NeuroImage 250 (2022) 118937

Contents lists available at ScienceDirect

NeuroImage

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

A new perspective on individual reliability beyond group effect for event-related potentials: A multisensory investigation and computational modeling

Zhenxing Hu [a,b,1], Zhiguo Zhang [a,b,c,d,1], Zhen Liang [a,b], Li Zhang [a,b], Linling Li [a,b], Gan Huang [a,b,*]

[a] School of Biomedical Engineering, Health Science Center, Shenzhen University, Shenzhen, Guangdong, 518060, China
[b] Guangdong Provincial Key Laboratory of Biomedical Measurements and Ultrasound Imaging, Shenzhen University, Shenzhen, Guangdong, 518060, China
[d] Marshall Laboratory of Biomedical Engineering, Shenzhen University, Shenzhen, 518060, China
[c] Peng Cheng Laboratory, Shenzhen, Guangdong, 518055, China

ARTICLE INFO

Keywords:
Event-related potentials
Multisensory
Group effects
Individual reliability
Computational Model

ABSTRACT

The dominant approach in investigating the individual reliability for event-related potentials (ERPs) is to extract peak-related features at electrodes showing the strongest group effects. Such a peak-based approach implicitly assumes ERP components showing a stronger group effect are also more reliable, but this assumption has not been substantially validated and few studies have investigated the reliability of ERPs beyond peaks. In this study, we performed a rigorous evaluation of the test-retest reliability of ERPs collected in a multisensory and cognitive experiment from 82 healthy adolescents, each having two sessions. By comparing group effects and individual reliability, we found that a stronger group-level response in ERPs did not guarantee higher reliability. A perspective of neural oscillation should be adopted for the analysis of reliability. Further, by simulating ERPs with an oscillation-based computational model, we found that the consistency between group-level ERP responses and individual reliability was modulated by inter-subject latency jitter and inter-trial variability. The current findings suggest that the conventional peak-based approach may underestimate the individual reliability in ERPs and a neural oscillation perspective on ERP reliability should be considered. Hence, a comprehensive evaluation of the reliability of ERP measurements should be considered in individual-level neurophysiological trait evaluation and psychiatric disorder diagnosis.

1. Introduction

Event-related potentials (ERPs) are noninvasive electrophysiological measures of indexing a range of sensory, cognitive, and motor processes involved in human brain activity. In clinical and translational applications of ERPs, a key challenge is to identify a reliable and valid mapping between individuals' brain activation and their perceptual or cognitive capacities (Nelson and Guyer, 2012). Measurement reliability is the prerequisite for clinical applications of ERPs, such as assessments of meditative practice using sensory-evoked potentials (Cahn and Polich, 2006) or diagnoses of psychiatric cognitive dysfunction by cognitive ERPs like P300 (Polich, 2004), and studies concerning reliability have received more attention recently (Dubois and Adolphs, 2016; Höller et al., 2017; Noble et al., 2019; Croce et al., 2020).

Originating from the field of psychometrics, reliability reflects the "trustworthiness" of a measure and denotes the extent to which a measure will yield a reproducible difference between individuals (Kraemer, 2014). The importance of reliability in the research of individual difference cannot be overstated, regardless of the data analytics approaches used (e.g., correlational analysis or machine learning). In correlational analysis, the ability to find correlations between brain activation and cognitive behavior depends on the reliability of these measures (Goodhew and Edwards, 2019). In other words, the maximum possible correlation is constrained by the reliability of the individual measures used to calculate the correlation (Spearman, 1910). In machine learning-based individualized prediction, reliability has been proved mathematically to provide a lower bound on predictive accuracy (Bridgeford et al., 2020).

Since the first systematic study on the reliability of ERPs (Segalowitz and Barnes, 1993), numerous studies have evaluated the test-retest reliability of ERP amplitude and the latency elicited from a variety of experimental paradigms (Cassidy et al., 2012; Cruse et al., 2014), but the primary focus has always been restricted to narrow time windows around ERP peaks (Thigpen et al., 2017; Cruse et al., 2014; Ip et al.,

* Corresponding author at: School of Biomedical Engineering, Health Science Center, Shenzhen University, Shenzhen, 518060, China.
E-mail address: huanggan1982@gmail.com (G. Huang).
[1] Zhenxing Hu and Zhiguo Zhang contributed equally to this paper.

https://doi.org/10.1016/j.neuroimage.2022.118937.
Received 5 August 2021; Received in revised form 14 January 2022; Accepted 24 January 2022
Available online 25 January 2022.

2018). Characteristic features, including latency, maximum amplitude, mean amplitude, and area under the window, are typically used to examine the reliability of ERPs. These ERP features are used in a machine learning model or correlation analysis to establish linkage between ERPs and cognitive/behavioral variables (Hu and Iannetti, 2019). However, such an analysis routine implicitly assumes that only the peak-related ERP measures reflect the subject-specific neurophysiological process to an external stimulus. This assumption is problematic because the entire ERP shapes (rather than latency and amplitude of ERP peaks) are physiologically meaningful and important (Gaspar et al., 2011). Taking the temporal evolution of facial emotion perception as an example, the temporal shape of ERP can provide valuable clues about processing dynamics beyond what can be inferred from data restricted to ERP peaks (Van Rijsbergen and Schyns, 2009).

ERP peaks represent the strongest group effects (i.e., group-level experimental effects among different conditions/cohorts). More specifically, by comparing the ERP response with its baseline activity, or contrasting two experimental conditions (i.e., the ERP difference wave), peak-related features of well-known ERP components, like N100, N200, and P300, were claimed to be closely associated with various perceptual and cognitive variables (Sur and Sinha, 2009). Here, the focus was on significant group effects responding to one condition versus another. As a representative example relevant to this research, the P300 was found to reflect the processes involved in stimulus evaluation or categorization as evidenced by experimental manipulation; thus, it is often reasonable to ask whether peak-related features of P300 reflect an individual's cognitive function. From this perspective, as the dominant approach in investigating individual differences in ERPs, peak-based analysis implicitly employs group-level prior information. However, from the perspective of individual difference, it remains unclear whether peak-related activity shows robustness or consistency in assessing between-subject variance(Brandmaier et al., 2018).

Indeed, the approach of identifying regions-of-interests (ROIs) by the strongest group effects and subsequently testing them for individual reliability was a common practice in evaluating individual differences in ERP studies, but recent studies have raised concerns that such a conventional approach may reduce the probability of detecting significant individual-level effects, especially in functional magnetic resonance imaging (fMRI) (Fröhner et al., 2019; Infantolino et al., 2018). For researchers interested in individual differences, between-subject variance in brain function is usually considered as the signal of interest rather than noise (Seghier and Price, 2018). For researchers interested in experimental effects, within-subject variance is treated as the signal of interest, and between-subject variance represents the noise that should be minimized. Those different views imply that regions eliciting greater activation (i.e., a peak at an electrode showing the strongest group-averaged activity) on group effects may not correspond to reliable individual effects, which has been thoroughly discussed in psychology recently (Hedge et al., 2018; Goodhew and Edwards, 2019; Fisher et al., 2018). To the best of our knowledge, the rationality of selecting individual difference variables based on group effects in ERP analysis has been seldom challenged. Whether and in which situation the group effects and individual reliability are consistent is still questionable. In real data, the underlying factors among different subjects are unmeasurable and cannot be adjusted at will, which makes it challenging to answer this question. Thus, a simulation model should be applied to investigate underlying factors of modulating the consistency between the group effect and individual reliability, but this investigation via computational modeling is still absent.

To address the abovementioned problems, the present study sought to examine the test-retest reliability of sensory-evoked potentials and cognitive ERPs based on the whole waveforms but not those restricted to narrow time windows around the peaks. More specifically, to test whether there is a spatial and temporal dissociation between group effects and the individual reliability result, the reliability of auditory-evoked potential (AEP), somatosensory-evoked potential (SEP), visual-evoked potential (VEP), and P300 were systematically examined by spatiotemporal decomposition and evaluation in a pointwise way (i.e., at each spatial-temporal EEG sample). Further, a dynamical system model was applied for the simulation of ERP generation to investigate the underlying mechanism explaining the real data results, in which key model parameters were varied to test their influences on the consistency between group effects and individual reliability. Data and code are available online (https://osf.io/v59qu).

2. Materials and methods

2.1. Data collection and preprocessing

2.1.1. Subject information

A total of 106 healthy subjects participated in this study, and 95 subjects ($Mean_{age}$ = 21.3 years; SD_{age} = 2.2 years, 73 males) among them attended two sessions, which were scheduled on different days, separated more than 6 days and 20 days apart on average. After removing 13 subjects whose data were corrupted with heavy artifacts, 82 subjects were included in subsequent reliability analyses. Ethical approval of the study was obtained from the Medical Ethics Committee, Health Science Center, Shenzhen University (No. 2,019,053). All subjects were informed of the experimental procedure, and they signed informed consent before the experiment.

2.1.2. Experimental paradigm

As illustrated in Fig. 1, the experimental paradigm was the same for the two sessions on different days. The experimental paradigm contained three types of sensory-evoked experiments (visual, auditory, and somatosensory) and a cognitive visual oddball experiment. Multiple sensory stimuli were arranged in two runs for each session. Each run consisted of 90 trials, including visual, auditory, and somatosensory vibration stimuli. These stimuli were delivered in a random order with inter-stimulus-interval (ISI) randomly distributed in the range of 2–4 s. Each stimulus lasted 50 ms. Hence, for each subject, there were a total of 180 trials of sensory stimulation in each session and 60 trials for each of the visual, auditory, and somatosensory stimuli. The P300 experiment was arranged between the two runs of multiple sensory stimuli for each session. The visual oddball experiment was performed with the red squares as the target stimuli and the white squares as the nontarget stimuli on the screen. Each square lasted 80 ms, with an ISI of 200 ms. Hence, a total of 600 trials were delivered within 2 min in a run, in which the target stimuli appeared with the possibility of 5%. A subject was asked to count the number of red squares and report the result at the end of the run to keep his/her attention on the screen.

2.1.3. Platform setup

During the experiment, the subjects were seated in a comfortable chair. For multisensory stimuli, an Arduino Uno platform was programmed to release the three types of stimuli, which communicated with the Matlab program (The MathWorks Inc., Natick, USA) on a PC through a serial port. Visual stimuli were delivered by a 3 W light-emitting diode (LED) with a 2 cm diameter circular light shield, which is placed 45 cm away from subjects' eyes. The LED intensity was 1074 Lux as measured by a light meter (TES-1332A, TES). Auditory stimuli were presented via a Nokia WH-102 headphone. The intensity is set at a comfortable level (75 dB SPL) for all subjects as measured by a digital sound level meter (Victor 824, Double King Industrial Holdings Co., Ltd. Shenzhen, China). Somatosensory stimuli were generated by a 1027 disk vibration motor with the rated power 3 W, efficiency 80%, and dimensions 10mm*2.7 mm). For the visual oddball P300 experiment, a 24.5-inch screen with a 240-Hz refreshing rate (Alienware AW2518H, Miami, USA) was used to present the visual stimuli. The 300×300 pixels red and white squares were delivered in sequence in the center of the 1920×1080 pixels screen with the background in black.

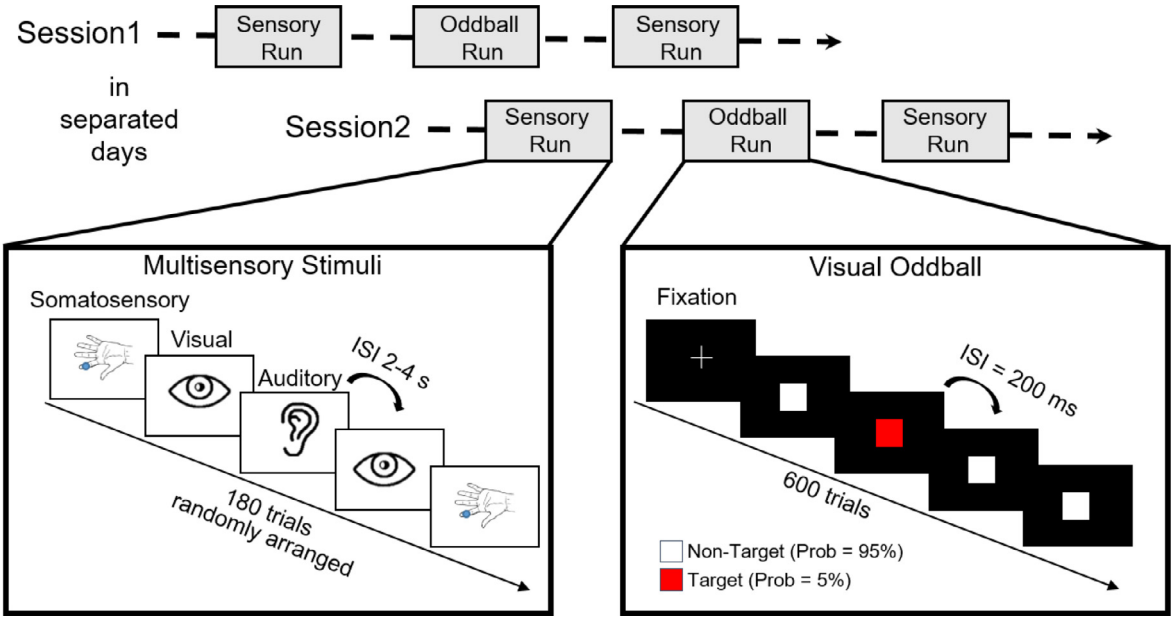

Fig. 1. Experiment procedure. Lower Left: Sensory-evoked potentials were elicited by a random sequence of somatosensory, auditory, and visual stimuli. Auditory stimuli were brief tones produced by a speaker; visual stimuli were brief flashes produced by an LED; somatosensory stimuli were applied to the index finger of the left hand by a vibrator. Lower Right: Cognitive ERPs were elicited by the classical visual oddball paradigm with the red squares as the target stimuli and white as the nontarget stimuli on the screen.

Table 1
Data preprocessing sheet.

Software	Matlab 2018b & Letswave7 (Letswave.cn)
Band-pass filtering	Butterworth filter, 0.01–200 Hz, 4th order, 24 dB/octave, zero-phase
Notch filtering	Butterworth filter, 49–51 Hz, 4th order, 24 dB/octave, zero-phase
Channel interpolation	Bad channels were identified manually and interpolated with the mean value of the three surrounding channels.
Re-reference	*Re*-reference to the mean value of TP9 and TP10
Artifacts removal by ICA	Eye movement related ICA components were identified by visual inspection of their scalp topographies, time courses, and spectra.
Band-pass filtering	Butterworth filter, 0.1–30 Hz, 4th order, 24 dB/octave, zero-phase
Segmentation	Segmentation from −0.5 to 1.0 s relative to the stimulus onset
Averaging	Average all trials for each subject
Baseline correction	Remove the DC offset (based on −0.5 to 0 s pre-stimulus)

For data collection, EEG signals were recorded via a multichannel EEG system (64 Channel, Easycap) and an EEG Amplifier (BrainAmp, Brain Products GmbH, Germany). The signals were recorded at a sampling rate of 1000 Hz by 64 electrodes, placed in the standard 10–20 positions. FCz was set to be the reference. Before data acquisition, the contact impedance between the EEG electrodes and the cortex was calibrated to be lower than 20 kΩ to ensure the quality of EEG signals during the experiments.

2.1.4. EEG preprocessing

For EEG preprocessing (Pernet et al., 2020), the sequence of steps, specific parameters for each step in preprocessing pipeline are shown in Table 1. After preprocessing, grand average ERP waveforms were computed for each participant and stimulus type (visual, auditory, somatosensory, and target stimuli of the visual oddball paradigm). All EEG pre-processing steps were carried out by Letswave7 (Huang, 2019) and Matlab.

2.2. Reliability analysis

2.2.1. Peak-based analysis and pointwise analysis

As the peak of each ERP component indicates the time point with a larger signal-to-noise ratio in the surrounding samples, peak amplitude is commonly used as a representative feature in ERP analysis. In this research, the most significant positive and negative peaks were detected by manually searching for the local maximum/minimum value in their corresponding time intervals for each subject. The mean amplitude around the peaks was not considered in this research because it is not fair to compare the reliability of pointwise analysis with the reliability of the mean amplitude, which is the average of multiple points.

Pointwise analysis was also used to examine the reliability of the ERP. More specifically, the ERP amplitude at each time point and each channel was taken as the variable for measuring the individual difference. Unlike the peak-based analysis, pointwise analysis is a fully data-driven method that is performed along with the temporal and spatial domain in a point-by-point way.

2.2.2. Metric of reliability: Intraclass correlation coefficient (ICC)

ICC is a commonly used metric for reliability analysis. In this study, the reliability was measured by using ICC(A, 1) of case 2A (McGraw and Wong, 1996) to represent the absolute agreement between repeated measurements for both the peak-based and pointwise analyses for both the peak-based and pointwise analyses. The subject-by-experiment matrix was modeled by a two-way ANOVA with random subject effects (row effects), random session effects (column effects), and residual effects, as shown in Eq. (1), and ICC(A, 1) is calculated as Eq. (2).

$$x_{ij} = \mu + r_i + c_j + e_{ij}, \tag{1}$$

$$ICC(\mathrm{A},1) = \frac{\sigma_r^2}{\sigma_r^2+\sigma_c^2+\sigma_e^2}, \tag{2}$$

in which, $i = 1, .., n\ (n = 82)$ is used as the subscript for subjects; $j = 1, \ldots, k\ (k = 2)$ is the subscript for multiple observations, which is the sessions in this work; x_{ij} denotes the observation in the j-th session from subject i; μ is the population mean for all observations; $r_i \sim N(0, \sigma_r)$ represents the row effects for subject i; $c_j \sim N(0, \sigma_c)$ represents the column effects for session j; $e_{ij} \sim N(0, \sigma_e)$ represents the residual effects as the error terms. According to ICC(A, 1), the reliability was defined in Eq. (2) as the proportion of the between-subject variation σ_r^2 over the total variation $\sigma_r^2 + \sigma_c^2 + \sigma_e^2$. In Eq. (1), we assume r_i, c_j, e_{ij} are independent random variables with normal distribution. The mean value and confidence interval of ICC(A, 1) for both peak-based and pointwise analyses were obtained 1600 times by bootstrap, which involved choosing random samples with replacement from a dataset and analyzing each sample in the same way. Two sample t-test was applied to compare the result between peak-based and pointwise analyses. Based on Eq. (1) and Eq. (2), the variance in an ERP measure comes from three parts, which are $Var(Trait) = \sigma_r^2$, $Var(State) = \sigma_c^2$, and $Var(Noise) = \sigma_e^2$ (Segalowitz and Barnes, 1993). Partitioning variance into these three parts, temporal and spatial variation of the $Var(Trait)$, $Var(State)$, and $Var(Noise)$ was analyzed in the spatial and temporal domains of the ERPs in a pointwise way.

2.2.3. Statistical analysis

By comparing group effects and individual reliability, correlation analysis was performed between the reliability and each group- and individual-level measure. Taking AEP as an example, a one-sample t-test was performed point wisely at the post-stimulus time points from the ERP signal of 82 subjects against zero at electrode Cz. The time points, which were significantly different from zeros ($p - value < 0.05/1000$ by Bonferroni correction; 1000 was the number of post-stimulus time points) were selected to reduce the influences of noisy background activity. At these selected time points, two group-level measures and one individual-level measure were extracted. Two group-level measures were (1) **abs(*t*-value)** calculated by the absolute value of the t-value and (2) **Hilbert envelope** calculated by the grand value of the ERP envelope via the Hilbert Transform. The individual-level measure was (3) **between-subject variance** estimated by the standard deviation across the 82 subjects. Then the linear trends were removed from the time series of each measure and reliability to avoid spurious correlation. The associations between different measures and reliabilities were quantified by Spearman's rank correlation coefficient, which is more robust to the non-linearity of changes and outliers than Pearson's correlation. For SEP, VEP, and P300, the same procedures were applied at electrodes Cz, Oz, and Pz, respectively, to explore the consistency between group effects and individual reliability because those electrodes showed the strongest group-level response.

2.3. Model simulation

For the given real EEG data, we can calculate its reliability, but the underlying factors affecting reliability are fixed and unknown, which limits the further study of reliability. Hence, a simulation model is needed to investigate the underlying mechanism behind the observations from real data. As a supplement to the real EEG data analysis, a dynamic model simulation allows us to further understand the internal mechanism of the brain. As rhythmic oscillations are the basic characteristics of an EEG signal and evoked changes of an EEG signal could be ascribed to transients that arise as the system's trajectory returns to its attractor (David et al., 2005), in this work, a second-order linear model was applied in this study for its simplicity to explain what we observed in real ERP data (i.e., the inconsistency between group effects and individual reliability).

$$x'(t) = Ax(t) + C * u(t) + e(t), \tag{3}$$

in which $A = \begin{bmatrix} c & d \\ -d & c \end{bmatrix}$ is the state-transition matrix with the corresponding eigenvalues $c \pm di$, $c < 0$ to ensure that the real part of the eigenvalue of A is negative which decides the decaying rate, d is the image part of the conjugate complex eigenvalues, which controls the natural oscillation frequency of this autonomous system. In this work, the elements of A were selected empirically as $c = -10$ and $d = 50$ to mimic the response of AEP at Cz. The influence of each parameter in A on model behavior was illustrated in the supplementary material (Fig. S6). The input strength, C, is formulated as $C_{sub} + C_{trial}$, where C_{sub} is a random variable representing the input strength for a given subject conformed to a Gaussian distribution (μ_{sub}, σ^2_{sub}), and C_{trial} is a random variable representing the input strength for a given trial conformed to a Gaussian distribution (μ_{trial}, σ^2_{trial}). According to Jansen and Rit's neural mass model (Jansen and Rit, 1995), the input of the system was simulated by using Eq. (4):

$$u(t) = \begin{cases} ate^{-bt} & t \geq jitter_{sub} \\ 0 & t < jitter_{sub} \end{cases} \tag{4}$$

in which $jitter_{sub}$ is a rounded random variable with a uniform distribution $[-\tau_{sub}, \tau_{sub}]$ relative to the onset time $t = 0$, and $e(t)$ is the pink Gaussian noise representing the input of the background EEG activity in the simulation. The core setting of this model was the additive term $C_{sub} + C_{trial}$, which coupled the input strength with the subject-level and the trial-level, thus allowing both $Var(Trait)$ and $Var(Noise)$ to co-vary with the signal amplitude. Considering the neglectable proportion of $Var(State)$ in real data results, there is no difference in simulation between sessions. To ensure consistency with real data, we set subject number $n = 82$ and session number $k = 2$ in the simulation. For each session, there are 60 trials. To mimic the real data preprocessing procedure, baseline correction was also applied to simulated ERP. With a sampling rate of 1000 Hz, there were 1500 time points for each trial, from −0.5 to 1 s.

In this study, two major parameters of this model potentially influencing the test-retest reliability were investigated: (1) inter-subject variability, τ_{sub}, for the latency jitter, $jitter_{sub}$, and (2) inter-trial variability, σ_{trial}, for the input strength, C_{trial}. Considering the oscillation of the ERP response as the trajectory in the 2-dimensional phase portrait in Fig. 2C, the observed ERP response is the projection of this trajectory on the axis of x_1. Hence, the peaks, troughs, and zero crossings have no special meaning, but some specific phases when the trajectory rotates along with the origin. The value of σ_{trial} affect the magnitude of the ERP trajectory. Hence, σ_{trial} determine the disturbance normal to the trajectory of the ERP response. While the value of τ_{sub} affect the time of trajectory of the ERP response. Hence, τ_{sub} determine the disturbance tangent to the trajectory of the ERP response. These two factors, τ_{sub} and σ_{trial}, were selected to investigate the test-retest reliability in the simulation because they provide disturbances in two directions orthogonal to each other. The different influence in different phases of the ERP response was expected for these two factors. Further, σ_{trial} is a trial-level factor, τ_{sub} is a subject-level factor, and the change in $Var(Trait)$ and $Var(Noise)$ could be investigated in the simulation. The simulation code is available online (https://osf.io/v59qu).

3. Results

3.1. Reliability of real data

3.1.1. Reliability for multisensory and cognitive ERPs

The grand average waveform of AEP at channel Cz, SEP at channel Cz, VEP at channel Oz, and P300 at channel Pz are shown in Fig. 3, where red and blue curves shaded by the standard deviation denoted the signals of the two sessions. The representative ERP peaks, including N1 at 90 ms and P2 at 180 ms for AEP, N2 at 150 ms and P2 at 245 ms for SEP, N1 at 64 ms, P2 at 185 ms for VEP, and P3 at 345 ms of P300, were selected for peak-based analysis. The negative and positive peaks are indicated by light gray and dark gray lines respectively and the gray shaded interval indicated the interval between the two peaks. For the pointwise analysis, the thick black lines indicated the maximal relia-

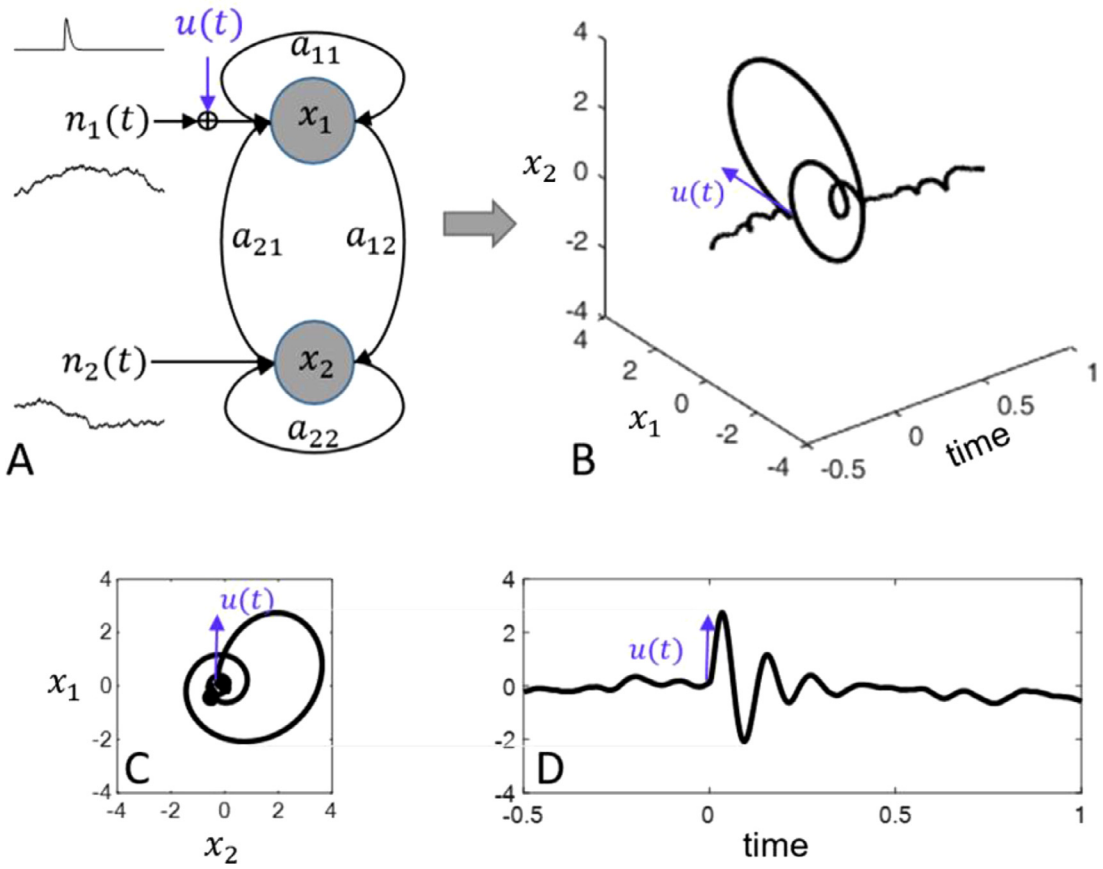

Fig. 2. The ERP simulation, which was generated by a second-order dynamic system model (3). (A) the framework of the dynamic model in Eq. (3); (B) the evolution of x_1 and x_2 over time, with its two-dimensional projection of the phase portrait in (C) and the evolution of x_2 over time in (D).

bility along with the ERP time courses. Since the subjects were more familiar with the experimental environment in the second session, the amplitudes of the ERP peaks were reduced. Importantly, the most reliable time point in the ERP did not correspond to the ERP peak. For P300 shown in Fig. 3D, the maximal reliability time point in the pointwise analysis appeared at 223 ms (thick black line), with a reliability of 0.61. This was much earlier than the well-known P3 component in the peak-based analysis, with the reliability of 0.44.

From each violin plot, the high reliability depended on (1) the consistency of individual ranks between two sessions and (2) the large between-subject variances. Taking the pointwise ICC results in AEP for example, the amplitude of ERP showed comparable inter-subject variance at time point 90 ms, 133 ms and 180 ms. But at 133 ms, the less interleaving between session 1 and 2 leads to a higher value of ICC (0.76) as compared with the other two time points. Taking the pointwise ICC results in VEP for another example, the inter-subject variances at 64 ms was relatively lower than time points at 183 ms and 185 ms, thus its reliability was much lower.

The comparison between the reliability results of the peak-based analysis and the pointwise analysis at corresponding time points were shown in Table 2, in which the peak amplitude was significantly ($p < 10^{-4}$) less reliable than the corresponding pointwise amplitudes at the latency of the grand average for all four types of ERPs.

Table 2
Comparisons between the reliability of peak amplitudes and corresponding pointwise amplitude for AEP, SEP, VEP, and P300.

ERPs	Measurements	Reliability		*p*-value
		Mean	Conf	
AEP(Cz)	N1	0.50	[0.31,0.68]	$< 10^{-12}$(***)
	90ms	0.54	[0.34,0.70]	
	P2	0.56	[0.39,0.68]	2.8×10^{-8}(**)
	180ms	0.58	[0.43,0.70]	
	133ms	**0.76**	**[0.68,0.85]**	
SEP(Cz)	N2	0.54	[0.38,0.68]	$< 10^{-12}$(**)
	150ms	0.57	[0.43,0.71]	
	P2	0.64	[0.52,0.78]	2.2×10^{-8}(*)
	245ms	0.66	[0.54,0.79]	
	110ms	**0.70**	**[0.60,0.83]**	
VEP(Oz)	N1	0.41	[0.22,0.58]	$< 10^{-12}$(***)
	64ms	0.48	[0.34,0.62]	
	P2	0.54	[0.33,0.69]	$< 10^{-12}$(***)
	185ms	0.73	[0.61,0.84]	
	183ms	**0.74**	**[0.62,0.84]**	
P300(Pz)	P3	0.44	[0.24,0.63]	$< 10^{-12}$(***)
	345ms	0.50	[0.31,0.66]	
	223ms	**0.61**	**[0.48,0.73]**	

3.1.2. *Spatiotemporal evaluation of reliability: a case study of AEP*

Next, AEP was used to further investigate the consistency between group effects and reliability with different exploratory analyses (results for other types of ERPs are provided in the supplementary material (Fig. S(1–4))). As illustrated in Fig. 4A, the *t*-value of significant regions ($p-$value $< 0.05/1000/64$ with Bonferroni correction, where 1000 is the number of post-stimulus time points, and 64 is the number of channels) were presented in the shaded region, which was consistent with the amplitude of the grand average waveform. Also, the post-stimulus AEP response behaved as a process of attenuating oscillations and finally approached the baseline. In contrast, the reliability of AEP after the stimulation shown in Fig. 4B increased greatly at the beginning of stimulation, lasting for a certain period, and then slowly returned to 0.

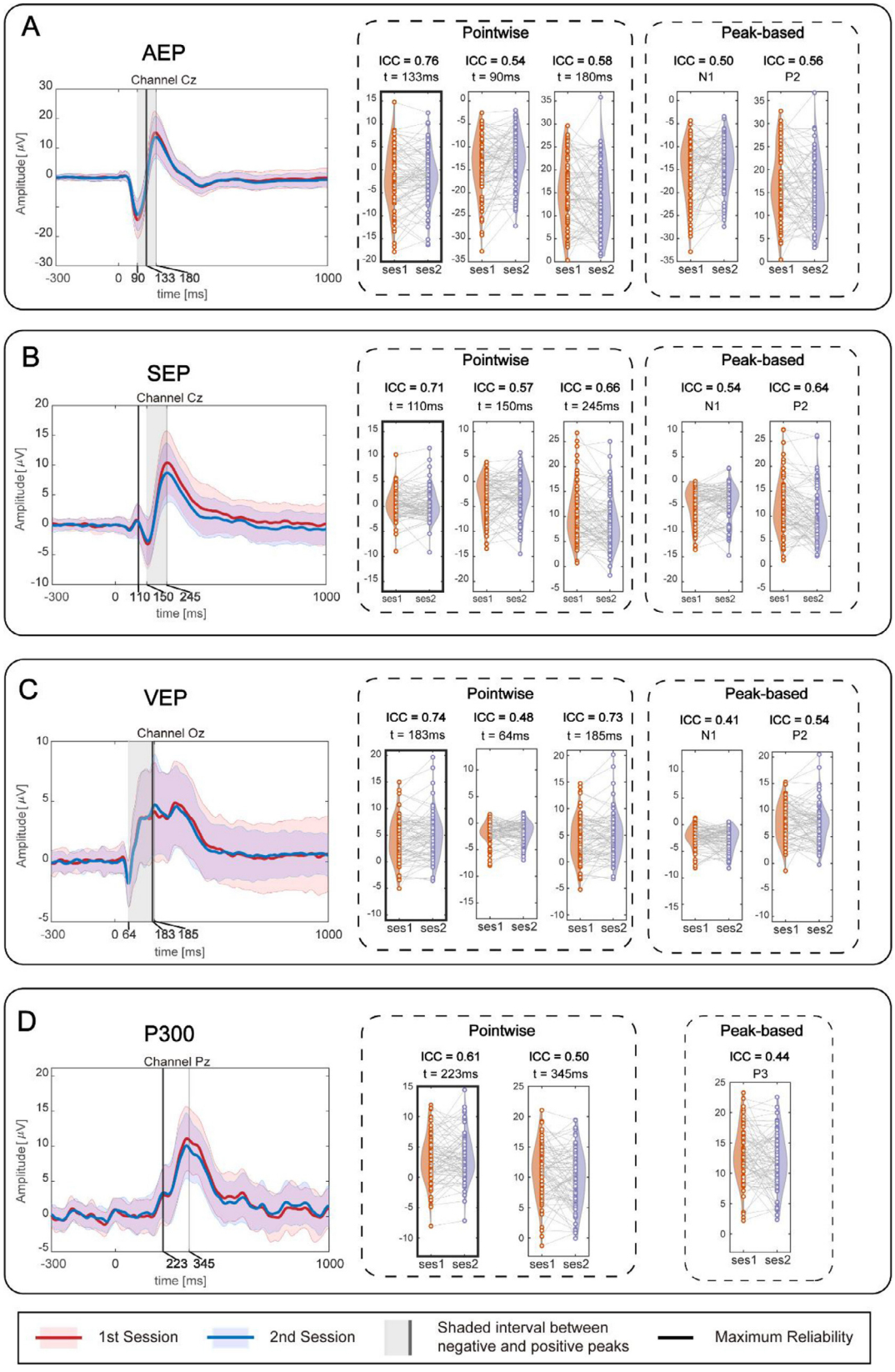

Fig. 3. Grand average waveform and the test-retest reliability result for AEP, SEP, VEP, and P300.The left part is the grand average waveform for the four types of ERP, shaded by the standard deviation in both 1st session (in red) and 2nd session (in blue). The gray shaded interval indicates the interval between the negative and positive peaks. The thick black line indicates the time point for the maximum ICC value. On the right part, the violin plot shows the amplitude distribution and change of amplitude for each subject between two sessions for both pointwise comparison and peak-based comparison.

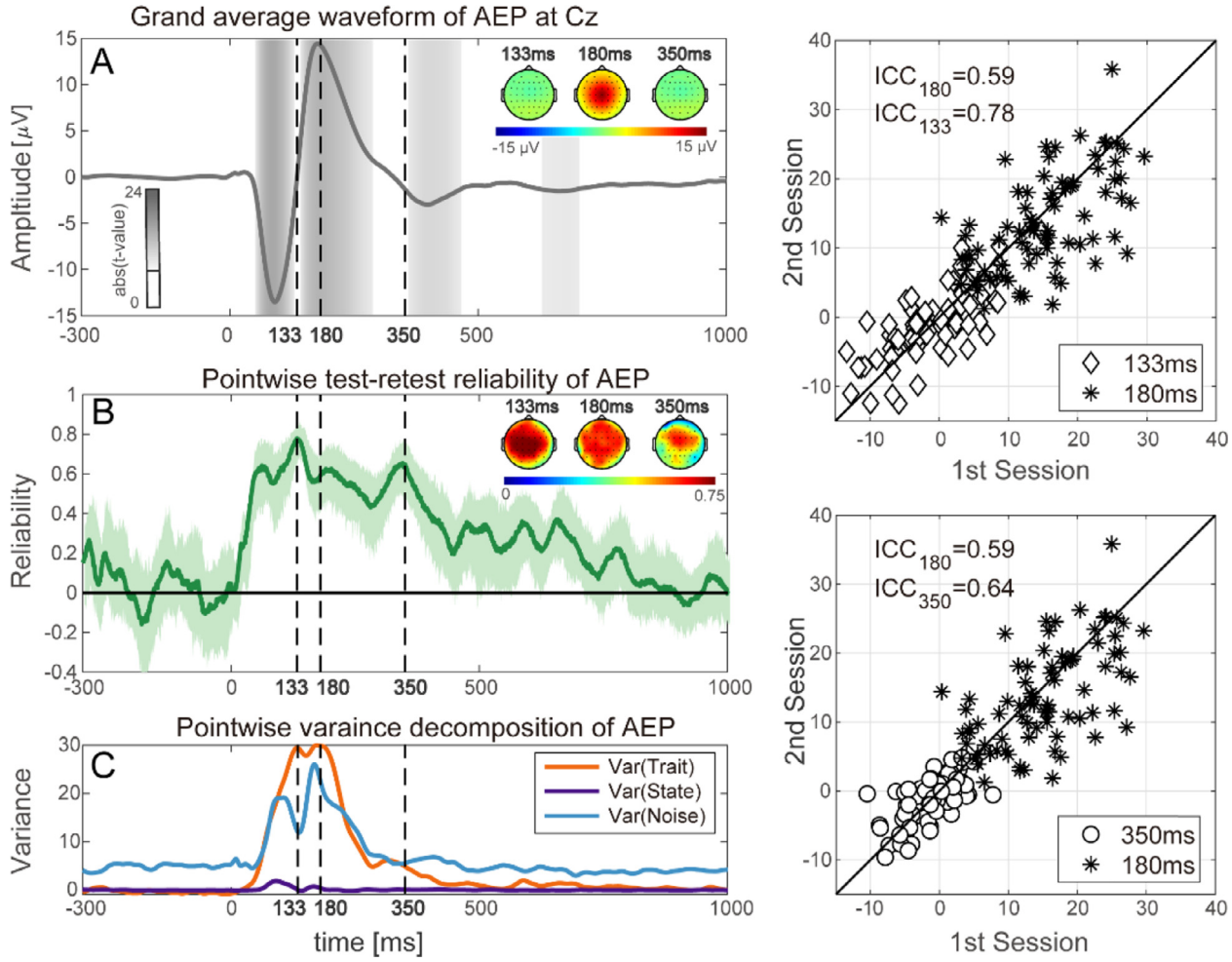

Fig. 4. The test-retest reliability analysis for AEP at electrode Cz. (A) The grand average waveform of AEP. (B) Pointwise test-retest reliability analysis considering the entire shape of the AEP time-course calculated by ICC(A,1). (C) The variance in the observation matrix with the size of subject × experiment was decomposed into three parts: $Var(Trait)$, $Var(State)$, and $Var(Noise)$ by a two-way random effects model along with the AEP time-course. Scatter plots of 82 subjects' amplitudes of AEP at Cz in the two experiments were compared between 180 and 133 ms in (D) and between 180 and 350 ms in (E).

Hence, the reliability of AEP along the ICC temporal profiles was not correlated with the amplitude of AEP. The maximal reliability of 0.78 appeared at the time point 133 ms, with the mean amplitude of AEP close to 0, which did not correspond to N1 at 90 ms or P2 at 180 ms with the minimal/maximal amplitude. The topographies of the grand average amplitude and the reliability at 90, 133, and 180 ms are illustrated in Figs. 4A and 4B.

Pointwise variance decomposition results based on a two-way ANOVA are shown in Fig. 4C. The magnitude of $Var(Trait)$ was close to 0 before the stimulation. At the beginning of stimulation, the magnitude of $Var(Trait)$ reached a peak at 180 ms and then returned to 0. The local maximum of $Var(Trait)$ did not correspond to the peak of AEP. The magnitude of $Var(Noise)$ showed similar trends as $Var(Trait)$, but the baseline was not 0. During the first 400 ms after stimulation, there was a certain correspondence between the waveform of AEP and $Var(Noise)$. The peak of AEP corresponded to the local maximum of $Var(Noise)$, while the zero-crossing point of AEP corresponded to the local minimum of $Var(Noise)$. Compared with $Var(Noise)$ and $Var(Trait)$, the magnitude of $Var(State)$ was too small and had little impact on reliability. Hence, the reliability was mainly determined by the ratio of $Var(Trait)$ to $Var(Noise)$. Next, the time points of 133, 180, and 350 ms were selected for the comparison, in which 180 ms corresponded to the peak of the grand average of AEP, while 133 and 350 ms corresponded to the local maximum of the reliability. It should be noted that, due to the insufficient session number ($n = 2$), the estimation of $Var(State)$ would possibly be negative, which was deeply elaborated in the supplementary material (Fig. S5).

Fig. 4D shows the comparison of the scatter plots of 82 subjects' amplitude of AEP between time points 133 ms (diamonds) and 180 ms (asterisks). As $Var(State)$ was close to 0, $Var(Trait)$ could be measured as the variance along the black diagonal line, and $Var(Noise)$ could be measured as the variance perpendicular to the black diagonal line. As shown in Fig. 4D, the mean amplitude of AEP at 133 ms (mean value of the diamonds) was much smaller than that at 180 ms (mean value of the asterisks), but the $Var(Trait)$ values at the two different time points were similar. Hence, the reliability at 133 ms was larger than that at 180 ms because of the smaller $Var(Noise)$ at 133 ms. Fig. 4E shows a different situation compared with that shown in Fig. 4D. The reliability at 180 ms (asterisks) and 350 ms (circles) were similar, but both $Var(Trait)$ and $Var(Noise)$ at 350 ms were smaller than that at 180 ms.

3.1.3. Statistical results

Spatiotemporal dissociation between group effects and individual reliability was revealed in Fig. 3 and Fig. 4. These findings went against our expectations, given the fact that extracting peak-based measures using group-level prior information was the most common approach in reliability analysis. Hence, Spearman's rank correlation analysis was further performed on AEP, SEP, VEP, and P300 to analyze the statistical

Table 3
Associations between group-level measures (abs(*t*-value), Hilbert envelope), individual-level measure (between-subject variance), and reliability.

ERPs	abs(*t*-value)		Hilbert envelope		between-subject variance	
	Spearman's ρ	*p*-value	Spearman's ρ	*p*-value	Spearman's ρ	*p*-value
AEP (Cz)	−0.19	3.04×10^{-5}	0.27	$< 10^{-12}$	0.36	$< 10^{-12}$
SEP (Cz)	0.38	$< 10^{-12}$	0.51	$< 10^{-12}$	0.71	$< 10^{-12}$
VEP (Oz)	0.17	0.002	0.12	0.017	0.75	$< 10^{-12}$
P300 (Pz)	0.54	$< 10^{-12}$	0.74	$< 10^{-12}$	0.84	$< 10^{-12}$

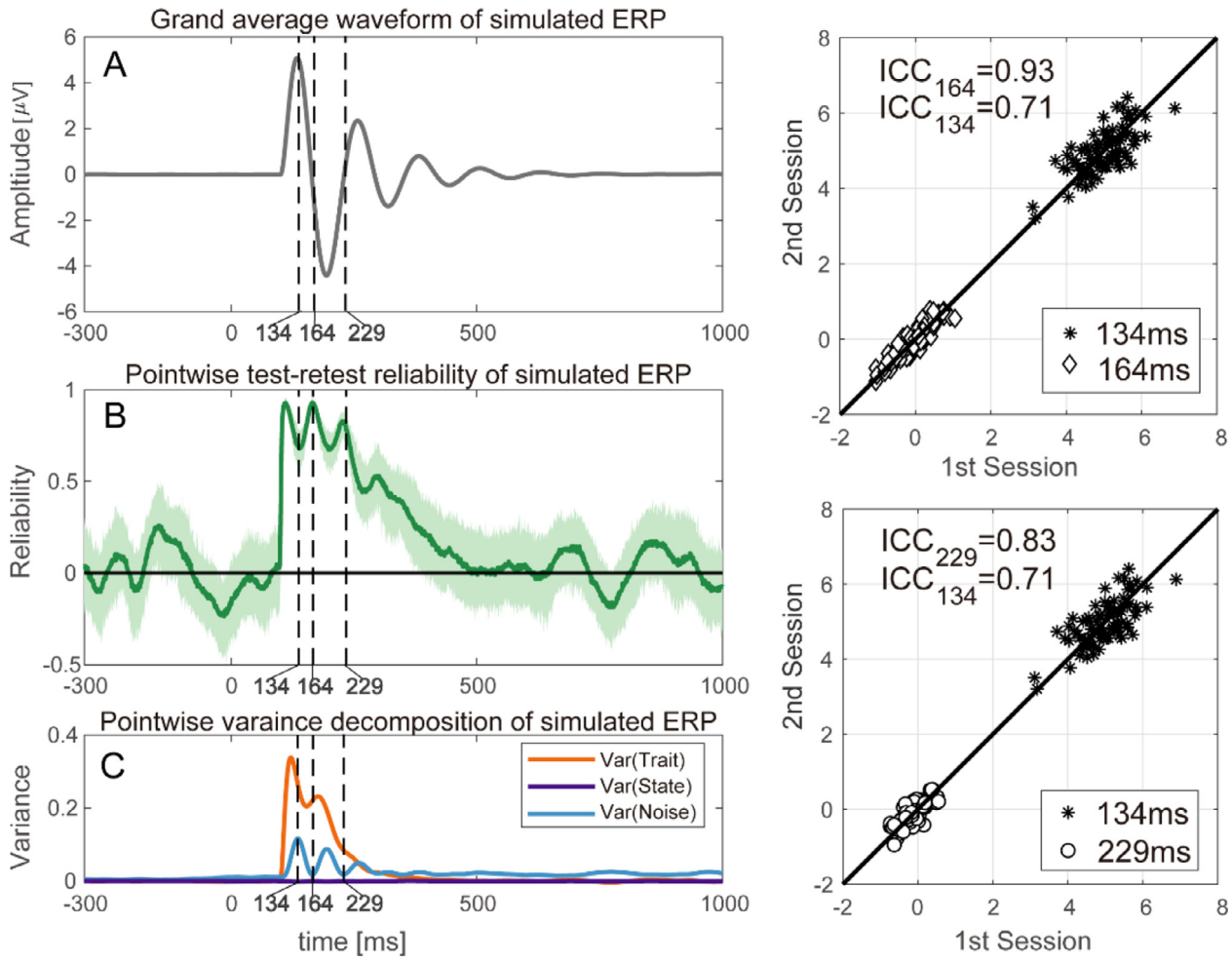

Fig. 5. The test-retest reliability analysis for simulated ERP. (A) The grand average waveform of simulated ERP for a given set of system parameters. (B) Pointwise test-retest reliability analysis along with the simulated ERP time course. (C) The variance of observation matrix with the size of the subject by experiment was decomposed into three parts: $Var(Trait)$, $Var(State)$, and $Var(Noise)$ by a two-way random effects model along with the simulated ERP time-course. Scatter plots of 82 subjects' amplitudes of AEP at electrode Cz in two experiments were compared between 134 and 164 ms in (D) and between 134 and 229 ms in (E).

relationships of reliability with the group-level measures (abs(*t*-value) and Hilbert envelope) and individual-level measure (between-subject variance). For group-level measures, it was observed that the Hilbert envelope showed a larger correlation coefficient with reliability than the abs(*t*-value) except for VEP. This result suggested an oscillation perspective on ERP reliability should be considered. For individual-level measure, Spearman's ρ between between-subject variance and reliability was greatly improved (Table 3).

3.2. Reliability of simulated data

3.2.1. Simulation results

To further understand the internal factor influencing the reliability in ERP analysis, a dynamic model in Eq. (3) was used for the simulation. The simulation results in Fig. 5 were consistent with the results from real ERP data in Fig. 4. Specifically, the grand average waveform of the simulated ERP is shown in Fig. 5A, with a peak of 134 ms and subsequent zero crossings appearing at 164 and 229 ms. The reliability curve across time is shown in Fig. 5B, and the corresponding variance decomposition is shown in Fig. 5C. In the simulation, the correlation between $Var(Noise)$ and the amplitude of the ERP was more obvious. $Var(State)$ was close to 0 because the systematic differences between the two sessions were not considered in this simulation. Hence, the reliability at the peak latency was the local minimum, and the reliability at the zero-crossing point was the local maximum. Similarly, the scatter plots in Figs. 5D and 5E show that the larger amplitude of the ERP may not necessarily lead to greater reliability, which was determined by the ratio of $Var(Trait)$ to $Var(Noise)$.

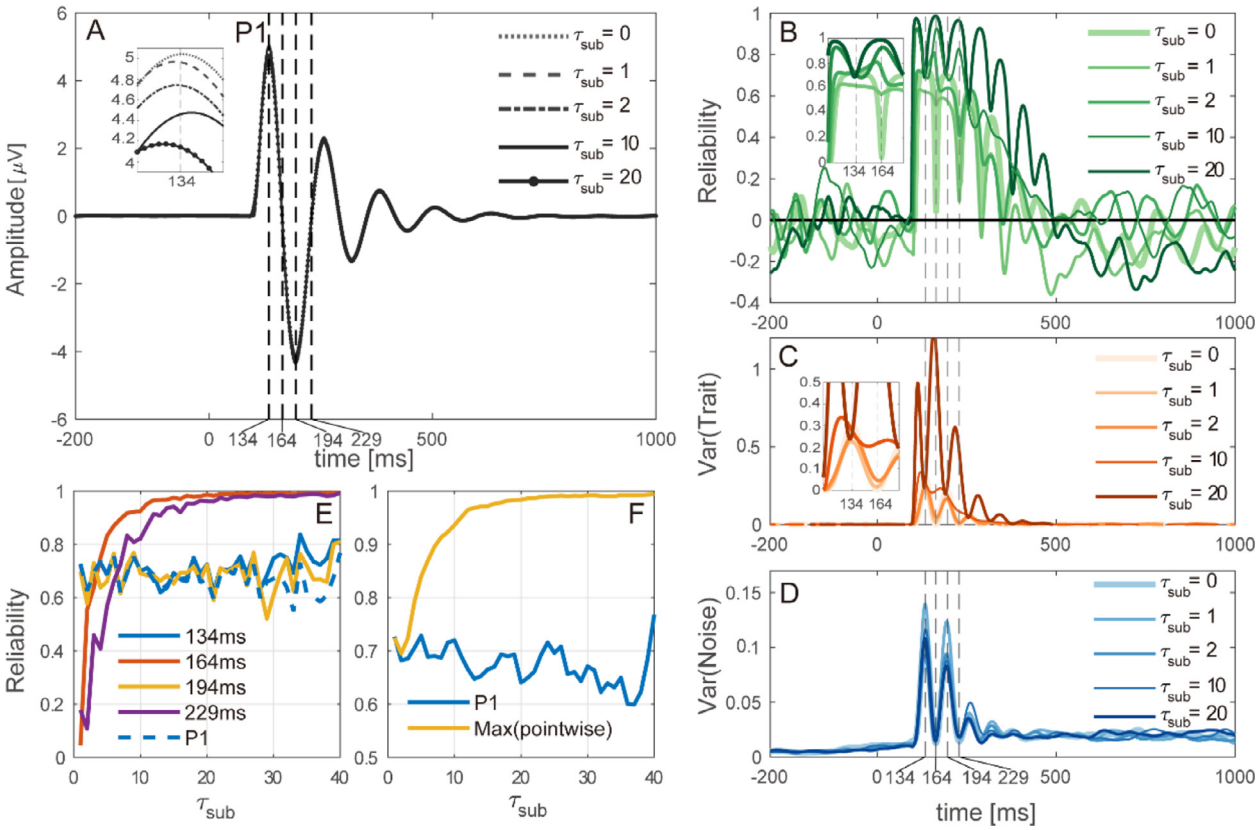

Fig. 6. The influence of increasing the variability of inter-subject latency jitter of the dynamic system at the subject-level on (A) Grand average waveform of simulated ERP. (B) Pointwise test-retest reliability along the time-course of simulated ERP. (C) $Var(Trait)$ along the time-course of simulated ERP. (D) $Var(Noise)$ along the time-course of simulated ERP. (E) Comparisons between peak-based reliability and pointwise reliability at group-level peak latencies. (F) Comparisons between the maximum value of pointwise reliability and peak-based reliability.

3.2.2. *The influence of the variability of jitter:* τ_{sub}

As a tangential disturbance in the phase portrait of Eq. (3) shown in Fig. 2, an increase in τ_{sub} did not make a large difference in the waveform of the grand average ERP in the simulation, but it made the peak of P1 and N2 smoother. As τ_{sub} increased from 0 to 20, the amplitude of the peak P1 reduced slightly, as shown in Fig. 6A. As illustrated in Fig. 6B, the reliability of the peaks at 134 and 194 ms remained around 0.7, while the reliability of the zero-crossing points at 164 and 229 ms increased greatly. As inter-subject latency jitter increased, the ICC values at 134 and 194 ms, which corresponded to the peaks of the grand average waveform, gradually shifted from the peaks of the ICC temporal profiles to their local minimum. The ICC values at 164 and 229 ms, which corresponded to the zero-crossing point, behaved conversely. The corresponding variance decomposition is shown with different values of τ_{sub} in Fig. 6(D–E). $Var(Trait)$ at the zero-crossing point (164 and 229 ms) of the grand average waveform increased as inter-subject latency jitter increased, while $Var(Noise)$ fluctuated randomly. In comparison with the reliability of the peak amplitude, there was a greater difference between the maximum values of the ICC temporal profiles and the reliability of the peak amplitude.

3.2.3. *The influence of the variability of input power:* σ_{trial}

For normal perturbation in the phase portrait of Eq. (3) in Fig. 2, it is shown in Fig. 6B that the overall magnitude of the ICC temporal profiles dropped because of increasing inter-trial variability in the dynamic systems' input, while the reliability in the response amplitude at 134 and 194 ms, which corresponded to the peak latencies of the grand average waveform, dropped more quickly compared with the reliability of the response amplitude at 164 and 229 ms, exhibiting an unbalanced influence. The above observations were further investigated by pointwise variance decomposition, in which the within-subject variation ($Var(Noise)$) increased systematically in proportion to the signal amplitude of the grand average waveform with increasing inter-trial variability, while the between-subject variation fluctuated randomly, thus explaining why the reliability of the signal with larger amplitude decreased more. Interestingly, it can be noted from Fig. 6F that there was a larger difference between the maximum values of the ICC temporal profiles and the reliability of the peak amplitude as the inter-trial variability increased.

4. Discussion

The purpose of this study was to investigate the relationships between group effects and individual reliability across different types of ERPs. By performing pointwise reliability analysis and rigorous simulation, we found inconsistency between individual reliability and group effects and provided potential explanations from the perspective of oscillations of ERP. The findings have implications for a series of questions that are of theoretical and practical relevance for ERP researchers, which will be discussed sequentially.

4.1. *Peak-based analysis versus pointwise analysis*

By briefly reviewing the ERP reliability research in the past decade in Table 4, we think it is necessary to re-emphasize that we should not restrict analysis in narrow time windows around peaks, especially for research about the individual difference. Until now, peak-related feature extraction (i.e., peak amplitude, area under the curve, mean amplitude) has been a dominant approach for examining the reliability of ERPs (Huffmeijer et al., 2014; Munsters et al., 2019; Devos et al.,

Table 4
Literature review on ERP reliability from 2011 to 2021.

ERP reliability research	
Gaspar et al., 2011	Grand average ERPs can be misleading because it does **NOT** reflect individual dynamics. Not only around the peak, but also entire temporal windows are reliable.
Cassidy et al., 2012	Reliability of peak and latency for a selected range of ERP components were evaluated.
Leue et al., 2013	Intra-individual N2 variability incorporated systematic variance.
Huffmeijer et al., 2014	ERP amplitudes generally showed adequate to excellent test-retest reliability. Averaging across several electrodes or trials improved the reliability of P3 amplitude.
Ip et al., 2018	ERP measures exhibited more variation and are less stable compared to continuous EEG.
Munsters et al., 2019	The face-sensitive ERP components (i.e. N290, P400, and Nc) in infants show adequate test-retest reliability
Our research	
• The peak-based analysis may not be sufficiently reliable to capture the individual difference.	
• A perspective of neural oscillations is more peak-based analysis to explain the inconsistency between group effects and individual reliability.	
• A simulation model is applied to investigate underlying factors of modulating the consistency between the group effect and individual reliability.	

2020). For the peak-based approach, researchers have found that the reliability of ERPs is influenced by the number of trials, channel selection, and various preprocessing strategies (Huffmeijer et al., 2014; Leue et al., 2013). The basic hypothesis behind the peak-based analysis is that the peak of the ERP indicates a higher signal-to-noise ratio, which produces results with a higher confidence level because of the relatively small interference from background EEG noise, yet this concept of signal-to-noise ratio may not generalize to the research area interested in individual difference, in which between-subject variance is treated as the signal, within-subject variance is treated as noise, as mentioned by (Brandmaier et al., 2018). Another limitation of the peak-based analysis is that the latency and amplitude of ERP peaks, as well as the entire ERP shapes, are physiologically meaningful and important (Gaspar et al., 2011).

Further, we also observed the phenomenon that ERP shapes in different sessions for the same subject were quite similar, while ERP shapes for different subjects were largely different, which may reflect that the information processing underlying ERP is unique to each subject. We believe future ERP research should not restrict its analysis in narrow time windows around peaks, but develop analysis techniques to characterize the ERP shape, which is useful in translational neuroscience.

4.2. Inconsistency between group effects and individual reliability

Stronger group effects do not guarantee higher individual reliability. In reliability analysis, the group effects are commonly used as prior information (Plichta et al., 2012; Aron et al., 2006; Fliessbach et al., 2010), which assumes that experimental manipulation eliciting greater activation at the group level should also show reliable between-subject variation. This conventional approach has been questioned in recent years, especially in the fMRI community (Fröhner et al., 2019; *Infantolino et al., 2018;* Yarkoni and Braver, 2010; Li et al., 2019). In line with these studies, our results also revealed inconsistency between group effects and individual reliability in ERP analysis. More specifically, concerning the temporal domain discrepancies illustrated in Fig. 3, the most reliable points in the four types of ERPs (Cz for AEP, SEP, Oz for VEP, and Pz for P300) did not all correspond to maximum or minimum points of group-level activations. For AEP, the most reliable point appeared at the zero-crossing point of ERP. The spatial domain discrepancies are illustrated in Fig. 4 for AEP, in which we did not find the topography of the AEP response corresponding to the topography of the reliability at 133 and 350 ms. Further analysis, presented in Table 1, indicated that, as an individual-level measure, the between-subject variance showed a higher correlation coefficient than the group-level measure (abs(*t*-value)) across all four types of ERPs. All these evidences suggest that it is not advisable to select peak-related features at the electrode showing the strongest group effects without carefully examining their reliability.

Intuitively, the spatial-temporal distribution of group-level analyses and individual-level analyses should tend to converge. In other words, increased activation by experimental manipulation at the group level should relate to individual-level analysis, given a large enough sample size and no confounding factors. However, few empirical pieces of evidence support this idea (Lee et al., 2006); more often, individual difference analyses simply fail to reveal any significant effects in regions that show a robust within-subject effect (Vetter et al., 2017; Raemaekers et al., 2007). In this research, the simulation results indicate that the consistency between group-level effects and individual reliability may be dynamically modulated by inter-subject latency jitter and inter-trial variability of dynamic system input, providing a dynamic view of the relationships between the two types of analysis in ERP analysis.

Spatiotemporal evaluation and decomposition of reliability are good for identifying the reproducible individual difference. In this research, it was found that the high signal-to-noise ratio assumption for the peak of ERP did not hold when considering individual difference research, which was also mentioned by Brandmaier et al. (2018). As illustrated in Fig. 4, the variance of the noise (blue curve in Fig. 4C) was highly correlated with the magnitude of the AEP response (absolute value of the black curve in Fig. 4A). Considering that the essence of EEG is neural oscillation, the peak in the ERP is just a certain phase (0 or π) during the oscillation. There is nothing more special about it compared to other phases. Hence, there is no reason why reliability analysis should be limited to peak-based features; pointwise analysis can bring us more comprehensive results. As compared with *t*-test or ANOVA, the pointwise analysis of test-retest reliability did not have the family-wise error rate problem, as we calculated the ICC values but did not judge whether there was a significant difference. Compared with peak-based analysis, the results from the pointwise analysis always had significantly higher ICC values at the time point of the peaks for all four types of ERP analysis in our investigation. In the test-retest reliability of AEP, SEP, VEP, and P300, the pointwise analysis consistently showed that the ICC value increased greatly after the stimulation, and after maintaining it for a while, decreased slowly to the baseline (Fig. 4). Hence, the peak of ERP may not relate to a higher ICC value. Even in AEP, the two local maximum points of the ICC value corresponded to the two zero-crossing points of the AEP. The findings suggest that reliability analysis restricted by the narrow time windows around the peaks is questionable. By performing pointwise analysis, dynamic changes in reliability in the spatial-temporal domain can be traced, given enough sample size, thus providing a new angle of ERP reliability analysis in a data-driven manner. Also, agreeing with the opinions of Gaspar et al. (2011), we believe that shape-based metrics rather than peak-based metrics may be more reliable for individual difference research.

To translate group effects into individual difference research, some issues must be reconsidered in ERP data analysis. ERP analysis focusing on individual difference often implicitly or explicitly uses prior information from group effects. For reliability analysis, electrodes showing the strongest stimulus-related activity by group-level analysis are often chosen for test-retest reliability analysis of ERPs (Gaspar et al., 2011). For constructing single-subject predictive models, it has been done in trans-

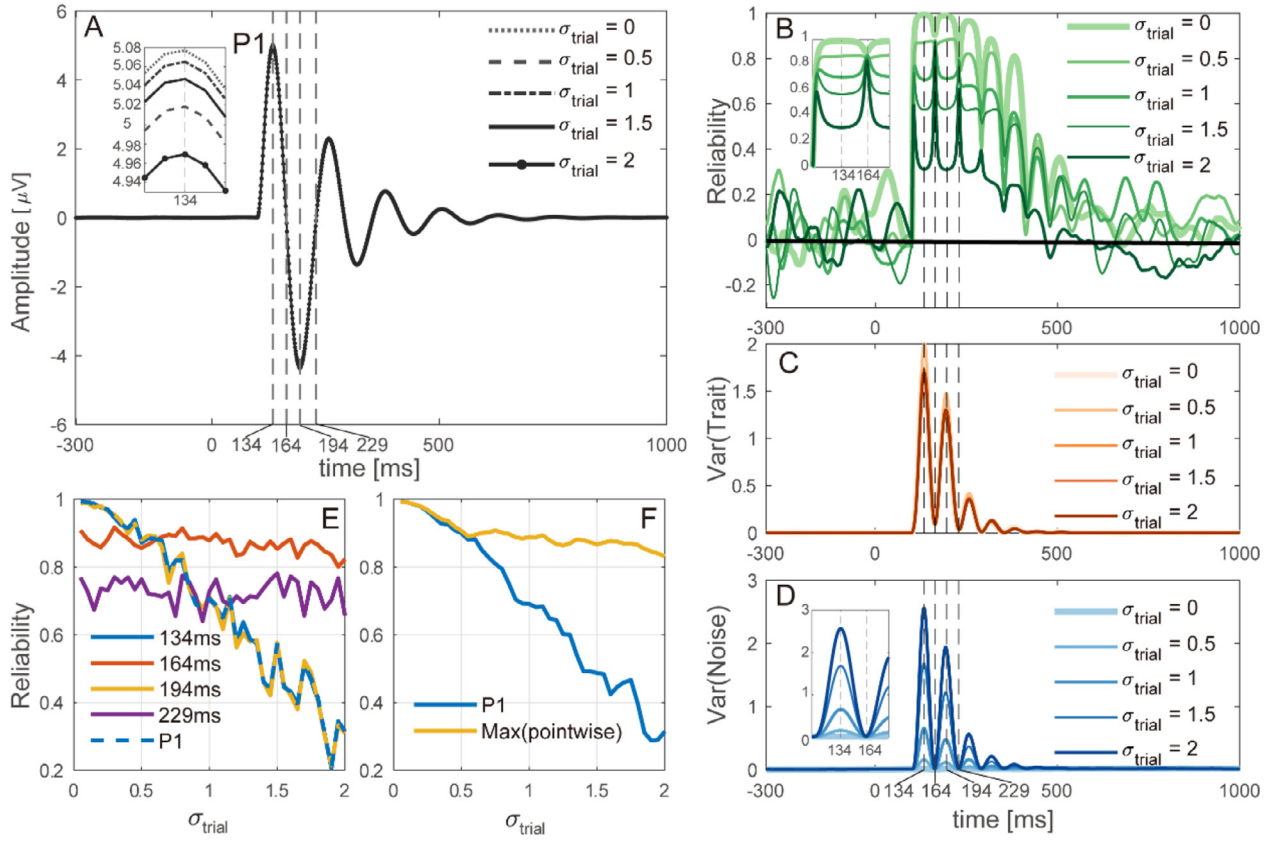

Fig. 7. The influence of increasing the variability of input power of the dynamic system at the trial-level on (A) Grand average waveform of simulated ERP. (B) Pointwise test-retest reliability along the time-course of simulated ERP. (C) $Var(Trait)$ along the time-course of simulated ERP. (D) $Var(Noise)$ along the time-course of simulated ERP. (E) Comparisons between peak-based reliability and pointwise reliability at group-level peak latencies. (F) Comparisons between the maximum value of pointwise reliability and peak-based reliability.

lating findings in group-level statistics of ERPs into a machine learning framework (Boshra et al., 2019). Hence, here we discuss some critical issues about ERP data analysis concerning the use of group effects in individual difference research.

- Tracing back to history, the original idea of ERPs was to index specific cognitive processes rather than distinguishing different individuals (i.e., the research interest was how brain activity responds to one condition versus the other). Individual differences were treated as measurement error that could not be explained by experimental manipulation, as *t*-test, omnibus ANOVA assumes. From this perspective, there is no reason to select regions based on strongest group effects and then feed them into the correlation or reliability analysis, except that this region also shows greater between-subject variation and smaller within-subject variation.
- For the data analysis pipeline of ERPs, it is very common to perform the subtraction operation (e.g. ERP difference waves) to minimize the impact of baseline individual differences. Such operations forcibly promote the activity of the baseline period at a constant rate across individuals, and the goal is to obtain a reliable experimental effect at the group level. This pervasive practice was inherited from research focusing on experimental effects, but few studies have noticed whether this approach is reasonable for individual difference analysis. Recently, the fMRI and psychology communities have argued that difference scores often exhibit a robust group-level effect but lower reliability (Infantolino et al., 2018; Onie and Most, 2017).

Considering statistical analysis, typically, ERP data are averaged within conditions and participants after preprocessing and then analyzed for the mean difference between conditions using paired *t*-test or repeated-measures ANOVA. This traditional approach implicitly assumes that experimental manipulation yields uniform effects across all participants. The random variance of individual difference in effect sizes is not taken into account. By adopting linear mixed-effect models, in which random effects are used to capture individual variability as a form of random slopes or random intercepts, fixed effects are estimated by the grand mean across all participants. Such an approach has been adopted to simultaneously capture both group effects and individual difference (Frömer et al., 2018; Tibon and Levy, 2015).

4.3. A neural oscillation perspective on ERP reliability

Both peak and zero-crossing points of ERPs just represent different phases of one unified oscillation process. To further understand spatial-temporal inconsistency between group-level effects and individual reliability in ERP analysis, a dynamic model was applied for the simulation. The simulation model was simplified to be a second-order linear attractor with noise to simulate the EEG oscillation. From the perspective of dynamic system theory (Jansen and Rit, 1995; Youssofzadeh et al., 2015), peaks in the EPR are just an observation of EPR from one dimension of the computational models of neural processes. The phase portrait of our simulations (Fig. 2) provides a more comprehensive perspective, in which the peaks are just some special phases during the neural oscillation. Consider the oscillation of the ERP response as the trajectory in the 2-dimensional phase portrait as illustrated in Fig. 2C, the ERP response we observed is the projection of this trajectory on the axis of x_1. Hence, the peaks, troughs, and zero crossings have no special meaning, but some specific phases when the trajectory rotates along with the origin. The value of σ_{trial} will affect the magnitude of the ERP trajectory. Hence σ_{trial} determined the disturbance normal to the trajectory of the ERP response. While the value of τ_{sub} will affect the time of trajec-

tory of the ERP response. Hence τ_{sub} determined the disturbance tangent to the trajectory of the ERP response. Owing to the different directions and different levels of the two factors, the simulation result showed that the changes of these two factors played very different roles in different phases of ERP, which can be illustrated by Fig. 6 and Fig. 7, especially at the peak and zero-crossing points of the ERP, with the changes of these two factors. In summary, With the similar wave form of the group-level ERP, the reliability would be determined by several factors. Measuring the peak-based features would not provide a comprehensive understanding about the oscillations in ERP.

Considering that stronger group effects do not guarantee higher individual reliability and the oscillation nature of ERP, the Hilbert transform was performed on trial-averaged data for each subject. The results in Table 3 suggest the Hilbert envelope is more consistent with reliability compared with the abs(t-value), which reflects the oscillation nature of ERP. The consistency between the grand averaged envelope of Hilbert transformed data and the reliability of ERP waveform for four kinds of ERPs without Bonferroni correction can be found in Fig. S7 in the supplementary material. For AEP, SEP, VEP, and P300, compared with the grand average ERP waveform, the grand average envelope of Hilbert transformed data is more consistent with individual reliability and shows a larger correlation coefficient, especially for AEP and SEP. These results further solidify the oscillation nature of ERP.

4.4. Limitations

Our research on reliability analyses had several limitations. First, higher reliability does not ensure higher validity. The fact that the response amplitude at some time points was more reliable than the peak amplitude may be explained by sacrificing validity. More specifically, each subject's response amplitude at a given time point may index different neurophysiological processes, leading to larger between-subject variance. Increasing reliability in this way is not desirable because the underlying process of this measure is different across subjects. However, we cannot verify this potential explanation without behavioral data. Second, the insufficient section number ($k = 2$) would leads to an inaccurate estimation of $Var(State)$, which make the negative value of $Var(State)$ possible in the practical calculation. Third, our analysis was restricted to univariate features; the relationship between group-level effects and individual reliability concerning multivariate analysis warrants further investigation in the future.

5. Conclusion

In summary, the purpose of this research was to investigate the consistency between group effects and individual reliability of ERPs. We performed spatiotemporal evaluation and decomposition of reliability in four different ERPs, and the findings indicate that the peak-based approach (i.e., selecting regions showing the strongest group-level response as individual difference variables) may be inappropriate for reliability analysis of ERPs. Without carefully examining reliability, this approach based on group-level prior information may fail to reliably capture individual differences, which is supported by spatiotemporal dissociation between group effects and individual reliability. The disadvantages of peak-based reliability analysis were illustrated by spatiotemporal evaluation and decomposition of reliability, statistical results, and the phase portrait in the simulation model. Further, the simulation results highlight the modulation role of inter-subject latency jitter and inter-trial variability in modulating the consistency between group-level effects and individual reliability. To conclude, all these results provide a new perspective beyond peak-based analysis in the ERP reliability studies. Furthermore, the findings deepen our understanding of ERP generation and the reliability of ERPs.

Data and code availability statement

Data and code are available online (https://osf.io/v59qu).

Credit author statement

Zhenxing Hu: Conceptualization, Formal analysis, Methodology, Writing- Original draft preparation; **Zhiguo Zhang:** Conception, Study organization, Writing- Reviewing and Editing; **Zhen Liang:** Consultation, Data analysis; **Li Zhang:** Initial analysis; **Linling Li:** Data curation; **Gan Huang:** Supervision, Conceptualization, Methodology, Data collection, Writing- Reviewing and Editing

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 81871443, 61906122, and 81901831), the Science, Technology, and Innovation Commission of Shenzhen Municipality Technology Fund (No. JCYJ20190808173819182), the Shenzhen Science and Technology Program (No. JSGG20210713091811038), the Shenzhen-Hong Kong Institute of Brain Science-Shenzhen Fundamental Research Institutions (No. 2021SHIBS0003).

None of the authors has potential conflicts of interest to be disclosed.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.neuroimage.2022.118937.

References

Aron, A.R., Gluck, M.A., Poldrack, R.A., 2006. Long-term test-retest reliability of functional MRI in a classification learning task. Neuroimage 29, 1000–1006. doi:10.1016/j.neuroimage.2005.08.010.

Boshra, R., Dhindsa, K., Boursalie, O., Ruiter, K.I., Sonnadara, R., Samavi, R., Doyle, T.E., Reilly, J.P., Connolly, J.F., 2019. From group-level statistics to single-subject prediction: machine learning detection of concussion in retired athletes. IEEE Trans. Neural Syst. Rehabil. Eng. 27, 1492–1501. doi:10.1109/TNSRE.2019.2922553.

Brandmaier, A.M., Wenger, E., Bodammer, N.C., Kühn, S., Raz, N., Lindenberger, U., 2018. Assessing reliability in neuroimaging research through intra-class effect decomposition (ICED). Elife 7, e35718. doi:10.7554/eLife.35718.001.

Bridgeford, E.W., Wang, S., Yang, Z., Wang, Z., Xu, T., Craddock, C., Dey, J., Kiar, G., Gray-Roncal, W., Coulantoni, C., 2020. Eliminating accidental deviations to minimize generalization error with applications in connectomics and genomics. Biorxiv 802629. doi:10.1101/802629.

Cahn, B.R., Polich, J., 2006. Meditation states and traits: EEG, ERP, and neuroimaging studies. Psychol. Bull. 132, 180–211. doi:10.1037/0033-2909.132.2.180.

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. doi:10.1111/j.1469-8986.2011.01349.x.

Croce, P., Quercia, A., Costa, S., Zappasodi, F., 2020. EEG microstates associated with intra- and inter-subject alpha variability. Sci. Rep. 10, 1–11. doi:10.1038/s41598-020-58787-w.

Cruse, D., Beukema, S., Chennu, S., Malins, J.G., Owen, A.M., McRae, K., 2014. The reliability of the N400 in single subjects: implications for patients with disorders of consciousness. NeuroImage Clin 4, 788–799. doi:10.1016/j.nicl.2014.05.001.

David, O., Harrison, L., Friston, K.J., 2005. Modelling event-related responses in the brain. Neuroimage 25, 756–770. doi:10.1016/j.neuroimage.2004.12.030.

Devos, H., Burns, J.M., Liao, K., Ahmadnezhad, P., Mahnken, J.D., Brooks, W.M., Gustafson, K., 2020. Reliability of P3 event-related potential during working memory across the spectrum of cognitive aging. Front. Aging Neurosci. 12, 1–8. doi:10.3389/fnagi.2020.566391.

Dubois, J., Adolphs, R., 2016. Building a science of individual differences from fMRI. Trends Cogn. Sci. 20, 425–443. doi:10.1016/j.tics.2016.03.014.

Fisher, A.J., Medaglia, J.D., Jeronimus, B.F., 2018. Lack of group-to-individual generalizability is a threat to human subjects research. Proc. Natl. Acad. Sci. U. S. A. 115, E6106–E6115. doi:10.1073/pnas.1711978115.

Fliessbach, K., Rohe, T., Linder, N.S., Trautner, P., Elger, C.E., Weber, B., 2010. Retest reliability of reward-related BOLD signals. Neuroimage 50, 1168–1176. doi:10.1016/j.neuroimage.2010.01.036.

Fröhner, J.H., Teckentrup, V., Smolka, M.N., Kroemer, N.B., 2019. Addressing the reliability fallacy in fMRI: similar group effects may arise from unreliable individual effects. Neuroimage 195, 174–189. doi:10.1016/j.neuroimage.2019.03.053.

Frömer, R., Maier, M., Rahman, R.A., 2018. Group-level EEG-processing pipeline for flexible single trial-based analyses including linear mixed models. Front. Neurosci. 12, 1–15. doi:10.3389/fnins.2018.00048.

Gaspar, C.M., Rousselet, G.A., Pernet, C.R., 2011. Reliability of ERP and single-trial analyses. Neuroimage 58, 620–629. doi:10.1016/j.neuroimage.2011.06.052.

Goodhew, S.C., Edwards, M., 2019. Translating experimental paradigms into individual-differences research: contributions, challenges, and practical recommendations. Conscious. Cogn. 69, 14–25. doi:10.1016/j.concog.2019.01.008.

Hedge, C., Powell, G., Sumner, P., 2018. The reliability paradox: why robust cognitive tasks do not produce reliable individual differences. Behav. Res. Methods 50, 1166–1186. doi:10.3758/s13428-017-0935-1.

Höller, Y., Uhl, A., Bathke, A., Thomschewski, A., Butz, K., Nardone, R., Fell, J., Trinka, E., 2017. Reliability of EEG measures of interaction: a paradigm shift is needed to fight the reproducibility crisis. Front. Hum. Neurosci. 11, 1–15. doi:10.3389/fnhum.2017.00441.

Hu, L., Iannetti, G.D., 2019. Neural indicators of perceptual variability of pain across species. Proc. Natl. Acad. Sci. U. S. A. 116, 1782–1791. doi:10.1073/pnas.1812499116.

Huang, G., 2019. EEG/ERP data analysis toolboxes. In: Li, H., Zhiguo, Z. (Eds.), EEG Signal Processing and Feature Extraction. Springer, pp. 407–434. doi:10.1007/978-981-13-9113-2.

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. doi:10.1016/j.physbeh.2014.03.008.

Infantolino, Z.P., Luking, K.R., Sauder, C.L., Curtin, J.J., Hajcak, G., 2018. Robust is not necessarily reliable: from within-subjects fMRI contrasts to between-subjects comparisons. Neuroimage 173, 146–152. doi:10.1016/j.neuroimage.2018.02.024.

Ip, C.T., Ganz, M., Ozenne, B., Sluth, L.B., Gram, M., Viardot, G., l'Hostis, P., Danjou, P., Knudsen, G.M., Christensen, S.R., 2018. Pre-intervention test-retest reliability of EEG and ERP over four recording intervals. Int. J. Psychophysiol. 134, 30–43. doi:10.1016/j.ijpsycho.2018.09.007.

Jansen, B.H., Rit, V.G., 1995. Electroencephalogram and visual evoked potential generation in a mathematical model of coupled cortical columns. Biol. Cybern. 73, 357–366. doi:10.1007/BF00199471.

Kraemer, H.C., 2014. The reliability of clinical diagnoses: state of the art. Annu. Rev. Clin. Psychol. 10, 111–130. doi:10.1146/annurev-clinpsy-032813-153739.

Lee, K.H., Choi, Y.Y., Gray, J.R., Cho, S.H., Chae, J.H., Lee, S., Kim, K., 2006. Neural correlates of superior intelligence: stronger recruitment of posterior parietal cortex. Neuroimage 29, 578–586. doi:10.1016/j.neuroimage.2005.07.036.

Leue, A., Klein, C., Lange, S., Beauducel, A., 2013. Inter-individual and intra-individual variability of the N2 component: on reliability and signal-to-noise ratio. Brain Cogn 83, 61–71. doi:10.1016/j.bandc.2013.06.009.

Li, M., Wang, D., Ren, J., Langs, G., Stoecklein, S., Brennan, B.P., Lu, J., Chen, H., Liu, H., 2019. Performing group-level functional image analyses based on homologous functional regions mapped in individuals. PLoS Biol 17, 1–27. doi:10.1371/journal.pbio.2007032.

McGraw, K.O., Wong, S.P., 1996. Forming inferences about some intraclass correlations coefficients": correction. Psychol. Methods 1. doi:10.1037//1082-989x.1.4.390, 390–390.

Munsters, N.M., van Ravenswaaij, H., van den Boomen, C., Kemner, C., 2019. Test-retest reliability of infant event related potentials evoked by faces. Neuropsychologia 126, 20–26. doi:10.1016/j.neuropsychologia.2017.03.030.

Nelson, E.E., Guyer, A.E., 2012. Beyond brain mapping: using neural measures to predict real- world outcomes. Curr. Dir. Psychol. Sci. 1, 233–245. doi:10.1177/0963721412469394.Beyond.

Noble, S., Scheinost, D., Constable, R.T., 2019. A decade of test-retest reliability of functional connectivity: a systematic review and meta-analysis. Neuroimage 203, 116157. doi:10.1016/j.neuroimage.2019.116157.

Onie, S., Most, S., 2017. Two roads diverged: distinct mechanisms of attentional bias differentially predict negative affect and persistent negative thought. Emotion 17. doi:10.1037/emo0000280.

Pernet, C., Garrido, M.I., Gramfort, A., Maurits, N., Michel, C.M., Pang, E., Salmelin, R., Schoffelen, J.M., Valdes-Sosa, P.A., Puce, A., 2020. Issues and recommendations from the OHBM COBIDAS MEEG committee for reproducible EEG and MEG research. Nat. Neurosci. 23, 1473–1483. doi:10.1038/s41593-020-00709-0.

Plichta, M.M., Schwarz, A.J., Grimm, O., Morgen, K., Mier, D., Haddad, L., Gerdes, A.B.M., Sauer, C., Tost, H., Esslinger, C., Colman, P., Wilson, F., Kirsch, P., Meyer-Lindenberg, A., 2012. Test-retest reliability of evoked BOLD signals from a cognitive-emotive fMRI test battery. Neuroimage 60, 1746–1758. doi:10.1016/j.neuroimage.2012.01.129.

Polich, J., 2004. Clinical application of the P300 event-related brain potential. Phys. Med. Rehabil. Clin. N. Am. 15, 133–161. doi:10.1016/S1047-9651(03)00109-8.

Raemaekers, M., Vink, M., Zandbelt, B., van Wezel, R.J.A., Kahn, R.S., Ramsey, N.F., 2007. Test-retest reliability of fMRI activation during prosaccades and antisaccades. Neuroimage 36, 532–542. doi:10.1016/j.neuroimage.2007.03.061.

Segalowitz, S.J., Barnes, K.l, 1993. The reliability of ERP components in the auditory oddball paradigm. Psychophysiology 30, 451–459. doi:10.1111/j.1469-8986.1993.tb02068.x.

Seghier, M.L., Price, C.J., 2018. Interpreting and utilising intersubject variability in brain function. Trends Cogn. Sci. 22, 517–530. doi:10.1016/j.tics.2018.03.003.

Spearman, C., 1910. Correlation calculated from faulty data. Br. J. Psychol. 3, 271–295. doi:10.1111/j.2044-8295.1910.tb00206.x, 1904-1920.

Sur, S., Sinha, V.K., 2009. Event-related potential: an overview. Ind. Psychiatry J. 18, 70–73. doi:10.4103/0972-6748.57865.

Thigpen, N.N., Kappenman, E.S., Keil, A., 2017. Assessing the internal consistency of the event-related potential: an example analysis. Psychophysiology 54, 123–138. doi:10.1111/psyp.12629.

Tibon, R., Levy, D.A., 2015. Striking a balance: analyzing unbalanced event-related potential data. Front. Psychol. 6, 1–4. doi:10.3389/fpsyg.2015.00555.

Van Rijsbergen, N.J., Schyns, P.G., 2009. Dynamics of trimming the content of face representations for categorization in the brain. PLoS Comput. Biol. 5. doi:10.1371/journal.pcbi.1000561.

Vetter, N.C., Steding, J., Jurk, S., Ripke, S., Mennigen, E., Smolka, M.N., 2017. Reliability in adolescent fMRI within two years - a comparison of three tasks. Sci. Rep. 7, 1–11. doi:10.1038/s41598-017-02334-7.

Yarkoni, T., Braver, T.S., 2010. Cognitive neuroscience approaches to individual differences in working memory and executive control: conceptual and methodological issues 87–107. https://doi.org/10.1007/978-1-4419-1210-7_6

Youssofzadeh, V., Prasad, G., Wong-Lin, K.F., 2015. On self-feedback connectivity in neural mass models applied to event-related potentials. Neuroimage 108, 364–376. doi:10.1016/j.neuroimage.2014.12.067.