Brain Stimulation 11 (2018) 536–544

Contents lists available at ScienceDirect

Brain Stimulation

journal homepage: http://www.journals.elsevier.com/brain-stimulation

Test-retest reliability of transcranial magnetic stimulation EEG evoked potentials

Lewis J. Kerwin [a, b, c, d, 1], Corey J. Keller [a, b, c, 1], Wei Wu [a, b, c, e], Manjari Narayan [a, b, c], Amit Etkin [a, b, c, *]

[a] Department of Psychiatry and Behavioral Sciences, Stanford University, Stanford, CA 94305, USA
[b] Stanford Neuroscience Institute, Stanford University, Stanford, CA 94305, USA
[c] The Sierra Pacific Mental Illness, Research, Education, and Clinical Center (MIRECC), Veterans Affairs Palo Alto Healthcare System, Palo Alto, CA 94394, USA
[d] Weill Cornell Medicine, Cornell University, New York, NY, USA
[e] School of Automation Science and Engineering, South China University of Technology, Guangzhou, Guangdong 510640, China

ARTICLE INFO

Article history:
Received 15 July 2017
Received in revised form
21 December 2017
Accepted 27 December 2017
Available online 29 December 2017

Keywords:
Transcranial magnetic stimulation (TMS)
Electroencephalogram (EEG)
Evoked potentials
Reliability
Plasticity

ABSTRACT

Background: Transcranial magnetic stimulation (TMS)-evoked potentials (TEPs), recorded using electroencephalography (TMS-EEG), offer a powerful tool for measuring causal interactions in the human brain. However, the test-retest reliability of TEPs, critical to their use in clinical biomarker and interventional studies, remains poorly understood.
Objective/Hypothesis: We quantified TEP reliability to: (i) determine the minimal TEP amplitude change which significantly exceeds that associated with simply re-testing, (ii) locate the most reliable scalp regions of interest (ROIs) and TEP peaks, and (iii) determine the minimal number of TEP pulses for achieving reliability.
Methods: TEPs resulting from stimulation of the left dorsolateral prefrontal cortex were collected on two separate days in sixteen healthy participants. TEP peak amplitudes were compared between alternating trials, split-halves of the same run, two runs five minutes apart and two runs on separate days. Reliability was quantified using concordance correlation coefficient (CCC) and smallest detectable change (SDC).
Results: Substantial concordance was achieved in prefrontal electrodes at 40 and 60 ms, centroparietal and left parietal ROIs at 100 ms, and central electrodes at 200 ms. Minimum SDC was found in the same regions and peaks, particularly for the peaks at 100 and 200 ms. CCC, but not SDC, reached optimal values by 60–100 pulses per run with saturation beyond this number, while SDC continued to improve with increased pulse numbers.
Conclusion: TEPs were robust and reliable, requiring a relatively small number of trials to achieve stability, and are thus well suited as outcomes in clinical biomarker or interventional studies.

Introduction

Transcranial magnetic stimulation (TMS) activates cortical neurons at the stimulation site, inducing action potentials which cause downstream effects throughout the brain [1]. These TMS-evoked Potentials (TEPs) can be recorded with electroencephalography (EEG) showing characteristic waveforms between 0 and ~300 milliseconds after the magnetic pulse [2], see Supplementary data. TEPs display neurocircuit causality since the source of stimulation is known and experimentally controlled [3,4] and, with the millisecond temporal resolution of EEG, linked neuronal events can be temporally distinguished [3]. These attributes give TEPs a unique perspective on pathological brain states. To this end they have been used to study diseases ranging from bipolar disorder [5] to schizophrenia [6] to epilepsy and neurodegenerative disease [7,8]. They hold potential as biomarkers of etiological subtyping [7], treatment selection [9] and tracking therapeutic progress [10,11].

While numerous studies have investigated the validity and responsiveness of TEPs across a range of diseases and interventions [5–8,12,13], they have paid less attention to TEP reliability. Reliability, meaning a test's ability to render the same value under unchanging conditions, is like validity, considered an essential feature for any clinical biomarker [14]. Prior TEP studies however have often probed for significant differences between participant groups by applying t-tests without prior assessment of signal noise

* Corresponding author. Stanford University, Department of Psychiatry and Behavioral Sciences, 401 Quarry Road, MC: 5797, Stanford, CA 94305-5797, USA.
E-mail address: amitetkin@stanford.edu (A. Etkin).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.brs.2017.12.010
1935-861X/

[14]. Such methods become problematic when considering the number of minute differences in methodology known to impact the TEP signal, including coil placement [15,16], coil orientation [17], stimulus intensity [15,16] and coil shape [15,18]. Given the space of possible parameters and the growing numbers of published results – which may only represent a fraction of attempted ones [19]– a test result whose p-value falls below 0.05 may in fact be likely to represent random chance [20]. Before attempting to interpret such a result, the signal noise must be quantified under unchanging conditions. In other words, reliability is a prerequisite to validity [21]. Failure to measure it may lead to false positive findings, as well as false negatives since studies may unknowingly lack the power to detect real changes [20].

To the best of our knowledge, two papers have discussed single-pulse TMS-EEG reliability. Lioumis et al. compared TEPs measured 2–5 min apart in seven healthy participants and report significant correlation between retested measurements without significant difference between them [22]. In a second study, Casarotto and colleagues report the divergence index that best separates TEPs of different participants (N = 4) while grouping together TEPs from the same participant [23]. While encouraging, these findings leave several important questions unresolved. First, the T-test and divergence index are better suited to detect a significant difference between putatively different groups (validity) than to confirm agreement between repeated iterations of the same one (reliability). Even a failure to reject the null hypothesis (i.e. a p value greater than 0.05) can only ensure that agreement (i.e. lack of difference) is more than 5% likely, which does not alone guarantee strong reliability [24]. Second, correlation between repeated TEPs is necessary but not sufficient to confirm measurement agreement; plausible scenarios could yield high Pearson correlation but low agreement, which could amplify future false positive rates [24,25]. For instance, if habituation led to systematically lower TEPs on day 2 compared to day 1, correlation might be near perfect while the actual values disagreed significantly. For these reasons the characterization of TEPs can benefit from statistical tests tuned specifically for reliability [15].

Here, we calculated two such metrics for TEPs: concordance correlation coefficient (CCC, quantifying the intersubject discriminability of TEPs) and smallest detectable change (SDC, quantifying the minimum change in TEP necessary to confirm underlying neural change rather than normal measurement variation). These statistics, defined further in Methods, are designed specifically to assess reliability between repeated tests and to our knowledge represent the first application of such methods in TMS-EEG literature. In short, CCC assesses the ability of a test to distinguish between different individuals. SDC meanwhile represents the likelihood of a test to detect change in the same individual [25–27]. We studied healthy participants, comparing TEPs across a variety of temporal intervals to locate the scalp regions and TEP peaks with the highest reliability. We analyzed how incremental increases in the number of pulses per run affects reliability to find the optimal balance between expedience and reliability. By quantifying these attributes, we aim to improve the power and confidence of future TMS-EEG studies, guide ongoing analyses to sites with the cleanest signal, and help forge the establishment of clinically useful EEG biomarkers.

Methods

Participants (N = 16) were: 18–65 years old (a large age range was chosen so as not to constrain findings to any a priori group), right-handed, and denied current psychiatric or neurological diagnoses. All data were collected at Stanford University under an approved institutional review board protocol after participants gave their written informed consent.

TEPs were collected on two separate days within one week of each other, a range chosen to capture normal variation in brain-state but likely not long enough for significant long-term plasticity to occur. On each day participants underwent two identical TMS-EEG recording blocks, or runs, separated by 5 min. Both runs consisted of 150 individual single pulse TMS-EEG trials, as described below.

Transcranial magnetic stimulation

TMS was performed with a MagPro R30 stimulator (MagVenture, Denmark) and an MCF-B65 figure-8 Coil (MagVenture, Denmark). Resting motor threshold (rMT) was obtained at the beginning of every session by stimulating the left motor cortex and defined as the intensity that produced a visible twitch in adductor policis brevis on 50% of stimulations [3], see Supplementary data. For TMS, the left dorsolateral prefrontal cortex (DLPFC) was located using neuronavigation. Specifically, we used a previously reported map of the fronto-parietal executive network derived from an independent components analysis of resting fMRI connectivity data from an independent group of healthy participants [28]. TMS coil placement was guided by Visor2 LT 3D neuro-navigation system (ANT Neuro, Netherlands) based on co-registration of the functionally defined target to each participant's structural MRI (T1 weighted, slice distance 1 mm, slice thickness 1 mm, sagittal orientation, acquisition matrix 256 × 256, 3T GE DISCOVERY MR750 scanner). TMS coil was positioned so handle pointed posterolaterally, 45° to the nasion-inion axis.

Each run entailed 150 pulses (biphasic pulses at 280μs pulse width) at an intensity of 120% rMT delivered with an inter-pulse interval of 1800–2200 ms (jittered to prevent entrainment or psychological expectation effects). Stimulation intensity was set to 120% of that session's rMT, a level used previously for both TMS-evoked potentials [29,30] and clinical rTMS [31]. Stimulator recharge was delayed to prevent recharge artifact on the EEG [15]. A thin (0.5 mm) foam pad was attached to the TMS coil to minimize electrode movement and bone-conducted auditory artifact. White noise matching the frequency of the TMS click [32] was provided with noise cancellation headphones in order to reduce the auditory artifact described previously [3,15]. Participants were instructed to keep their eyes open and gaze relaxed throughout each run.

Electroencephalography

64-channel EEG data were obtained using two 32-channel TMS-compatible BrainAmp DC amplifiers (5 kHz sampling rate; ± 16.384 mv measurement range; analog low pass filter 1 kHz). These were attached to the Easy EEG cap (EasyCap GmbH, Germany) with extra flat, freely rotatable, sintered, interrupted disk, Ag-AgCl electrodes designed to record high-quality data without overheating with simultaneous TMS [33]. Equidistant electrode distribution surveyed frontal, temporal, parietal and occipital scalp locations. Electrode impedances were kept below 5 kΩ. The reference electrode was affixed to the nose. EEG data was recorded using BrainVision Recorder software (Brain Products GmbH, Germany).

Data organization and processing

All data analysis was carried out on MATLAB (R2016b, The Mathworks Inc, MA) with custom scripts. Data was preprocessed using a fully automated artifact rejection algorithm developed in our lab for TMS-EEG [34]. The algorithm proceeds as follows: 1) The initial 10 ms data segment following each TMS pulse is discarded to

remove the large stimulation-induced electric artifact; 2) The EEG data are downsampled to 1 kHz; 3) Big decay artifacts are automatically removed using ICA based on thresholding; 4) The 60 Hz AC line noise artifact is identified via the Thompson F-statistic and removed by a multi-taper regression technique. 5) Non-physiological slow drifts in the EEG recordings are removed using a 0.01 Hz high-pass filter; 6) The spectrally filtered EEG data is then re-referenced to the common average and epoched with respect to the TMS pulse (−500 to 1000 ms); 7) Bad trials are automatically rejected by thresholding the magnitude of each trial. Bad channels are rejected based on thresholding the spatial correlations among channels. The rejected bad channels are then interpolated from the EEG of adjacent channels. 8) Remaining artifacts are automatically removed using ICA. ICs related to the scalp muscle artifact, ocular artifact, ECG artifact, are rejected using a pattern classifier trained on expert-labeled ICs from other TMS/EEG data sets. The algorithm was validated against manual rejection by EEG experts on data not used for training. The results revealed that ARTIST classified artifact ICs at 96% accuracy, significantly outperforming a state-of-the art automated algorithm for conventional EEG data.

The above preprocessing algorithm resulted in a set of EEG time series vectors (from −500 to +1000 ms, relative to the TMS pulse) representing every electrode (64), trial (140 per run, the highest multiple of ten shared by all runs after bad trials were removed from the original 150), run (2), day (2) and participant (16). We first analyzed only the first ten trials per run. Then to measure the effects of trial number on reliability, we repeated analysis while increasing trial number by increments of ten trials up to 140 per run. In keeping with the relatively low spatial resolution of EEG [4], seven sensor regions of interest (ROIs) were defined as particular electrode groups whose signals were averaged (see Supplementary Material). Analysis proceeded for each combination of (i) trial number and (ii) ROI in four main steps:

Grand average latency recording (Fig. 1A–B)

For each ROI and trial number (i.e. number of trials averaged together), the grand average TEP waveform across all participants was calculated (Fig. 1A). Peak latencies of the four largest peaks in the grand average were found (Fig. 1B). We chose this number because prior literature typically identifies four characteristic TEP peaks (N40, P60, N100, P200) [3,33,35]. The variation in exact latency in these prior studies led us to define peaks in an unguided process that identified the most prominent peaks between 30 and 250 ms for each ROI (Tables 1 and 2).

Trial sorting (Fig. 1C)

All trials considered were sorted into two partitions per participant, yielding 2 × 16 partitions total. Trial sorting was performed with four parallel methods: 1) odd trials versus even trials (odd vs. even), 2) the first half of a run versus the second half (split 1 vs. 2), 3) the first run in the day versus the second run (run 1 vs. 2), and 4) the first day versus the second (day 1 vs. 2). These parallel comparisons were chosen in order to capture the different sources of temporal variance that may arise when clinical populations are studied.

Peak quantification (Fig. 1D)

The average waveform of each partition was quantified by calculating its amplitude at the grand-average peak latencies found in step 1. For comparisons of trials within the same run (i.e. odd vs. even and split 1 vs. 2), we quantified the waveform and computed reliability metrics within each run separately, then averaged these results across runs (Fig. 1C). This prevented the introduction of unwanted inter-run variance in the results. Analogously, for run 1 vs. 2, we quantified each day separately then averaged across days. We explored alternative methods of peak quantification including calculating area under the curve and finding peaks near the grand-average peak latencies in each

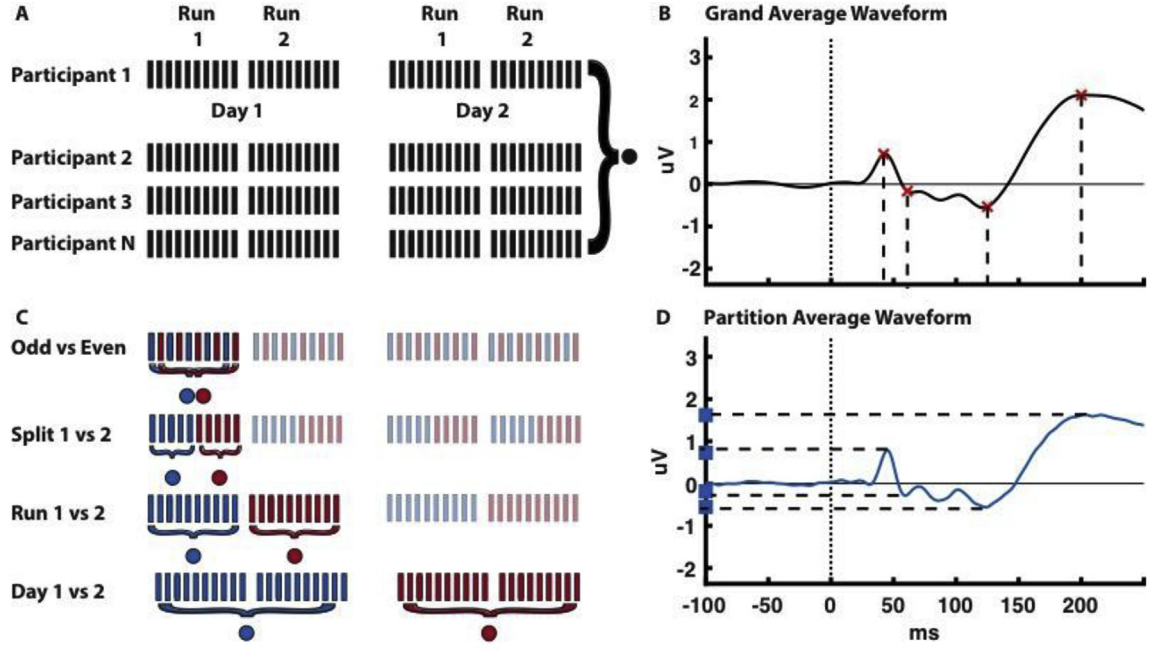

Fig. 1. Method for TEP Quantification. For each ROI: (A) Each participant underwent two separate days of TEP measurements, each with two runs containing 150 trials (vertical bars). To obtain grand-average peak latencies, all trials from all participants were averaged together into a *grand-average waveform* (black circle). (B) The four most prominent peaks (red crosses) of the grand-average waveform were identified. (C) For quantification of reliability, trials were grouped by four parallel comparison methods into two *partitions*, each producing a *partition average waveform* (blue and red circles). (D) Each *partition* average waveform was quantified by measuring at each grand-average peak latency (dashed lines) the waveform's amplitude (blue squares).

Table 1
Peak latencies identified from grand average waveforms.

	N40	P60	N100	P200
Left PFC	58	77	126	219
Right PFC	43	64	122	200
Left Parietal	42	73	127	180
Right Parietal	42	63	95	162
Central	55	71	86	195
Centroparietal	48	74	129	167
Occipital	42	61	125	200

For each region, a grand average waveform across all participants and trials was determined. Peaks were then found and the latencies of the largest four peaks were recorded (in milliseconds).

Table 2
Peak prominences identified from grand average waveforms.

	N40	P60	N100	P200
Left PFC	1.4308	0.8764	0.7648	1.274
Right PFC	1.291	0.7315	1.6252	2.7718
Left Parietal	0.9699	1.0068	1.7564	1.0962
Right Parietal	1.2866	0.7695	1.3381	1.3312
Central	0.8343	0.6517	2.6708	4.8725
Centroparietal	0.9174	0.9592	1.1562	0.6904
Occipital	1.0787	0.7024	1.3151	2.4423

Prominence (in μV) was calculated as the average amplitude difference between a peak and its two neighboring peaks of opposite polarity. All prominences correspond to peak latencies in Table 1.

partition. However, as these methods depended on a priori parameters (e.g. integration width, maximum distance between grand-average peak and partition peak etc.) we elected to follow the unguided algorithm described above.

Statistical analysis

A unique combination of (i) trial number (ii) ROI (iii) peak (e.g. N100) and (iv) comparison method (e.g. split 1 vs. 2) were selected. Each such parameterization produced a 2×16 table of TEP amplitudes corresponding to both partitions and all participants. For each comparison, reliability metrics (described further in the following section) were calculated. Confidence intervals were calculated by repeating steps 1 to 4 on 100 successive iterations that randomly half-sampled each partition's trials, then calculating the standard error.

Reliability metrics

We used two reliability statistics, each targeting a potential future clinical application of TMS-EEG:

Concordance correlation coefficient

The ability of a measurement to distinguish different individuals, for instance to subcategorize a brain disorder, is best assessed by an intraclass correlation coefficient (ICC), which represents the ratio of the inter-participant variance to the total variance [24,36]. We applied the concordance correlation coefficient (CCC), a form of ICC optimally tuned to assess test-retest reliability [25,26].

$$CCC = \frac{2\sigma_{12}}{\sigma_1^2 + \sigma_2^2 + (\mu_1 - \mu_2)^2}$$

where σ_{12} is the covariance between two partitions (e.g. day 1 vs. 2) across all 16 participants, σ_x^2 is the variance within partition x across all participants and μ_x is the average of partition x.

Smallest detectable change

If instead of distinguishing between different participants we wish to detect significant change in a single participant over time, for instance to measure the effect of a therapy, we must first know the variation in TEPs when repeated in the same individual when no significant change has occurred. We therefore computed the smallest detectable change (SDC). Unlike CCC, this metric depends only on the dispersion of measurements within an individual, not on that between individuals [24,27].

$$SDC_{Indiv} = \sqrt{\sigma^2_{Intraparticipant}} * \sqrt{2} * 1.96$$

By determining how much change can be expected from chance or measurement error, SDC informs future research efforts of the minimum change in a biomarker that is needed to be 95% confident of a change in the "real," underlying construct. In combination with a future study's measured difference in mean TEP (e.g. before and after an intervention), SDC produces a more robust form of the T-test [27]. SDC values computed in this study were normalized by dividing by average peak amplitude (nSDC) and are therefore unitless. We furthermore calculated group SDC, the change in mean TEP amplitude required across a group of future participants to ensure significance:

$$SDC_{Group} = SDC_{Indiv} * \sqrt{N}$$

where N is the number of participants whose mean TEP change is being calculated. Notably, our choice of CCC and SDC parallels that of Schambra et al. (2015), who apply these tests to the reliability of motor-evoked potentials (MEPs) [24]. While MEPs and TEPs have different artifact profiles and therefore reliability, this study serves as a useful model for the current analysis.

Results

Participant demographics

16 participants (8 Female, mean age 31.6 ± 10.6) were included in this study. All completed both days. Separation of sessions ranged from 1 to 7 days (mean separation = 3.4 ± 2.6 days). Grand-average waveforms for each ROI demonstrated peak latencies at approximately 50 (42–58), 70 (61–77), 120 (86–127) and 200 (162–219) ms in every ROI (Table 1), consistent with previous research [26,28,30]. In keeping with prior literature, these will be referred to as N40, P60, N100 and P200.

Inter-participant variance consistently achieves significance and exceeds other sources

Prior to analyzing reliability, we performed an ANOVA to establish sources of variance. Variance between participants, days, runs and split-halves were each compared to the residual variance of the data set (Fig. 2). Inter-participant variance (averaged across all ROIs) was significant in each peak (Fig. 2; $F_{15,109} = 3.97 \pm 0.97$, $p < .001$ to $F_{15,109} = 10.92 \pm 7.63$, $p < .001$). Other sources of variance did not reach significance for any peak when averaged across ROIs ($F_{2,109}$ ranging from 0.22 ± 0.43, $p = .78$ to 2.34 ± 2.74 0, $p = .07$). On examination of individual ROIs, only two peak-ROI combinations showed significance without correction for multiple comparisons, both for between-day variance ($F_{1,109} = 8.71$, $p = .004$ for left parietal at N100; $F_{1,109} = 5.09$, $p = .026$ for occipital at P200).

TEP peaks achieve substantial concordance

Based on CCC, all peaks demonstrated substantial (>0.8) concordance [21] in at least one comparison level ($CCC_{maxN40} = 0.94 \pm 0.1$, $CCC_{maxP60=}0.84 \pm 0.2$, $CCC_{maxN100=}0.88 \pm 0.2$, $CCC_{maxP200=}0.95 \pm$

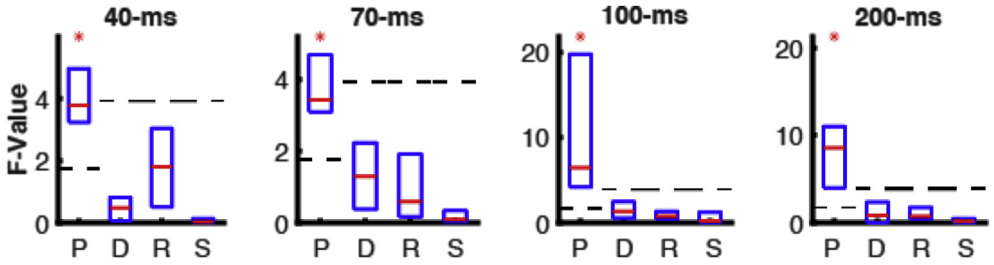

Fig. 2. Majority of variance in peak features lies between rather than within participants. Amplitude for of each of the four peaks was analyzed for variance across participants (P), days (D), runs (R) and split-halves (S). F-values (y axis) represent the ratio of variance within each variable to the variance due to noise in each peak feature. F-values were computed at all ROIs to produce a distribution (box denotes middle two quartiles) for each peak and source of variance. Horizontal dashed lines denote thresholds of significance for each F value ($p < .05$), which are lower for inter-participant variance because more degrees of freedom exist between participants. Stars denote sources of variance that reach statistical significance when averaged across ROI. Here only, inter-participant variance surpassed this threshold for statistically significance and did so in all four peaks.

0.2). At 40 ms CCC_{max} was found in the prefrontal ROI nearest the stimulation site (Fig. 3A) and spread to the central, centroparietal and left parietal ROIs in later peaks (Fig. 3B–D). Concordance was highest when comparing odd trials to even ones (Fig. 3A) and diminished somewhat with increasing comparison interval (e.g. for P200, $CCC_{maxP200odd\ vs.\ even} = 0.91 \pm 0.2$, $CCC_{maxP200split} = 0.95 \pm 0.4$, $CCC_{maxP200} = 0.86 \pm 0.2$, $CCC_{maxP200split} = 0.82 \pm 0.2$). Nevertheless, the peaks with highest concordance, N100 and P200, maintained substantial concordance across all levels of comparison, including comparison of separate days (Fig. 3C–D, $CCC_{maxN100} = 0.89 \pm 0.2$; $CCC_{maxP200=}0.82 \pm 0.2$ respectively). Notably the regions with highest concordance for both N100 (left parietal and centroparietal) and P200 (central) remained higher than other regions across all comparisons. The number of trials needed to achieve substantial concordance however, increased with increasing temporal retest interval. For instance, the P200 required only 20 trials per distribution for CCC to exceed 0.8 when comparing odd vs. even ($n = 20$, $CCC = 0.85 \pm 0.3$), but required 100 trials per distribution to meet this threshold when comparing separate TEPs from separate days ($n = 100$, $CCC = 0.80 \pm 0.2$). As expected, regions with high CCC also demonstrated high Pearson's correlation coefficient with a low difference between coefficients ($4.16 \pm 1.07\%$ difference at 60 trials per run).

To summarize the temporal patterns of TEP reliability, we examined the effect on reliability of increasing the interval between repeated tests regardless of region (Fig. 4). Maximum concordance remains substantial (>0.8) on all comparisons for the N100 and P200, with the N40 and P60 achieving substantial concordance across some of the intervals and otherwise largely remaining significant (>0.6). Additionally, to better characterize the spatial

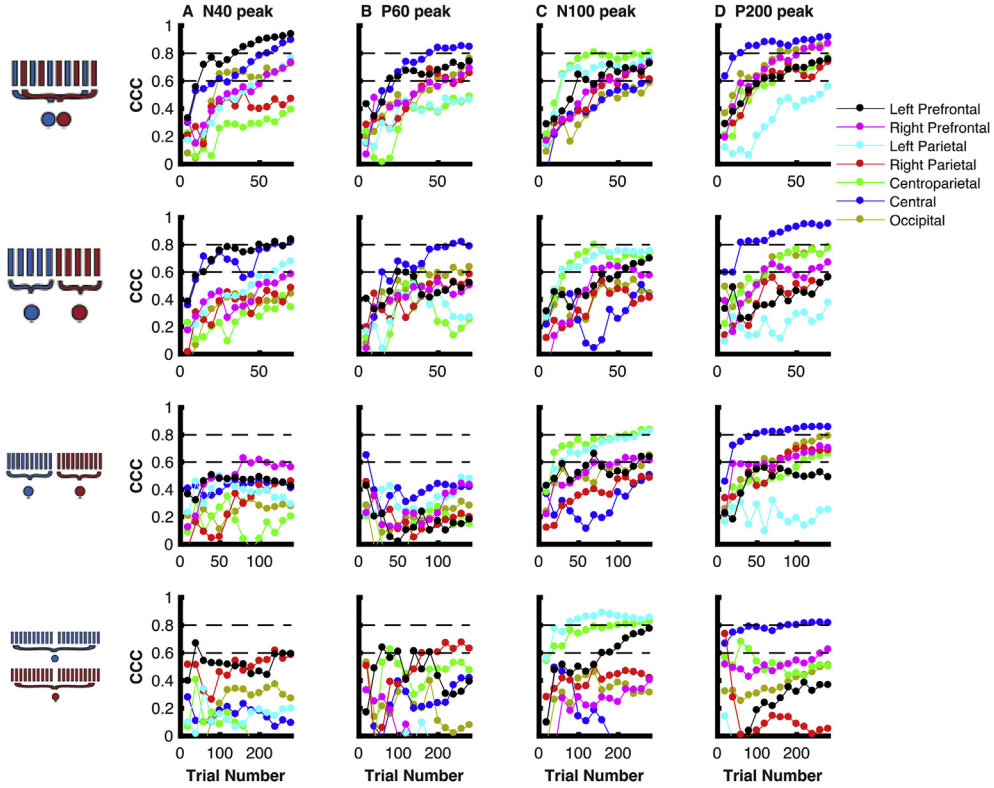

Fig. 3. Concordance across ROI, trial number and comparison methods. *Concordance correlation coefficient* (CCC) is plotted across comparison methods (rows: odd vs. even; split 1 vs. 2; run 1 vs. 2; day 1 vs. 2) and peaks of interest (columns). Each subplot displays CCC (y axis) as a function of total trial number (x axis) in all regions of interest (colors). Horizontal lines at 0.6 and 0.8 denote significant and substantial CCC respectively.

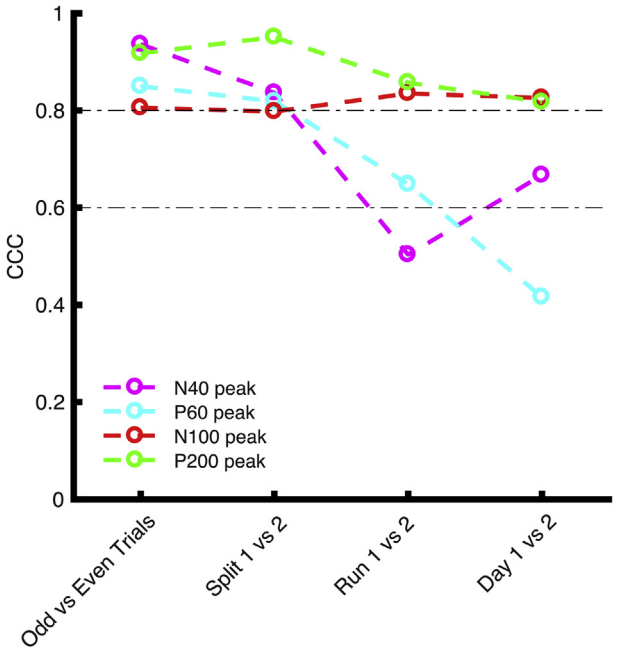

Fig. 4. Maximum concordance remains stable for most TEPs and runs. CCC for each peak window (see legend) is plotted across the four levels of comparison. For each line, the region with highest overall concordance was chosen and its maximum concordance across trial numbers was plotted.

patterns of TEP reliability, we examined reliability on the level of individual electrodes (Fig. 5). These results are in line with the ROI results. Specifically, the most reliable TEP electrodes were determined to be in the centroparietal region for both the N100 and the P200 ($CCC_{MaxN100} = 0.89 \pm 0.02$; $CCC_{MaxP200=}0.95 \pm 0.2$). The N40 was most reliable for electrodes near the left dorsolateral prefrontal site of stimulation ($CCC_{max40} = 0.73 \pm 0.1$) while the P60 was most reliable in frontocentral and right frontal electrodes ($CCC_{max60} = 0.71 \pm 0.2$).

P60, N100 and P200 achieve smallest detectable change below 100% of peak amplitude

To identify the regions with smallest TEP measurement error, normalized smallest detectable change (nSDC) was also calculated for each region-peak combination (see Methods). All four peaks demonstrate an nSDC below 1 in at least one region (Fig. 6). This means that a 100% change in a single participant's TEP amplitude would ensure the statistical significance of an intervention for that individual. The lowest, i.e. optimal, nSDC was found in the P200, where a 42% reduction in amplitude in the central scalp electrodes would meet the threshold for significance ($nSDC_{minP200} = 0.42 \pm 0.03$). Consistent with CCC, left parietal and centroparietal ROIs showed the lowest SDC for N100 ($nSDC_{minN100} = 0.91 \pm 0.03$) and central ROI for the P200 ($nSDC_{minP200} = 0.42 \pm 0.03$). Unlike concordance, additional trials up to approximately 150 per distribution appear to benefit nSDC. For instance, $nSDC_{P200}$ continues to improve up to 130 trials per run when comparing both runs and days ($nSDC_{minP200run1vs2} = 0.72 \pm 0.02$, $nSDC_{minP200Day1vs2} = 0.78 \pm 0.04$). Group SDC ($nSDC_{group}$) meanwhile falls below 0.2 (a 20% change in amplitude significance threshold) by 18 participants in the P200 ($SDC_{GroupMinP200Day} = 0.198 \pm 0.1$) and by 21 participants in the N100 ($SDC_{GroupMinP200Day} = 0.199 \pm 0.1$). In summary SDC, like CCC, identifies the N100 and P200 as the most reliable peaks, and suggests that an intervention must induce at least a 42% change in their amplitudes to ensure significance in a single participant with lower thresholds corresponding to larger multi-participant comparisons.

Discussion

Summary of findings

A biomarker can be evaluated in terms of its reliability, validity and responsiveness [14]. In this study, we investigated the first of these, the reliability of EEG potentials evoked from left dorsolateral prefrontal TMS. We found high concordance and low measurement error in characteristic peaks. Specifically, our findings show that (i) the peaks with best reliability in general lie at 100 and 200 ms; (ii) that these peaks' reliabilities are strongest in the left parietal, centroparietal and central regions; (iii) that substantial reliability is found for the N40 and P60 peaks within a run but not between separate runs or separate days, potentially limiting their clinical usefulness; (iv) that high reliability (>0.8) is seen across intervals ranging from 5 min to multiple days; and (v) that CCC, but not SDC, saturates after about 80 trials, with SDC continuing to benefit from additional pulses.

Interpretation of reliability metrics

CCC assesses the agreement between repeated measures, or the degree to which one participant's TEPs distinguish that participant from the cohort. Exceeding 0.8 across all comparison levels, the N100 and P200 peaks meet previously delineated standards for "substantial" concordance and therefore may be sufficiently reliable for clinical application [21]. These results also complement the findings of prior studies on TEP reliability, which determined divergence to be much greater between dissimilar TEPs than between repeated ones [23] and found high correlation (>0.8) between repeated measures several minutes apart [22].

SDC meanwhile assesses the degree to which a participant's TEP must change to confirm a significant change in underlying brain state, thus setting a helpful standard for biomarker responsiveness. Prior investigations have studied the responsiveness of TEPs to clinical interventions [5,9–11]. However, as the neural mechanisms underlying TEPs remain uncertain [2,3,15], the number of

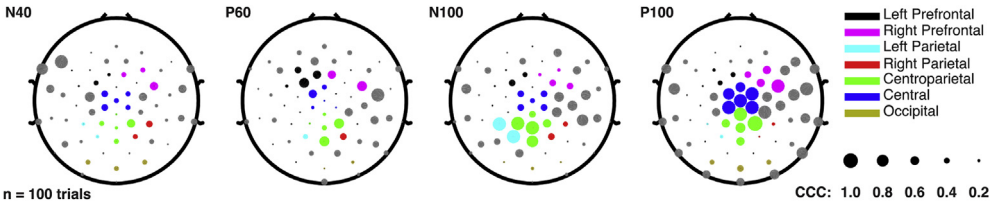

Fig. 5. Topographical plots show electrodes with maximum reliability at 100 trials. Concordance correlation coefficient (CCC) was calculated for individual electrodes at 100 trials between consecutive runs. Colors designate ROIs while gray disks represent electrodes not specific to an ROI. Radius of each disk represents CCC of corresponding electrode (ranging from 0 to 1).

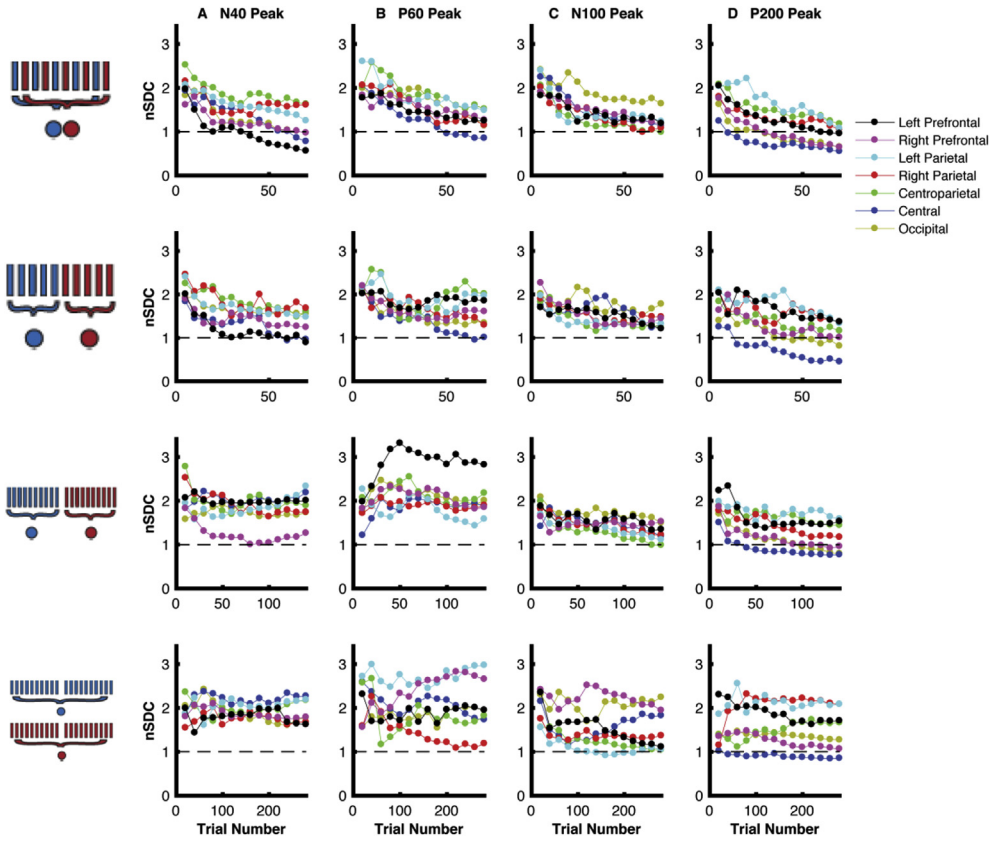

Fig. 6. Regions of smallest detectable change agree with those of optimal concordance. Normalized SDC (nSDC; divided by average peak amplitude) is plotted in four comparison intervals (rows) and peaks of interest (columns). Each subplot displays SDC (y axis) as a function of total trial number (x axis) in all regions of interest (colors).

participants and trials per participant needed to prevent alpha or beta error is difficult to predict [20]. Our findings on TEP smallest detectable change aid such efforts by showing the minimum change in each peak needed to ensure significance for a given number of participants. Specifically, SDC is higher than 0.5 in all windows and regions of interest between different days, meaning that greater than 50% change in amplitude would be required to confirm a significant change in one participant over multiple days. These findings echo those previously described for MEPs [24]. The SDC is however much lower for multi-participant studies; SDC for both N100 and P200 falls to 20% of average peak amplitude with approximately 20 participants. Moreover the SDC will likely improve with future methodology since the optimal stimulation site, intensity, coil orientation and coil shape have yet to be codified for TEPs [3,15]. These findings also help inform power calculations for TMS/EEG studies of clinical interventions.

Temporal and spatial patterns of reliability

TEP peaks are generally more reliable after 100 ms than before. One possibility for this is that later peaks may represent that of auditory potentials (found to contribute to N100 and P200 [37]) despite our efforts to control for this with auditory masking. Alternatively, the broader spatial distribution and magnitude of these potentials may promote reliability. Indeed, reliability within prefrontal regions is high for earlier potentials, consistent with the expectation that TMS-induced activity changes are initially localized to the site of stimulation, and only later reach other sites across the brain. Thus, it seems likely that the greater the magnitude of the evoked response, the more physiologically meaningful and reliable it is.

Interestingly, however, these results differ from the only other reported comparison of repeated measures of these specific peaks. Lioumis and colleagues found earlier peaks to have slightly higher Pearson's correlation coefficients ($r_{N40} = 0.88$; $r_{P60} = 0.922$; $r_{N100} = 0.867$; $r_{P200} = 0.644$) [22]. This discrepancy may be explained by the fact that this earlier study did not use the auditory masking, since auditory evoked potentials seem to influence the TEP most between 100 and 200 ms [37]. Additionally, Lioumis et al. quantified EEG peaks at only the left and right prefrontal regions, whereas we observe highest reliability in the central and parietal electrodes for later potentials.

Trial number and reliability

Our results show that optimal concordance can be achieved by 60–100 trials, fewer than the 150–300 recommended by prior guidelines [3]. Thus, future investigations with similar stimulation parameters may be able to limit trial number to <100 to balance optimal reliability with greatest experimental efficiency. Strikingly, for the N100 and P200, multiple days of separation did not introduce more variance in the TEP than existed between consecutive measurements five minutes apart, further securing their use as targets for clinical studies.

Limitations and future studies

One potential limitation in our study lies in the preprocessing method used. All data from this study was cleaned using an automated artifact rejection algorithm that identified components from all 140 non-rejected trials of each run. Therefore, the substantial

reliability measured using only the first 80 trials might be slightly inflated since these trials benefit from independent component analysis across 140 measurements. To address this possibility future studies might preprocess and analyze a fixed number of trials. At the same time, the ICA-based artifact rejection may have removed biologically relevant information, especially at latencies known to have artifacts. If heterogeneous between runs, this effect could have potentially lowered reported reliability. This effect might also potentially limit the remaining signal's ability to accurately track disease state or predict response, which can be determined only through future investigation of biomarker validity.

Additionally, our results quantify a reliability profile specific to healthy control participants with stimulation to the left DLPFC and encompassed a large age range (21–59). This reliability profile may change with other stimulation sites and in diseased populations or ones constrained to more specific age ranges. Thirdly, we measure resting motor threshold using visual inspection of twitch in the policis brevis muscle. While this is the method used in most rTMS clinics and thus may have external validity to future clinical applications, visual inspection may also overestimate rMT compared to measurement with electromyography [31].

This study identified the most reliable TMS-EEG peaks and the regions where they are most reliably measured. Future studies can further guide TEP application using the same statistical methods to optimize coil angle, stimulus intensity, coil type and artifact rejection parameters. Other efforts quantifying time-frequency effects of stimulation would also be valuable. In sum, our findings of high TEP reliability with a limited number of stimulations encourage future investigation for potential clinical and biomarker utility.

Funding

This work was funded by Big Idea in Neuroscience research funds from the Stanford Neurosciences Institute. LJK was funded by a Alpha Omega Alpha Carolyn L. Kuckein Student Research Fellowship. CJK was funded by the Alpha Omega Alpha Postgraduate Research Award. WW was supported by the National Natural Science Foundation of China under Grants 61403144 and the Tip-Top Scientific and Technical Innovative Youth Talents of Guangdong Special Support Program (No. 2015TQ01X361).

Appendix A. Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.brs.2017.12.010.

References

[1] Camprodon JA, Pascual-Leone A. Multimodal applications of transcranial magnetic stimulation for circuit-based psychiatry. JAMA Psychiatry 2016;73(4):407–8.
[2] Siebner HR, Hartwigsen G, Kassuba T, Rothwell JC. How does transcranial magnetic stimulation modify neuronal activity in the brain? Implications for studies of cognition. Cortex 2009;45(9):1035–42.
[3] Rosanova M, Casarotto S, Pigorini A, Canali P, Casali AG, Massimini M. Combining transcranial magnetic stimulation with electroencephalography to study human cortical excitability and effective connectivity. In: Fellin T, Halassa M, editors. Neuronal network analysis: concepts and experimental approaches. Totowa, NJ: Humana Press; 2012. p. 435–57.
[4] O'Shea J, Taylor PC, Rushworth MF. Imaging causal interactions during sensorimotor processing. Cortex 2008;44(5):598–608.
[5] Canali P, Sferrazza Papa G, Casali AG, Schiena G, Fecchio M, Pigorini A, et al. Changes of cortical excitability as markers of antidepressant response in bipolar depression: preliminary data obtained by combining transcranial magnetic stimulation (TMS) and electroencephalography (EEG). Bipolar Disord 2014;16(8):809–19.
[6] Ferrarelli F, Sarasso S, Guller Y, Riedner BA, Peterson MJ, Bellesi M, et al. Reduced natural oscillatory frequency of frontal thalamocortical circuits in schizophrenia. Arch Gen Psychiatr 2012;69(8):766–74.
[7] Kimiskidis VK. Transcranial magnetic stimulation (TMS) coupled with electroencephalography (EEG): biomarker of the future. Rev Neurol 2016;172(2):123–6.
[8] Pascual-Leone A, Freitas C, Oberman L, Horvath JC, Halko M, Eldaief M, et al. Characterizing brain cortical plasticity and network dynamics across the age-span in health and disease with TMS-EEG and TMS-fMRI. Brain Topogr 2011;24(3–4):302–15.
[9] Sun Y, Farzan F, Mulsant BH, Rajji TK, Fitzgerald PB, Barr MS, et al. Indicators for remission of suicidal ideation following magnetic seizure therapy in patients with treatment-resistant depression. JAMA Psychiatry 2016;73(4):337–45.
[10] Romero Lauro LJ, Rosanova M, Mattavelli G, Convento S, Pisoni A, Opitz A, et al. TDCS increases cortical excitability: direct evidence from TMS-EEG. Cortex 2014;58:99–111.
[11] Casarotto S, Canali P, Rosanova M, Pigorini A, Fecchio M, Mariotti M, et al. Assessing the effects of electroconvulsive therapy on cortical excitability by means of transcranial magnetic stimulation and electroencephalography. Brain Topogr 2013;26(2):326–37.
[12] Julkunen P, Jauhiainen AM, Westeren-Punnonen S, Pirinen E, Soininen H, Kononen M, et al. Navigated TMS combined with EEG in mild cognitive impairment and Alzheimer's disease: a pilot study. J Neurosci Meth 2008;172(2):270–6.
[13] Brodbeck V, Thut G, Spinelli L, Romei V, Tyrand R, Michel CM, et al. Effects of repetitive transcranial magnetic stimulation on spike pattern and topography in patients with focal epilepsy. Brain Topogr 2010;22(4):267–80.
[14] Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, et al. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. J Clin Epidemiol 2010;63(7):737–45.
[15] Siebner HR, Bergmann TO, Bestmann S, Massimini M, Johansen-Berg H, Mochizuki H, et al. Consensus paper: combining transcranial stimulation with neuroimaging. Brain Sci 2009;2(2):58–80.
[16] Brasil-Neto JP, Cohen LG, Panizza M, Nilsson J, Roth BJ, Hallett M. Optimal focal transcranial magnetic activation of the human motor cortex: effects of coil orientation, shape of the induced current pulse, and stimulus intensity. J Clin Neurophysiol 1992;9(1):132–6.
[17] Bonato C, Miniussi C, Rossini PM. Transcranial magnetic stimulation and cortical evoked potentials: a TMS/EEG co-registration study. Clin Neurophysiol 2006;117(8):1699–707.
[18] Cohen LG, Roth BJ, Nilsson J, Dang N, Panizza M, Bandinelli S, et al. Effects of coil design on delivery of focal magnetic stimulation. Technical considerations. Electroencephalogr Clin Neurophysiol 1990;75(4):350–7.
[19] Dickersin K. The existence of publication bias and risk factors for its occurrence. JAMA 1990;263(10):1385–9.
[20] Button KS, Ioannidis JP, Mokrysz C, Nosek BA, Flint J, Robinson ES, et al. Power failure: why small sample size undermines the reliability of neuroscience. Nat Rev Neurosci 2013;14(5):365–76.
[21] Shrout PE. Measurement reliability and agreement in psychiatry. Stat Meth Med Res 1998;7(3):301–17.
[22] Lioumis P, Kicic D, Savolainen P, Makela JP, Kahkonen S. Reproducibility of TMS-Evoked EEG responses. Hum Brain Mapp 2009;30(4):1387–96.
[23] Casarotto S, Romero Lauro LJ, Bellina V, Casali AG, Rosanova M, Pigorini A, et al. EEG responses to TMS are sensitive to changes in the perturbation parameters and repeatable over time. PLos One 2010;5(4), e10281.
[24] Schambra HM, Ogden RT, Martínez-Hernández IE, Lin X, Chang YB, Rahman A, et al. The reliability of repeated TMS measures in older adults and in patients with subacute and chronic stroke. Front Cell Neurosci 2015;9:335.
[25] Lin LI. A concordance correlation coefficient to evaluate reproducibility. Biometrics 1989;45(1):255–68.
[26] King TS, Chinchilli VM, Carrasco JL. A repeated measures concordance correlation coefficient. Stat Med 2007;26(16):3095–113.
[27] Beckerman H, Roebroeck ME, Lankhorst GJ, Becher JG, Bezemer PD, Verbeek ALM. Smallest real difference, a link between reproducibility and responsiveness. Qual Life Res 2001;10(7):571–8.
[28] Chen AC, Oathes DJ, Chang C, Bradley T, Zhou ZW, Williams LM, et al. Causal interactions between fronto-parietal central executive and default-mode networks in humans. Proc Natl Acad Sci U S A 2013;110(49):19944–9.
[29] Farzan F, Vernet M, Shafi MM, Rotenberg A, Daskalakis ZJ, Pascual-Leone A. Characterizing and modulating brain circuitry through transcranial magnetic stimulation combined with electroencephalography. Front Neural Circ 2016;10:73.
[30] Ferreri F, Pasqualetti P, Maatta S, Ponzo D, Ferrarelli F, Tononi G, et al. Human brain connectivity during single and paired pulse transcranial magnetic stimulation. Neuroimage 2011;54(1):90–102.
[31] McClintock SM, Reti IM, Carpenter LL, McDonald WM, Dubin M, Taylor SF, et al. National network of depression centers r TMSTG, American psychiatric association council on research task force on novel B, treatments. Consensus recommendations for the clinical application of repetitive transcranial magnetic stimulation (rTMS) in the treatment of depression. J Clin Psychiatr 2017.
[32] Rosanova M, Casali A, Bellina V, Resta F, Mariotti M, Massimini M. Natural frequencies of human corticothalamic circuits. J Neurosci 2009;29(24):7679–85.
[33] Ilmoniemi RJ, Kicic D. Methodology for combined TMS and EEG. Brain Topogr 2010;22(4):233–48.

[34] Wu CK W, Rogasch N, Longwell P, Shpigel E, Rolle C, Etkin A. ARTIST: a fully automated artifact rejection algorithm for single-trial TMS-EEG data. Hum Brain Mapp 2017 (under review).
[35] Rogasch NC, Daskalakis ZJ, Fitzgerald PB. Cortical inhibition of distinct mechanisms in the dorsolateral prefrontal cortex is related to working memory performance: a TMS-EEG study. Cortex 2015;64:68–77.
[36] Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychol Bull 1979;86(2):420–8.
[37] Nikouline V, Ruohonen J, Ilmoniemi RJ. The role of the coil click in TMS assessed with simultaneous EEG. Clin Neurophysiol 1999;110(8):1325–8.