hz EEG auditory steady-state response. PLoS One 9, 59–61. https://doi.org/10.1371/journal.pone.0085748.

Mcgraw, K.O., Wong, S.P., 1996. Forming inferences about some intraclass correlation coefficients. Psychol. Methods 1, 30–46. https://doi.org/10.1037/1082-989X.1.1.30.

Mulert, C., Juckel, G., Brunnmeier, M., Karch, S., Leicht, G., Mergl, R., Möller, H., Hegerl, U., Pogarell, O., 2007. Prediction of treatment response in major depression: integration of concepts. J. Affect. Disord. 98, 215–225. https://doi.org/10.1016/j.jad.2006.07.021.

Näpflin, M., Wildi, M., Sarnthein, J., 2007. Test-retest reliability of resting EEG spectra validates a statistical signature of persons. Clin. Neurophysiol. 118, 2519–2524. https://doi.org/10.1016/j.clinph.2007.07.022.

Olbrich, S., Arns, M., 2013. EEG biomarkers in major depressive disorder: discriminative power and prediction of treatment response. Int. Rev. Psychiatry 25, 604–618. https://doi.org/10.3109/09540261.2013.816269.

Olvet, D.M., Hajcak, G., 2009a. Reliability of error-related brain activity. Brain Res. 1284, 89–99. https://doi.org/10.1016/j.brainres.2009.05.079.

Olvet, D.M., Hajcak, G., 2009b. The stability of error-related brain activity with increasing trials. Psychophysiology 46, 957–961. https://doi.org/10.1111/j.1469-8986.2009.00848.x.

O'Reilly, M.A., Cunningham, C.J., Lawlor, B.A., Walsh, C.D., Rowan, M.J., 2004. The effect of the menstrual cycle on electrophysiological and behavioral measures of memory and mood. Psychophysiology 41, 592–603. https://doi.org/10.1111/j.1469-8986.2004.00194.x.

Poyraz, B.Ç., Sakallı Kani, A., Aksoy Poyraz, C., Öcek Baş, T., Arıkan, M.K., 2017. Cognitive psychophysiological substrates of affective temperaments: a P300 study. Clin. EEG Neurosci. 48, 96–102. https://doi.org/10.1177/1550059416650112.

Rentzsch, J., Jockers-Scherübl, M.C., Boutros, N.N., Gallinat, J., 2008. Test-retest reliability of P50, N100 and P200 auditory sensory gating in healthy subjects. Int. J. Psychophysiol. 67, 81–90. https://doi.org/10.1016/j.ijpsycho.2007.10.006.

Ruchsow, M., Spitzer, M., Grön, G., Grothe, J., Kiefer, M., 2005. Error processing and impulsiveness in normals: evidence from event-related potentials. Cogn. Brain Res. 24, 317–325. https://doi.org/10.1016/j.cogbrainres.2005.02.003.

Ruchsow, M., Walter, H., Buchheim, A., Martius, P., Spitzer, M., Kächele, H., Grön, G., Kiefer, M., 2006. Electrophysiological correlates of error processing in borderline personality disorder. Biol. Psychol. 72, 133–140. https://doi.org/10.1016/j.biopsycho.2005.08.006.

Sandman, C.A., Patterson, J.V., 2000. The auditory event-related potential is a stable and reliable measure in elderly subjects over a 3 year period. Clin. Neurophysiol. 111, 1427–1437. https://doi.org/10.1016/S1388-2457(00)00320-5.

Segalowitz, S.J., Santesso, D.L., Murphy, T.I., Homan, D., Chantziantoniou, D.K., Khan, S., 2010. Retest reliability of medial frontal negativities during performance monitoring. Psychophysiology 47, 260–270. https://doi.org/10.1111/j.1469-8986.2009.00942.x.

Semlitsch, H.V., Anderer, P., Schuster, P., Presslich, O., 1986. A solution for reliable and valid reduction of ocular artifacts, applied to the P300 ERP. Psychophysiology 23, 695–703. https://doi.org/10.1111/j.1469-8986.1986.tb00696.x.

Tenke, C.E., Kayser, J., Manna, C.G., Fekri, S., Kroppmann, C.J., Schaller, J.D., Alschuler, D.M., Stewart, J.W., Mcgrath, P.J., Bruder, G.E., 2011. Current source density measures of electroencephalographic alpha predict antidepressant treatment response. Biol. Psychiatry 70, 388–394. https://doi.org/10.1016/j.biopsych.2011.02.016.

Tenke, C.E., Kayser, J., Alvarenga, J.E., Abraham, K.S., Warner, V., Talati, A., Weissman, M.M., Bruder, G.E., 2018. Temporal stability of posterior EEG alpha over twelve years. Clin. Neurophysiol. 129, 1410–1417. https://doi.org/10.1016/j.clinph.2018.03.037.

Towers, D.N., Allen, J.J.B., 2009. A better estimate of the internal consistency reliability of frontal EEG asymmetry scores. Psychophysiology 46, 132–142. https://doi.org/10.1111/j.1469-8986.2008.00759.x.

Van Deursen, J.A., Vuurman, E.F.P.M., Van Kranen-Mastenbroek, V.H.J.M., Verhey, F.R.J., Riedel, W.J., 2011. 40-Hz steady state response in Alzheimer's disease and mild cognitive impairment. Neurobiol. Aging 32, 24–30. https://doi.org/10.1016/j.neurobiolaging.2009.01.002.

Walhovd, K.B., Fjell, A.M., 2002. One-year test-retest reliability of auditory ERPs in young and old adults. Int. J. Psychophysiol. 46, 29–40. https://doi.org/10.1016/S0167-8760(02)00039-9.

Walpurger, V., Pietrowsky, R., Kirschbaum, C., Wolf, O.T., 2004. Effects of the menstrual cycle on auditory event-related potentials. Horm. Behav. 46, 600–606. https://doi.org/10.1016/j.yhbeh.2004.07.002.

Weinberg, A., Hajcak, G., 2011. Longer term test-retest reliability of error-related brain activity. Psychophysiology 48, 1420–1425. https://doi.org/10.1111/j.1469-8986.

2011.01206.x.
Williams, L., Simms, E., Clark, C.R., Paul, R.H., Rowe, D., Gordon, E., 2005. The test-retest reliability of a standardized neurocognitive and neurophysiological test battery: "Neuromarker.". Int. J. Neurosci. 115, 1605–1630. https://doi.org/10.1080/00207450590958475.
Yener, G.G., Güntekin, B., Öniz, A., Başar, E., 2007. Increased frontal phase-locking of event-related theta oscillations in Alzheimer patients treated with cholinesterase inhibitors. Int. J. Psychophysiol. 64, 46–52. https://doi.org/10.1016/j.ijpsycho.2006.07.006.