Test-Retest Reliability of EEG: A Comparison Across Multiple Resting-State and Task-State Experiments

Authors: Qin Huiyi, Ding Lihong, Duan Wei, Lei Xu, Lei Xu

Date: 2023-04-12T19:21:07+00:00

Abstract

Investigating the cognitive neural mechanisms of psychological processes based on EEG requires that the signals themselves possess good test-retest reliability. This study comprehensively compared the test-retest reliability of two resting-state EEG conditions (eyes-open and eyes-closed) and two task-state event-related potentials (the psychomotor vigilance task and Oddball task). It was found that the test-retest reliability of resting-state was generally superior to that of task-state, with eyes-closed resting-state showing higher reliability than eyes-open resting-state, and the alpha frequency band demonstrating the highest reliability among all frequency bands. For both task states, higher test-retest reliability was observed in the time domain around 200 ms post-stimulus onset. Spatially, results from all five conditions indicated that central regions exhibited higher test-retest reliability than peripheral regions, which may be related to peripheral regions being more susceptible to artifacts. This study involved multiple resting-state and task-state EEG experiments, comprehensively comparing test-retest reliability across three dimensional features (frequency domain, time domain, and spatial domain) and analyzing potential reasons, providing recommendations for selecting appropriate experimental paradigms and metrics for future research on EEG signal test-retest reliability, and holding important reference value for EEG applications in both basic and clinical fields.

Full Text

Test-retest Reliability of EEG: A Comparison Across Multiple Resting-State and Task-State Experiments

QIN Huiyi[1], DING Lihong[1], DUAN Wei[1,2], LEI Xu[1]

[1]Faculty of Psychology, Southwest University, Chongqing 400715, China
[2]Institute of Psychology, Chinese Academy of Sciences, Beijing 100101, China

Abstract

Investigating the cognitive and neural mechanisms of psychological processes using EEG requires the signal itself to possess robust test-retest reliability. This study comprehensively compared the test-retest reliability of two resting-state EEG conditions (eyes-open and eyes-closed) and two task-state event-related potentials (psychomotor vigilance task and Oddball paradigm) from multiple perspectives. We found that resting-state EEG generally exhibited superior test-retest reliability compared to task-state ERPs, with the eyes-closed condition showing higher reliability than the eyes-open condition. The alpha frequency band demonstrated the highest reliability across all bands. For both task states, the highest reliability was observed around 200 ms post-stimulus onset. Spatially, all five conditions showed higher reliability in central regions than peripheral regions, likely due to fewer artifacts in central areas. By examining multiple resting-state and task-state EEG experiments across frequency, time, and spatial domains, this study provides a comprehensive comparison of test-retest reliability and analyzes potential underlying causes. These findings offer recommendations for selecting appropriate experimental paradigms and metrics in future EEG reliability studies and hold important reference value for both basic and clinical EEG applications.

Keywords: EEG, event-related potential, test-retest reliability, resting-state EEG

Introduction

The development and application of neuroimaging technologies constitute an important foundation for brain science research. Among various brain imaging techniques—including functional magnetic resonance imaging (fMRI), electroencephalography (EEG), and magnetoencephalography (MEG)—EEG has become the most commonly used tool in cognitive neuroscience due to its high temporal resolution, simple operation, non-invasiveness, portability, and low cost. To investigate the cognitive and neural mechanisms of psychological processes using scalp EEG signals, the signals must exhibit good test-retest reliability (Elliott et al., 2020).

Resting-state EEG (rsEEG) and task-state EEG (typically event-related potentials, ERP) represent the two most common forms of scalp EEG research. rsEEG refers to the rhythmic spontaneous neural electrical activity recorded from the scalp during a wakeful resting state, reflecting intrinsic brain connectivity and providing information about individual differences in cognition and personality (Deco, Jirsa, & McIntosh, 2011). Consequently, many studies have extended rsEEG to clinical research as a biomarker for neuropsychiatric disorders, including Alzheimer's disease (Bonanni et al., 2008), epilepsy (Rotondi et

al., 2016), schizophrenia (Siebenhühner et al., 2013), and insomnia (Zhao et al., 2021). In contrast, task-state EEG typically involves recording brain activity during cognitive tasks, often resulting in event-related potentials (ERP) after averaging multiple trials (Wang et al., 2022). ERPs reflect neurophysiological changes during cognitive processes, enabling investigation of individual cognitive characteristics and dynamic processes, and have been applied in brain-computer interfaces (Lugo et al., 2020).

EEG analysis techniques include waveform analysis, spectral analysis, time-frequency analysis, and source localization (赵文瑞 et al., 2020). Spectral analysis typically employs frequency band divisions: delta (0.5–4 Hz), theta (4–8 Hz), alpha (8–13 Hz), beta (13–30 Hz), and gamma (30–80 Hz) (Tatum et al., 2016). This study focuses on spectral information from rsEEG and waveform characteristics from ERPs—the former providing power spectral information in specific frequency bands, while the latter examines amplitude and latency of ERP components at specific time points.

The stability and consistency of EEG across different contexts and time intervals have long been concerns. Theoretical calculations indicate that reliability determines the maximum detectable validity, yet the interplay among reliability, sample size, and effect size is often underestimated in psychological research (Zuo et al., 2019). Test-retest reliability studies are therefore particularly important. The commonly used statistic for assessing test-retest reliability is the intra-class correlation coefficient (ICC), typically defined as the ratio of between-subject variance to total variance. Generally, ICC values less than 0.5 indicate poor reliability, 0.5–0.75 indicate moderate reliability, 0.75–0.9 indicate good reliability, and values greater than 0.9 indicate excellent reliability (Koo & Li, 2016).

Recent studies have compared the test-retest reliability of rsEEG and ERP across different contexts, revealing respective advantages. Compared to ERP, rsEEG features simpler experimental designs and is easier to implement in special patient populations or specific cognitive states. ERP, however, allows for more targeted behavioral tasks and provides more accurate electrophysiological indices of cognitive processes. Nevertheless, most existing EEG reliability studies have focused on only one category. For instance, some researchers have primarily examined resting-state conditions, investigating reliability under different preprocessing conditions and data lengths, finding that eyes-closed conditions generally show higher reliability than eyes-open conditions (Duan et al., 2021), and that specific frequency bands such as alpha and theta exhibit better test-retest reliability (Corsi-Cabrera et al., 2007; McEvoy et al., 2000). Other studies have focused on task-state reliability, reporting ERP test-retest correlations above 0.9 in auditory Oddball tasks, with correlations slightly lower but still reaching 0.8 over 12–16 week intervals (Salinsky et al., 1991). In psychomotor vigilance tasks (PVT), ERP reliability has been found to be excellent ($r > 0.9$) across intervals of several hours to several days (McEvoy et al., 2000).

Overall, previous studies exhibit substantial differences in reliability metrics,

experimental design, data processing, interval duration, and participant populations, reducing comparability across experimental paradigms. For example, although most studies have found higher reliability in eyes-closed rsEEG, they have not compared it with ERP under identical testing conditions. When searching for optimal reliability metrics in ERP, most studies have focused horizontally on identifying optimal components within individual tasks, lacking longitudinal comparisons across multiple tasks to obtain more generalizable recommendations. Spatially, few studies have examined reliability across electrode locations. While resting-state MEG research has found that parietal regions show the highest reliability (Lew et al., 2021), it remains uncertain whether this generalizes to scalp EEG.

Despite numerous studies examining EEG reliability, systematic horizontal comparisons of EEG reliability across different experimental conditions are still lacking (Bennett & Miller, 2010; Elliott et al., 2020). There is an urgent need for experimental designs that enable direct comparisons of different tasks and analysis methods within the same participant sample. Furthermore, reliability depends on preprocessing steps and the selection of brain functional metrics; horizontal comparisons facilitate identification of the most suitable metrics for investigating individual differences (Zuo et al., 2019). In this study, we discuss the test-retest reliability of resting-state and task-state EEG across three dimensions—frequency, time, and space—from a more comprehensive perspective. We recorded rsEEG and ERP from 42 healthy adult participants across three testing sessions with intervals of 1.5 hours and one month, conducting comprehensive comparisons of test-retest reliability across these EEG types to identify overall reliability distributions, optimal frequency bands, optimal time points, and optimal scalp regions. Based on previous research, we hypothesized that: (1) rsEEG would demonstrate better overall test-retest reliability than ERP; (2) EEG signal-to-noise ratio (SNR) would influence reliability, with higher SNR yielding better reliability, thus the alpha band in rsEEG, task-relevant components (time windows, brain regions) in ERP, and parietal regions less affected by ocular artifacts would show higher reliability. This study attempts to compare test-retest reliability across different dimensional features and propose possible explanations, thereby providing recommendations for selecting appropriate experimental paradigms and metrics in future EEG studies.

Methods

Participants

This study recruited 42 healthy participants (14 males), all right-handed, aged 18–26 years (mean age = 19.5 ± 1.4 years). All participants had normal or corrected-to-normal vision, no history of neurological disease or head trauma, no sleep or psychological disorders, and reported good sleep quality in the previous week. On the day of EEG recording, participants were prohibited from consuming alcohol or caffeine-containing foods and beverages. Participants received compensation after completing the experiment. All participants provided

informed consent after the experimental procedure was explained in detail by the experimenter. This study was approved by the Ethics Committee of the authors' institution, and all experiments complied with the Declaration of Helsinki.

Experimental Design

Each participant was invited to the laboratory twice for three EEG recording sessions. During the first visit, participants underwent two EEG tests with a 90-minute interval between them. One month later, participants returned for the second visit and completed the third test. During each EEG recording session, all participants completed four conditions: two resting states (eyes-open and eyes-closed) and two task states (PVT and Oddball tasks). In each session, participants first completed the resting-state conditions, followed by the task states, with the two tasks presented in random order across participants. Specifically, during resting-state EEG recording, participants were instructed to watch a white fixation cross at the center of the screen for 5 minutes (eyes-open, EO) and then close their eyes and rest for 5 minutes (eyes-closed, EC), remaining as still, quiet, and relaxed as possible. During the PVT task, participants were required to respond as quickly as possible to a timer appearing at the center of the screen; ERP analysis segmented EEG signals with the timer onset as the baseline (time 0). During the Oddball task, participants counted high-frequency or low-frequency sounds that were randomly presented as high-probability (75%) standard stimuli and low-probability (25%) deviant stimuli; ERP analysis segmented EEG signals with sound onset as the baseline.

EEG Signal Acquisition

EEG was recorded from 63 Ag/AgCl electrodes mounted in an elastic cap according to the extended international 10–20 system (Brain Products GmbH, Steingrabenstr, Germany). Two additional electrodes recorded ocular activity, and FCz served as the online reference electrode. The sampling frequency was 500 Hz, and electrode impedance was maintained below 5 kΩ using conductive gel. Five conditions were included in the analysis: two resting states (eyes-open and eyes-closed) and three task-state ERPs (Oddball standard, Oddball deviant, and PVT).

Resting-State EEG Preprocessing and Spectral Analysis

Preprocessing of raw rsEEG data consisted of four steps implemented using EEGLAB (version 2019_1, http://sccn.ucsd.edu). First, raw EEG data were bandpass filtered between 0.3–45 Hz. Second, bad channels were identified through visual inspection, and removed channel signals were reconstructed using linear interpolation (replaced with the average of neighboring channels). The average number of bad channels per participant per session was 1.61 ($±$1.96, range 0–14). Third, data were re-referenced to the global average and segmented into 4-second epochs, with bad epochs identified and removed through visual inspection. The final number of retained epochs was 72.56 ± 4.30 for eyes-closed

and 71.44 ± 5.47 for eyes-open conditions. Fourth, independent component analysis (ICA) was performed to remove ocular artifact components.

For preprocessed rsEEG datasets, we calculated the power spectrum for each electrode using Welch's method. Absolute power values for each electrode were log-transformed to compute power spectra (1 dB = 10×log V^2).

ERP Preprocessing and Analysis

ERP data preprocessing steps and software were identical to those used for rsEEG. ERP epochs were 1500 ms in length, including a 500 ms pre-stimulus baseline and 1000 ms post-stimulus period. The average number of bad channels across all participants was 1.82 ($±$1.73, range 0–9). The –200 to 0 ms interval was used for baseline correction, and three ERP types were extracted: PVT, Oddball deviant, and Oddball standard. The final number of retained epochs for the three conditions was 36.13 ± 8.71, 12.83 ± 4.40, and 51.13 ± 12.62, respectively. ERP waveforms were obtained by averaging across epochs for each condition.

Data Analysis

We used intra-class correlation coefficients (ICC) to assess test-retest reliability of signals across the five experimental conditions. ICC compares between-subject variability to total variability to evaluate the reliability of repeated measurements (Koo & Li, 2016). Higher ICC values indicate lower within-subject variability and greater variability attributable to between-subject differences. Following spectral analysis, we calculated ICC values for five conditions (two resting states, three task states). ICC was computed based on:

In this study, d (=3) represents the number of test sessions, n (=42) represents the number of participants, MSp represents the mean square between subjects, and MSe represents the mean square error. Yij represents the measurement value for the ith participant in the jth session, Ȳi represents the mean across sessions for the ith participant, Ȳj represents the mean across participants for the jth session, and Ȳ represents the grand mean across all measurements and participants.

We used analysis of variance (ANOVA) to verify the interaction between electrode location and experimental condition. P-values less than 0.05 were considered statistically significant. Shapiro-Wilk tests were used in all analyses, and Greenhouse-Geisser correction was applied to adjust for any violations of sphericity.

Results

Power Spectrum and ERP Waveform Analysis

We first averaged signals from the Cz electrode and calculated the mean and standard deviation across all participants. Figure 1 [Figure 1: see original paper]A displays the spectral activity for eyes-open and eyes-closed rsEEG across the 0.3–45 Hz range during the three measurement sessions. Absolute power was generally higher in the eyes-closed state than in the eyes-open state, with maximum power for both conditions concentrated in the low-frequency range (approximately 0.3–13 Hz) and peaking around 10 Hz, corresponding to the alpha frequency band—consistent with previous research findings.

Figure 1B shows the amplitude patterns for the three ERP conditions. The PVT task evoked larger ERP components compared to the Oddball task. The PVT task elicited an initial positive wave around 100 ms and maximum amplitude between 300–700 ms. For the Oddball task, the two stimulus types overlapped substantially before 250 ms, both eliciting an N1 wave around 130 ms (with slightly larger amplitude for deviant stimuli) and a P2 wave around 250 ms (also slightly larger for deviant stimuli). After 250 ms, the ERP waveforms diverged considerably, with deviant stimuli showing larger amplitudes than standard stimuli. A prominent P3 wave emerged for deviant stimuli between 300–400 ms. Comparison across the three test sessions revealed high similarity in both rsEEG spectral plots and ERP waveforms, with greater overlap between the first two sessions. This indicates good test-retest reliability across all five EEG signals, with short-term reliability superior to long-term reliability.

Figure 1. Spectral plots (A) and waveform plots (B) for two resting-state EEG and three task-state ERP conditions across three experimental sessions. Electrode position is Cz; solid lines represent means, and light shading represents standard error. (A) Spectral plots for two resting-state conditions; the x-axis represents frequency, the y-axis represents absolute power; purple line indicates eyes-open resting state (EO), green line indicates eyes-closed resting state (EC). (B) Waveform plots for three task-state conditions; the x-axis represents time, the y-axis represents amplitude; black line indicates PVT paradigm, yellow line indicates Oddball deviant stimulus (Oddball-D), blue line indicates Oddball standard stimulus (Oddball-S).

Note: EO: eyes-open; EC: eyes-closed; PVT: psychomotor vigilance task; Oddball-D: deviant stimulus of oddball task; Oddball-S: standard stimulus of oddball task.

Overall Test-Retest Reliability

Following preliminary data quality checks, we computed intra-class correlation coefficients (ICC) for two rsEEG and three ERP conditions. Results revealed that rsEEG generally demonstrated superior test-retest reliability compared to

ERP (Figure 2 [Figure 2: see original paper]). ICC values for both resting states clustered around 0.5, with the eyes-open condition showing concentrations at 0.45 and 0.6, and the eyes-closed condition at 0.4 and 0.6. The overall ICC ranges were similar between conditions (Figure 2A), though the eyes-closed condition outperformed the eyes-open condition in mean, maximum, and median ICC values (Table 1), with a significant difference in paired samples t-test ($t = –2.71$, $p = 0.007$). For ERPs, the PVT condition showed superior reliability to the Oddball paradigm across mean, maximum, and median ICC values (Table 1). PVT ICC values clustered around 0.5 with a more concentrated distribution, while Oddball ICC values clustered around 0.35 (Figure 2B). One-way ANOVA of ICC values across the three ERP conditions revealed significant differences ($F = 753.16$, $p < 0.001$), with post-hoc tests showing significant differences between PVT and both Oddball conditions, but no significant difference between the two Oddball conditions.

Figure 2. Distribution of ICC values across power (A) and amplitude (B) for the five conditions, including all electrodes, frequency points (rsEEG), and time points (ERPs). The x-axis represents condition, the y-axis represents ICC value. Black reference line indicates mean, red reference line indicates median. (A) ICC value distribution for two resting-state conditions. (B) ICC value distribution for three task-state conditions.

Note: EO: eyes-open; EC: eyes-closed; PVT: psychomotor vigilance task; Oddball-D: deviant stimulus of oddball task; Oddball-S: standard stimulus of oddball task.

Table 1. Overall ICC distribution across five conditions

Condition	Mean ± SD
EO	0.54 ± 0.16
EC	0.55 ± 0.19
PVT	0.50 ± 0.13
Oddball-D	0.33 ± 0.13
Oddball-S	0.33 ± 0.16

Statistical comparisons: EC > EO, $t = –2.71$, $p = 0.007$; PVT > Oddball-D / Oddball-S, $F = 753.16$, $p < 0.001$

Note: Statistical tests were conducted separately for resting-state and task-state conditions. Paired samples t-test was used for resting-state conditions, and one-way within-subjects ANOVA was used for task-state conditions. “>” indicates significant difference ($p < 0.05$); “/” indicates no significant difference ($p \geq 0.05$).

Time-Frequency Domain Reliability

We averaged ICC values across all electrodes within the frequency domain for resting-state and time domain for task-state conditions to evaluate reliability trends. For resting-state conditions, maximum ICC values for both rsEEG appeared in the alpha band (8–12 Hz) (ICC > 0.8), with ICC values for other bands clustering around 0.6 and 0.4. Notably, EC showed larger ICC values in the 0–23 Hz range, while EO showed larger ICC values in the 23–45 Hz range (Figure 3 [Figure 3: see original paper]A). For task-state conditions, maximum ICC values for both paradigms appeared around 200 ms (200 ± 20 ms) (ICC_{PVT} > 0.6, ICC_{Oddball}-S > 0.5, ICC_{Oddball}-D > 0.4), with ICC values at other time points slightly lower. Given that ERP component analysis typically examines peak amplitudes, we additionally calculated ICC values for the N1 component in PVT (130–200 ms) (Hoedlmoser et al., 2011) and the P200 component in Oddball (150–275 ms) (赵文瑞 et al., 2020). These component-specific ICC calculations yielded results consistent with our 200 ms findings: ICC values were 0.67 for PVT, 0.40 for Oddball-D, and 0.28 for Oddball-S. Notably, PVT showed larger ICC values than Oddball across all time ranges. Within the Oddball paradigm, standard stimuli elicited larger ICC values in the 0–400 ms range, while deviant stimuli elicited larger ICC values after 400 ms (Figure 3B).

Figure 3. ICC value variations across all electrodes for five conditions. Solid lines represent means, light shading represents standard error. (A) ICC value variations for two resting-state conditions; x-axis represents frequency, y-axis represents ICC value; purple line indicates eyes-open (EO), green line indicates eyes-closed (EC). (B) ICC value variations for three task-state conditions; x-axis represents time, y-axis represents ICC value; black line indicates PVT, yellow line indicates Oddball deviant (Oddball-D), blue line indicates Oddball standard (Oddball-S).

Note: EO: eyes-open; EC: eyes-closed; PVT: psychomotor vigilance task; Oddball-D: deviant stimulus of oddball task; Oddball-S: standard stimulus of oddball task.

Spatial Distribution of Reliability

The preceding analyses revealed that maximum ICC values for resting-state appeared in the alpha band (8–12 Hz), while maximum ICC values for task-state appeared around 200 ms ($±$20 ms). To compare spatial locations while minimizing confounding factors, we selected the ranges where maximum ICC values occurred (alpha band, ~200 ms) for each of the five signals. As shown in Figure 4 [Figure 4: see original paper], for resting-state conditions, higher ICC values in EO were concentrated in central and parietal regions, while EC showed generally larger ICC values with broader distribution. Both resting-state conditions exhibited high ICC value clusters in central regions. For task-state conditions, PVT showed high ICC value regions in frontal, central, and parietal areas, while Oddball standard stimuli showed larger ICC values; both stimulus

types exhibited similar high ICC regions in frontal, central, and temporal areas.

Figure 4. Scalp topographical distribution of ICC values corresponding to the alpha band for rsEEG and the P2 component for ERP.

Note: EO: eyes-open; EC: eyes-closed; PVT: psychomotor vigilance task; Oddball-D: deviant stimulus of oddball task; Oddball-S: standard stimulus of oddball task.

Comparing the topographical distributions of maximum ICC values across the five conditions revealed that the first three conditions showed maximum ICC values primarily in central regions. For Oddball conditions, however, temporal regions also showed higher ICC values in addition to central regions. Based on SNR theory, we hypothesized that central regions showed larger ICC values in the first three conditions because this region is less affected by noise artifacts (ocular, cardiac, muscular) and thus has higher SNR. For Oddball conditions, greater activation in temporal regions resulted in larger EEG amplitudes recorded by temporal electrodes, thereby increasing SNR and ICC values. To test this hypothesis, we divided scalp electrodes into a central region less affected by noise and a peripheral region more affected by noise. As shown in Figure 5 [Figure 5: see original paper], the peripheral region included 30 electrodes (AFz, FPz, FP1, etc.), with the remainder constituting the central region. Mean ICC values were calculated for both regions (Figure 5), and a two-way ANOVA was conducted with electrode location and experimental condition as factors, followed by independent samples t-tests. ANOVA results showed a significant main effect of region ($F(1, 23) = 18.58$, $p < 0.001$), with central region ICC values 0.068 higher than peripheral region ($p < 0.001$). The condition factor also significantly affected ICC values ($F(4, 92) = 260.01$, $p < 0.001$). Given five condition levels, pairwise comparisons revealed statistically significant differences in ICC across all conditions ($p < 0.001$). ANOVA results confirmed significant main effects of electrode region and experimental condition, though the interaction effect was not significant. Post-hoc independent samples t-tests revealed significant differences between central and peripheral region ICC values for EO, EC, PVT, and Oddball-D conditions, but not for Oddball-S (Table 2).

Figure 5. Mean ICC values for two regions of interest (central and peripheral) across five conditions (EO, EC, PVT, Oddball-D, Oddball-S). The x-axis represents condition, the y-axis represents ICC value. Blue bars represent central region, orange bars represent peripheral region. Inset shows electrode division into central (blue) and peripheral (orange) regions.

Note: EO: eyes-open; EC: eyes-closed; PVT: psychomotor vigilance task; Oddball-D: deviant stimulus of oddball task; Oddball-S: standard stimulus of oddball task.

Table 2. ICC values for two regions of interest across five conditions and independent samples t-test results

Condition	Central Region	Peripheral Region	t-value	p-value
EO	0.58 ± 0.15	0.51 ± 0.16	3.42	<0.001
EC	0.60 ± 0.18	0.51 ± 0.19	3.68	<0.001
PVT	0.54 ± 0.12	0.46 ± 0.13	4.71	<0.001
Oddball-D	0.36 ± 0.12	0.31 ± 0.13	2.89	0.004
Oddball-S	0.35 ± 0.16	0.32 ± 0.16	1.52	0.130

Note: EO: eyes-open; EC: eyes-closed; PVT: psychomotor vigilance task; Oddball-D: deviant stimulus of oddball task; Oddball-S: standard stimulus of oddball task.

Discussion

By employing multiple task-state and resting-state experiments, this study compared the test-retest reliability of five EEG conditions across time, frequency, and spatial domains. We found that rsEEG generally showed higher ICC values than ERP. The eyes-closed resting state exhibited higher reliability than the eyes-open state, consistent with previous research (Corsi-Cabrera et al., 2007; Duan et al., 2021). Both eyes-open and eyes-closed states showed superior reliability in the alpha frequency band compared to other bands. Under task-state conditions, PVT showed higher ICC values than Oddball, with maximum reliability occurring around 200 ms post-stimulus for both paradigms. Spatially, maximum ICC values were distributed in central regions across all conditions except Oddball standard stimuli. To our knowledge, this is the first study to compare EEG test-retest reliability across multiple dimensions—time, frequency, and space—and to identify several conditions that enhance reliability.

Reliability of Resting-State Power Spectra

The spectral plots revealed that both resting-state conditions showed maximum power in the alpha band, which primarily reflects inhibition of task-irrelevant neural activity (Klimesch et al., 2007; Uusberg et al., 2013), consistent with the high reliability observed in eyes-closed resting state dominated by alpha activity.

Comparing the two rsEEG conditions revealed similar ICC distributions with large ranges, both showing maximum values above 0.9. Studies of 30-day quantitative EEG have also found good test-retest reliability in resting states, with ICC averages reaching 0.81 in eyes-closed conditions (Cannon et al., 2012). Our results showing higher reliability in eyes-closed versus eyes-open conditions align with previous findings (Corsi-Cabrera et al., 2007). Based on SNR theory (Zuo et al., 2019), we analyzed two important factors affecting reliability beyond inherent differences in EEG activity between conditions. First, from a signal perspective, alpha band activity is larger in eyes-closed than eyes-open conditions, resulting in greater signal energy in the power spectrum. Additionally, alpha rhythms are associated with cognitive activity across the whole brain and

can be detected across widespread electrode regions, whereas high-frequency activity (e.g., beta and gamma rhythms) is more associated with local brain activity (Martín-Buro et al., 2016), weakening its whole-brain impact. Second, from a noise perspective, these two conditions differ in their susceptibility to ocular artifacts, with significantly fewer eye movement artifacts in eyes-closed states, resulting in lower noise energy (Ding et al., 2022).

Reliability of Task-State ERP Components

ERP component analysis revealed that in the PVT paradigm, P1 (positive component peaking around 100 ms post-stimulus) and N1 (negative component peaking around 160 ms) are closely related to selective, reflexive, visuospatial attention and feature detection (Hoedlmoser et al., 2011). In the Oddball paradigm, standard stimuli elicited N1 and P2 waves, while deviant stimuli elicited N1, P2, N2, P3, and N4 waves (Figure 1). Our ERP waveforms are largely consistent with previous research (Cassidy et al., 2012), initially confirming data reliability.

Based on previous EEG reliability research, we initially hypothesized that components with larger amplitudes would show better test-retest reliability. However, results did not fully support this hypothesis. For all three ERP signals, the most reliable components appeared around 200 ms. As shown in Figure 1, all three signals exhibited task-relevant components around 200 ms (N1 for PVT, P2 for Oddball), supporting the SNR-based hypothesis that greater proportion of valid information yields higher reliability (Zuo et al., 2019). Previous studies of multiple tasks have also found that early-latency components (P1, N1, N170, ERN) show the most stable peak amplitudes (Cassidy et al., 2012). Therefore, based on EEG acquisition methods and SNR theory, early ERP components reflecting more synchronized neural activity from the whole brain represent fewer cognitive processes. This signal synchronization is weakened as multiple cognitive processes are engaged later in the ERP, making later components more complex and potentially less consistent across test sessions.

Spatial Characteristics

Two primary approaches exist for identifying electrodes with optimal reliability: selecting locations less affected by artifacts (ocular, muscular) or selecting regions more relevant to the task. Both methods improve reliability by increasing SNR—either by reducing noise or enhancing signal. We hypothesized based on SNR theory to compare the relative impact of information and noise on reliability. From a noise reduction perspective, central regions should have higher SNR because EEG recorded from central regions is less affected by eye movements, head movements, heartbeat, and poor electrode contact than peripheral regions.

We found that maximum ICC values were concentrated in central regions across all experimental conditions, with central electrodes showing higher reliability than peripheral electrodes (except for Oddball standard stimuli), initially con-

firming that artifact noise substantially impacts reliability. The Oddball standard condition was exceptional: as shown in Figure 1B, this condition only elicited a significant peak around 200 ms, while the P3a component (220–280 ms) reflects auditory-driven, bottom-up anterior attention mechanisms activating primary auditory cortex and prefrontal regions (Vanhaudenhuyse et al., 2008), meaning Oddball standard ERP signals were more distributed in peripheral regions. Therefore, in addition to selecting electrodes less affected by artifacts, considering ERP components elicited by the experimental task to enhance signal strength is also an effective means of improving reliability. Notably, with improved electrode performance and preprocessing techniques, the current disadvantage of peripheral regions relative to central regions may be mitigated in the future.

Comparison Across Multiple Task and Resting States

To obtain more generalizable indicators, this study selected commonly used resting-state eyes-open/eyes-closed conditions and two task paradigms with minimal practice effects to ensure consistency of cognitive and psychological activities across repeated measurements. Although results showed higher ICC values for rsEEG than ERP (Figure 2), generalization requires caution. Some scholars have proposed that rsEEG reliability is lower than ERP because attention and vigilance levels vary more during rest than during task execution, whereas mental states during task performance are typically more uniform, stabilizing EEG signals (Burgess & Gruzelier, 1993). However, increasing recent research has found that specific rsEEG frequency bands show higher reliability than ERP (Shirk et al., 2017). We speculate this may be because certain frequency bands represent relatively constant cognitive-neural activity (Feyissa & Tatum, 2019). Additionally, advances in rsEEG preprocessing techniques have substantially reduced artifact interference (Duan et al., 2021).

Notably, PVT ICC values clustered around 0.5 with maximum values around 0.8, results similar to resting-state conditions (Figure 2). Early studies indicated high PVT reliability, even exceeding resting-state reliability at some electrodes and frequency bands (McEvoy et al., 2000). Although some studies measuring Oddball ERP reliability reported high stability (Salinsky et al., 1991), our findings showed the poorest ICC results for Oddball, with median values around 0.35. From a practical standpoint, researchers should strive to measure and eliminate non-interest signals (e.g., movement artifacts) to improve SNR and employ optimal methods to enhance reliability and validity (Zuo et al., 2019).

Furthermore, this study found that trial number appears correlated with reliability: eyes-closed rsEEG had the most trials and the largest ICC, while Oddball deviant had the fewest trials and the smallest ICC. Consistent with previous research (Cassidy et al., 2012; Ding et al., 2022; Duan et al., 2021), we speculate that trial number may indirectly affect reliability by influencing SNR. Increasing trial numbers in both task-state and resting-state EEG experiments may be an effective method for improving test-retest reliability.

Limitations

Although this study extensively compared EEG reliability characteristics across multiple task and resting states, several limitations remain. First, we employed conventional data processing methods to enhance comparability with other studies. However, given inconsistent conclusions about how preprocessing methods and parameters affect EEG reliability (Duan et al., 2021; Suarez-Revelo et al., 2016), our conclusions have limited reference value for substantially different preprocessing strategies and many new methods (e.g., functional connectivity, brain networks). Future research should focus more on new strategies and methods to promote standardization of EEG data processing and analysis pipelines (Cohen, 2017). Second, our EEG analysis was limited to the scalp level without source-level analysis, which may weaken theoretical support from a cognitive neuroscience perspective; source-level reliability can be referenced in recent reports (Ding et al., 2022; Duan et al., 2021). Third, when comparing resting-state and task-state conditions, duration and trial number should be considered, as experimental duration and bad-epoch proportions typically differ between states, resulting in non-uniform trial numbers. Fourth, the two state types differ in difficulty, and this study was limited to resting-state and two attention/vigilance-related task states, requiring caution when generalizing to broader naturalistic stimulation experiments and task paradigms. Fifth, our participant sample was limited to healthy young adults; validation in more diverse populations is needed.

Conclusion

In summary, EEG test-retest reliability demonstrates considerable comparability across multiple experimental conditions, time-frequency domain indices, and spatial locations, with average ICC reliability ranging from 0.33 to 0.55. Resting-state EEG showed higher reliability than task-state EEG, with eyes-closed rsEEG demonstrating the best reliability, and PVT task reflecting mental vigilance also showing high stability—both explainable by increased trial numbers and stable cognitive states producing high SNR. In the frequency domain, the alpha band showed highest reliability for resting-state; in the time domain, components around 200 ms were most stable for task-state; in the spatial domain, central regions generally outperformed peripheral regions. Evidence from these three dimensions initially supports the determining role of SNR in test-retest reliability.

Overall, our findings discuss test-retest reliability across a broader range of factors. These results provide recommendations for researchers selecting experimental paradigms and metrics and hold important reference value for EEG applications in both basic and clinical fields.

References

Bennett, C. M., & Miller, M. B. (2010). How reliable are the results from functional magnetic resonance imaging? *Annals of the New York Academy of Sciences, 1191*(1), 133–155.

Bonanni, L., Thomas, A., Tiraboschi, P., Perfetti, B., Varanese, S., & Onofrj, M. (2008). EEG comparisons in early Alzheimer's disease, dementia with Lewy bodies and Parkinson's disease with dementia patients with a 2-year follow-up. *Brain, 131*(3), 690–705.

Burgess, A., & Gruzelier, J. (1993). Individual reliability of amplitude distribution in topographical mapping of EEG. *Electroencephalography and Clinical Neurophysiology, 86*(4), 219–223.

Cannon, R. L., Baldwin, D. R., Shaw, T. L., Diloreto, D. J., Phillips, S. M., Scruggs, A. M., & Riehl, T. C. (2012). Reliability of quantitative EEG (qEEG) measures and LORETA current source density at 30 days. *Neuroscience Letters, 518*(1), 27–31.

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*(5), 659–664.

Cohen, M. X. (2017). Rigor and replication in time-frequency analyses of cognitive electrophysiology data. *International Journal of Psychophysiology, 111*, 80–87.

Corsi-Cabrera, M., Galindo-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. *Clinical Neurophysiology, 118*(1), 9–21.

Deco, G., Jirsa, V. K., & McIntosh, A. R. (2011). Emerging concepts for the dynamical organization of resting-state activity in the brain. *Nature Reviews Neuroscience, 12*(1), 43–56.

Ding, L., Duan, W., Wang, Y., & Lei, X. (2022). Test-retest reproducibility comparison in resting and the mental task states: A sensor and source-level EEG spectral analysis. *International journal of psychophysiology: official journal of the International Organization of Psychophysiology, 173*, 20–28.

Duan, W., Chen, X., Wang, Y. J., Zhao, W., Yuan, H., & Lei, X. (2021). Reproducibility of power spectrum, functional connectivity and network construction in resting-state EEG. *Journal of neuroscience methods, 348*, 108985.

Elliott, M., Knodt, A., Ireland, D., Morris, M., Poulton, R., Ramrakha, S., … Hariri, A. (2020). What is the test-retest reliability of common task-fMRI measures? New empirical evidence and a meta-analysis. *Biological Psychiatry, 87*(9), S132–S133.

Feyissa, A. M., & Tatum, W. O. (2019). Adult EEG. *Handbook of Clinical Neurology, 160*, 103–124.

Hoedlmoser, K., Griessenberger, H., Fellinger, R., Freunberger, R., Klimesch, W., Gruber, W., & Schabus, M. (2011). Event-related activity and phase locking during a psychomotor vigilance task over the course of sleep deprivation. *Journal of Sleep Research, 20*(3), 377–385.

Klimesch, W., Sauseng, P., & Hanslmayr, S. (2007). EEG alpha oscillations: the inhibition-timing hypothesis. *Brain Research Reviews, 53*(1), 63–88.

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of chiropractic medicine, 15*(2), 155–163.

Lew, B. J., Fitzgerald, E. E., Ott, L. R., Penhale, S. H., & Wilson, T. W. (2021). Three-year reliability of MEG resting-state oscillatory power. *Neuroimage, 243*, 118516.

Lugo, Z. R., Pokorny, C., Pellas, F., Noirhomme, Q., Laureys, S., Müller-Putz, G., & Kübler, A. (2020). Mental imagery for brain-computer interface control and communication in non-responsive individuals. *Annals of Physical and Rehabilitation Medicine, 63*(1), 21–27.

Martín-Buro, M. C., Garcés, P., & Maestú, F. (2016). Test-retest reliability of resting-state magnetoencephalography power in sensor and source space. *Human Brain Mapping, 37*(1), 179–190.

McEvoy, L. K., Smith, M. E., & Gevins, A. (2000). Test-retest reliability of cognitive EEG. *Clinical Neurophysiology, 111*(3), 457–463.

Newson, J. J., & Thiagarajan, T. C. (2019). EEG frequency bands in psychiatric disorders: a review of resting state studies. *Frontiers In Human Neuroscience, 12*, 521.

Rotondi, F., Franceschetti, S., Avanzini, G., & Panzica, F. (2016). Altered EEG resting-state effective connectivity in drug-naïve childhood absence epilepsy. *Clinical Neurophysiology, 127*(2), 1130–1137.

Salinsky, M. C., Oken, B. S., & Morehead, L. (1991). Test-retest reliability in EEG frequency analysis. *Electroencephalography and Clinical Neurophysiology, 79*(5), 382–392.

Shirk, S. D., McLaren, D. G., Bloomfield, J. S., Powers, A., Duffy, A., Mitchell, M. B., … Atri, A. (2017). Inter-Rater Reliability of Preprocessing EEG Data: Impact of Subjective Artifact Removal on Associative Memory Task ERP Results. *Frontiers In Neuroscience, 11*, 322.

Siebenhühner, F., Weiss, S. A., Coppola, R., Weinberger, D. R., & Bassett, D. S. (2013). Intra-and inter-frequency brain network structure in health and schizophrenia. *Plos One, 8*(8), e72351.

Suarez-Revelo, J., Ochoa-Gomez, J., & Duque-Grajales, J. (2016). Improving test-retest reliability of quantitative electroencephalography using different preprocessing approaches. *Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2016*, 961–964.

Tatum, W. O., Olga, S., Ochoa, J. G., Munger Clary, H., Cheek, J., Drislane, F., & Tsuchida, T. N. (2016). American Clinical Neurophysiology Society Guideline 7: Guidelines for EEG Reporting. *Journal Of Clinical Neurophysiology, 33*(4), 328–332.

Uusberg, A., Uibo, H., Kreegipuu, K., & Allik, J. (2013). EEG alpha and cortical inhibition in affective attention. *International Journal of Psychophysiology, 89*(1), 26–36.

Vanhaudenhuyse, A., Laureys, S., & Perrin, F. (2008). Cognitive event-related potentials in comatose and post-comatose states. *Neurocritical Care, 8*(2), 262–270.

Wang, Y., Duan, W., Dong, D., Ding, L., & Lei, X. (2022). A test-retest resting, and cognitive state EEG dataset during multiple subject-driven states. *Scientific data, 9*(1), 566.

Zhao, W. R., Li, C. Y., Chen, J. J., & Lei, X. (2020). Insomnia disorder and hyperarousal: evidence from resting-state and sleeping EEG. *Scientia Sinica (Vitae), 50*(3), 270–286.

Zhao, W., Van Someren, E. J. W., Li, C., Chen, X., Gui, W., Tian, Y., … Lei, X. (2021). EEG spectral analysis in insomnia disorder: A systematic review and meta-analysis. *Sleep Medicine Reviews, 59*, 101457.

Zuo, X. N., Xu, T., & Milham, M. P. (2019). Harnessing reliability for neuroscience research. *Nature Human Behaviour, 3*(8), 768–771.

Note: Figure translations are in progress. See original paper for figures.